SENSITIVITY TO DETAILED ACOUSTIC INFORMATION IN WORD RECOGNITION

ISBN: 90-76203-23-7 / 978-90-76203-23-2

Cover design: Linda van den Akker, Inge Doehring

Cover illustration: Petra van Alphen

Printed and bound by Ponsen & Looijen bv, Wageningen

SENSITIVITY TO DETAILED ACOUSTIC INFORMATION IN WORD RECOGNITION

een wetenschappelijke proeve

op het gebied van de Sociale Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom,

volgens besluit van het College van Decanen

in het openbaar te verdedigen

op donderdag 6 April 2006

des namiddags om 3:30 uur precies

door

Keren Batya Shatzman

geboren op 15 september 1972

te Jerusalem, Israël

Promoter:	Prof. dr. A. Cutler
Co-promoter:	Dr. J.M. McQueen
Manuscriptcommissie:	Prof.dr. A.C.M. Rietveld
	Prof. dr. G.T.M. Altmann (University of York)
	Dr. H. Quené (Universiteit Utrecht)

The research reported in this thesis was supported by a grant from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany.

To my parents

In memory of my brother

Acknowledgements

In many dissertations, it seems that people really look forward to writing the acknowledgement part. Not me. This is a moment I have dreaded for a long time: The book is finished and now I have to try and write something genuinely original...A difficult task given the long series of dissertations preceding this one. I guess the best way to express my gratitude to the many people that have helped me during my Ph.D. and made it an unforgettable period is to say: Thanks everybody!!

There are some people, however, who deserve a special note.

First, of course, my promoter, Prof. Anne Cutler, who gave me a job as a student-assistant some six years ago and later asked me to stay as a Ph.D. student. In my opinion, the head of the group is crucial in determining the group's character. In the case of the Comprehension Group this means hard-working and professional yet informal, critical but not dogmatic, ambitious but cooperative. It has been a real pleasure to be part of this.

My beacon of stability in this project has been my supervisor, James McQueen. I couldn't have wished for a better supervisor. Thank you for your patience, your meticulousness, the open door, always being there with helpful comments, listening patiently, the optimistic mood, the high standards, the bright ideas and generally for showing me the ropes of academic work (did I mention patience?). Most importantly, though, thank you for being a mensch.

I would further like to thank all the members of the Comprehension Group. It was a wonderful experience to work with so many people sharing their expertise. I would like to thank in particular Delphine Dahan for introducing me to the nitty-gritty of eye-tracking.

My eternal gratitude goes to all the people who made the practical part of my research possible. First and foremost, John Nagengast and Johan Weustink for helping me with NESU and rescuing my crashing experiments; Herbert for maintaining the network and restoring my files when I accidentally deleted them; Gert Klaas for support with the eye-tracking lab;

Reiner for not tiring of cleaning my sick profile; Kees for the participants' lists and, last but not least, Ad, for keeping it all together. Special thanks goes to Loulou Edelman, Petra van Alphen and Hanke van Buren for lending their voices to my stimuli. I'd also like to thank Agnes and Jose for opening the door, the administration for keeping the money coming, the library staff and in particular Karin who always managed to find the book on the shelf when I failed to see it and, of course, Pim 'boter-groenten?' for keeping me fed.

A warm thank you to all my friends at the Max Planck with whom I shared many memorable moments: The volleyball, the freezing and thawing of the MaxKrant, the Tubbs, the dancing, Sint Pannenkoek, the movies, the karaoke and so much more… I have no intention of revealing the details of these memories here. I just hope each of you knows I mean YOU when I say: thanks for sharing this time with me, thanks for laughing with me. I would like to specifically mention Kerstin, Martijn, Anita, Mirjam, Heidrun, Claudia and Annelie for those unforgettable lunches. I am also exceptionally indebted to Petra for putting me in the corner office, for making the cover of this book, for translating the summary and for being a really good friend. And speaking of which, I am extremely grateful to Anita and Heidrun for being my paranimphs (and looking after my cat).

Finally, I would like to thank my family for their support. Vetoda lehorai shehaviuni ad halom. Ik wil mijn schoonfamilie ook bedanken voor hun steun tijdens de jaren. And lastly Femke, without whom I would not be here at all.

Thanks everybody!!

Table of Contents

Introduction

CHAPTER 1

Language is all around us. Most of us spend almost every waking hour doing something that is language related: talking, listening, reading, memorizing a groceries list, watching TV, writing a dissertation...and the list goes on. Language not only serves for communicating with other people but also forms the basis of our thoughts and the sense we make of the world around us. Due to the ubiquity of language we tend to take it for granted. We are not aware that we use it all the time.

Spoken language is the most common means of communication. In most cases, in order to understand the meaning of a spoken utterance we need to recognize the words that were said. This seems trivial. At the conscious level, when someone talks to us, words simply appear to emerge from what they are saying. However, the speech signal itself is nothing more than a continuous stream of acoustic information caused by changes in air pressure, produced by the speaker. We start appreciating this when we hear an unknown language – spoken language then seems to be a long, uninterrupted babble.

That we experience speech in our own language as if there were clear breaks between the words is due to the fact that, at a level we are unaware of, the acoustic information in the speech signal is analyzed and compared to stored knowledge about what words sound like. The words that resemble the input the most are those that will be recognized. Not only are we unaware of this process, we also have no control over it. If we hear a language we know, we can not help but recognize the words. We may not always understand the meaning of an utterance, but we will recognize the words. For example, if someone says “this book does not constitute a rabbit” you will recognize the words they say, but probably not understand what they mean. Extracting the words from the speech signal is an automatic, uncontrolled process. It is this process that is at the focus of my thesis: How do listeners recognize the words in the speech they hear?

Investigating this question requires taking into account both the listener and the speaker. As mentioned before, the recognition of spoken words occurs as the acoustic information in the speech signal that is reaching our ears is analyzed and compared to stored knowledge about what words sound like. Thus, both the information in the speech signal (which is produced by the speaker) and the stored knowledge (held by the listener) are crucial for this process. This thesis examines the acoustic information and the stored knowledge that are involved in the process of spoken-word recognition.

The speech signal is a very rich source of acoustic information. When we hear someone speak we are not just aware of the words they are saying, but also of other characteristics, for example, whether the speaker is a man or a woman. The recognition of words, however, is not influenced by information such as the speaker's gender. What matters is that the words are pronounced clearly. However, not all mispronunciations are equal. For example, imagine a person pronouncing a [p]-sound instead of a [b] in the sentence "I'm going to get that book". Although the word "book" would be mispronounced, there would be little problem in recognizing it. In fact, the chances are that you would not even notice that it was mispronounced. Now imagine the same mispronunciation in the sentence "I'm going to watch that bear". In this case, you might think that the speaker intended to say "pear" (van Alphen & McQueen, 2006). Thus, the same mispronunciation has different consequences in different contexts. In the first example, word recognition would not be hindered; in the second example it would be. This demonstrates how stored linguistic knowledge about the existence of words (e.g., that "pook" is not a word but "pear" is) influences speech perception.

The word recognition process is quite tolerant to mispronunciations if these do not result in other words. But it can be extremely sensitive to very fine-grained acoustic details if these differentiate between two alternatives. Therefore, investigating which kind of acoustic information influences the perception of spoken words is often done in contexts that can be interpreted in two alternative ways, that is, in contexts that have a certain degree of ambiguity.

In this thesis I studied the influence of fine-grained acoustic information on the recognition of spoken words in ambiguous contexts of various types. Fully ambiguous contexts are those in which a phrase can be interpreted in two ways. Take, for example, the sentence “Because the expert witness had one slide, the jury did not know whether to believe him”. If this sentence were spoken, the phrase “one slide” could be understood as “once lied”. Although these two phrases sound alike, acoustic-phonetic research has shown that people tend to produce them in slightly different ways. For example, the [s] in the word “slide” tends to have a longer duration than the [s] in “once”. The influence of fine acoustic detail in Dutch versions of such contexts was examined in Chapter 2 and Chapter 3. Spoken sentences containing the ambiguous phrases were recorded and then manipulated. Splicing is a common manipulation: the recording of a word in one of the contexts is replaced by a recording from the other context (e.g., replacing the word “slide” with the recording of the [s] and the word “lied” from “once lied”). I tested whether listeners’ interpretation of the ambiguous phrase was influenced by this manipulation. I also examined whether manipulating the duration of the [s] alone would influence word recognition.

Temporarily ambiguous contexts are those in which the uncertainty regarding which words are spoken is only momentary. For example, if you hear the sentence “He had once met the Queen” it is very clear at the end of it which words were spoken. However, before hearing all of “met”, what you heard could also be the start of “He had one smell”. In Chapter 4, the manipulation of [s] duration was performed in such temporarily ambiguous sentences.

Another kind of temporary ambiguity is word embedding. There is a “pie” in the beginning of every “pirate” we hear, and a “seat” in the end of “receipt”. These are examples of onset- and offset-embedded words. Previous research on onset-embedded words has shown that there are small durational differences between, for example, the monosyllabic word “pie” and the first syllable of “pirate”. Furthermore, listeners use these acoustic differences to recognize the intended word (Davis, Marslen-Wilson & Gaskell, 2002; Salverda, Dahan, & McQueen,

2003). In Chapter 5, I looked at whether such fine-grained acoustic details are used when listeners hear words containing offset-embedded words. Using splicing, the recording of the carrier-word (e.g., “receipt”) was manipulated so that the embedded sequence was replaced either by the monosyllabic word (e.g., “seat”) or by the embedded sequence from another recording of the carrier-word (e.g., the second syllable of “receipt”). I examined whether the recognition of the carrier-words and the offset-embedded words would be influenced by this manipulation.

Chapter 6 reports an experiment investigating listeners’ sensitivity to the small durational differences that distinguish onset-embedded words from monosyllabic words, in their recognition of newly-learned words. Using new words allows a high degree of control on the exact information that listeners are exposed to. Participants learned new spoken words by associating them to novel shapes. The new words were either bisyllables (e.g., baptoe) or onset-embedded monosyllabic words (e.g., bap). The ambiguity in this case is therefore analogous to that in pie/pirate. The experiment examined whether listeners’ recognition of newly-acquired words is determined only by the experience they have had with those words, or whether recognition is also determined by prior experience with similar-sounding real words.

Eye-tracking

Because the process of extracting the words from the speech stream is automatic, we can have no subjective insights into how it operates. It is no use asking people whether they use this or that kind of acoustic information. Therefore, psycholinguists have devised methods to infer from listeners’ behavior which acoustic information they use in the process of spoken-word recognition. The eye-tracking paradigm, which I use throughout my thesis, is one such method.

In the eye-tracking paradigm, participants hear a sentence and are shown four objects presented as pictures on a computer screen. Their task is to click on and move the object referred to in the sentence with the computer mouse. This task exploits the fact that people make eye movements to objects as the names of the objects are spoken. Because eye movements can be monitored continuously, it is possible to examine the recognition process as the speech stream unfolds over time. For example, you might hear the sentence “click on the pirate” and among the four objects on the screen are a picture of pirate and a picture of a pie. At the end of the word “pirate” it is very likely that you would be looking at the picture of the pirate. But before you hear the second syllable of “pirate”, it is very likely that you would also look at the picture of the pie (more than at a picture of an unrelated word; Salverda et al., 2003). By monitoring eye movements we can detect the consequences of the momentary ambiguity. Moreover, the probability of looking at an object on the screen is a function of, amongst other things, the match between the acoustic information in the speech signal and the stored knowledge about what the name of that object sounds like. This allows us to create subtle modifications in the acoustic information in the speech signal and examine whether that has an effect on the recognition of the words, as reflected by the fixations that people make to the pictures on the screen. Thus, the sensitivity of the eye-tracking method allows the investigation of the sensitivity of the speech recognition system to very minute details in the speech signal. As the following chapters will show, this method can reveal the fine characteristics of the word recognition process.

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

CHAPTER 2

Keren B. Shatzman and James M. McQueen (2006), *Perception & Psychophysics*, **68**, 1-16.

Abstract

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"). 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, 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.

Introduction

"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. The competition process resolves these ambiguities, however, such that the result of the recognition process is a parse of non-overlapping 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 like 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 (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 (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 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 to say, 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 which 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 where the speaker intended the longer word (e.g., captain) and in sentences where 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 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, compared to sentences in which there was no local lexical ambiguity. However, there was no such delay if 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 on-line segmentation of continuous speech.

The influence of individual segment duration on listeners' off-line segmentation judgements 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 judgements. 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 judgements 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 on-line 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 which 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 [raIp] berries and a genuine [p] in ripe berries), even for tokens that were judged to be highly ambiguous in an off-line perceptual rating task.

Gow and Gordon (1995) examined recognition of lexically-ambiguous sequences which could either be interpreted as 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 non-initial 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 this study was to examine the degree to which different acoustic cues to word boundaries are used by listeners in their on-line 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 either [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 current study, therefore, we used phrases such 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 on-line 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 off-line measures (i.e., forced-choice tasks) they do not show that listeners use such differences during normal speech processing.

In this 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 our 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 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 either replaced 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 participants' fixations to the target picture differed across the splicing conditions. Subsequently, we examined which acoustic information 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 twenty 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 participants to adopt a strategy to deal with those items. Each target was instead paired with a picturable noun which 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 either 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 Art Explosion Library, 1995, Cycowicz, Friedman, Rothstein, & Snodgrass, 1997, and Snodgrass & Vanderwart, 1980)[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).

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

A female speaker of Dutch who was naïve 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 re-digitized at a sample rate of 16kHz 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.

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	

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 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, 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 which included (a) cluster-initial targets preceded by the carrier phrase (e.g., ze heeft wel een sleutel gezegd [she did say one key]); (b) 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]); (c) 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]); (d) 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 (e) 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 filler items (bak [bowl], klok [clock], scepter [scepter]) proved to be problematic to cross-splice without creating acoustic artefacts 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 Centre 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 (http://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 ms 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. Participants were tested individually. To ensure that they identified the pictures as intended, 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.

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.

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 a 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 presented over headphones using NESU software (http://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 ms. A spoken sentence was then presented and, simultaneously, a 5x5 grid with pictures appeared on the screen (see Figure 1). 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. Participants' fixations for the entire trial were completely unconstrained and participants 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 eye-tracker 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 participants were instructed to look at it. The experimenter could then correct potential drifts in the calibration of the eye tracker.

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. 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 participants with the task and procedure.

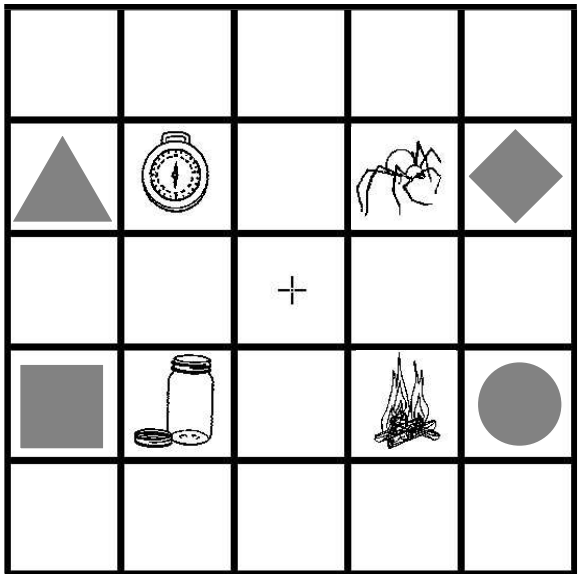


Figure 1. *Example of stimulus display presented to participants. The geometric shapes (triangle, diamond, circle aned square) and the central fixation cross were in fixed positions across trials. The pictures objects, and their positions, varied over trials. In this example, these were, clockwise from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

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 participants had clicked on the target picture with the mouse, which was taken to reflect the participants' identification of the referent. In most cases, participants were fixating the target picture when clicking on it. In the rare cases where 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, 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 ms 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 ms time slices from the splice point (the onset of the [s] preceding the target word) to 1200 ms thereafter.

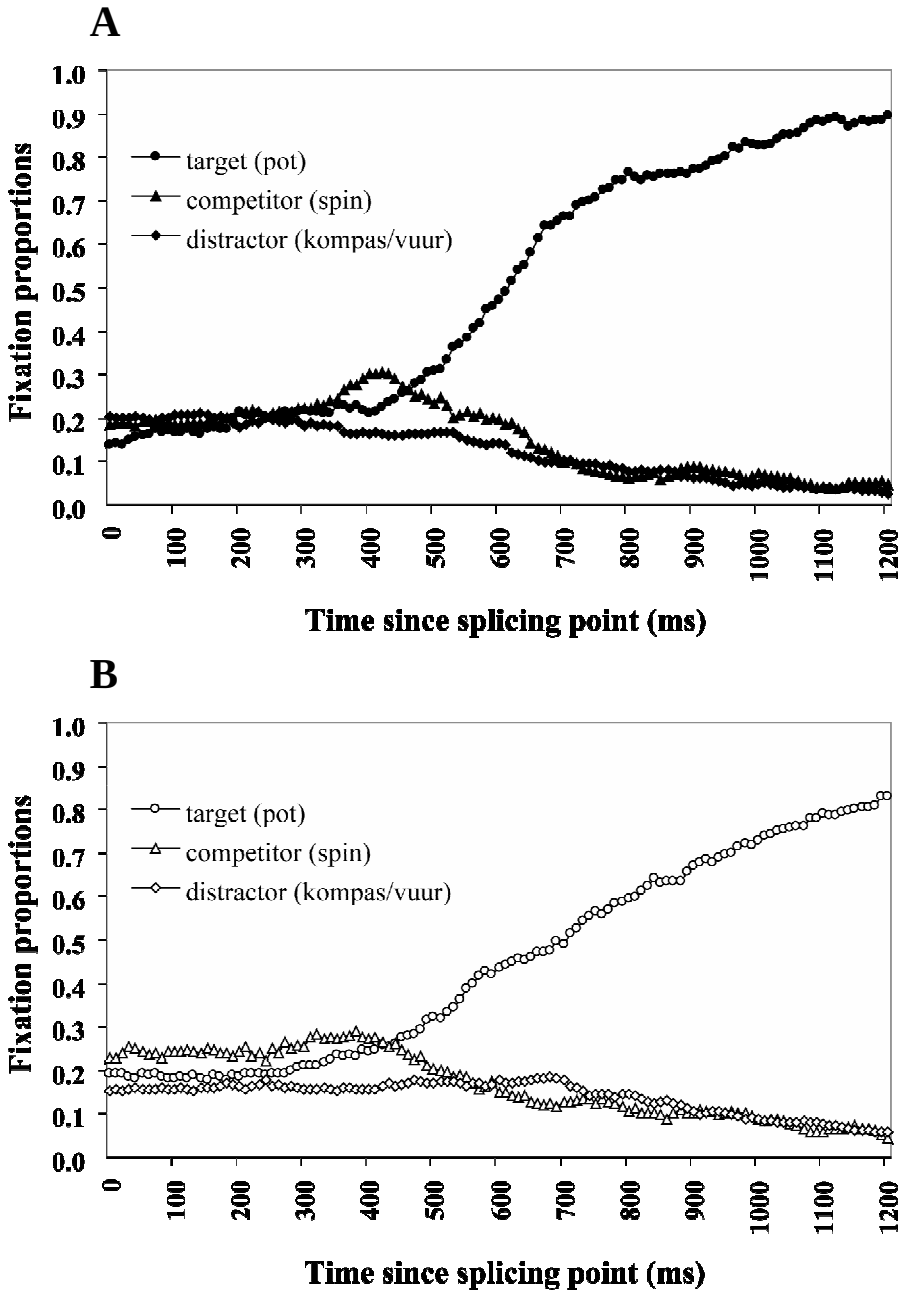


Figure 2. *Fixation proportions over time to the target, the competitor, and averaged distractors, in the identity-spliced condition (A) and the cross-spliced condition (B), in Experiment 1.*

As is apparent from Figure 2, fixation proportions to the competitor pictures began to rise in both conditions at around 300 ms. It is estimated that an eye movement is typically programmed about 200 ms before it is launched (e.g., Fischer, 1992; Hallett, 1986; Matin, Shao, & Boff, 1993; Saslow, 1967), so that 300 ms after the splicing point is approximately the moment at which fixations driven by the first 100 ms 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 ms on[2]. In both conditions, fixation proportions to the competitor pictures remained higher than those to the distractor pictures until around 600 ms, where they merged again.

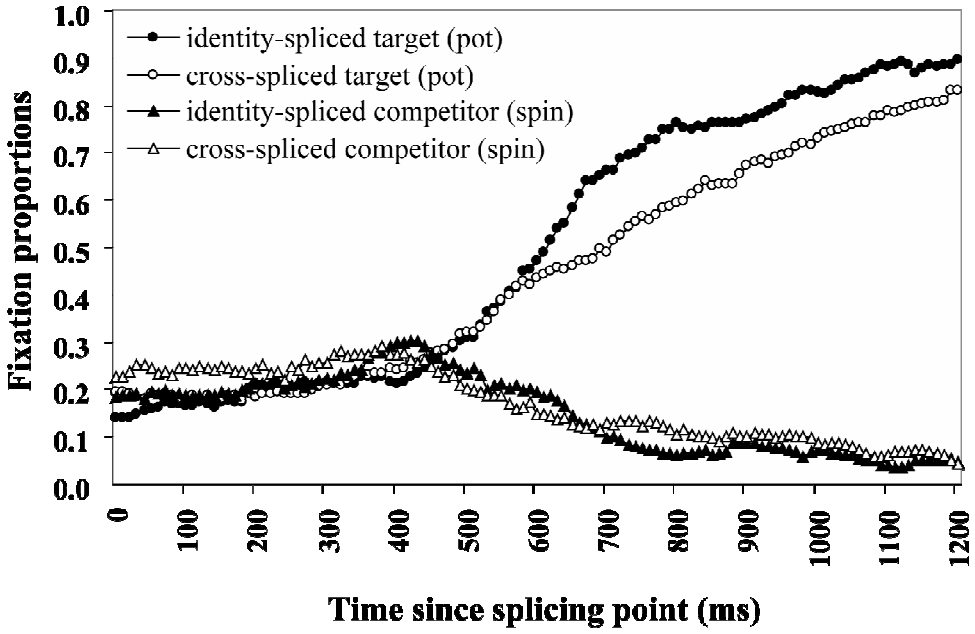


Figure 3. *Fixation proportions to the target and competitor pictures in the identity-spliced and the cross-spliced conditions in Experiment 1.*

Fixation proportions to the target pictures also began to increase around 300 ms, in both conditions, and rose above fixation proportions to the competitor at around 450 ms. The fixation proportions in the two conditions increased initially with a similar slope, but at

[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 ms, it is not uncommon to find values closer to 300 ms. A reviewer suggested that analysis based on fixation times (as in the current 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, re-analyze 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 ms after the splicing point (i.e., the average duration of saccades preceding fixations was 50 ms). This estimate is still longer than the 200 ms reported in some studies. Note, however, that Altmann and Kamide (2004) have argued that the estimation of 200 ms 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.

around 600 ms the proportion 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 1200 ms. 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 analysis of variance (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 = 0.38; F_2 (1,19) = 6.95, p < .05, eta^2 = 0.27). Additionally, fixations to the competitor and distractor pictures were compared over the time window extending from 300 to 600 ms. 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 = 0.36; F_2 (1,19) = 19.04, p < .001, eta^2 = 0.5), but there was no effect of splicing (average fixation proportions: .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 which listeners were sensitive to, resulting in modulation of their lexical interpretation. To examine these fine-

[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.

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: (a) duration of the [s] in the Target word context was shorter than that in the Cluster-initial word context; (b) Closure duration was longer in the Target context than in the Cluster-initial context; (c) duration of the target words (measured from after the release of the stop) was longer than duration of the cluster-initial words; (d) RMS energy of [s] in the Target context was lower than in the Cluster-initial context and (e) RMS energy of the stop in the target words was lower than in the cluster-initial words.

Table 2. *Mean segmental duration (in ms), RMS energy (in dB), spectral centre of gravity SCG; (in Hz) an standard deviations (SDs) 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	.14	3.37	.11	6.85	< .05
stop	3.11	.16	3.21	.13	4.1	= .057
SCG						
[s]	5322	372	5458	392	1.73	n.s.
stop	1231	995	1487	1207	3.23	n.s.

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 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 1200 ms).

Table 3. *Correlation of the difference in the acoustic measurements with the perceptual effect.*

Measurement	*r*(20)	*r*(19)*
Duration of [s]	.263	.60
Closure duration	.184	.142
Word Duration (excluding stop)	.243	.148
RMS energy of [s]	.381	.261
RMS energy of stop	.176	.237

* Outlier excluded.

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 ($r(19) = 0.60$, $p < 0.01$). 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 $r^2 = 0.32$). Thus, the data suggest that, although the ambiguous sequences differed on several measurements, listeners used only the duration of the [s] as a cue for the word boundary.

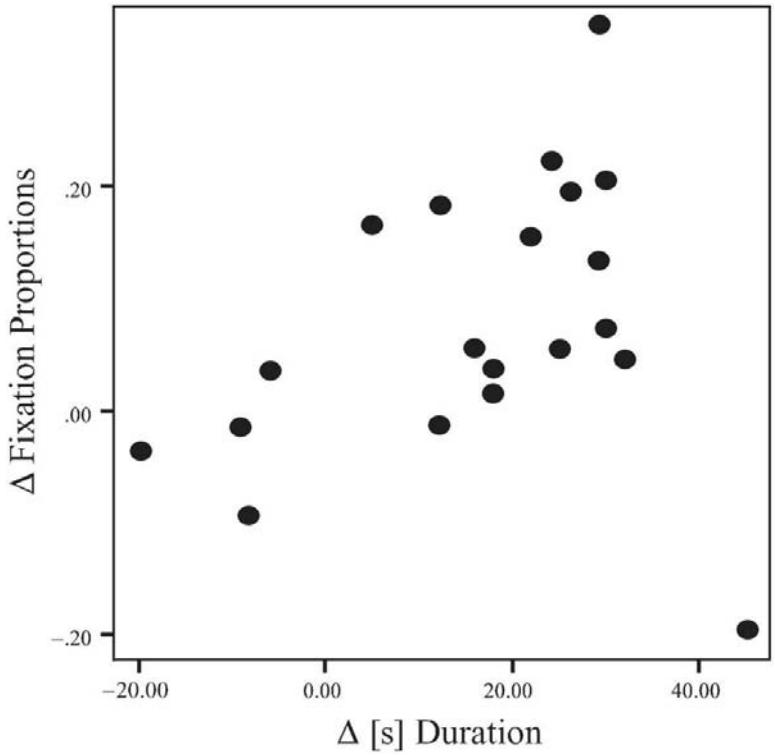


Figure 4. *Scatter plot of the difference between the identity-spliced and cross-spliced conditions of Experiment 1 in fixation proportions to the target against the difference between the conditions in the duration of the [s].*

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 ms after the splicing point. Indeed, statistical analyses across the 300-1200 ms time frame in 100 ms 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 600-700 ms time bin ($\underline{F_1}$ (1,23) = 12.78, $\underline{p} < .01$; $\underline{F_2}$ (1,19) = 6.93, $\underline{p} < .05$). If one assumes a 200-ms delay in programming and launching an eye movement, this would mean that the difference started to appear after 400 ms of the post-splice portion of the ambiguous sequence had been heard. Given that the spliced portion of the stimulus was, on average, 380 ms 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 that either 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. This issue was examined 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 non-specific 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 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 one standard deviation below the mean of the word-final distribution (short version) or one standard deviation 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 ms (SD=15) when it was in word-final position, and 107 ms (SD=14) when it was in word-initial position. Based on these numbers, for the version with the short [s] duration, the duration of the [s] was selected to be approximately one standard deviation lower than the mean duration of the [s] in word-final position, resulting in a value of 70 ms. For the long version, the duration of the [s] was approximately one standard deviation higher than the mean duration of the [s] in word-initial position, resulting in a value of 121 ms. 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 ms 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 ms long or 80 ms long, resulting in fricatives that had duration of, respectively, 70 ms (short version) or 120 ms (long version). Care was taken to avoid any acoustic artifacts, such as clicks or other distortions.

Procedure and Design. Procedure and design were identical to Experiment 1.

Results and Discussion

On two trials, 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 ms after the splicing point. At the same time, fixations to the target picture were also rising. At around 450 ms, fixations to the target rose above those to the competitor. From that point on, the probability of fixating the competitor started to drop and around 600 ms it was indistinguishable from the probability of fixating a distractor.

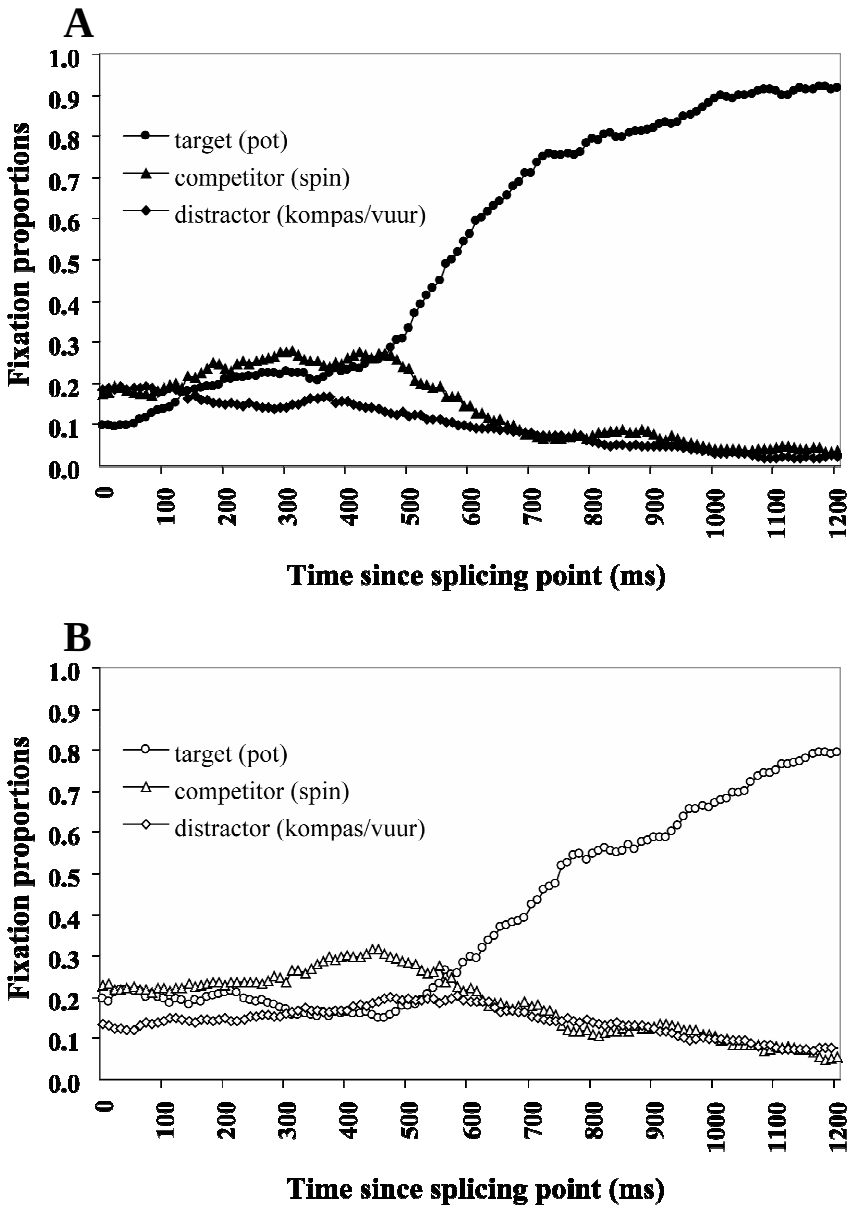


Figure 5. *Fixation proportions over time to the target, the competitor, and averaged distractors, in the short-version condition (A) and the long-version condition (B), in Experiment 2.*

Figure 5B shows a somewhat different pattern of fixation proportions in the long-version condition. Fixations to the competitor gradually rose and reached a peak at about 450 ms after the splicing point, while the probability of fixating the target picture remained indistinguishable from the probability of fixating the unrelated distractors. At around 450 ms, fixations to the target picture started to increase while fixations to the competitor were decreasing, until around 600 ms, where they merged again with the fixation proportions of the unrelated distractors.

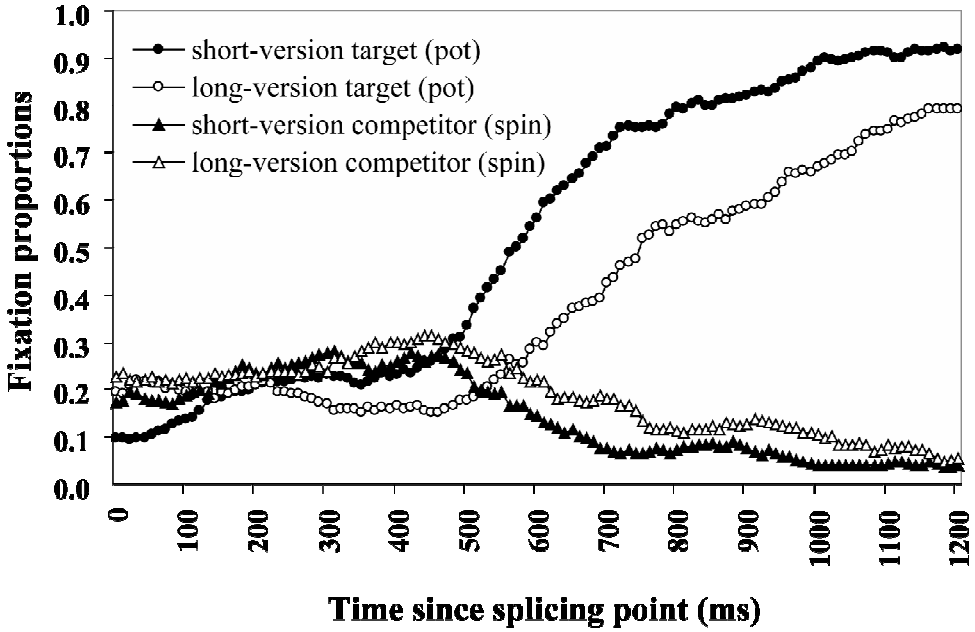


Figure 6. *Fixation proportions to the target and competitor pictures in the short-version and the long-version conditions in Experiment 2.*

Figure 6 presents the proportion 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 ms after the splicing point, participants tended to fixate the target picture less when they heard the long [s] version of the sentence. Over the 300-1200 ms 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 ($\underline{F_1}$ (1,23) = 91.55, $\underline{p} < .001$, $eta^2 = 0.80$; $\underline{F_2}$ (1,19) = 44.17, $\underline{p} < .001$, $eta^2 = 0.70$). Given that the duration of the [s] was longer in one condition, it could be argued that the effect is partly due to the delay in the onset of the target in the signal. The data was 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 300-1200 ms time window, the average proportion of fixations to the target picture was .70 in the short-version condition and .54 in the long-version condition, yielding a significant effect ($\underline{F_1}$ (1,23) = 50.81, $\underline{p} < .001$, $eta^2 = 0.69$; $\underline{F_2}$ (1,19) = 23.68, $\underline{p} < .001$, $eta^2 = 0.55$).

Fixations to the competitor and distractor pictures were compared over the time window extending from 300 to 600 ms. 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, $p < .001$, $eta^2 = 0.42$; F_2 (1,19) = 19.21, $p < .001$, $eta^2 = 0.5$). 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: .23 and .18 in the short and long-version conditions, respectively; F_1 (1,23) = 5.89, $p < .05$, $eta^2 = 0.20$; F_2 (1,19) = 24.81, $p < .001$, $eta^2 = 0.57$). 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, participants looked more often at the competitors (compared to 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 (compared 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, 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 300-1200 ms time frame in 100 ms bins confirmed this difference in the timing of the effect. In the 300-400 ms 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 which was reliable in the participants analysis (F_1 (1,23) = 5.07, $p < .05$, $eta^2 = 0.18$), though not in the items analysis (F_2 (1,19) = 2.45, $p = .13$). In the 400-500 ms time bin, this difference was significant (F_1 (1,23) = 12.32, $p < .01$, $eta^2 = 0.35$; F_2 (1,19) = 11.61, $p < .01$, $eta^2 = 0.38$). 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 300-1200 ms interval was then conducted to compare directly the results of the two experiments. Experiment was treated as a between-subjects 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, $p < .001$, $eta^2 = 0.66$; F_2 (1,19) = 28.28, $p < .001$, $eta^2 = 0.6$), no main effect of Experiment, and, critically, a significant interaction between Condition and Experiment (F_1 (1,46) = 17.21, $p < .001$, $eta^2 = 0.27$; F_2 (1,19) = 14.20, $p < .01$, $eta^2 = 0.43$). 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 on-line 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 or a cluster-initial word (e.g., spot). 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, 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. 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 current 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 either replaced 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. 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 current 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 (Quené, 1992; Kemps, 2004) which have demonstrated in off-line tasks that listeners use phoneme duration to segment ambiguous sequences. Using the eye-tracking paradigm, the current study extends those findings by showing that listeners use phoneme duration in the on-line segmentation of ambiguous phrases. Furthermore, unlike previous studies of on-line 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 current study has shown that the perceptual effect correlated with the duration of the [s], and that manipulating the [s] duration alone can bias subjects' 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 current 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 current 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 study of Waals (1999), in which the duration of various consonants in Dutch was measured, in various word positions. In Waals' study, the duration of [s] in a word-initial cluster (followed by a [t] or a [p]) was 107 ms, while a word-final [s] (in a cluster, preceded by [n], as in the word eens) was on average 76 ms. Hence, the duration of word-initial [s] in Waals' study was virtually identical to that of the speaker used for the recording of our materials (108 ms). In our study, word-final [s] duration tended to be longer than in Waals' study (91 ms in the stimuli used in Experiment 1 and 87 ms 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 can not 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 can not predict with any confidence whether the current 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,b), 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), and 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 current 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, while 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 to say, 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, and 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 (1992) listeners appeared to use the length of an [s] as a cue to the voicing contrast between [s] and [z].

The results of the current 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 to say, 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 towards 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 (e.g., Allen &Miller, 2001). Duration, independent of speaking rate, can also be used in segmental interpretation (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 which 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.

The experiments reported in this article 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 listeners' 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.

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)

Segmenting ambiguous phrases using phoneme duration

CHAPTER 3

A slightly adapted version of this chapter was published in *Proceedings of the Eighth International Conference on Spoken Language Processing* (Shatzman, 2004)

Abstract

The results of an eye-tracking experiment are presented in which Dutch listeners' eye movements were monitored as they heard sentences and saw four pictured objects. Participants were instructed to click on the object mentioned in the sentence. In the critical sentences, a stop-initial target (e.g., "pot") was preceded by an [s], thus causing ambiguity regarding whether the sentence refers to a stop-initial or a cluster-initial word (e.g., "spot"). 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 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. Thus, in this context, the [s] duration information is an important factor guiding word recognition.

Introduction

In order to understand a spoken utterance, the continuous speech signal must be segmented into lexical units. Current models of spoken-word recognition such as TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994) and DCM (Gaskell & Marslen-Wilson, 1997) agree that, as speech unfolds over time, words that are fully or partially consistent with the input become activated and compete among one another. The outcome of the competition is that the input is parsed into a sequence of non-overlapping words.

The lexical competition process is determined, to a large extent, by the information in the acoustic signal. An important line of research has therefore focused on identifying acoustic cues that may mark word boundaries. One consistent finding in the acoustic-phonetic literature is that phoneme duration varies as a function of the position of the phoneme with respect to word boundaries. Thus, for example, phonemes in word-initial position tend to be longer than in word-medial or word-final position (e.g., Klatt, 1974; Oller, 1973; Umeda, 1977).

The influence of phoneme duration on segmentation has been demonstrated in a study with ambiguous two-word sequences, such as the Dutch phrases *diep in* (“deep in”) and *die pin* (“that pin”; Quené, 1992). Using a forced-choice task, it was shown 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 judgements.

Other studies have shown the influence of phoneme duration on segmentation using on-line measures. A study in French (Spinelli, McQueen & Cutler, 2003) examined segmentation in a liaison environment, that is, in phrases where resyllabification occurs across word boundaries. In contexts like *dernier oignon* (last onion), the final [ʁ] of *dernier* is produced and resyllabified with the following syllable, making the phrase homophonous with *dernier rognon* (last kidney). The results of the study suggest that French listeners’ segmentation of

an ambiguous liaison phrase (e.g., *dernier oignon / rognon)* is influenced by the length of the liaison consonant: the consonants in the liaison environments were shorter than the word-initial consonants (e.g., [ʁ] in *dernier oignon* vs. *rognon*). Similarly, in an English study (Gow & Gordon, 1995) evidence was found for priming of words in two-word sequences (e.g., of *lips* in *two lips*) but not when the words were pronounced as part of single-word sequences (e.g., *tulips*). Given the fact that the word-initial consonants (e.g., the [l] in *two lips*) were longer than the non-initial consonants (e.g., the [l] in *tulips*), the authors concluded that listeners were using this acoustic marker of word onset in lexical access and segmentation. However, neither study demonstrated that the duration of the critical consonants is what affected listeners' segmentation. In other words, the link between the acoustic cue of word-initial phoneme duration and the perceptual effect remains inferential. In addition, other cues to word boundaries were not examined.

The main aim of the present study was thus to explore the degree to which listeners use various acoustic cues to word boundaries in segmentation of continuous speech. To this end, ambiguous Dutch phrases were constructed, containing either a stop-initial word (e.g., the word *pot* in *ze heeft wel eens pot gezegd*, "she said once jar") or a cluster-initial word that matched the stop-initial word together with the preceding [s] phoneme (e.g., the word *spot* in *ze heeft wel een spot gezegd*, "she did say a spotlight"). The sentences were, thus, phonemically identical but differed in their precise acoustic-phonetic realization. The sentences were manipulated by cross-splicing, such that the initial stop of the target word and the preceding phoneme [s] (e.g., the [s] and the [p] in *ze heeft wel eens pot gezegd*) were either replaced by a cluster from the cluster-initial word, or by an initial stop and preceding [s] from another recording of the sentence. Acoustic measurements of the ambiguous sequences (e.g., the [sp] in *eens pot* vs. *een spot*) were performed to assess the differences between them. 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. The eye-tracking

paradigm was used to evaluate listeners' ability to distinguish between the two ambiguous sentences. In the eye-tracking paradigm, participants generally hear a sentence and are then shown four objects presented as pictures on a computer screen. Their task is to click on and move the object referred to in the sentence with the computer mouse. Of primary interest was whether participants' fixations to the target picture would differ across the splicing conditions. Subsequently, the acoustic information which participants might be using was determined by correlating their performance in the eye-tracking task with the differences found in the acoustic analyses.

Method

Participants. Twenty-four members of the Max Planck Institute subject panel, native speakers of Dutch, were paid to take part.

Materials. Twenty stop-initial Dutch nouns referring to picturable objects (e.g., *pot*) were selected, such that the addition of an initial [s] to each word would result in an existing Dutch noun. For example, the addition of an [s] to the Dutch word *pot* results in the word *spot*. Note that the cluster-initial counterpart words were not necessarily picturable nouns. Each target was paired with a cluster-initial picturable noun (the competitor) which overlapped with the first two phonemes of the target's cluster-initial counterpart. For example, for the target *pot,* the competitor *spin* (spider) was selected, overlapping with the first two phonemes of *spot*. Two semantically and phonologically unrelated distractors were assigned to each target and competitor pair (e.g., *vuur* [fire] and *kompas* [compass]). Line-drawing pictures associated with the items were selected from various picture databases. In addition to the 20 experimental item sets, 50 filler sets were constructed. Pictures for the filler trials were selected from the same databases as were used for the experimental trials.

For each experimental item, two recording contexts were constructed. In one of the contexts the target word was mentioned and in the other the target's cluster-initial counterpart was mentioned. The recording contexts were constructed such that the sequences containing the target or its counterpart were identical and, therefore, fully ambiguous. For example, the sentence *ze heeft wel eens pot gezegd* is phonemically identical to the sentence *ze heeft wel een spot gezegd*.

All sentences were read aloud in random order by a female speaker of Dutch in a sound-attenuated booth and recorded on a DAT tape (sampling at 48 kHz with 16-bit resolution). Each sentence was recorded at least four times. The sentences were then re-digitized at a sample rate of 16kHz and edited using Xwaves speech-editor software. For each target word, two spliced versions of the sentence were created. The carrier phrase for both versions consisted of the initial portion of the target sentence (up to the [s], e.g., *ze heeft wel een*) taken from the target recording context, and the final portion of that context (e.g., *gezegd*). For one version (hereafter, the identity-spliced version) the stop (e.g., the [p] in *pot*) and the preceding [s] were taken from another token of the target recording context and spliced onto the carrier phrase. In the other version (hereafter, the cross-spliced version) the stop and the preceding [s] originated from the cluster-initial recording context (e.g., the [sp] from *spot*) (see Table 1). The cross-spliced sentences were thus lexically identical to the identity-spliced sentences, but differed in the origin of the [s] and the following stop (i.e., whether this sequence was taken from the target or the cluster-initial recording context). Cross-spliced sentences were constructed for 19 filler items. Three of the fillers items proved to be problematic and had to be excluded from the experiment. All splicing points were taken at zero-crossings and the splicing manipulation was done very carefully so as to prevent any acoustic artifacts, such as clicks or other distortions.

Table 1. *Stimulus example of the conditions in the experiment.*

Origin of Recording	Spliced version
Identity-spliced condition	
1a. Ze heeft wel eens pot gezegd	Ze heeft wel een**s** **p**ot gezegd
1b. **Ze heeft wel eens pot gezegd**	
Cross-spliced condition	
1a. Ze heeft wel eens pot gezegd	Ze heeft wel een**s** **p**ot gezegd
2. **Ze heeft wel een spot gezegd**	

Acoustic analyses. Acoustic measurements of the spliced portions of the stimuli (e.g., the [sp] in *eens pot* or in *een spot*) were carried out to evaluate the extent to which their acoustic realization was influenced by the intended meaning. The following measurements were taken: [s] duration, stop closure duration, and Voice Onset Time. Duration measurements were taken from the spectrograms and the waveforms combined, using Xwaves. In addition, RMS energy and Spectral Centre 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 was measured using the built-in function in Praat speech-editor software, which calculates the average frequency from an FFT spectrum over a frequency range from 0 to 10000 Hz. To measure the SCG of [s], the segment was divided into 15 ms intervals, an FFT spectrum was made for each interval (filtering out the frequency range below 1000 Hz to remove any spurious low frequency components) and the SCG of each interval was taken. The maximal SCG was taken as the SCG for the segment.

Procedure and design. Participants were tested individually. They were first familiarized with the 268 pictures. The pictures appeared on a computer screen, one at a time, along with their printed name, and participants pressed a response button to proceed to the next picture. After this part of the experiment, the eye-tracking system was set up.

Participants were seated at a comfortable distance from the computer screen. Eye movements were monitored using a SMI EyeLink head-mounted eye-tracking system, sampling at 250 Hz. Pictures were presented on a ViewSonic 17PS screen, and the auditory stimuli were presented over headphones using NESU software. Both eyes were monitored, but only the data from the right eye were analyzed.

The structure of each trial was as follows. First, a central fixation dot appeared on the screen for 500 ms. After that, a spoken sentence was presented to the participants and simultaneously a 5x5 grid with pictures appeared on the screen (see Figure 1). Prior to the experiment, participants 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 across four fixed positions of the grid while the geometric shapes appeared in fixed positions on every trial. Once the picture had been moved, the experimenter pressed a button to initiate the next trial.

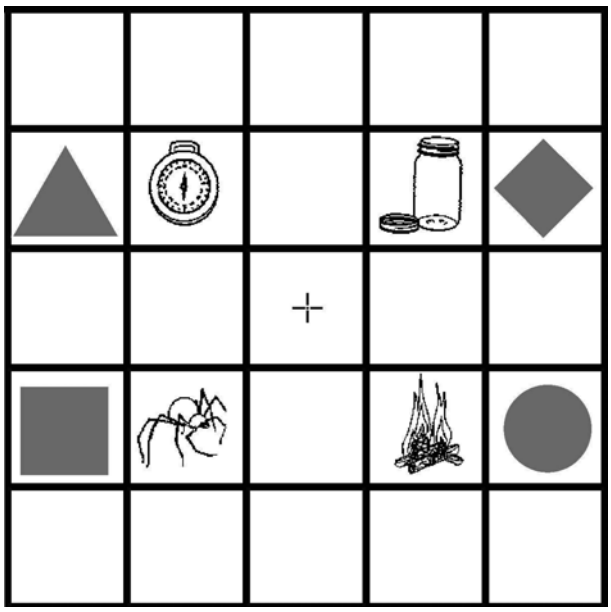


Figure 1. *Example of stimulus display presented to participants.*

Two lists were created, containing the filler and the experimental items. The lists varied on which of the two sentences (i.e., the identity-spliced or the cross-spliced sentence) was

presented for each of the experimental items. Within each list, 10 experimental items were assigned to each condition. Twelve random orders were created for the lists, with the constraints that there was always at least one filler item between two experimental items. Participants were randomly assigned to one list.

Results

Graphical software was used to display the locations of the participants' fixations as dots superimposed on the four line drawings for each trial and each participant. The timing of the fixations was established relative to the onset of the [s] preceding the target word. Fixations on the line drawings were coded as pertaining to the target object, the competitor, or one of the two unrelated distractors, or to anywhere else on the screen. For each trial, fixations were coded from the onset of the [s] until the subject had clicked with the mouse cursor on the target picture. Four trials had to be removed from the analysis, because participants clicked on an object other than the target object. The proportions of the fixations were analyzed in 10 ms slices to provide fine-grained information about the time course of lexical activation as the speech unfolded.

Figure 2 presents the proportions of fixations to the target, competitor and distractor pictures, averaged over participants, in the identity-spliced condition and the cross-spliced condition. Fixation proportions to the two unrelated distractors were averaged. Fixation proportions are shown from the splice point (the onset of the [s] preceding the target word) to 1200 ms thereafter.

As is apparent from Figure 2, starting around 350 ms, fixation proportions in the identity-spliced condition rose faster and remained higher than those in the cross-spliced condition. The difference between conditions was statistically tested over a time window extending from 350 to 1200 ms. Over this time interval, the average fixation proportion to the target picture was .61 in the identity-spliced condition and .53 in the cross-spliced condition

(F_1 (1,23) = 17.32, $p < .001$; F_2 (1,19) = 7.78, $p < .05$). This demonstrates that the spliced sequences (i.e., the [s] and the following stop) contained fine-grained differences, modulating listeners' lexical interpretation.

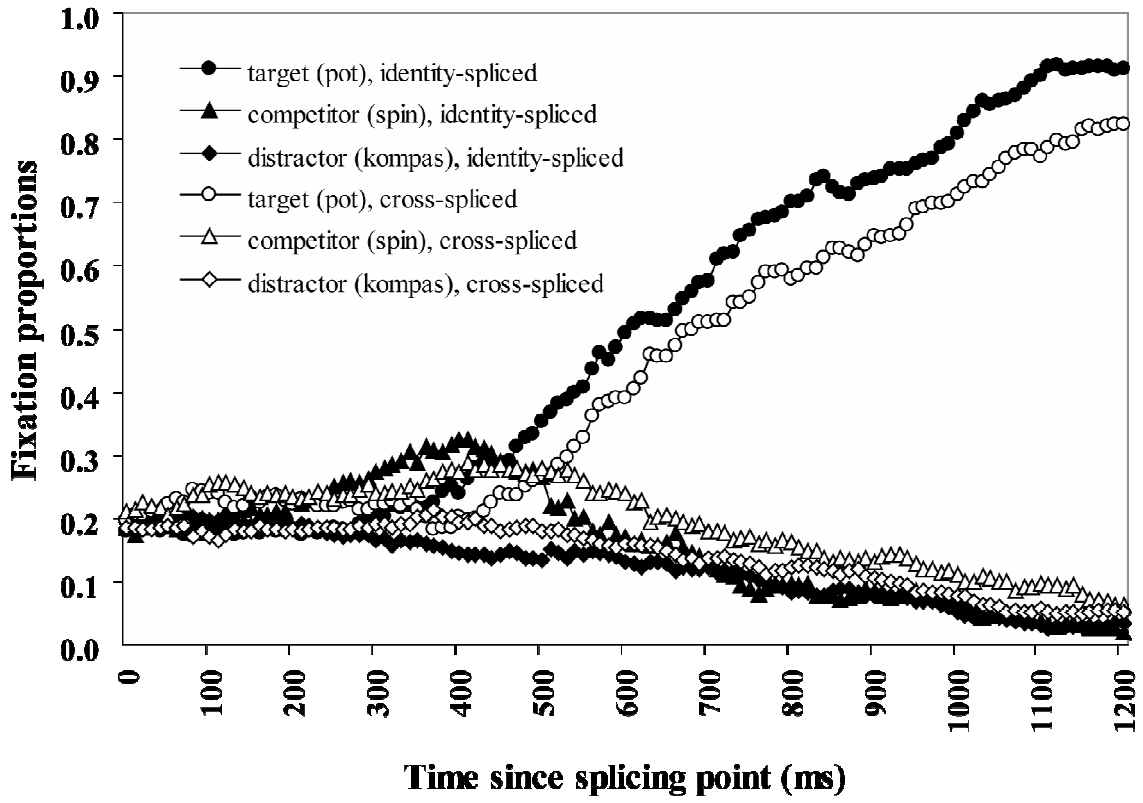


Figure 2. *Fixation proportions over time for identity-spliced and cross-spliced targets, averaged over participants.*

To examine these fine-grained differences, acoustic analyses were conducted on the spliced portions in both versions. The results of the acoustic measurements are displayed in Table 2. The results of one-way analyses of variance (ANOVAs) performed on these data are presented in the same table.

The analyses revealed significant differences between the two versions on the following measures: (a) the duration of the [s] in the Target context was significantly shorter than that in the Cluster-initial context; (b) Closure duration was longer in the Target context than in the Cluster-initial context; (c) RMS energy of [s] in the Target context was lower than in the

Cluster-initial context and (d) RMS energy of the [t] in the target words was lower than in the cluster-initial words (see Table 2)

Table 2. *Mean segmental duration (ms), RMS energy (dB), spectral centre of gravity (SCG; Hz) and standard deviations of the spliced portions of the experimental stimuli.*

	Target Context		Cluster Context			
	eens pot		*een spot*		ANOVA	
Duration	*M*	*SD*	*M*	*SD*	*F*(1,19)	*p*
[s]	87	12	108	14	24.35	< 0.001
Closure	90	23	59	22	41.20	< 0.001
Voice Onset Time	23	7	21	7	1.31	n.s.
RMS Energy						
[s]	3.29	0.10	3.37	0.11	4.62	< 0.05
stop	3.10	0.13	3.20	0.14	6.36	< 0.05
SCG						
[s]	5328	316	5458	392	1.33	n.s.
stop	1572	1252	1409	1159	< 1	n.s.

The acoustic differences between the Target and Cluster-initial sequences are effective cues only to the extent that listeners can perceive these differences and use them in segmentation and word activation. Therefore, for the acoustic measurements for which a significant difference was found, the difference in the measurements for each item was correlated with the perceptual effect for that item (i.e., the difference in the average fixation proportions to the item between the identity-spliced and the cross-spliced conditions in the time window extending from 350 to 1200 ms). These correlations are displayed in Table 3.

Table 3. *Correlation of the difference in the acoustic measurements with the perceptual effect.*

Measurement	Correlation with perceptual effect
Duration of [s]	*r(20)* = .454*
Closure duration	*r(20)* = .285
RMS energy of [s]	*r(20)* = -.119
RMS energy of stop	*r(20)* = .051

*NOTE: * = p < 0.05*

The correlational analysis showed that out of all the measurements for which a significant difference was found between the two versions, only the duration of the [s] significantly correlated with the perceptual effect. Thus, the data suggest that listeners were using the duration of the [s] as a word boundary cue.

Discussion

The main finding of this experiment is that listeners can use phoneme duration as a signal to the location of word boundaries. In the materials used in the experiment, the duration of the [s] was biasing listeners' lexical interpretation of the ambiguous sequence. Other measurements that differentiated the acoustic-phonetic realization of the ambiguous sequences did not correlate with the perceptual data. This suggests that although more cues were available, listeners were attending only to the duration of the [s]. It is however also possible that it is not duration per se that is used by the speech recognition system, but rather some other factor which has not been measured and is highly correlated with duration.

In another eye-tracking experiment (Shatzman & McQueen, 2006) the influence of phoneme duration on segmentation has been demonstrated by manipulating the duration of the [s] in the same ambiguous sentences as were used here. The results showed that listeners

were slower to identify the stop-initial target object when the duration of the [s] in the spoken signal was lengthened.

These findings add to a growing body of research (e.g., Gow & Gordon, 1995; Quené, 1992; Spinelli et al., 2003) showing that fine-grained information in the speech signal can modulate lexical activation. This poses a challenge to current models of spoken-word recognition, none of which take into account the contribution of this type of information.

The modulation of lexical competition by segment duration

CHAPTER 4

Keren B. Shatzman and James M. McQueen (submitted), *Psychonomic Bulletin & Review*

Abstract

An eye-tracking study examined how fine-grained phonetic detail, such as segment duration, influences the lexical competition process during spoken-word recognition. Dutch listeners' eye movements to pictures of four objects were monitored as they heard sentences in which a stop-initial target word (e.g., pijp, [pipe]) was preceded by an [s]. Participants made more fixations to pictures of cluster-initial words (e.g., spijker, [nail]) when they heard a long [s] (mean duration 103 ms), compared to when they heard a short [s] (mean duration 73 ms). Conversely, participants made more fixations to pictures of the stop-initial words when they heard a short [s], compared to when they heard a long [s]. Lexical competition between stop- and cluster-initial words is therefore modulated by segment duration differences of only 30 ms.

Introduction

One of the major objectives of current psycholinguistic research is to unravel the processes involved in understanding spoken language. More specifically, how do people recognize words from the speech they hear? Speech is a continuous signal; explicit physical cues to word boundaries are not always available. Nevertheless, even in the absence of explicit word-boundary markers, listeners will rapidly identify the discrete words in the speech stream.

Current models of spoken-word recognition seek to explain this behavior. An established finding is that as the speech signal unfolds over time, words that are fully or partially consistent with the available acoustic-phonetic information become activated and compete among one another (see McQueen, 2005, for review). The activation of a given word is thus determined by both its goodness of fit with the input and the activation of other competitors. The outcome of the competition is a parse of the spoken utterance in which each speech sound is attributed to only one word, yielding a sequence of non-overlapping words. The aim of the present study was to examine whether fine-grained phonetic detail in the speech signal can modulate this lexical competition process.

A growing body of evidence suggests that even very subtle acoustic information can have an impact on lexical activation levels (see McQueen, 2005, for review). One type of such fine-grained information is segment duration. This has been shown by both off- and on-line measures to influence lexical interpretation. For example, listeners' off-line segmentation judgements of ambiguous sequences are influenced by individual segment duration (Quené, 1992; Kemps, 2004). Using on-line priming measures, Gow and Gordon (1995) observed evidence for the activation of both tulips and lips when listeners heard two lips, but no evidence for the activation of lips when listeners heard tulips. The word-initial consonants (e.g., the [l] in two lips) had longer durations than the non-initial consonants (e.g., the [l] in tulips); the authors concluded therefore that segment duration was guiding listeners' segmentation. Similarly, Spinelli, McQueen and Cutler (2003) have shown in a cross-modal

priming task that, even though French sequences such as dernier oignon (last onion) and dernier rognon (last kidney) are phonemically identical, French listeners appear to segment such ambiguous phrases correctly, that is, as intended by the speaker. The consonants in liaison environments (e.g., [ʁ] in dernier oignon) were shorter than genuine word-initial consonants (e.g., [ʁ] in dernier rognon), suggesting once more that fine-grained acoustic details bias the lexical competition in the correct direction.

These studies did not in fact show that individual segment duration and not some other acoustic information influenced ambiguity resolution. A recent eye-tracking study (Shatzman & McQueen, 2006) found specific effects of segment duration, however. Dutch listeners' eye movements were monitored as they heard sentences and saw four pictured objects. Participants were instructed to click on the object mentioned in the sentence. In the critical sentences, a stop-initial target (e.g., pot, jar) was preceded by an [s], thus causing ambiguity regarding whether the sentence referred to a stop-initial or a cluster-initial word (e.g., spot, spotlight). In these trials, the visual display contained, in addition to the target object, a cluster-initial object, which overlapped with the first two phonemes of the target's cluster-initial counterpart (e.g., spin, spider). 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 target's cluster-initial counterpart than when they were spliced from another token of the target-bearing sentence. Acoustic analyses revealed several differences between the two recordings, but only [s] duration correlated with listeners' fixations (more target fixations for shorter [s]'s). In a second experiment, participants made more fixations to target pictures when the [s] was shortened than when it was lengthened. However, the long [s] did not elicit more cluster-initial word (e.g., spin) interpretations: Although participants in the long [s] condition looked more often at the cluster-initial objects (compared to the short [s] condition), they also looked more often at the distractors. A long [s] before a stop was thus a poorer match for the stop-initial target, but not a better match for the cluster-initial competitor. Because the effect was

only observed with the stop-initial words, these data do not show that segment duration directly influences the competition between stop-initial and cluster-initial words.

In the present study, a similar design was used as in Shatzman and McQueen (2006), except that the critical stimuli were no longer fully ambiguous (e.g., eens pot/een spot) and that the overlap of the cluster-initial object with the signal included the vowel following the cluster. For example, the Dutch word pijp (pipe) was preceded by an [s], so that it was temporarily congruent with the cluster-initial word spijker (nail). The duration of the [s] was manipulated such that it was either short (one standard deviation shorter than the mean duration of [s] in, e.g., eens pijp [once pipe]) or long (one standard deviation longer than the mean natural duration of [s] in, e.g., een spijker [one nail]). If [s] duration modulates the lexical competition process, the proportion of looks to spijker (but not to the distractors) should increase, and looks to pijp should decrease, for the longer relative to the shorter [s]. Additionally, we examined the influence of segment duration on the time course of word recognition.

Method

Participants. Thirty Max-Planck-Institute subject pool volunteers, all Dutch native speakers, were paid for their participation.

Materials. Twenty-six stop-initial Dutch nouns referring to picturable objects (e.g., pijp [pipe]) were selected as targets. Each target was paired with a cluster-initial picturable noun (henceforth, the competitor). The competitor always started with an [s], and the following stop and vowel overlapped with the target word's onset (e.g., spijker [nail]). Two additional picturable nouns that were phonologically unrelated to the target and competitor were assigned to each target/competitor pair. There were no semantic or morphological

relationships between the words within each quadruple. The full set of items is presented in the Appendix.

Recording contexts were constructed such that the target was always preceded by an [s] (always in the word eens), and the sequences preceding the target or the competitor were otherwise identical (e.g., ik zou ooit eens pijp willen roken, “I would like to smoke a pipe some time” and ik zou ooit een spijker willen kopen, “I would like to buy a nail”). All sentences were produced by a female native Dutch speaker in a sound-attenuated booth and recorded directly onto computer (sampling at 44.1 kHz with 16-bit resolution). Acoustic measurements showed that the average duration of the [s] was 88 ms (SD=15) when it was in word-final position (target context), and 95 ms (SD=9) when it was in word-initial position (competitor context). These values are comparable, albeit with a somewhat smaller difference, to those reported in Shatzman and McQueen (2006), in which a different speaker was recorded. In that study the average duration of the [s] was 87 ms and 108 ms in word-final and word-initial position, respectively. In the present study, stimuli were created by manipulating the duration of the [s] in the target sentences. For each sentence, two spliced versions were created: In the short-[s] version, the duration of the [s] was approximately one standard deviation lower than its average duration in word-final position, while in the long-[s] version, [s] duration was one standard deviation higher than its average duration in word-initial position.

The stimuli were edited using Xwaves speech-editing software. In each sentence, the steady-state phase of the fricative was excised, leaving approximately 20 ms 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 ms long or 60 ms long, resulting in fricatives that had average durations of, respectively, 73 ms (short version) or 103 ms (long version). Average duration of the ambiguous sequence (i.e., the [s], and the following stop

and vowel) was 260 ms in the short-[s] condition and 289 ms in the long-[s] condition. Average target duration was 360ms.

Forty-four filler trials were constructed, in which the target was phonologically unrelated to all three distractors. Sentences mentioning the filler targets were produced by the same speaker, and recorded at the same time as the experimental sentences. Line-drawing pictures associated with the experimental and filler items were selected from various picture databases[1].

Procedure and Design. Participants were tested individually. They were first familiarized with the 280 pictures to ensure that they identified them as intended. The pictures appeared on a computer screen in a randomized order, one at a time, along with their printed name. Participants pressed a response button to proceed to the next picture. After familiarization, the eye-tracker (an SMI Eyelink system, sampling at 250 Hz) was mounted and calibrated. The experiment was controlled by a Compaq 486 computer. Pictures were presented on a ViewSonic 17PS screen, and the auditory stimuli were presented over headphones using NESU software (http://www.mpi.nl/world/tg/experiments/nesu.html).

Each trial was structured as follows. A central fixation dot appeared on the screen for 500 ms. Then a spoken sentence was presented and simultaneously a 5x5 grid with pictures appeared on the screen (see Figure 1). Prior to the experiment, participants were instructed to use the computer mouse to move the object mentioned in the spoken sentence above or below the geometrical shape adjacent to it. Once the picture had been moved, the experimenter pressed a button to initiate the next trial.

Two lists were created, each containing 26 experimental and 44 filler trials. The lists varied in which of the two versions (i.e., the short-[s] or the long-[s] version) was presented for each of the experimental trials. Within each list, 13 experimental trials were assigned to each

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

condition. Fifteen random orders were created for the lists. There was always at least one filler trial between two experimental trials. Five filler trials were presented at the beginning of the experiment to familiarize participants with the task. Participants were randomly assigned to one list.

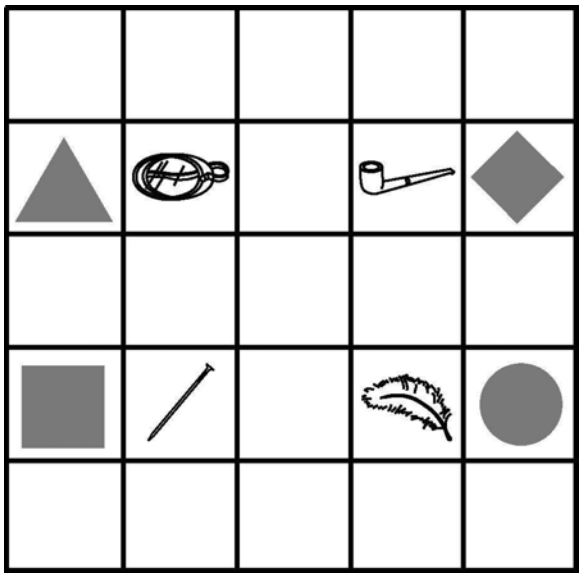


Figure 1. *Example of stimulus display presented to participants. Clockwise from top left corner: duikbril [diving mask], pijp [pipe], veer [feather] and spijker [nail].*

Results

Using graphical software, the locations of the participants' fixations were displayed as dots superimposed on the four line drawings for each trial and each participant. The timing of the fixations was established relative to the acoustic onset of the [s] preceding the target word. Fixations on the line drawings were coded as pertaining to the target, the competitor, one of the two distractors, or to anywhere else on the screen. For each trial, fixations were coded from the onset of the [s] until the subject had clicked on the target picture. Three trials were removed from the analysis because participants erroneously selected a non-target picture. The proportion of fixations to each picture type were computed by summing trials in which each

type was fixated (in each 10 ms slice, in each condition), and dividing the sum by the total number of trials (in the same time interval) in which any picture or location was fixated.

Figure 2 presents the average proportion of fixations to the target, competitor and distractor pictures, averaged over participants. Fixation proportions to the two distractors were averaged. Fixation proportions are shown in 20 ms time slices from the onset of the [s] preceding the target word to 1200 ms thereafter.

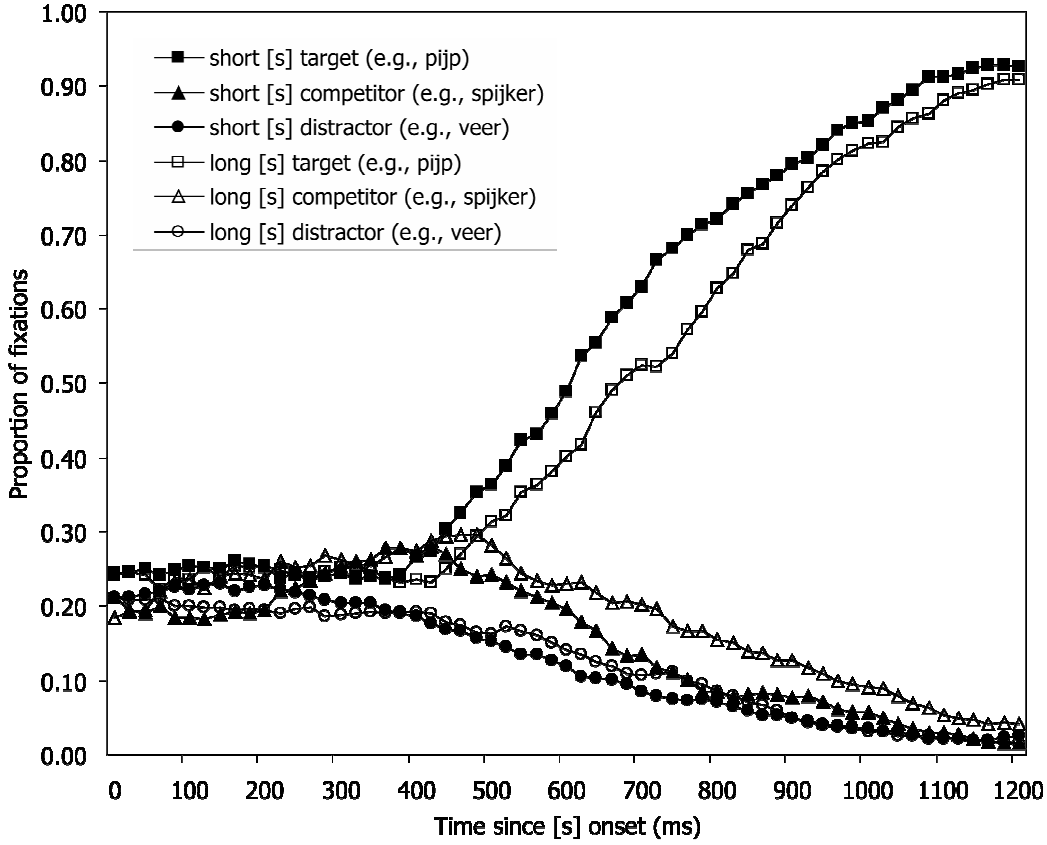


Figure 2. *Fixation proportions over time to the target, the competitor, and averaged distractors, in the short-[s] condition and the long-[s].*

In both conditions, fixation proportions to the target and competitor pictures began to differ from those to the averaged distractors at around 250 ms after the onset of the [s]. The proportion of fixations to the target began to increase rapidly at 400 ms in the short-[s] condition and 450 ms in the long-[s] condition. Fixation proportions to the competitor rose

until 400 ms in the short-[s] condition and 500 ms in the long-[s] condition, and then started decreasing, with a shallower slope in the long-[s] condition.

Fixation proportions to each picture type were calculated over a time window extending from 200 to 1100 ms after the onset of the [s]. Fixation proportions start reflecting significant events in the speech stream after approximately 200 ms (Saslow, 1967). The duration of the window corresponds to the time interval during which fixation proportions to the competitor were higher than fixation proportions to the distractors. Analyses of variance (ANOVAs) on the fixation proportions to the targets were computed by subjects (F_1) and by items (F_2). The average fixation proportion to the target picture was 54% in the short-[s] condition and 49% in the long-[s] condition (F_1 (1,29) = 12.05, $p < .005$; F_2 (1,25) = 7.63, $p < .05$). To compare fixation proportions to the competitors and distractors without violating the assumption of independence between observations, the mean difference between fixation proportions to the competitor and distractor pictures – averaged over participants or items – was computed and compared to zero in one-sample t tests. Over the 200–1100 interval, participants looked more at the competitors than at the distractors, both in the short [s] condition (mean difference: 4%, $t_1(29) = 3.72$, $p < .005$; $t_2(25) = 3.06$, $p < .01$) and the long [s] condition (mean difference: 8%, $t_1(29) = 7.04$, $p < .001$; $t_2(25) = 3.93$, $p < .005$). ANOVAs on the mean difference values showed that the difference between the conditions was significant only by participants (F_1 (1,29) = 5.1, $p < .05$; F_2 (1,25) = 2.69, $p = .11$).

To study the time course of the influence of segment duration on the competitor's activation, fixation proportions were examined in the interval in which fixations reflect the processing of the target word and in the interval after word offset. The average offset of the target word was roughly 450 ms (433 and 463 ms in the short-[s] and long-[s] conditions, respectively). Allowing 200 ms for saccadic latency, the analyses were therefore performed in the 200–650 and 650–1100 ms intervals. As the target word unfolded (interval 200–650) participants fixated the competitor more than the distractor in both conditions (6% in the

short-[s] condition, $t_1(29) = 2.23$, $p < .05$; $t_2(25) = 2.68$, $p < .05$, and 8% in the long-[s] condition, $t_1(29) = 4.39$, $p < .001$; $t_2(25) = 3.07$, $p < .01$), but the difference between the conditions was not significant (Fs < 1). In the period after word offset (650–1100 ms), the competitor was fixated more than the distractor in the long-[s] condition (mean difference 7%, $t_1(29) = 5.00$, $p < .001$; $t_2(25) = 2.61$, $p < .05$), but only slightly so in the short-[s] condition (mean difference 2%, $t_1(29) = 2.23$, $p < .05$; $t_2(25) = 1.59$, $p = .13$), yielding a significant difference between the conditions (F_1 (1,29) = 7.67, $p < .05$; F_2 (1,25) = 4.91, $p < .05$).

These analyses confirm the impression given by Figure 2 that the activation of the competitor is longer-lasting in the long-[s] condition than in the short-[s] condition. As [s] duration was 30 ms longer in the long-[s] condition, it could be argued that the differences between the conditions are due to the delay in the onset of the target in that condition. However, a reanalysis of the data corrected for this durational difference did not change the pattern of results.

Discussion

This study demonstrates that fine-grained acoustic detail, such as segment duration, differentially favors lexical candidates, thereby biasing the competition process. Participants' eye-movements to four displayed objects were tracked as they listened to sentences in which stop-initial target words (e.g., pijp, [pipe]) were preceded by an [s]. Participants made more fixations to pictures of cluster-initial words, which partially overlapped with the signal (e.g., spijker, [nail]), when they heard a long [s], compared to when they heard a short [s]. Conversely, participants made more fixations to pictures of the stop-initial words when they heard a short [s], compared to when they heard a long [s].

The present study extends the findings of previous studies (Shatzman, 2004; Shatzman & McQueen, 2006) which indicated that the interpretation of a fully ambiguous sequence

involving an [s] followed by a stop can be influenced by the duration of the [s]. While those studies showed that the interpretation containing a stop-initial word is disfavored when the [s] is long, the current results also show that when the [s] is long the cluster-initial word interpretation is favored. Furthermore, by using temporarily ambiguous phrases, the current study revealed that the influence of segment duration was detectable long after disambiguating information has been heard. By modulating the competition process, segment duration winnows down the set of candidate words, thus affecting how and when the competition is resolved.

The fact that the effect of segment duration lasted for a relatively long time period can be attributed to two reasons. First, segment duration, being a temporal cue, is likely to be interpreted relative to other durational information in the signal. That is, segment duration is not evaluated in absolute terms, but in relation to both the preceding context and the unfolding signal. Consequently, the effects of a segment's duration may occur only after a sufficient amount of information has accrued for its relative duration to be evaluated. A second, not mutually exclusive, possibility is that the long-lasting effect is due to the dynamics of the lexical competition process. Before disambiguating information is heard, the signal is a better match for the cluster-initial words in the long-[s] than in the short-[s] condition. Lexical activation levels for the cluster-initial words are therefore likely to be higher in the long-[s] condition. Once disambiguating information is heard, its effect is immediate in that fixation proportions to the cluster-initial picture begin to drop in both conditions. However, it takes time for lexical activation levels to return back to baseline, and the higher the activation levels were, the more time is required (assuming that the decay rate is independent of activation levels). Consequently, the difference between the conditions is spread over time, resulting in a long-lasting effect.

One way to incorporate the accumulating evidence regarding listeners' sensitivity to fine-grained acoustic detail involves the notion of prosodic hierarchy: The view that spoken

utterances are hierarchically organized, with large prosodic constituents, or domains, consisting of smaller constituents (e.g., Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986; see Shattuck-Hufnagel & Turk, 1996, for a review). Acoustic-phonetic research has shown that initial segments and syllables in higher-level constituents are different, articulatorily and acoustically, from initial segments and syllables in lower domains. Consequently, within a prosodic domain, a domain-initial segment or syllable has different fine-grained phonetic properties from a domain-medial one (e.g., Fougeron, 2001; Fougeron & Keating, 1997; Turk and Shattuck-Hufnagel, 2000). For example, segments in word-initial position tend to be longer than in word-medial or word-final position (e.g., Klatt, 1974; Oller, 1973; Umeda, 1977). This finding is a result of the fact that a word-initial segment tends to be articulatorily stronger than a word-medial segment, a phenomenon known as domain-initial strengthening.

In a recent cross-modal priming experiment examining domain-initial strengthening (Cho, McQueen & Cox, in press), listeners heard sentences containing a temporary lexical ambiguity (e.g., the phrase bus tickets, containing the embedded word bust) and had to make lexical decision to visual targets. The onset of the phrase's second word (e.g., /tɪ/) was spliced either from word-initial position or from Intonational-Phrase-initial position. When the target was the first word in the phrase (e.g., bus), there was more priming when the second word came from an Intonational-Phrase-initial position than from a word-initial position, demonstrating that listeners may use the acoustic consequences of prosodic strengthening during word recognition. The current results could also be interpreted as indicating that listeners are sensitive to the acoustic correlates of domain-initial strengthening, and use these in the segmentation process: A longer [s] provides the listener with evidence that a word boundary is more likely to occur before that [s].

Recent research also suggests that listeners exploit other aspects of the prosodic structure of utterances in the on-line segmentation of continuous speech (e.g., Davis, Marslen-Wilson

& Gaskell, 2002; Salverda, Dahan & McQueen, 2003; Salverda, 2005). For example, in a French study (Christophe, Peperkamp, Pallier, Block & Mehler, 2004), listeners were presented with sentences containing a local lexical ambiguity, such as the phrase chat grincheux (grumpy cat) containing the word chagrin (sorrow). Listeners were delayed in recognizing the word chat in these sentences, compared to sentences in which there was no lexical ambiguity, but not if there was a phonological phrase boundary between the two words containing the ambiguity.

The principal finding of the present research is the direct evidence that segment duration modulates the lexical competition process. Furthermore, this effect is quite long-lasting: Fixations to the competitor remained higher in the long-[s] condition for a considerable amount of time. Thus, a 30 ms difference in segment duration resulted in an ongoing effect on lexical competition. The effect of prosodically-dependent acoustic detail is therefore congruent with the view that spoken-word recognition is cascaded: Information percolates continuously to higher-level representations and modulates the activation levels of those representations.

Appendix

Target	Competitor	Distractor	Distractor
paard (horse)	spaarvarken (piggy bank)	bezem (broom)	mand (basket)
pak (package)	spar (fir)	kasteel (castle)	tomaat (tomato)
paling (eel)	spatel (spatula)	auto (car)	zaag (saw)
peddel (paddle)	speld (pin)	bok (goat)	radijs (radish)
pen (pen)	specht (woodpecker)	varen (fern)	wekker (alarm)
perzik (peach)	sperzieboon (green bean)	trap (stairs)	vleermuis (bat)
piano (piano)	spion (spy)	gieter (watering can)	rok (skirt)
pijp (pipe)	spijker (nail)	duikbril (diving mask)	veer (feather)
pinguin (penguin)	spinneweb (spider web)	lamp (lamp)	raket (rocket)
pizza (pizza)	spier (muscle)	cactus (cactus)	bel (bell)
poes (cat)	spoel (spool)	tulp (tulip)	ei (egg)
poort (gate)	spook (ghost)	anker (anchor)	hooivork (hayfork)
python (python)	spiegel (mirror)	knuppel (bat)	emmer (bucket)
tafel (table)	stadion (stadium)	kikker (frog)	radio (radio)
tas (bag)	stang (rod)	ballon (balloon)	kam (comb)
teckel (sausage dog)	stempel (stamp)	fee (fairy)	piramide (pyramid)
tennisracket (tennis racket)	stekker (plug)	boterham (sandwich)	jurk (dress)
thermometer (thermometer)	ster (star)	pan (pot)	rits (zipper)
toekan (toucan)	stoel (chair)	penseel (paintbrush)	mijter (miter)
ton (barrel)	stofzuiger (hoover)	chocolade (chocolate)	bus (bus)

tor (beetle)	stoplicht (traffic light)	zadel (saddle)	wortel (carrot)
trein (train)	strijkplank (ironing board)	moer (nut)	fopspeen (pacifier)
trommel (drum)	stropdas (necktie)	giraf (giraffe)	kerk (church)
troon (throne)	stroopwafel (syrup waffle)	beker (mug)	koffer (suitcase)
trui (sweater)	struisvogel (ostrich)	fototoestel (camera)	kraan (tap)
tuba (tuba)	stuur (handlebar)	kleed (rug)	libel (dragonfly)

The activation of offset-embedded words:

Evidence from eye-tracking and identity priming

CHAPTER 5

Keren B. Shatzman and James M. McQueen (in preparation)

Abstract

Previous research on the activation of offset-embedded words (e.g., bone in trombone) has yielded contradictory findings. The current study sought to resolve this controversy by comparing performance on such materials in two different tasks. One experimental series used eye-tracking to examine activation of conceptual representations. Dutch listeners looked more at pictures of embedded words than at unrelated distractors, but only when the embedded word coincided with a syllable onset and was unstressed. A similar pattern emerged when the offset-embedded sequences were replaced by recordings of the monosyllabic words themselves, indicating that acoustic marking of the word is not sufficient for conceptual activation. The second series used cross-modal identity priming to examine activation of phonological representations. Responses to embedded words were slower after related primes (the carrier words). The activation of offset-embedded words' phonological representations appears therefore to be obligatory, but does not automatically lead to activation of their conceptual representations.

Introduction

Understanding spoken language involves recognizing the words that comprise a spoken utterance and accessing the meaning of those words. Despite the seemingly effortless speed and efficiency of this process, this is no small feat. The speech signal is a continuous stream of acoustic-phonetic information and the individual words in it are not unambiguously marked. Furthermore, as if to complicate matters, words often contain spuriously embedded words. For example, the word straight contains the words stray, tray, ray, trait, rate and eight. The majority of polysyllabic English words have at least one shorter word embedded within them (McQueen, Cutler, Briscoe & Norris, 1995). Psycholinguistic research has therefore endeavored to explain how listeners deal with these two related problems – segmenting the correct words from the speech signal and rejecting the spurious embeddings.

According to current theories of spoken-word recognition, listeners evaluate simultaneously multiple word candidates that are fully or partially consistent with the speech signal (e.g., Allopenna, Magnuson & Tanenhaus, 1998; Marslen-Wilson, 1987; McQueen, Norris & Cutler, 1994; Zwitserlood, 1989). As the signal unfolds, the active lexical candidates compete among one another for recognition. The level of activation of a particular candidate is a function of its goodness of fit with the current signal, the number of other active lexical candidates and their goodness of fit with the signal. The outcome of the competition is that each section of the speech signal is assigned to one word and the whole utterance is parsed as a sequence of non-overlapping words. The parallel activation of lexical candidates and the competition between them are the general mechanisms by which the segmentation and embedding problems are solved. In addition, researchers have identified a number of factors (see McQueen, 2005 for review) that influence the goodness of fit between a lexical candidate and the signal and thus modulate the lexical activation and competition process.

Several studies have investigated the recognition of words with initial embeddings, such as cap in captain. Davis, Marslen-Wilson and Gaskell (2002), using the gating paradigm and

cross-modal identity priming, showed that the shorter word (e.g., cap) was more active when the ambiguous sequence /kæp/ came from a recording of the monosyllabic word than when it came from the carrier word (e.g., captain). Vice versa, there was more activation for the carrier word when the sequence came from a recording of the carrier word than when it came from the monosyllabic word. Another study (Salverda, Dahan & McQueen, 2003) used eye movement data to demonstrate that the duration of the initially-embedded sequences (e.g., ham in hamster) modulated the amount of transitory fixations to pictures representing the monosyllabic embedded words. Longer sequences generated more monosyllabic-word interpretations and shorter sequences generated more polysyllabic-word interpretations (see also Salverda, 2005). These studies thus showed that listeners are sensitive to fine-grained acoustic differences between onset-embedded words and monosyllabic words. These differences influence the amount of support for particular lexical candidates, thereby biasing the competition in the direction of the correct interpretation.

The goal of the present study was to investigate the activation of offset-embedded words and whether fine-grained acoustic detail plays a role in this process, as it does with onset-embedded words. Offset-embedded words (e.g., bone in trombone) have an inherent disadvantage relative to the longer word, as the latter receives more support from the signal (i.e., it matches more portions of the speech signal). It is possible that the "head start" that the longer word enjoys will effectively eliminate the competition from the embedded words, making the usage of fine-grained information superfluous.

Indeed, while there is substantial evidence showing the activation of onset-embedded words, the empirical evidence regarding the activation of offset-embedded words is inconclusive. A number of studies using associative priming have found evidence of activation of offset-embedded words. Shillcock (1990) reported that offset-embedded words (e.g., bone in trombone) facilitated lexical decisions to a semantically related word (e.g., RIB) relative to an unrelated word (e.g., BUN). Luce and Cluff (1998) found similar results with

carrier words in which the first part was also a word (e.g., lock in hemlock primed KEY). Similarly, a study in French showed that the word car embedded at the offset of brancard (stretcher) primed the visual target AUTOBUS (Isel & Bacri, 1999). Vroomen and de Gelder (1997) extended these findings by observing facilitatory priming effects when the offset-embedded words corresponded to the final syllable of the carrier word, but not when the embedded word did not coincide with the onset of a syllable (e.g., wijn [wine] in zwijn [swine] did not prime ROOD [red]).

But not all studies using associative priming have found evidence for activation of offset-embedded words. Pitt (1994) reported the activation of offset-embedded words when the carrier was a nonword (e.g., light embedded in the nonword trolite primed DARK), but not when the carrier was a word (e.g., polite). Gow and Gordon (1995) showed that a phrase such as two lips primes words related to tulips and to lips, but there was no priming from tulips to words related to lips. Norris, Cutler, McQueen and Butterfield (submitted) used a similar set of items to Shillcock (1990), but did not find any evidence of associative priming. They did, however, observe that when the embedded word itself (e.g., bone) was presented as the visual target, lexical decision was slowed after a related prime (e.g., trombone) than after an unrelated prime. This result is reminiscent of an earlier study (Marslen-Wilson, Tyler, Waksler & Older, 1994), which also presented offset-embedded words as targets in a cross-modal identity-priming task. The inhibitory effect in that study, however, was not significant. Luce and Lyons (1999) attempted to detect the activation of offset-embedded words using methodologies other than the cross-modal priming task. Carrier words with word-initial embeddings (e.g., witness) were responded to more quickly in both lexical decision and shadowing tasks than were carrier words with nonword-initial syllables (e.g., harness), but no difference was found between carrier words with word-final embeddings (e.g., chloride) and carrier words with nonword-final syllables (e.g., chlorine).

Research on the activation of offset-embedded words has therefore yielded contradicting results. One worrying aspect concerning this body of evidence is that the activation of offset-embedded words has only been detected using the cross-modal priming paradigm (sometimes using associative priming and sometimes identity priming). An inherent weakness of the priming paradigm is that it involves the successive and rapid presentation of two stimuli (prime and target). Conclusions regarding how the prime has been processed are made based on the responses to the visual target. However, one can never be certain that the presentation of the visual target does not introduce some effects that cause the prime to be processed in a different way than it would have been if the visual target had not been presented. Primed presentation can cause different processing than if the stimuli are presented in isolation (Prather & Swinney, 1988; Shelton & Martin, 1992). Furthermore, lexical decision latencies to the target can be facilitated by an association from the target to the prime even in the absence of an association from the prime to the target ('backward priming', Koriat, 1981). Thus, if one finds that responses to the visual target RIB were faster after the prime trombone, one can not rule out the possibility that at least part of the effect is due to the association between the visual target and bone (see also Glucksberg, Kreuz & Rho, 1986). It is therefore crucial to have converging evidence from different methodologies.

The present study uses the eye-tracking paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Tanenhaus & Spivey-Knowlton, 1996 for an overview of the paradigm) as an alternative to the cross-modal associative priming task. Eye-tracking provides a measure of lexical processing that is closely time-locked to the information in the signal. During a trial, participants hear a sentence and are shown four objects presented as pictures on a computer screen. Their task is to click on and move the object referred to in the sentence with the computer mouse. It is assumed that eye-movements to the objects depicted on the screen are a function of the similarity between those objects and the conceptual information associated with the spoken input (e.g., Dahan & Tanenhaus, 2005; Huettig & Altmann, 2005).

If the conceptual representations of offset-embedded words are activated during lexical access then we would expect participants to look at pictures of the embedded words more than at pictures with names that do not overlap with the signal.

In a subsequent series of experiments we used the cross-modal identity priming task. The study by Norris et al. (submitted) yielded results suggesting that lexical access involves different types of representations. The study was a series of nine experiments, including cross-modal associative priming and identity priming of non-embedded and embedded words, when the primes appeared either in isolation or in sentence contexts. In none of their experiments did Norris et al. find associative priming from embedded words (e.g., date in sedate did not prime TIME). However, in the identity priming task with the embedded word as the visual target (e.g., the prime sedate followed by DATE) an inhibitory effect was observed. The dissociation between the results obtained with the associative priming task and those obtained with the identity priming task suggests that the two tasks tap into different types of representations: associative priming depends on activating the conceptual representation of the word, while identity priming reflects the activation of phonological representations. In the current study, we use eye-tracking to examine the activation of conceptual representations and identity priming to index phonological activation.

As the studies on onset-embedded words (e.g., ham in hamster) have shown, even very subtle acoustic details in the signal can influence lexical activation and competition. Listeners' remarkable sensitivity to fine-grained acoustic detail has also been suggested as a factor bearing on the contradictory findings regarding the activation of offset-embedded words. For example, Gow and Gordon (1995) argued that their failure to find priming indicates that listeners are sensitive to the fine-grained acoustic differences between word-initial consonants and non-initial consonants (e.g., the [l] of lips tends to be slightly longer than the [l] in tulips). To examine this issue, we manipulated the spoken input by splicing. For each target word (i.e., carrier word containing an offset-embedded monosyllabic word) two

versions were created. In one version, the word embedded in the target word was spliced from another token of the target word. In the other version, the embedded word in the target word was replaced by a recording of the monosyllabic word. If listeners are sensitive to the subtle acoustic difference between the realization of an embedded word and the realization of a monosyllabic word, they should look more at the picture of the embedded word when the target word contains the recording of the monosyllabic word than when the embedded word is spliced from another token of the target word.

Another issue which might have contributed to the conflicting results in the literature concerns the presence of stress on the embedded word in items that were tested. All of the items used by Shillcock (1990) and Vroomen and de Gelder (1997) had primary stress on the embedded-word syllable (e.g., bone in trombone). In contrast, more than half of the items used by Gow and Gordon had stress on the first syllable (e.g., tulips). In the present study we therefore investigated the role of stress by including items in which the embedded word was either stressed or unstressed (e.g., the Dutch word pipet (/piː-ˈpɛt/ [pipette]) containing the stressed syllable pet [cap] at its offset and the word bizon (/ˈbiː-zɔn/ [bison]) containing the unstressed syllable zon [sun]). Our prediction was that there will be more looks to a picture of the embedded word when the second syllable is stressed than when it is unstressed, because a stressed syllable is more like a monosyllabic word than an unstressed syllable is (i.e., the second syllable of pipet is more like the monosyllabic word pet than the second syllable of bizon is like the word zon, as confirmed in acoustic measurements of our stimuli, which we report below).

In addition to these items, in which the embedded word was always aligned with the syllable boundary, we included items in which the embedded word did not coincide with the onset of a syllable. The results of Vroomen and de Gelder (1997), as well as other studies (e.g., Dumay, Frauenfelder & Content, 2002; McQueen, 1998; Weber & Cutler, 2006; see Cutler, McQueen, Norris & Somejuan, 2001, for review), suggest that embedded words are

strongly activated only if they are aligned with a syllable boundary. If syllable boundaries play a role in lexical access then we would predict that there will be fewer looks to the pictures of the embedded words when the word is misaligned with the syllable boundary (e.g., the word speen (/speːn/ [pacifier]) containing the word peen [carrot]) than when it is aligned.

The outline of the study is therefore as follows. In Experiment 1, the eye-tracking paradigm was used to examine the conceptual activation of offset-embedded words. We tested whether participants would make more fixations to pictures of embedded words than to pictures that are phonologically unrelated to the target word. We examined words in which the embedded word was either aligned with the syllable boundary (stressed and unstressed embedded words) or misaligned with the syllable boundary. In Experiment 2, we investigated the activation of the offset-embedded words' phonological representations. Listeners performed a cross-modal identity priming task, with either the carrier words or the embedded words as visual targets. In both experiments, the contribution of fine-grained acoustic detail was explored by creating spliced versions of the carrier words, in which the origin of the embedded word was either another token of the carrier word or a recording of the monosyllabic embedded word.

Experiment 1

In Experiment 1 we used the eye-tracking paradigm to examine the activation of offset-embedded monosyllabic words. Participants heard short Dutch sentences and saw four pictures of objects on the screen. They were instructed to click on the object mentioned in the sentence. In the critical trials the name of the target object was a word containing an offset-embedded word. In these trials, the visual display included pictures of both the target object and the embedded-word object. We tested whether people would look more at the embedded-word object than at a phonologically unrelated object. The spoken input that listeners heard was manipulated such that the word embedded at the offset of the target word was either spliced from another token of the target word or from a matched recording where the speaker

intended the monosyllabic word itself. We expected more looks to the picture of the embedded word when the embedded sequence originated from the monosyllabic word.

The role of syllable boundaries in the activation of offset-embedded words was examined by presenting listeners with target words in which the embedded word was either aligned with the syllable boundary (e.g., pipet; Experiment 1A) or misaligned with the syllable boundary (e.g., speen; Experiment 1B). In Experiment 1A, the embedded word either carried the primary stress (e.g., pipet) or did not (e.g., bizon), allowing us to investigate whether lexical stress influences the activation of offset-embedded words.

Method

Participants. Sixty-two student volunteers from the Max Planck Institute subject pool took part in this experiment (36 in Experiment 1A and 26 in Experiment 1B). They were all native speakers of Dutch. They were paid for their participation.

Materials. The target words for Experiment 1A were twenty-four polysyllabic Dutch words referring to picturable objects and containing a picturable embedded monosyllabic word at their offset. Thirteen of the words had lexical stress on the second syllable (e.g., the word pipet containing the word pet at its offset). The other eleven words had lexical stress on the first syllable (e.g., the word bizon containing the word zon). In all 24 words the embedded words were aligned with the syllable boundary. We will refer to these as the carrier words with aligned stressed embeddings and the carrier words with aligned unstressed embeddings, respectively. In Experiment 1B, the target words were eighteen words referring to picturable objects and containing a picturable word which was misaligned with the syllable boundary (e.g., the word speen containing the word peen). These will be referred to as the carrier words with misaligned embeddings. In both experiments, there were no semantic or morphological relationships between the carrier word and the embedded word within each pair. The lemma frequencies of the carrier words and the embedded words were computed using the CELEX lexical database (Baayen, Piepenbrock & Gulikers, 1995), for each embedding type (i.e.,

aligned stressed, aligned unstressed and misaligned). The average frequencies (per million) were 4.75, 7.02 and 12.32 for the carrier words with aligned stressed, aligned unstressed and misaligned embeddings, respectively. The average frequencies for the corresponding embedded words were 21.7, 34.53 and 24.05. A mixed two-factor ANOVA with Word Type (with the two levels 'carrier' and 'embedded') as within-items factor and Embedding Type (with the three levels 'aligned stressed', 'aligned unstressed' and 'misaligned') as between-items factor indicated that the frequency of the embedded words was significantly higher than that of the target words ($F(1,39) = 6.52$, $p < .05$). There were no other significant main effects or interactions ($Fs < 1$).

Two additional picturable nouns (e.g., vlag [flag] and asperge [asparagus], see Figure 1) were assigned to each pair of target and embedded word (henceforth, the competitor), to serve as distractors in the eye-tracking experiment. These distractors were phonologically unrelated to both the target and the competitor. The full set of items is presented in Appendix A. Line-drawing pictures associated with the items were selected from various picture databases (among which the Snodgrass and Vanderwart, 1980, and Cycowicz, Friedman, Rothstein, & Snodgrass, 1997, picture sets, as well as the Art Explosion library, 1995).

For each target-competitor pair two sentences were constructed: one sentence referring to the target word and another sentence referring to the competitor. The sentences were constructed such that the context surrounding the critical word was identical in both sentences (e.g., Ze kon de grotere pipet niet vinden [she could not find the bigger pipette] and Ze kon de grote hippie pet niet vinden [she could not find the big hippy cap]). The two sentences were always matched on the number of syllables. Further, we attempted to keep the two sentences as similar as possible. The sentences are listed in Appendix B.

All sentences were read aloud in a sound-attenuated booth in random order by a female speaker of Dutch, naïve to the purpose of the experiment, and recorded on a DAT tape (sampling at 48 kHz with 16-bit resolution). The target words, and the competitor words in

the matched sentences, were marked on the script by the use of capital letters, and the speaker was instructed to produce these words as the focus of the sentence. Each sentence was recorded at least four times. The sentences were then re-digitized at a sample rate of 16kHz and edited using Xwaves speech-editing software. Two spliced versions of the sentences mentioning the target words were created. In one version (hereafter, the identity-spliced version) the word embedded in the target word (e.g., pet) was spliced from another token of the target sentence (e.g., by splicing the second syllable of the word pipet). In the other version (hereafter, the cross-spliced version) the embedded word originated from the sentence mentioning the competitor word (e.g., pet) (see Table 1). The two versions of each sentence differed therefore only in the origin of the embedded word (i.e., whether it was recorded in the context of the target word or as a monosyllabic word). All splicing points were at zero-crossings and care was taken to avoid any acoustic artifacts, such as clicks or other distortions.

Table 1. *Example of the sentences used to produce the identity-spliced and cross-spliced stimuli for Experiment 1.*

(A) Target context (1)	Ze kon de grotere pipet niet vinden (she could not find the bigger pipette)
(B) Target context (2)	**Ze kon de grotere pipet niet vinden**
Identity-spliced version created from (A) and (B)	Ze kon de grotere pi**pet** niet vinden
A. Target context (1)	Ze kon de grotere pipet niet vinden
B. Competitor context	**Ze kon de grote hippie pet niet vinden** (she could not find the big hippy cap)
Cross-spliced version created from (A) and (B)	Ze kon de grotere pi**pet** niet vinden

On average, the duration of the context preceding the target word was 815 ms. The average duration of the pre-spliced portions of the carrier words, the embedded words and the carrier words are shown in Table 2, for the carrier words with aligned stressed, aligned unstressed

and misaligned embeddings, for each splicing condition. These measurements show that the cross-spliced words were slightly longer than the identity-spliced words, as expected given that monosyllabic words tend to have longer durations than the same syllables in a polysyllabic word (e.g., Lehiste, 1972; Turk & Shattuck-Hufnagel, 2000). Two-tailed matched-pairs t tests indicated that the durational difference between the identity- and cross-spliced versions was statistically significant in the carrier words with aligned unstressed embeddings ($t(10) = -2.70$, $p < .05$) and in the carrier words with misaligned embeddings ($t(17) = -3.23$, $p < .001$), but not in the carrier words with aligned stressed embeddings ($t < 1$). These analyses show that the durational differences between the monosyllabic word (e.g., pet) and the embedded sequence (e.g., the second syllable of the word pipet) were smaller when the embedded sequence was stressed. This simply reflects the fact that a stress-bearing monosyllabic word is more like an embedded stressed syllable than like an embedded unstressed syllable (see also Lindblom, Lyberg & Holmgren, 1981).

Table 2. *Average duration of the pre-spliced portions, the embedded words and the carrier words, for the carrier words with aligned stressed, aligned unstressed and misaligned embeddings, for each splicing condition.*

	Pre-spliced portion	Embedded word		Total duration	
		Splicing version		Splicing version	
		Identity	Cross	Identity	Cross
Aligned stressed (pipet)	160	264	275	424	435
Aligned unstressed (bizon)	210	191	217	401	427
Misaligned (speen)	130	255	276	385	406

In addition to the experimental items, 43 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. Sentences mentioning the filler items were produced by the

same speaker, and recorded at the same time as the experimental sentences. Twenty-two cross-spliced filler sentences were created by excising the filler word from one token of the filler sentence and splicing it into another token of the sentence.

Procedure and Design. Participants were tested individually. The experiment began with a familiarization phase. Participants were exposed to the pictures, which appeared on the computer screen, one at a time in a randomized order, along with their printed name. Participants were instructed to familiarize themselves with each picture and then pressed a response button to proceed to the next picture. After this part of the experiment, the eye-tracking system was set up.

Participants were seated at a comfortable distance from the computer screen. The eye-tracking system was mounted and calibrated. Eye movements were monitored using a 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, and the auditory stimuli were presented over headphones. The presentation of stimuli was controlled using NESU software (http://www.mpi.nl/world/tg/experiments/nesu.html).

After the eye tracker was calibrated, participants were presented with the experimental and filler trials. The structure of each trial was as follows. First, a central fixation dot appeared on the screen for 500 ms. After that, a spoken sentence was presented to the participants and simultaneously a 5x5 grid with pictures appeared on the screen (see Figure 1). Prior to the experiment, participants 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. Once the picture had been moved, the experimenter pressed a button to initiate the next trial. The positions of the target object and its competitor were randomized across four fixed positions of the grid. The geometric shapes appeared in fixed positions on every trial. 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 participants were instructed to look at it. The experimenter could then correct potential drifts in the calibration of the eye tracker.

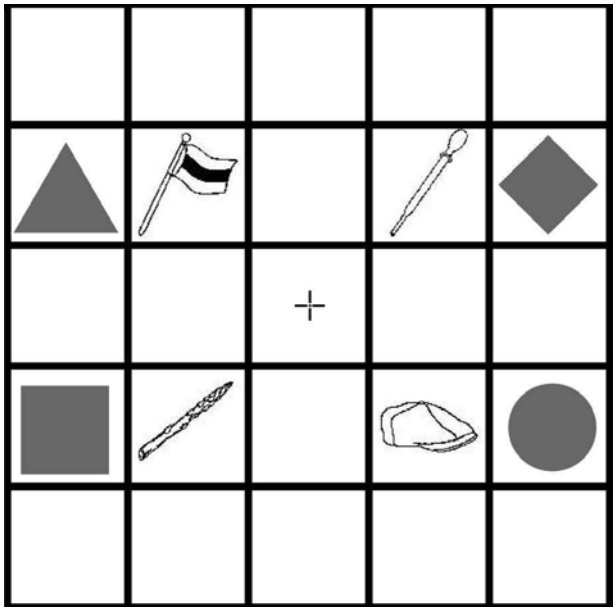


Figure 1. *Example of stimulus display presented to participants. Clockwise from top left corner: vlag [flag], pipet [pipette], pet [cap] and asperge [asparagus].*

Within each experiment (Experiment 1a or 1b), two lists were created, each containing the experimental items (24 items in Experiment 1a; 18 items in Experiment 1b) and 43 filler items. The lists varied on which of the two sentences (i.e., the identity-spliced or the cross-spliced sentence) was presented for each of the experimental items. Within each list, half of the experimental items were assigned to the identity-spliced condition and the other half to the cross-spliced condition. In Experiment 1a, approximately half of the items in each list were stressed and the other half unstressed. The two lists were randomized, such that a different random order was created for each participant, but the lists were yoked so that the same randomizations were used for both lists. The random orders 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 participants with the task and procedure.

Coding Procedure. The stream of sampled eye-position data were analyzed and coded in terms of fixations, saccades, and blinks, using the algorithm provided in the Eyelink software. The data from each participant's right eye were analyzed. However, for ten participants, the data from the left eye was used due to calibration problems with the right eye. Graphical software was used to display 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. The timing of the fixations was established relative to the onset of the target word in the spoken utterance. Fixations that fell within the cell of the grid in which a picture was presented were coded as fixations to that picture. Fixations were coded as pertaining to the target object, to the competitor, to one of the two unrelated distractors, or to anywhere else on the screen. For each experimental trial, fixations were coded from the onset of the target word until the last fixation to the target picture before the participant clicked on it.

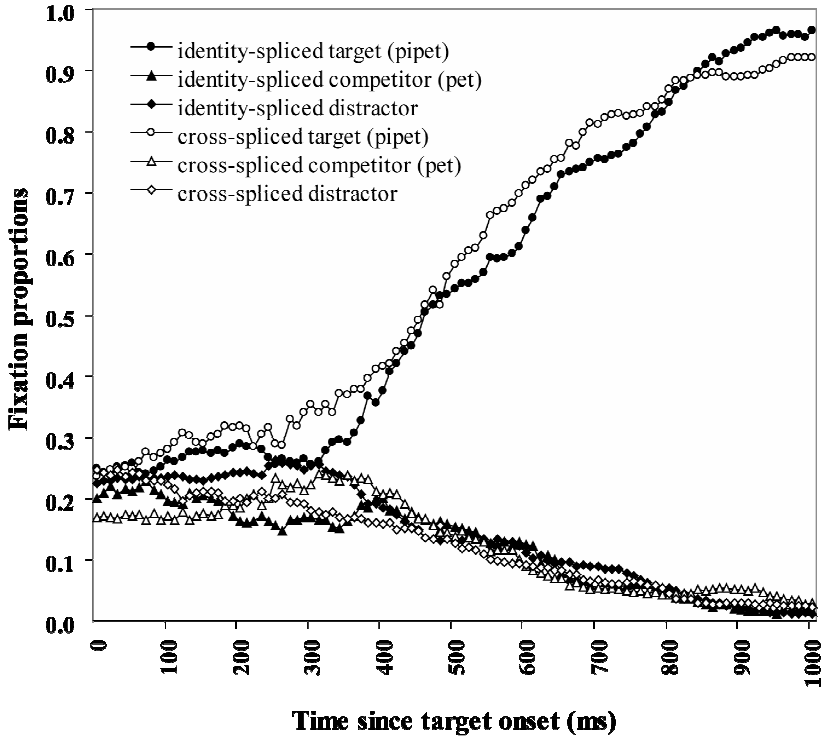
Results

Experiment 1A. On seven trials, participants erroneously selected an object other than the target picture without correcting their choice. These trials were excluded from further analyses. For each participant, the proportion of fixations to each type of picture or location (i.e., target picture, competitor picture, distractor pictures, or elsewhere) were computed for each 10 ms slice, separately for each condition. This was done by summing the number of trials in which a particular type of picture was fixated (in each 10 ms slice, in each condition), and dividing it by the total number of trials (in the same time interval) in which any picture or location was fixated. Blinks and saccades were not included in this calculation. A similar analysis was done for each item, averaging across participants.

The proportions of fixations to the target, the competitor and the averaged distractor in the identity-spliced and cross-spliced conditions are shown in Figure 2, separately for the aligned

stressed words (e.g., pipet; Figure 2a) and the aligned unstressed words (e.g., bizon; Figure 2b). Fixation proportions are shown in 10 ms time slices from target onset to 1000 ms thereafter.

(a)



(b)

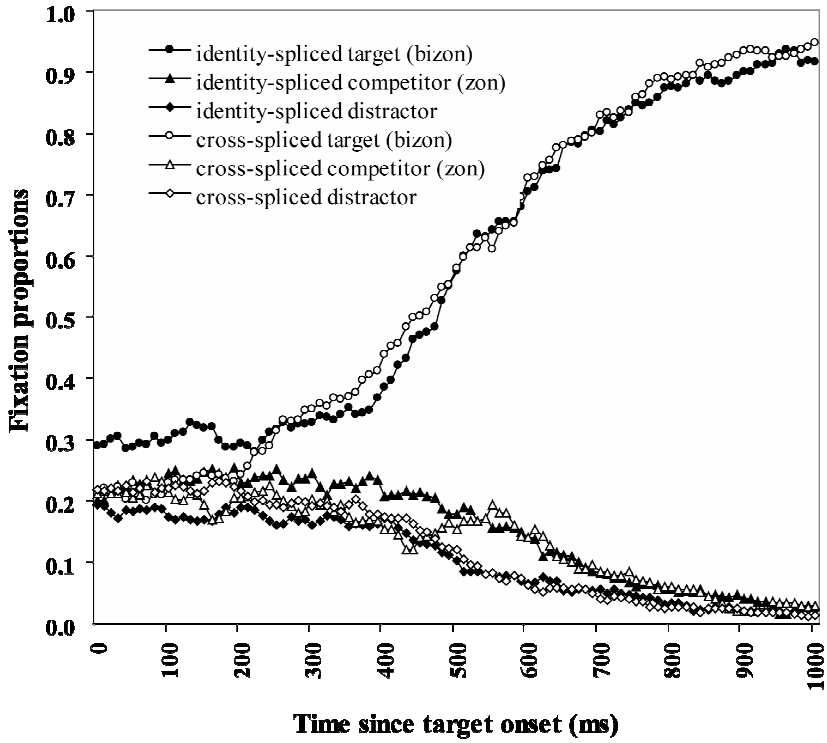


Figure 2. *Fixation proportions over time to the target, the competitor, and averaged distractors, for the carrier words with aligned stressed embeddings (Fig. 2a) and aligned unstressed embeddings (Fig. 2b), in the identity- and the cross-spliced conditions.*

Target fixations. Figure 2 shows that target fixation proportions began to rise around 200 ms after target onset. This is consistent with the standard estimate of 200 ms for planning and launching a saccade (e.g., Fischer, 1992; Hallett, 1986; Matin, Shao, & Boff, 1993; Saslow, 1967; see also Altmann & Kamide, 2004). Recall that the average durations of the pre-splice portions were 160 ms and 210 ms for the carrier words with stressed and unstressed embeddings, respectively. Therefore, differences in the fixation proportions due to the splicing manipulation can be observed, at the earliest, 360 ms after target onset in the stressed condition, and 410 ms after target onset in the unstressed condition. In the stressed condition, participants fixated the targets slightly more when they heard the cross-spliced version than when they heard the identity-spliced version, until about 800 ms after target onset. There seems to be no difference in target fixation proportions between the identity- and cross-spliced versions in the unstressed condition.

The effect of the splicing manipulation was statistically tested by computing the average fixation proportion to the target picture over a time window extending from the average onset of the embedded word with the 200 ms delay (i.e., 360 ms in the stressed condition, and 410 ms in the unstressed condition) until 800 ms after target onset[1]. Average fixation proportions were submitted to a one-way analysis of variance (ANOVA), with Splicing (identity-spliced vs. cross-spliced) as a within-participants factor or within-items factor. With the stressed words, the average fixation proportion to the target picture over this time interval was 61% in the identity-spliced condition and 66% in the cross-spliced condition. This difference was not significant ($\underline{F}_1$ (1,35) = 2.07, $\underline{p}$ = .16; $\underline{F}_2$ < 1). The average fixation proportion to the

[1] Because of the difference in the average onset time of the embedded word between the stressed and unstressed words, a different time window was applied to each stress condition and the analyses were run separately. One could also define the time window as starting at the mean onset of the embedded word (averaged over both conditions) and adding the 200 ms delay. This definition would allow both types of words to be included in the same analysis, adding stress as a factor. Analyses using this definition yielded very similar results to those reported here.

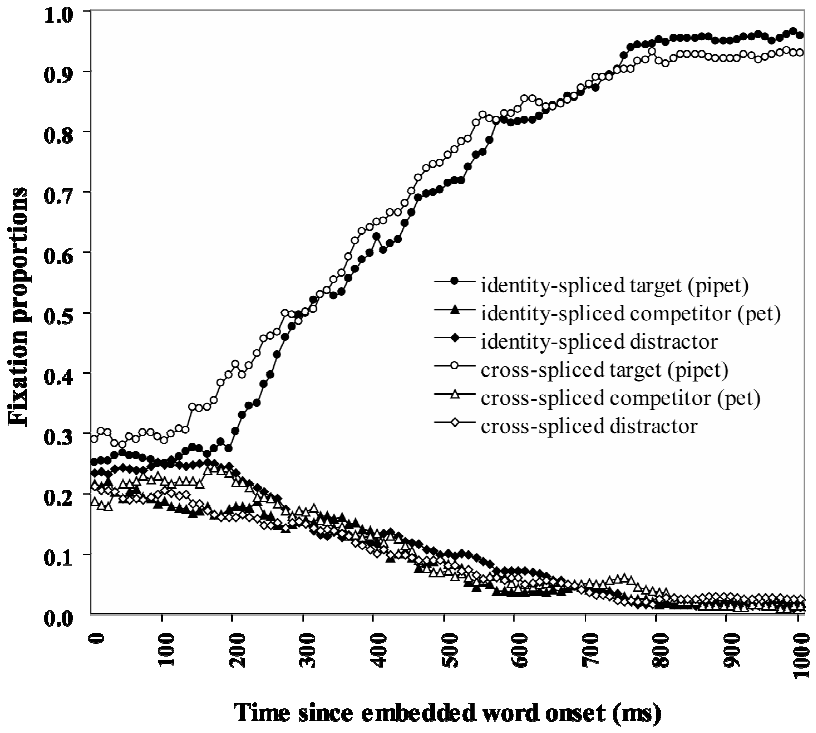
unstressed target pictures was 67% in the identity-spliced condition and 65% in the cross-spliced condition (both Fs < 1).

Competitor and distractor fixations. The pattern in Figure 2 seems to suggest that in the stressed condition participants did not fixate the competitor pictures (i.e., pictures of the embedded words) more than they did the distractors, while in the unstressed condition there was a small difference starting about 400 ms after target onset and extending until about 800 ms after target onset. In order to examine this pattern more accurately, the data was realigned to the onset of the embedded word. The realigned data is presented in Figure 3.

The results in Figure 3 show more clearly that the competitor pictures were fixated more than the distractor pictures in the unstressed, but not in the stressed, condition. Fixations to the competitor and distractor pictures were compared over the time window extending from 200 to 400 ms in the unstressed condition, and from 200 to 470 ms in the stressed condition. This window corresponds to the time interval at which we expect to observe differences in fixation proportions due to the processing of the embedded word (i.e., from 200 ms after embedded word onset until the average word offset plus 200 ms delay). Average fixation proportions were submitted to a two-way ANOVA, with Picture (competitor vs. distractor) and Splicing (identity-spliced vs. cross-spliced) as the within-participants or within-items factors.

In the unstressed condition, the average proportion of fixations to the competitors (18%) was higher than to the distractors (12%) (F_1 (1,35) = 13.24, $p < .01$, $eta^2 = 0.27$; F_2 (1,10) = 5.80, $p < .05$, $eta^2 = 0.37$). There was no difference between the splicing conditions (both 15%; Fs < 1) and no interaction between the factors ($F_1 < 1$; F_2 (1,10) = 1.03, $p = .33$). In the stressed condition, there was virtually no difference in the average fixation proportions to the competitor and the distractors (15% and 14%, respectively). There were no significant main effects or interactions in this condition (all Fs < 1).

(a)



(b)

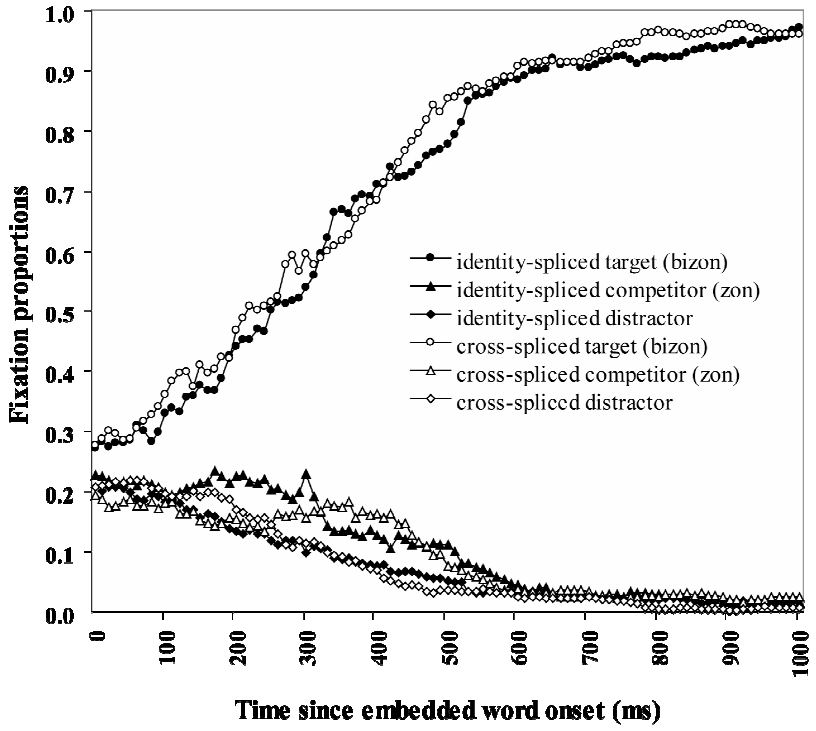


Figure 3. *Fixation proportions from the onset of the embedded word over time to the target, the competitor, and averaged distractors, for the carrier words with aligned stressed embeddings (Fig. 3a) and aligned unstressed embeddings (Fig. 3b), in the identity- and the cross-spliced conditions.*

In sum, we found evidence for the activation of the offset-embedded words in the unstressed, but not in the stressed, condition. The splicing manipulation did not seem to have an effect on listeners' fixations.

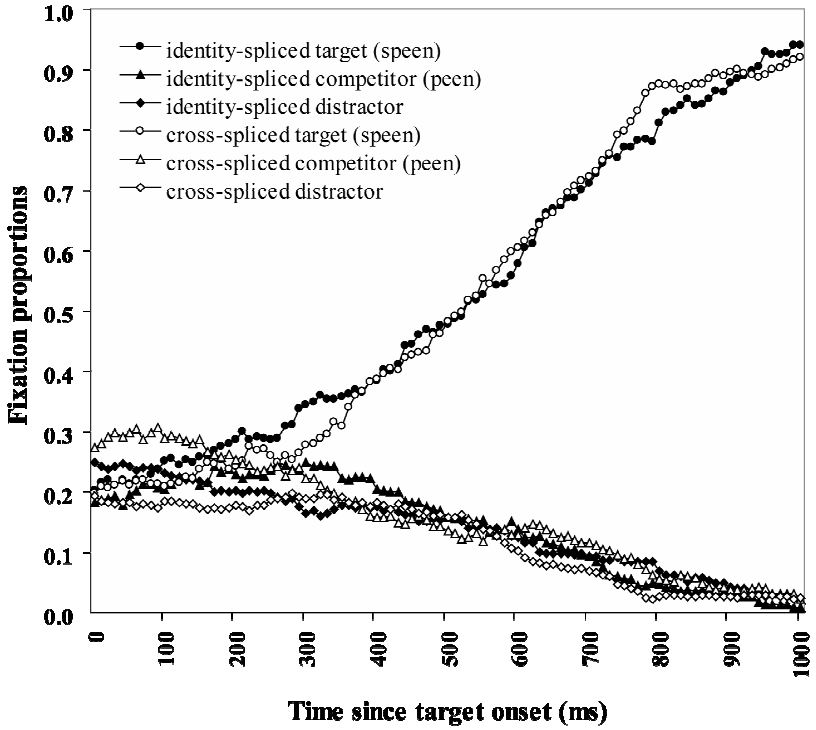


Figure 4. *Fixation proportions over time to the target, the competitor, and averaged distractors, for the carrier words with misaligned embeddings, in the identity- and the cross-spliced conditions.*

Experiment 1B. On two trials, participants erroneously selected an object other than the target picture without correcting their choice. These trials were excluded from further analyses. Figure 4 presents the fixation proportions to the target picture, the competitor picture (the embedded word), and to the averaged distractors in the identity-spliced and cross-spliced conditions, in 10 ms time slices from target onset to 1000 ms thereafter. Around 200 ms after target onset, target fixation proportions began to increase, rising with a similar slope in both splicing conditions. The average onset of the embedded word was 130 ms. Average fixation proportions were therefore computed over a time window extending from that point with the 200 ms delay added (i.e., from 330 ms) until 600 ms after target onset (average offset of the target word plus the 200 ms delay). In this time interval there was no difference between the

splicing conditions in the proportion of fixations to the target (both 45%; Fs < 1). Moreover, listeners looked as often to the pictures of the embedded words as they did to the distractors (both 16%). In a two-way ANOVA (Picture X Splicing) there was no main effect of picture (Fs < 1) or splicing (F_1 (1,25) = 1.18, p = .29; F_2 < 1), and there was no interaction (F_1 (1,25) = 1.23, p = .28; F_2 < 1)[2].

Discussion

The eye-tracking data showed evidence for the activation of the offset-embedded words only when the embedded words were aligned with the syllable boundary and unstressed. This result is surprising. Our prediction was that the similarity between an offset-embedded word and the corresponding monosyllabic word would be bigger when the second syllable of the carrier word is stressed than when it is unstressed, and that this similarity would lead to a higher activation and more fixations to the picture of the embedded word in the case of the carrier words with stressed embeddings. This prediction was based on previous studies that have used the cross-modal associative priming paradigm (e.g., Isel & Bacri, 1999; Shillcock, 1990; Vroomen & de Gelder, 1997) and found that stressed and aligned offset-embedded words primed an associatively related word. We therefore sought an explanation for our failure to detect the activation of such words in the aligned stressed condition. Note first that the fact that we found activation of the embedded words in the unstressed condition indicates that the method we are using is sensitive enough to detect the activation of offset-embedded words. In other words, the eye-tracking paradigm can not be the sole culprit for the failure to find activation of the embedded words in the stressed condition.

[2] Analyses of the data realigned to the onset of the embedded word yielded an identical statistical pattern to the one reported.

One inherent difference between the carrier words with aligned stressed embeddings and carrier words with aligned unstressed embeddings concerns their cohort sizes: As the majority of words in Dutch have lexical stress on the first syllable (Schreuder & Baayen, 1994), it is plausible that there would be fewer cohort competitors for the carrier words with stressed embeddings (where stress is on the second syllable) than for the carrier words with unstressed embeddings (i.e., where stress is on the first syllable). We examined this possibility by calculating for each word in the experiment the number of cohort words, using the CELEX lexical database. A cohort word was defined as beginning with the same sequence as the target word – up to the point of the onset of the embedded word – and having the same stress pattern. For example, for the word pipet, the cohort set included all words beginning with an unstressed /piː/, while for the word bizon, the cohort set included all words beginning with a (stressed) /ˈbiː/. In addition to counting the number of words in the cohort of each experimental item, we calculated the cohort's frequency-weighted density. For each word in the cohort set, the raw frequency in CELEX (based on a corpus of approximately 42.4 million words) was multiplied by 10. The log-transformations of the product of all words in the cohort set were summed to yield the frequency-weighted density (henceforth, cohort density). Words which had a frequency of 0 were excluded. In a second analysis, the number of cohort members and the cohort density of each experimental item were calculated again, but a cohort word was defined as beginning with the same sequence as the target word up to one segment after the onset of the embedded word (e.g., for the word pipet, the cohort set included all words beginning with /piːˈp/, while for the word bizon, the cohort set included all words beginning with a (stressed) /ˈbiːz/).

The results of the first analysis showed that up to the point of the embedded words' onset the average cohort number and density did not differ much between the carrier words with stressed embeddings and those with unstressed embeddings. On average, there were 246 cohort words in the cohorts of the carrier words with stressed embeddings and 210 in the

cohorts of the carrier words with unstressed embeddings. Cohort densities were 514 and 443, respectively. Two-tailed t tests indicated that these differences were not statistically significant, given a Type I error rate of 0.05 ($t < 1$). However, the second analysis showed that one segment after the onset of the embedded word there were, on average, as predicted, fewer cohort competitors in the stressed embedding cohorts than in the unstressed embedding cohorts (11 and 38 respectively; $t(22) = -2.77$, $p < .05$). Furthermore, the cohort density of the carrier words with stressed embeddings was lower than that of the carrier words with unstressed embeddings (23 and 78 respectively; $t(22) = -2.75$, $p < .05$).

These findings show that by the time listeners heard one segment of the embedded word, there were fewer words in the cohort of the carrier words with stressed embeddings, and its density was lower than that of the carrier words with unstressed embeddings. Other things being equal, words with a sparse cohort and low cohort density are recognized faster. This means that the level of activation of the carrier words with stressed embeddings may have increased faster than the level of activation of the carrier words with unstressed embeddings. Consequently, the embedded words in the stressed condition may have offered very little competition and this may be why their activation was not detected. The increase in the activation levels of the carrier words with unstressed embeddings was possibly slower (due to the higher cohort density) allowing the embedded words to compete long enough for their activation to be detected. Note, however, that this explanation is tentative – the size and density of the carrier words' cohorts did not correlate significantly with the activation of the embedded words (as measured by calculating the fixation proportions to the embedded words' pictures minus the fixation proportions the distractors).

In Experiment 1B, no evidence was found for the activation of the misaligned embedded words. Listeners fixated pictures of these words as much as they did pictures of phonologically unrelated words. This finding is consistent with the results of a cross-modal associative priming experiment (Vroomen & de Gelder, 1997), in which misaligned offset-

embedded words did not facilitate lexical decision reactions to associatively related words (e.g., wijn [wine] in zwijn [swine] did not prime ROOD [red]). Word-spotting studies (e.g., Dumay et al., 2002; McQueen, 1998; Weber & Cutler, 2006) have also elicited results that suggest that syllable boundaries play an important role in lexical access. The results of Experiment 1 are, however, moot with regard to the role of syllable boundaries, because there was no evidence of the activation of the aligned and stressed embedded words either.

The splicing manipulation we employed did not seem to influence listeners' fixation patterns. Acoustic analyses indicated that the realizations of the words recorded in monosyllabic contexts were different than those recorded in carrier-word context (i.e., durational differences). In both sub-experiments, however, participants reacted to words embedded in the carrier words' offset as they did to words that were recorded in a monosyllabic context and spliced into the carrier words. Even in the case of the carrier words with aligned unstressed embeddings, participants fixated the pictures of the embedded words as much when they heard the identity-spliced versions as when they heard the cross-spliced versions. This result thus suggests that the activation of the offset-embedded words was not modulated by the fine-grained acoustic details in the signal. This finding contrasts with evidence from studies investigating word-initial embeddings (e.g., Salverda et al., 2003; Salverda, 2005), which showed that the duration of the embedded sequence (e.g., ham in hamster) influenced the interpretation of the sequence as either a monosyllabic word or the first syllable of a longer word. Furthermore, in the case of the carrier words with aligned stressed embeddings and with misaligned embeddings there was no evidence of activation of the offset-embedded words even when the embedded sequence originated from the monosyllabic word. Activation of the offset-embedded word therefore does not depend solely on the acoustic match with the signal.

Experiment 2

In Experiment 1, evidence was found for the conceptual activation of carrier words with aligned unstressed embeddings, but not for carrier words with aligned stressed or misaligned embeddings. In order to interpret this result, however, we need to have a measure of the embedded words' phonological activation. It is possible, for example, that the results of Experiment 1 are due to differences in the activation of phonological representations. Therefore, in Experiment 2, the cross-modal identity-priming paradigm was used to tap into the phonological activation of the offset-embedded words. Unlike the eye-tracking paradigm, which provides a measure of the activation of multiple words simultaneously, the priming paradigm allows the activation of only one word (i.e., the visual target) to be measured in a given trial. Due to constraints on the number of stimuli that we had and in order that participants be exposed to each prime only once, three sub-experiments were designed. In Experiment 2A the primes were carrier words with aligned stressed embeddings and carrier words with misaligned embeddings. The visual targets were the carrier words themselves. In Experiment 2B, the same primes were used as in Experiment 2A, but the visual targets were the embedded words. In Experiment 2C, the primes were the carrier words with aligned stressed and unstressed embeddings, and the visual targets were the embedded words.

Method

Participants. One hundred and twenty students from the Max-Planck-Institute subject pool, all native speakers of Dutch, took part in this experiment (48 in Experiment 2A; 48 in Experiment 2B; 24 in Experiment 2C). They were paid for their participation. None of them had participated in Experiment 1.

Materials. The auditory stimuli for Experiment 2 were recorded at the same time and in the same way as those for Experiment 1. Experiments 2A and 2B had as critical items twenty

polysyllabic Dutch words with an aligned stressed embedding at their offset (e.g., the word pipet containing the word pet at its offset) and twenty-eight carrier words with a misaligned embedding (e.g., the word speen containing the word peen). In Experiment 2C, the critical items were eleven carrier words with an aligned unstressed embedding (e.g., the word bizon containing the word zon at its offset) and twenty carrier words with an aligned stressed embedding (the same as those used in 2A and 2B). Identity and cross-spliced sentences were created using the same procedure as in Experiment 1. Most of the words used in Experiment 1 were also used here, but as the constraint of picturability was removed more items could be included in Experiment 2. In Experiment 2A, the visual targets in the experimental trials were the carrier words. In Experiments 2B and 2C, the visual targets were the embedded words. One aligned stressed item (harpoen [harpoon]) had to be discarded from both sub-experiments because the embedded word (poen [cash]) appeared in another carrier word (pompoen [pumpkin]). Finally, in Experiment 2C, the item kerstboom [Christmas tree] was used as a prime and the word boom [tree] as a target, for counterbalancing purposes. The item was not included in the analyses. The items are listed in Appendix C.

Experiments 2A and 2B had, in addition to the experimental trials, 10 filler trials in which the visual target was a word that did not contain an embedded word and 58 filler trials in which the auditory prime was followed by a nonword target. In Experiment 3B there were 10 word fillers and 48 nonword fillers. In order ensure that form overlap was not a cue to whether the visual target was a word or not, half the nonword targets in all three sub-experiments were phonologically similar to the auditory primes (e.g., sok-sog). Sixteen of the filler primes had embedded words in their offset to prevent participants from developing the expectation that a prime with an embedded word will be followed by a word target. Finally, there were twelve practice trials.

Design. For all three sub-experiments, a within-item design was created by constructing four lists in which the splicing version (identity-spliced vs. cross-spliced) and the relatedness

of the visual target to the prime (related (identical) vs. unrelated) were counterbalanced. The target words were arranged in pairs (e.g., pipet [pipette] was paired with karkas [carcass]). The combination of each of the primes (both identity- and cross-spliced) with each of the targets (related or unrelated) generated the four conditions. For example, in Experiment 2A, the visual target pipet was paired with the identity- and cross-spliced versions of the related prime pipet in two of the lists, and with the two spliced versions of the unrelated prime karkas in the other two lists. Each list thus contained all experimental targets, with an equal number of items in each condition. Each list also contained all the filler trials. There were three pseudo-random orders for each list, with the constraints that there were at least five intervening items between related experimental items, no more than five items of the same status (word or nonword) in a row and at least 30 intervening items between paired items (e.g., between pipet and karkas). Participants were randomly assigned to one list and presentation order.

Procedure. Participants were tested in a quiet room. They were told that they would hear spoken sentences, followed by visual stimuli on a computer screen and that their task was to decide whether the visual target was a word or not, by responding as quickly and as accurately as possible by pressing either the “yes” or the “no” button. The “yes” response was always assigned to the participant’s dominant hand. The spoken sentences were presented at a comfortable listening level through headphones. The visual targets were displayed in lower case Arial 48-point typeface on the center of the computer screen. The target appeared on the screen at the offset of the spoken prime and remained on the screen for 1 second. The target display stopped when the participant pressed one of the two buttons. However, if no response was registered within 2.5 seconds the next trial began. Before the main part of the experiment participants received the block of 12 practice trials.

Results

Experiment 2A. Reaction times were calculated from the onset of visual target presentation to response onset. All incorrect responses (4.5% of the data) and latencies exceeding 1200 ms (0.6%) were treated as errors. Mean reaction times (RTs), standard errors (SEs) and error rates for word targets in the four priming conditions are given in Table 3, separately for the aligned and misaligned words. Note that the data were collapsed over presentation orders.

Since the error rates were consistently very low across conditions no analysis of variance was run on errors. The latency results were submitted to a three-way repeated measures ANOVA with the factors Splicing (identity-spliced vs. cross-spliced), Relatedness (related vs. unrelated) and Embedding Type (aligned vs. misaligned). There was a strong main effect of relatedness: Participants reacted, on average, 87 ms faster to related targets than to unrelated targets (F_1 (1,47) = 197.11, $p < .001$; F_2 (1,46) = 115.3, $p < .001$). That is to say, responses to the visual targets (e.g., PIPET) were faster after related primes (e.g., pipet) than after an unrelated prime (e.g., karkas). There was no main effect of Splicing ($F_1 < 1$; F_2 (1,46) = 1.53, $p = .22$), nor an interaction with the Relatedness factor ($Fs < 1$), indicating that the priming effect was not modulated by the acoustic detail in the spoken primes. There was a difference between the average RT in the aligned and the misaligned conditions (599 and 571 ms respectively; F_1 (1,47) = 31.23, $p < .001$; F_2 (1,46) = 5.10, $p < .05$) and an interaction between Relatedness and Embedding Type (F_1 (1,47) = 20.57, $p < .001$; F_2 (1,46) = 6.85, $p < .05$), reflecting the fact that the effect of relatedness was smaller for the carrier words with misaligned embeddings than for those with the aligned embeddings (69 and 105 ms, respectively). However, as the RTs in Table 3 show, this difference is mostly due to the fact that in the unrelated conditions, participants reacted faster to carrier words with misaligned embeddings than they did to those with aligned embeddings. The carrier words with aligned embeddings were mostly bisyllabic and the carrier words with misaligned embeddings mostly

monosyllabic, so the former tended to be longer (in terms of number of phonemes and letters). The difference in word length might have elicited the interaction because the unrelated target words in the aligned condition were longer and required longer reading times.

Table 3. *Mean Reaction Times (RT, in milliseconds), Standard Errors (SE) and Percentage of Errors (Errors) in Experiment 2, in each Priming Condition, for each Embedding Type.*

Splice	Identity-spliced		Cross-spliced	
Relatedness	Related	Unrelated	Related	Unrelated
Experiment 2A				
Aligned	*pipet-PIPET*	*karkas-PIPET*	*pi**pet**-PIPET*	*kar**kas**-PIPET*
RT	549	659	544	644
SE	14	14	14	14
Errors	0.39%	0.91%	0.39%	0.91%
Misaligned	*speen-SPEEN*	*stang-SPEEN*	*s**peen**-SPEEN*	*s**tang**-SPEEN*
RT	541	602	532	609
SE	13	13	10	12
Errors	0.3%	0.82%	0.52%	0.87%
Experiment 2B				
Aligned	*pipet-PET*	*karkas-PET*	*pi**pet**-PET*	*kar**kas**-PET*
RT	653	640	664	616
SE	18	18	18	14
Errors	0.7%	0.4%	0.3%	0.2%
Misaligned	*speen-PEEN*	*stang-PEEN*	*s**peen**-PEEN*	*s**tang**-PEEN*
RT	689	654	695	656
SE	14	17	18	16
Errors	0.8%	0.8%	0.8%	1.0%
Experiment 2C				
Aligned Stressed	*pipet-PET*	*karkas-PET*	*pi**pet**-PET*	*kar**kas**-PET*
RT	679	628	674	635
SE	29	27	27	24
Errors	0.5%	0.7%	1.0%	1.1%
Aligned Unstressed	*bizon-ZON*	*cola-ZON*	*bi**zon**-ZON*	*co**la**-ZON*
RT	632	613	668	621
SE	24	28	27	27
Errors	0.7%	0.3%	0.4%	0.3%

Note. The visual targets were the carrier words in Experiment 2A and the embedded words in Experiments 2B and 2C.

Experiment 2B. RTs were again calculated from the onset of visual target presentation to response onset. Incorrect responses (2.6% of the data) and latencies exceeding 1200 ms (2.6%) were removed. Mean RTs, SEs and error rates for word targets in the four priming conditions are given in Table 3, separately for the carrier words with aligned and misaligned embeddings. Once again, no error analyses were carried out because performance was very accurate.

Overall, responses to related targets were slower than responses to unrelated targets (mean difference: -34 ms; F_1 (1,47) = 24.78, $p < .001$; F_2 (1,45) = 11.90, $p < .01$). Thus, responses to the visual targets (e.g., PET) were slower after a related prime (e.g., pipet) than after an unrelated prime (e.g., karkas). This inhibitory effect of relatedness was smaller with identity-spliced primes than with the cross-spliced primes (24 ms and 44 ms, respectively), but the interaction was not statistically significant at the 0.05 level (F_1 (1,47) = 2.96, $p = .09$; F_2 (1,45) = 1.70, $p = .20$). As the RTs in Table 3 indicate, the inhibitory effect of identity-spliced primes in the aligned condition was smaller than the effect of cross-spliced primes (-13 ms and -48 ms, respectively), while the identity- and cross-spliced primes produced approximately the same amount of inhibition in the misaligned condition (-35 ms and -39 ms). However, the three-way interaction (Embedding Type x Splicing x Relatedness) was not statistically significant (F_1 (1,47) = 2.13, $p = .15$; F_2 (1,45) = 1.74, $p = .19$). On average, responses to carrier words with aligned embeddings were faster than to those with misaligned embeddings (643 and 674 ms respectively; F_1 (1,47) = 22.02, $p < .001$; F_2 (1,45) = 3.80, $p = .057$). Note that in this experiment, in contrast to Experiment 2A, the visual targets in the misaligned condition were not shorter than in the aligned condition.

Experiment 2C. Incorrect responses (3.5% of the data) and latencies exceeding 1200 ms (1.5%) were removed. Mean RTs, SEs and error rates are given in Table 3, separately for the carrier words with stressed and unstressed embeddings. Analyses of RTs revealed a main effect of relatedness: Participants reacted on average 39 ms slower to related targets (e.g., the

target ZON after hearing bizon) than to unrelated targets (F_1 (1,23) = 20.90, $p < .001$; F_2 (1,28) = 8.11, $p < .01$). There was neither a main effect of Splicing ($F_1 = 1.02$, $p = .32$; $F_2 < 1$), nor an interaction with the Relatedness factor ($Fs < 1$). Responses to carrier words with stressed embeddings tended to be slower than to words with unstressed embeddings (654 ms vs. 633 ms), but this difference was significant only by subjects (F_1 (1,23) = 14.22, $p < .01$; $F_2 < 1$). No error analyses were carrier out because the error rates were again very low.

Discussion

The results of Experiment 2 showed that when the embedded word was presented as the visual target, lexical decision latencies were slower after a related prime (i.e., the carrier word) than after an unrelated prime. This result replicates the findings of Norris et al. (submitted), who found a similar inhibitory effect in English. Their proposed explanation was that the embedded word is initially activated, enters the competition process, and is subsequently suppressed when it loses the competition to the carrier word. The suppression of the embedded word causes responses to the visually presented embedded words to be slower. The inhibitory effect is thus taken as evidence for the activation of the offset-embedded words. All three types of embedded words (aligned stressed, aligned unstressed and misaligned) were found to exhibit this inhibitory effect. Furthermore, the inhibitory effect found for the words in the aligned stressed condition was of a similar magnitude in Experiments 2B and 2C, indicating the robustness of this effect.

As in Experiment 1, we did not observe an effect of the splicing manipulation. When the carrier words were presented as targets, listeners responses were facilitated to the same degree by the cross-spliced primes as by the identity-spliced primes. This suggests that the carrier words were activated to the same extent by both types of primes. When the embedded words were presented as targets, the inhibitory effect was of the same magnitude following the identity- and cross-spliced primes. This could indicate that the embedded words were

activated to the same degree by both versions of the prime. Alternatively, it is possible that losing the competition with the carrier word causes the embedded word to be suppressed to a certain level, regardless of how activated it was by the prime.

General Discussion

The conceptual and phonological activation of offset-embedded words was investigated using, respectively, eye-tracking and cross-modal identity priming. In Experiment 1, participants' eye-movements to pictures of the embedded words were monitored as they heard sentences mentioning the carrier words. In the subsequent experiments, listeners heard the carrier-word sentences and made lexical decisions to either the carrier words (Experiment 2A) or the embedded words (Experiments 2B and 2C). We examined words in which the embedded word was either aligned with the internal syllable boundary and stressed (e.g., pet in pipet), aligned with the internal syllable boundary and unstressed (e.g., zon in bizon) or misaligned with the syllable boundary (e.g., peen in speen). The contribution of fine-grained acoustic detail was explored by creating spliced versions of the carrier words, in which the origin of the embedded word was either another token of the carrier word or a recording of the monosyllabic embedded word.

The results of Experiment 1 showed that when the embedded word was aligned with the syllable boundary and unstressed, listeners looked more at the picture of the embedded word compared to a picture of a phonologically unrelated word. The current study thus makes a novel contribution by providing evidence for the conceptual activation of offset-embedded words (at least in the aligned unstressed condition) with a task that does not involve priming. This result constitutes converging evidence with previous priming studies that reported the conceptual activation of offset-embedded words (Isel & Bacri, 1999; Luce & Cluff, 1998; Shillcock, 1990; Vroomen & de Gelder, 1997).

In Experiment 2, listeners' lexical decision responses to the carrier words were faster after a related prime (i.e., the carrier word) than after an unrelated prime, while responses to the embedded words were slower after a related prime than after an unrelated one. The inhibitory effect found with the embedded words as visual targets indicates that the phonological representations of the embedded words must have been activated at some point by the carrier word prime (or else responses to them would not be different than those to targets after an unrelated prime). Furthermore, the inhibitory effect was as large for the three types of embeddings that were examined. Nevertheless, activation of the embedded words' conceptual representations, as reflected by fixations to a picture of that word (Experiment 1), was only found for the carrier words with aligned unstressed embeddings. Taken together, the results from the two experiments are in line with the suggestion of Norris et al. (submitted), who argued that spoken language comprehension involves processing at (at least) two distinct levels of representation: The mental lexicon contains separable phonological and semantic representations for each word, and activation of a word's phonological form does not necessarily entail the activation of the word's conceptual representation. Similar to the Norris et al. study, we find that phonological activation of an embedded word (reflected by the inhibitory priming effect) does not necessarily cause activation of that word's conceptual representation (no effect for the aligned stressed and misaligned embedded words in Experiment 1).

Contrary to our prediction, conceptual activation was detected only for the carrier words with unstressed embeddings. The acoustic analyses indicated that the realization of the embedded sequences in these carrier words was less like the realization of the monosyllabic words themselves, compared to the offset-embedded words in the carrier words with stressed embeddings. Thus, despite the high acoustic/phonetic similarity between the second syllable of the carrier words with stressed embeddings and their monosyllabic counterpart words, listeners did not fixate the pictures of the embedded words more than the distractors in this

condition. One possible explanation for this finding is based on the fact that the lexical stress pattern of the carrier words (i.e., whether stressed is on the first or second syllable) is confounded with the lexical neighbourhood of the carrier word, most notably the carrier word's cohort and its density. Generally, the level of activation of a particular candidate is not just a function of its goodness of fit with the current signal but also of the number of other active lexical candidates and their goodness of fit. Our analyses showed that at the onset of the embedded word, the cohort size and density of the carrier words with stressed embeddings was lower than that of the carrier words with unstressed embeddings. A sparse cohort with low density would result in a higher level of activation of the carrier word, relative to a carrier word with a dense cohort. A highly activated carrier word will win the lexical competition (with other word candidates, amongst which the offset-embedded word) more rapidly than a carrier word which is less highly activated. In other words, it is likely that due to the difference in cohort sizes between the carrier words with stressed and unstressed embeddings, the stressed embedded words lose the competition with the carrier words faster than the unstressed embedded words. This explanation hinges on the idea that phonological activation needs to subsist for a minimal period of time for conceptual activation to take place. If the activation of the carrier word is very high, the competition process could resolve too fast for the phonological activation of the embedded word to spread to the conceptual level. Indeed, the only trace of the activation of the stressed embedded words was the inhibitory effect in Experiment 2, indicating that the embedded word has lost the competition with the carrier word. While this explanation is tentative – it is not certain that the cohort densities are causing the obtained pattern of results – it is plausible that cohort density plays an indirect role in the conceptual activation of offset-embedded words, by influencing the dynamics of the competition process. It is important to note that the difference in cohort sizes and densities between the carrier words with first-syllable and second-syllable stress is a natural pattern in Dutch and not a peculiarity of the items we used in this study. The overwhelming majority of

Dutch words (87%) begin with a strong syllable (Schreuder & Baayen, 1994). It is therefore likely that stress-initial words will have larger and denser cohorts than words with an unstressed first syllable. Separating the effects of stress from the effects of cohort size and density is an arduous task. Given the constraints on the items used in the current study, it was certainly not possible to disentangle the two factors.

The inhibitory effect we observed in Experiment 2 with the misaligned embedded words indicates the phonological activation of these words. In Experiment 1, we observed no evidence for the activation of the misaligned embedded words. Vroomen and de Gelder (1997) found that offset-embedded words that were misaligned with the syllable boundary did not facilitate lexical decision reactions to associatively related words. Together, these findings suggest that the phonological representations of misaligned embedded words are activated by carrier words but their semantic representations are not.

The importance of syllable boundaries for lexical access has been demonstrated by many studies (e.g., Content, Kearns & Frauenfelder, 2001; Dumay et al., 2002; McQueen, 1998; Weber & Cutler, 2006), all indicating that word recognition is delayed when words are misaligned with a syllable boundary. It has been suggested (Norris, McQueen, Cutler & Butterfield, 1997) that misalignment with a syllable boundary such that there are no vowels between the onset of the candidate word and the syllable boundary causes that candidate word to be disfavoured. According to this proposal, phonological representations of lexical candidates that are consistent with the input are activated, but if they leave a stretch of the input consisting solely of consonants then their activation is reduced. Syllable boundaries thus help segmentation indirectly by biasing the competition process against misaligned words. While the data that we have obtained in the current study do not speak to whether the level of activation of misaligned lexical candidates is reduced, they do indicate that misaligned embedded words are phonologically activated, enter the competition process and lose that competition.

In neither experiments did we find an effect of the splicing manipulation. In Experiment 1, there was no evidence for the activation of the aligned stressed and misaligned embedded words, even when listeners heard the cross-spliced stimuli, in which the origin of the embedded word was the monosyllabic word. Gow and Gordon (1995) argued that offset-embedded words do not prime related words (e.g., no priming from tulips to words associated with lips) due to the fact that "listeners do not access the meanings of words that begin at syllable boundaries but lack the special acoustic marking of word onsets" (p. 352). The results of Experiment 1 indicate, however, that the acoustic marking of a word onset at a syllable boundary does not automatically lead to the activation of the meaning of that word. An acoustic match with the signal is not a sufficient condition for the conceptual activation of an offset-embedded word. Furthermore, in the case of the carrier words with the aligned unstressed embeddings, the splicing manipulation did not modulate the amount of fixations that listeners made to the pictures of the embedded words. Studies investigating onset-embedded words have shown that listeners are sensitive to fine-grained acoustic differences between onset-embedded words and monosyllabic words (Davis et al., 2002; Salverda et al., 2003; Salverda, 2005). In contrast to these findings with onset-embedded words, the subtle acoustic differences between offset-embedded words and monosyllabic words do not appear to influence the lexical competition. This is probably because onset-embeddings and offset-embeddings do not pose comparable problems to the listener: onset-embedded words are strongly supported by the signal, with only fine-grained acoustic detail to bias the competition between the onset-embedded word and the actual intended word; offset-embedded words are always at a disadvantage relative to the intended word, because the intended word matches more of the signal, making the use of fine-grained acoustic information redundant.

The issue of whether offset-embedded words are activated during spoken-word recognition has received considerable attention because different models of continuous speech recognition have made different claims regarding these words. In the early Cohort model (Marslen-

Wilson & Welsh, 1978), one of the first and most influential models, an offset-embedded word would not be activated because its initial mismatch with the signal would prevent it from entering the cohort. Other models, and more recent versions of Cohort, have been more lenient regarding the activation of offset-embedded words. In addition, different studies have resulted in different and contradictory findings. The present state of our knowledge about spoken-word recognition suggests that asking whether offset-embedded words become activated or not is too general a question. We are now in a position to ask which representations of the embedded word are activated, under which conditions does this occur, and which factors influence this process. The findings we present here suggest that offset-embedded words are automatically activated at the phonological level and enter the lexical competition with the carrier word. Activation of the embedded word's conceptual representation can occur under certain circumstances, but is by no means obligatory.

Appendix A

Stimulus sets used in Experiment 1.

Carrier words with aligned stressed embeddings

Target	Competitor	Distractor	Distractor
cognac (cognac)	jak (yak)	rasp (grater)	potlood (pencil)
diamant (diamond)	mand (basket)	bom (bomb)	radio (radio)
galei (galley)	lei (slate)	schroef (screw)	flacon (flask)
kameel (camel)	meel (flour)	bruid (bride)	piano (piano)
kanon (cannon)	non (nun)	mus (sparrow)	frambozen (raspberry)
karkas (carcass)	kas (greenhouse)	wieg (cradle)	fontein (fountain)
kornet (cornet)	net (net)	riem (belt)	longen (lungs)
lakei (lackey)	kei (boulder)	tent tent)	vliegtuig (airplane)
libel (dragonfly)	bel (bell)	pijp (pipe)	tijger (tiger)
olijf (olive)	lijf (body)	bier (beer)	strijkplank (ironing board)
pipet (pipette)	pet (cap)	vlag (flag)	asperge (asparagus)
pupil (pupil)	pil (pill)	helm (helmet)	weegschaal (scale)
sate (satay)	thee (tea)	dak (roof)	pistool (pistol)

Carrier words with aligned unstressed embeddings

aambeeld (anvil)	beeld (statue)	kaars (candle)	mijter (miter)
aardbei (strawberry)	bij (bee)	pop (doll)	vrachtwagen (truck)
bizon (bison)	zon (sun)	kast (closet)	pleister (plaster)

circus (circus)	kus (kiss)	rok (skirt)	asbak (ashtray)
cola (cola)	la (drawer)	gesp (buckle)	sambaballen (maracas)
haring (herring)	ring (ring)	jas (jacket)	varen (fern)
kubus (cube)	bus (bus)	mes (knife)	radijs (radish)
motor (motorbike)	tor (beetle)	spuit (syringe)	brandkraan (fire hydrant)
python (python)	ton (barrel)	klok (clock)	sigaar (cigar)
robot (robot)	bot (bone)	kroon (crown)	perzik (peach)
toekan (tucan)	kan (pitcher)	bloem (flower)	kasteel (castle)

Carrier words with misaligned embeddings

clip (clip)	lip (lip)	appel (apple)	fakkel (torch)
fruit (fruit)	ruit (rhombus)	koffer (suitcase)	cadeau (gift)
kaas (cheese)	aas (ace)	gieter (watering can)	iglo (iglo)
kegel (skittle)	egel (hedgehog)	zaag (saw)	sla (lettuce)
kluis (safe)	luis (louse)	trechter (funnel)	scepter (scepter)
kraam (stall)	raam (window)	sleutel (key)	vogelhuis (birdhouse)
krat (crate)	rat (rat)	bliksem (lightning)	avocado (avocado)
snavel (beak)	navel (bellybutton)	muts (woolen hat)	zweep (whip)
speen (pacifier)	peen (carrot)	nijlpaard (hippo)	bezem (broom)
speer (javelin)	peer (pear)	fornuis (oven)	cocktail (cocktail)
spier (muscle)	pier (worm)	beitel (chisel)	meloen (mellon)
spin (spider)	pin (pin)	hersenen (brain)	bureau (desk)
spion (spy)	pion (pawn)	mier (ant)	tak (branch)

spray (spray)	ree (roe deer)	kano (canoe)	graf (grave)
staart (tail)	taart (cake)	anker (anchor)	ratel (rattle)
stang (rod)	tang (pliers)	vleermuis (bat)	strijkijzer (iron)
steen (stone)	teen (toe)	viool (violin)	lepel (spoon)
stempel (stamp)	tempel (temple)	lamp (lamp)	das (tie)

Appendix B

Carrier-word and monosyllabic-word sentences used for the stimuli in Experiment 1.

Carrier words with aligned stressed embeddings

1. Ik geloof dat de duurdere COGNAC er niet meer is

   Ik geloof dat de dure Ikon JAK er niet meer is

2. Hij had de DIAMANT meegenomen

   Hij had de dia MAND meegenomen

3. Ik had nog nooit een mooiere GALEI gezien

   Ik had nog nooit een grote fuga LEI gezien

4. Ze wilde een KAMEEL in de woestijn zien

   Ze wilde mokka MEEL in het deeg doen

5. Hij zei dat een klein KANON indrukwekkender was

   Hij zei dat een Inka NON indrukwekkender was

6. Hij zag dat een stokoud KARKAS verdwenen was

   Hij zag dat de touringcar KAS verdwenen was

7. Je mag die gewone KORNET nog niet bespelen

   Je mag het toneeldecor NET nog niet weghalen

8. Ze zagen de LAKEI op en neer lopen

   Ze zag de lila KEI op de grond liggen

9. Wij dachten dat die LIBEL kon vliegen

   Hij dacht dat die olie BEL kon bevriezen

10. Hij vond een grote OLIJF in zijn drankje

Hij vond een farao LIJF in de woestijn

11. Ze kon de grotere PIPET niet vinden

    Ze kon de grote hippie PET niet vinden

12. Ze zei dat die donkere PUPIL te groot was

    Ze zei dat die witte braadjus PIL te zout was

13. Wij vonden die SATE niet zo lekker

    Ik vond die mensa THEE niet zo lekker

Carrier words with aligned unstressed embeddings

1. Hij wilde een AAMBEELD voor me halen

   Hij wilde een NAAM-BEELD voor me maken

2. Hij zei dat die AARDBEI kon groeien

   Hij zei dat die HAARDBIJ kon steken

3. Hij vertelde dat de oude BIZON nog leefde

   Hij vertelde dat de AMFIBIE ZON nog op ging

4. Ze zei dat een CIRCUS haar heel leuk leek

   Ze zei dat een BIER KUS haar heel vies leek

5. Ik wist dat die nieuwe COLA heel raar smaakte

   Ik wist dat die ROCOCO LA heel zeldzaam was

6. Wij dachten dat die HARING niet vers was

   Zij dacht dat die BEHA RING niet dicht was

7. Met de KUBUS kun je niet zo lang spelen

Met de Q BUS kun je naar het station rijden

8. Ze zeiden dat die MOTOR in de weg stond

   Ze zei dat die PLUMEAU TOR in haar kleding zat

9. Hij zei dat die grote PYTHON zelden doodt

   Hij zei dat die THERAPIE TON zelden valt

10. Ze dachten dat de ROBOT hen zou bedienen

    Ze dacht dat het TAROT BOT haar geluk zal brengen

11. Ik vertelde dat deze TOEKAN kon praten

    Ik vertelde dat de TAHOE KAN kon breken

Carrier words with misaligned embeddings

1. Ze zeiden dat die CLIP van hen was

   Ze zei dat die hoek LIP veel pijn deed

2. Ik wilde graag nieuw FRUIT proberen

   Ik wilde een duif-RUIT proberen

3. Hij had die oude KAAS meegenomen

   Hij had een oud lok AAS meegenomen

4. Hij dacht dat die grotere KEGEL zou vallen

   Hij dacht dat die buitenwijk-EGEL zou lopen

5. Ik wilde graag die KLUIS van hem kopen

   Ik wilde die boek-LUIS van de plank weghalen

6. Ze dacht dat die nieuwe KRAAM niet zo goed stond

Ze dacht dat een strodak-RAAM niet zou helpen

7. Ik probeerde een groter KRAT op te tillen

   Ik probeerde de waterdijk-RAT op te sporen

8. Wij zagen dat de SNAVEL open ging

   Hij zag dat de buis NAVEL open ging

9. Ze vertelde dat de SPEEN op de grond was gevallen

   Ze vertelde dat bos PEEN op het eiland kon groeien

10. Ze dacht dat die kleinere SPEER licht zou zijn

    Ze dacht dat die kleine was PEER licht zou zijn

11. Ze zeiden dat die SPIER niet verrekt was

    Ze zei dat die bos PIER niet meer leefde

12. Ze hadden die SPIN buiten gezet

    Ze had die vlees-PIN buiten gezet

13. Ik dacht dat die oude SPION niet mocht komen

    Ik dacht dat die speelhuis PION niet kon breken

14. Wij dachten dat die SPRAY heel goedkoop was

    Hij dacht dat die wesp-REE heel zeldzaam was

15. Hij zag die kleine STAART op en neer bewegen

    Hij zag die roomkaas-TAART op de tafel staan

16. Hij wilde die grote STANG aan mij geven

    Hij wilde die verlos TANG aan mij geven

17. Hij zei dat een grote STEEN veel beter was

Hij zei dat een steunkous TEEN veel pijn zou doen

18. Zij dacht dat die nieuwe STEMPEL mooier zou zijn

Zij dacht dat die Venus TEMPEL mooier zou zijn

Appendix C

Carrier-word and monosyllabic-word items used in Experiment 2.

Carrier words with aligned stressed embeddings

Carrier word	Embedded word	Unrelated Target
diamant (diamond)	mand (basket)	poen
fazant (pheasant)	zand (sand)	pil
geweer (rifle)	weer (again)	maat
gewei (antlers)	wei (meadow)	poen
gezin (family)	zin (sense)	meel
harpoen (harpoon)	poen (cash)	mand
kameel (camel)	meel (flour)	zin
kanon (cannon)	non (nun)	lijf
karkas (carcass)	kas (greenhouse)	pet
kompas (compass)	pas (just)	bel
lakei (lackey)	kei (boulder)	thee
libel (dragonfly)	bel (bell)	pas
olijf (olive)	lijf (body)	non
pipet (pipette)	pet (cap)	kas
piraat (pirate)	raad (advice)	vet
pompoen (pumpkin)	poen (cash)	wei
pupil (pupil)	pil (pill)	zand
sate (satey)	thee (tea)	kei

servet (napkin)	vet (fat)	raad
tomaat (tomato)	maat (size)	weer

Carrier words with aligned unstressed embeddings

aambeeld (anvil)	beeld (statue)	kan
aardbei (strawberry)	bij (bee)	ton
bizon (bison)	zon (sun)	la
circus (circus)	kus (kiss)	bot
cola (cola)	la (drawer)	zon
haring (herring)	ring (ring)	boom
kubus (cube)	bus (bus)	tor
kerstboom	boom	ring
motor (motorbike)	tor (beetle)	bus
python (python)	ton (barrel)	bij
robot (robot)	bot (bone)	kus
toekan (tucan)	kan (pitcher)	beeld

Carrier words with misaligned embeddings

clip (clip)	lip (lip)	rand
fles (bottle)	les (lesson)	taart
friet (French fries)	riet (straw)	tempel
fruit (fruit)	ruit (rhombus)	teen

gras (grass)	ras (race)	lof
kleed (rug)	leed (grief)	lak
kluis (safe)	luis (louse)	pier
kraam (stall)	raam (window)	navel
krat (crate)	rat (rat)	peer
kruk (stool)	ruk (jerk)	pion
prei (leek)	rij (row)	lot
schil (peel)	gil (scream)	pin
slak (snail)	lak (varnish)	leed
slang (snake)	lang (long)	ree
slof (slipper)	lof (praise)	ras
slot (lock)	lot (fate)	rij
snavel (beak)	navel (bellybutton)	raam
speen (pacifier)	peen (carrot)	tang
speer (javelin)	peer (pear)	rat
spier (muscle)	pier (worm)	luis
spin (spider)	pin (pin)	gil
spion (spy)	pion (pawn)	ruk
spray (spray)	ree (roe deer)	lang
staart (tail)	taart (cake)	les
stang (rod)	tang (pliers)	peen
steen (stone)	teen (toe)	ruit

stempel (stamp)	tempel (temple)	riet
strand (beach)	rand (edge)	lip

Note. The carrier words were used as primes throughout the series and as related targets in Experiment 2A. The embedded words were used as related targets in Experiments 2B and 2C. The unrelated targets in Experiments 2B and 2C were the embedded words, as listed; in Experiment 2A, the corresponding carrier words served as the unrelated targets.

Prosodic knowledge affects the recognition of newly-acquired words

CHAPTER 6

An adapted version of this chapter will appear in *Psychological Science* (Shatzman & McQueen, in press).

Abstract

An eye-tracking study examined the involvement of prosodic knowledge – specifically, that monosyllabic words tend to have longer durations than the first syllables of polysyllabic words – in the recognition of newly-learned words. Participants learned new spoken words (by associating them to novel shapes): bisyllables and onset-embedded monosyllabic competitors (e.g., baptoe and bap). In the learning phase, the duration of the ambiguous sequence (e.g., bap) was held constant. In the test phase, it was longer, shorter or equal to the learning phase duration. Listeners' fixations indicated that short syllables tended to be interpreted as the first syllables of the bisyllables, while long syllables generated more monosyllabic word interpretations. The real-word neighbourhoods of the newly-acquired words modulated this effect. Recognition of newly-acquired words is influenced by prior prosodic knowledge – in the absence of such prosodic information in the exposure – and is therefore not determined solely on the basis of episodes of those words.

Introduction

Are newly-acquired words processed like well-known words? Throughout our lives we continue learning new words. This ability is fundamental to our language capacity. But how does word learning relate to the existing word-recognition system? Specifically, does prior knowledge about words influence the recognition of new words?

The present study addresses these questions by examining ambiguity resolution in speech comprehension, using newly-learned words. Current models agree that, during word recognition, multiple lexical candidates consistent with the acoustic-phonetic information in the speech signal become activated and compete among one another (e.g., Allopenna, Magnuson & Tanenhaus, 1998; Marslen-Wilson, 1987; McQueen, Norris & Cutler, 1994; Zwitserlood, 1989). Therefore, a certain degree of ambiguity resolution is required in all sentences.

Onset-embedded words, such as ham in hamster, are a critical case. The phonemic overlap between such words suggests that recognition of the embedded word could only occur after its offset. Fine-grained acoustic information can, however, bias the lexical competition in favour of the correct interpretation (Davis, Marslen-Wilson & Gaskell, 2002). Using eye-movement data, Salverda, Dahan and McQueen (2003) demonstrated that the duration of the ambiguous sequence (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, Salverda et al. showed that longer sequences generated more monosyllabic-word interpretations, while shorter durations generated more polysyllabic-word interpretations (see also Salverda, 2005). Salverda et al. argued that these durational differences reflect the prosodic structure of the utterance and that listeners compute this prosodic structure during the perception process.

We examined here whether listeners would display sensitivity to such prosodic information in their recognition of newly-learned words. Participants learned new spoken words (by

associating them to novel shapes) along with new onset-embedded competitors (e.g., baptoe and bap). In the learning phase, the duration of the ambiguous sequence (e.g., bap) was held constant. In the test phase, we examined whether manipulating the sequence's duration would modulate the amount of monosyllabic and polysyllabic-word interpretations, as it does with existing words (Salverda et al., 2003).

The answer to this question is critical for an ongoing debate regarding the format of lexical representations. Some models of spoken-word recognition assume that words are represented in the lexicon in some phonologically-abstract form (e.g., McClelland & Elman, 1986; Gaskell & Marslen-Wilson, 1997; Norris, 1994). Other authors have suggested that the lexicon contains multiple exemplars, in the form of detailed acoustic traces of specific episodes of each word (e.g., Goldinger, 1998; Johnson, 1997a,b; Pierrehumbert, 2001, 2002).

We investigated whether listeners' recognition of newly-acquired words is determined only by the experience they have had with those words, that is, based on the stored episodes of those words, or whether recognition is also determined by prior experience with similar-sounding real words. We employed the eye-tracking paradigm (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995), which has recently been used with an artificial lexicon (Magnuson, Tanenhaus, Aslin & Dahan, 2003). In that study, participants learned to associate novel shapes with novel words, some of which were cohorts and rhymes of each other (e.g., pibo, pibu and dibo). Participants' eye movement patterns when hearing the novel words showed cohort and rhyme effects similar to those observed with existing words (Allopenna et al., 1998). Furthermore, manipulating the new words' occurrence frequency during the learning phase elicited a differential fixation pattern to targets of high and low frequency, replicating frequency effects found with real words (Dahan, Magnuson & Tanenhaus, 2001).

The current study examined whether recognition of newly-acquired words is influenced by prosodic knowledge that was not present in the learning phase. If recognition is based purely on the episodes that listeners are exposed to, the newly-acquired words should be recognized

fastest when listeners hear the exact recordings they heard in the learning phase. If, however, there is transfer of knowledge about the relative duration of syllables in existing words (i.e., that monosyllabic words tend to have longer durations than the initial syllables of polysyllabic words), we would expect more monosyllabic-word interpretations when the ambiguous sequence (e.g., bap) is long, while shorter durations should generate more polysyllabic-word interpretations.

A crucial question is the extent to which the listener's native lexicon is involved. Although the critical stimuli in the experiment are all new words, there are reasons to assume that existing words will also be temporarily activated. Several eye-tracking studies with bilinguals have shown that the native lexicon is active in second-language listening, that is, even when it is irrelevant to performing the task (Spivey & Marian, 1999; Weber & Cutler, 2004). Thus, we investigated whether certain aspects of the novel words' lexical neighbourhood correlated with the eye-tracking results.

Our primary question, therefore, is whether phonological knowledge – about prosodic structure and about the form of known words – is used in processing newly-acquired words. Is recognition of such words determined largely (or solely) by experience with those words, as the episodic view would predict, or is abstract phonological knowledge also brought to bear?

Method

Participants. Twenty-four Max-Planck-Institute subject pool volunteers, all Dutch native speakers, were paid for their participation.

Materials. Twenty line-drawings of nonsense objects were randomly selected from a database of non-objects (see Figure 1 for examples). Ten CVC Dutch nonwords (e.g., bap) were selected as monosyllabic novel words. Ten bisyllabic novel words were constructed by

adding a second syllable to these monosyllables (e.g., baptoe). The items are listed in Table 1. The nonsense-object pictures were randomly assigned to the novel words.

Table 1. *Mean duration (in ms) of the monosyllabic words, the first syllables of the bisyllabic words, and their average used in the training version.*

Item Pair	Phonetic Transcription	Monosyllabic	Bisyllabic	Training Version
bap-baptoe	bɑp-bɑptu	316	266	292
fiem-fiemser	fiːm-fiːmsɛʀ	358	252	305
jom-jomtie	jɔm-jɔmtiː	320	237	278
kes-keste	kɛs-kɛstə	322	257	289
kuin-kuinwes	kœyn-kœynwɛs	312	288	300
nim-nimsel	nɪm-nɪmsəl	329	221	275
soer-soerket	suːʀ-suːʀkɛt	386	281	333
taaf-taafpag	taːf-taːfpɑx	335	273	304
tuik-tuikfom	tœyk-tœykfɔm	331	228	280
zaf-zafkes	zɑf-zɑfkəs	362	277	313
AVERAGE		362	276	

The auditory stimuli for the learning and the test phases were spoken instructions to click on the picture of the novel word and then on one of four geometric forms appearing on the screen (see Figure 1). The novel words appeared in sentence-medial position, preceded by the phrase "Klik op de___" (e.g., Klik op de bap [Click on the bap]) and followed by "en dan op de___" (e.g., en dan op de driehoek [and then on the triangle]). Each novel word appeared in four sentences (once with each geometric form). Twenty feedback sentences were constructed (e.g., Hier zie je de bap nog een keer [Here you can see the bap again]).

All sentences were produced by a female native Dutch speaker in a sound-attenuated booth and recorded directly onto computer (sampling at 44.1 kHz with 16-bit resolution). The durations of the monosyllabic words and the first syllables of the bisyllabic words were measured and averaged across all four recordings of the sentences in which they appeared (see Table 1). As expected, the monosyllabic words were longer than the first syllables of the bisyllabic words. Three versions of the monosyllabic words, varying in their duration, were

created using the PSOLA (Pitch-Synchronous Overlap and Add) resynthesis method in the Praat speech editor (http://www.praat.org). For the training version the first syllable of the bisyllabic word was excised from the sentential context and resynthesized such that its duration was halfway between the average of that syllable in the monosyllabic and bisyllabic sentences. For the long version the same token of the syllable was resynthesized such that its duration was the average of that syllable in the monosyllabic sentences. The short version's duration was its average duration in the bisyllabic sentences.

The syllables were then spliced back into the sentential contexts. The phrase preceding the manipulated token ("Klik op de___") was always taken from an utterance with a bisyllable. In the monosyllabic sentences, the manipulated syllables were followed by the phrase "en dan op de___", and the name of each of the four geometric forms. In the bisyllabic sentences, the manipulated syllables were followed by the second syllable of the word, the phrase "en dan op de___", and each of the four geometric forms. In total, 240 instruction sentences were created. The feedback sentences were created in a similar fashion, except that only the training versions of the syllables were used.

Procedure. Participants were tested individually. The learning phase consisted of six training blocks with feedback. The eye-tracker (an SMI Eyelink system, sampling at 250 Hz) was then mounted and calibrated, and the test phase, comprising two blocks without feedback, followed. The experiment was controlled by a Compaq 486 computer. Pictures were presented on a ViewSonic 17PS screen, and the auditory stimuli were presented over headphones using NESU software (http://www.mpi.nl/world/tg/experiments/nesu.html).

Each trial was structured as follows. A central fixation dot appeared on the screen for 500 ms. Then a spoken sentence was presented and simultaneously a 5x5 grid with the pictures of the novel objects and the geometric forms appeared on the screen (see Figure 1). In the first three training blocks participants had to choose one of two pictures. In the last three training blocks and in the test phase they had to choose from four pictures. In a training trial, as soon

as participants clicked on the geometric form, the distractor pictures disappeared, leaving only the correct referent on display. At the same time, a sentence was played indicating to the participants whether their response was correct or incorrect (Dat was goed/fout [that was right/wrong]), followed by the feedback sentence. The trials in the test phase were identical to the trials in the last three training blocks, except that no feedback was given.

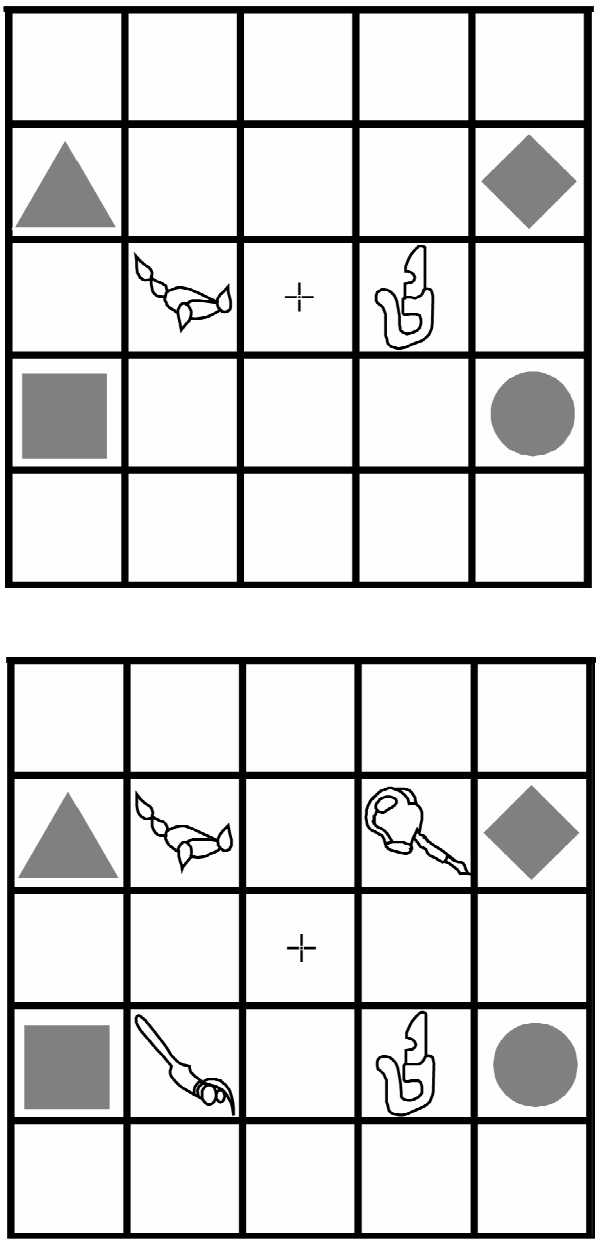


Figure 1. *Example of stimulus display presented to participants in a two-alternative forced choice trial (top) and a four-alternative forced choice trial (bottom).*

The learning phase consisted of six training blocks, in each of which each word was presented three times (in total, 360 trials). For each trial, one or three items were selected randomly from the set of nonsense objects to serve as distractors. A random order was created for each block, with the constraint that at least five items intervened between two presentations of the same item, or between paired items (e.g., bap and baptoe). The order of presentation in the training phase was identical for all participants.

In the test phase, there were 60 experimental trials. Each word was presented three times as a target (i.e., with pictures of the target and competitor on the screen). On these trials the participants heard either the training, long or short versions. Each participant heard all three versions. Additionally, there were 120 filler trials in which the three distractors were unrelated to the target. In these trials participants heard the training versions. The distractors for both filler and experimental trials were randomly selected. A random order was created for the test block, with the constraints that at least one filler item intervened between two experimental items and at least five intervening items between paired items (e.g., bap and baptoe). Six lists were created, each containing 180 trials. The lists varied on the order in which the training, long or short versions of each item were presented. Participants were randomly assigned to one list.

Results and Discussion

Analysis of fixations over time. On 73 experimental trials (5%), participants erroneously selected an object other than the target picture. These trials were excluded from further analyses. For each participant and each trial, fixations were coded as pertaining to the target object, to the competitor, to one of the two unrelated distractors, or to anywhere else on the screen. Fixations were coded from the onset of the target word until the last fixation to the target picture before the participant clicked on it. The proportion of fixations to each type of picture were computed by summing the number of trials in which a particular type of picture

was fixated (in each 10 ms slice, in each condition), and dividing it by the total number of trials (in the same time interval) in which any picture or location was fixated.

The proportions of fixations to the target, the competitor and the averaged distractor are shown in Figure 2, separately for the long, training and short versions. Figure 3 shows the proportions of fixations for the three versions, separately for the targets and the competitors. All figures show fixation proportions in 20 ms time slices from target onset to 1400 ms thereafter.

In all conditions the competitor was fixated more than the averaged distractor. The pattern differs, however, between the bisyllabic and monosyllabic conditions: while in the bisyllabic condition the competitor was fixated most in the long version, in the monosyllabic condition it was fixated the most in the short version. This can be clearly seen in Figure 3. Figure 3B shows that participants looked more at the picture of the monosyllabic competitor (e.g., bap) when they heard the long version of the bisyllabic target (e.g., baptoe) than when they heard the short version. The training version was intermediate. When the target was monosyllabic, participants looked more at the bisyllabic competitor (Figure 3D) when they heard the short version than when they heard the long version. In this condition, the training version did not seem to differ from the long version.

The target fixation proportions in the bisyllabic condition (Figure 3A) reveal a pattern that is the mirror image of the competitor fixations: participants fixated the bisyllabic target most when they heard the short version, least when they heard the long version, and intermediate when they heard the training version. The pattern of fixations in the monosyllabic condition (Figure 3C), however, is not entirely complementary to the pattern of competitor fixations: participants fixated the target most when they heard the training version, less when they heard the long version and least when they heard the short version.

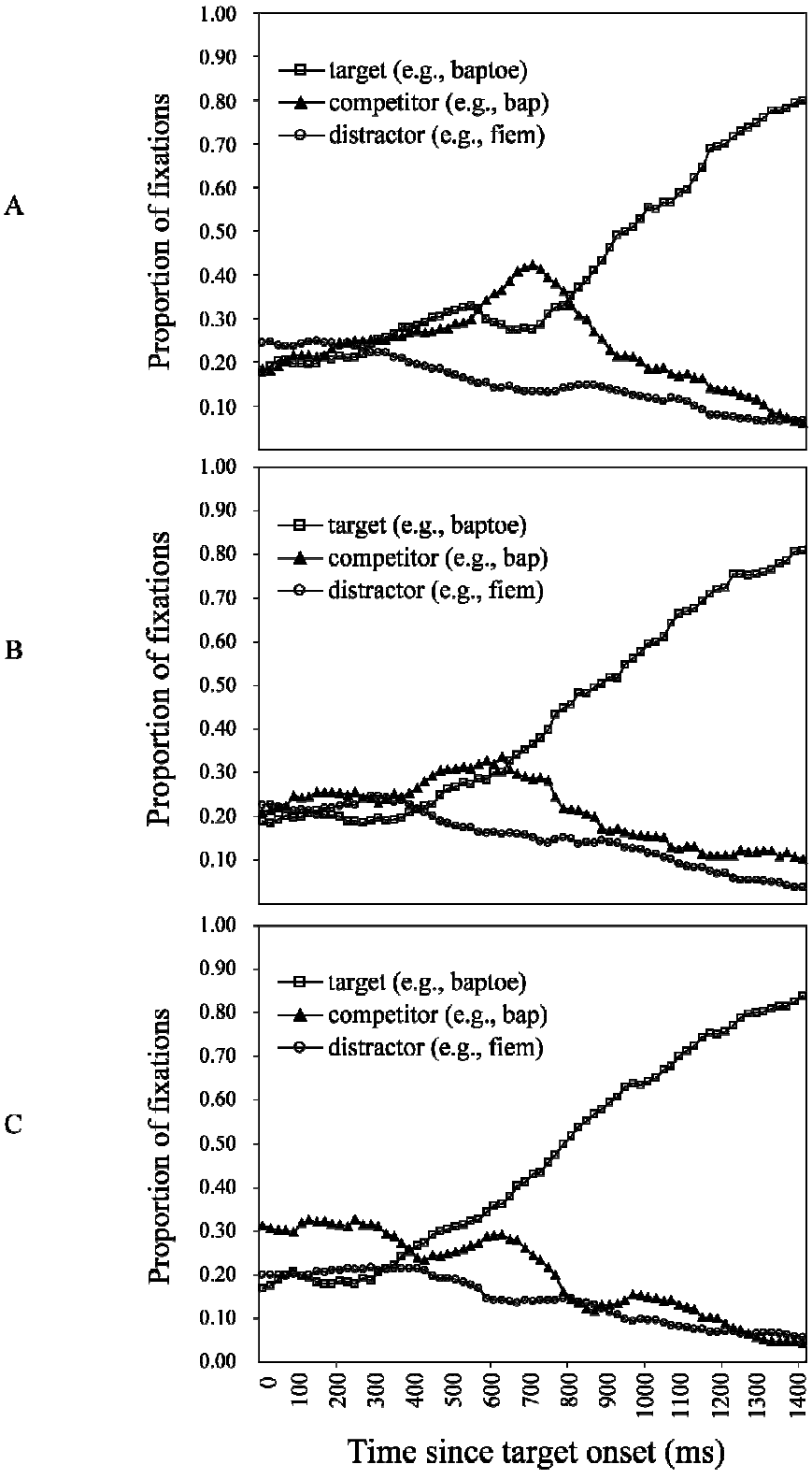
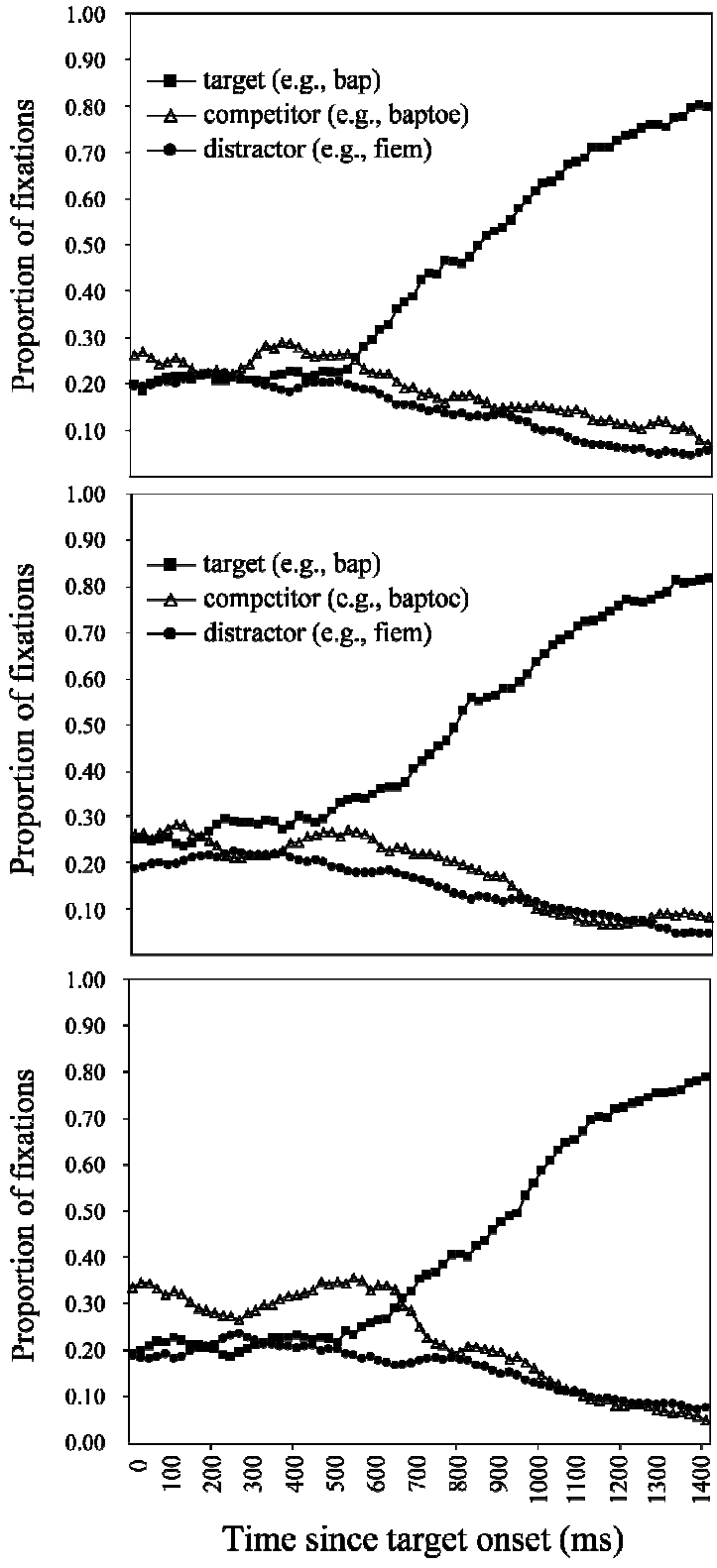


Figure 2. *Fixation proportions over time to the target, the competitor and the averaged distractors, as a function of the target's number of syllables (bisyllabic: A-C; monosyllabic: D-F) and the target's duration (long version: A and D; training version: B and E; short version: C and F).*

D

E

F


target (e.g., bap)
competitor (e.g., baptoe)
distractor (e.g., fiem)
Proportion of fixations
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Time since target onset (ms)

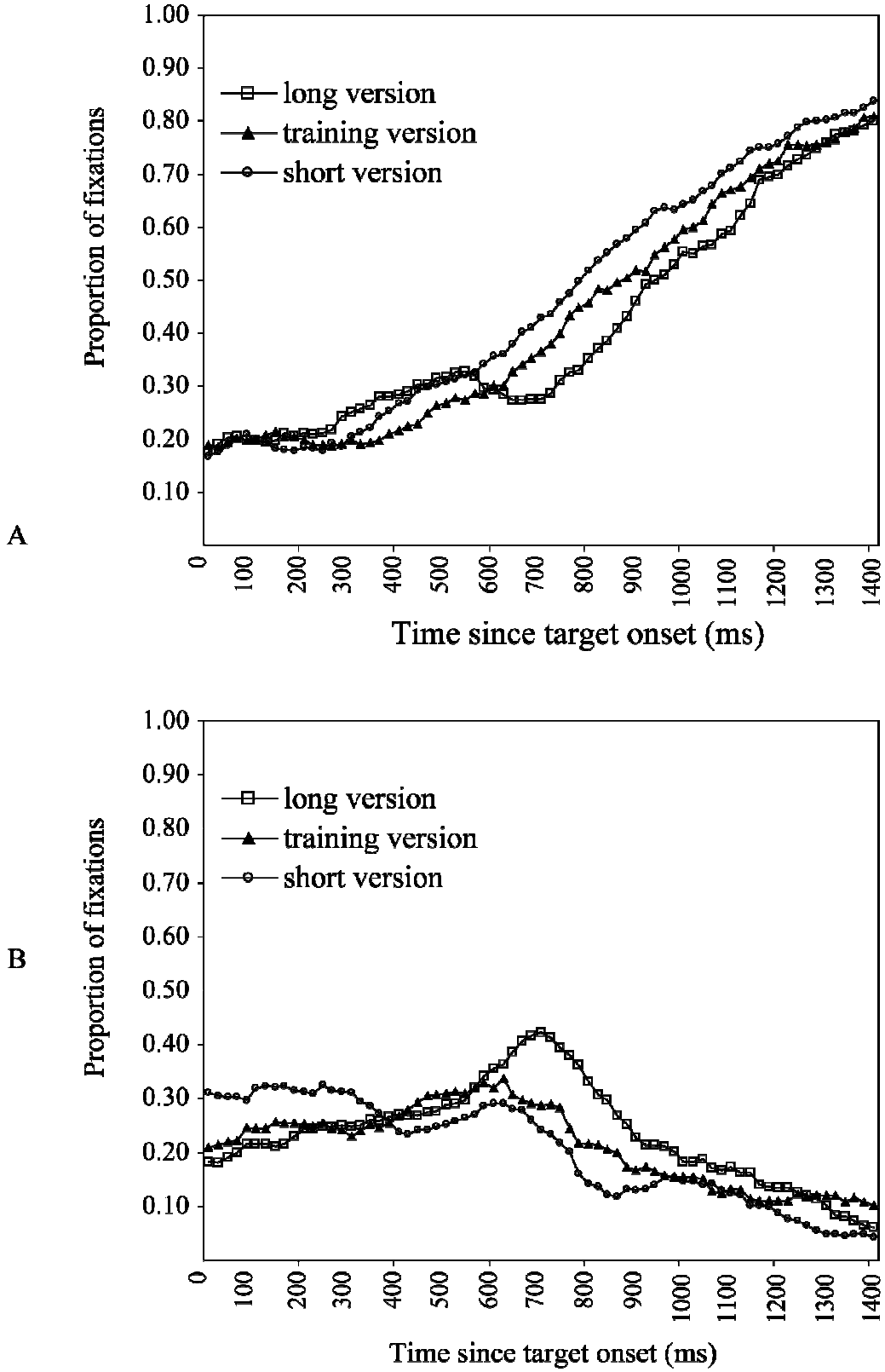
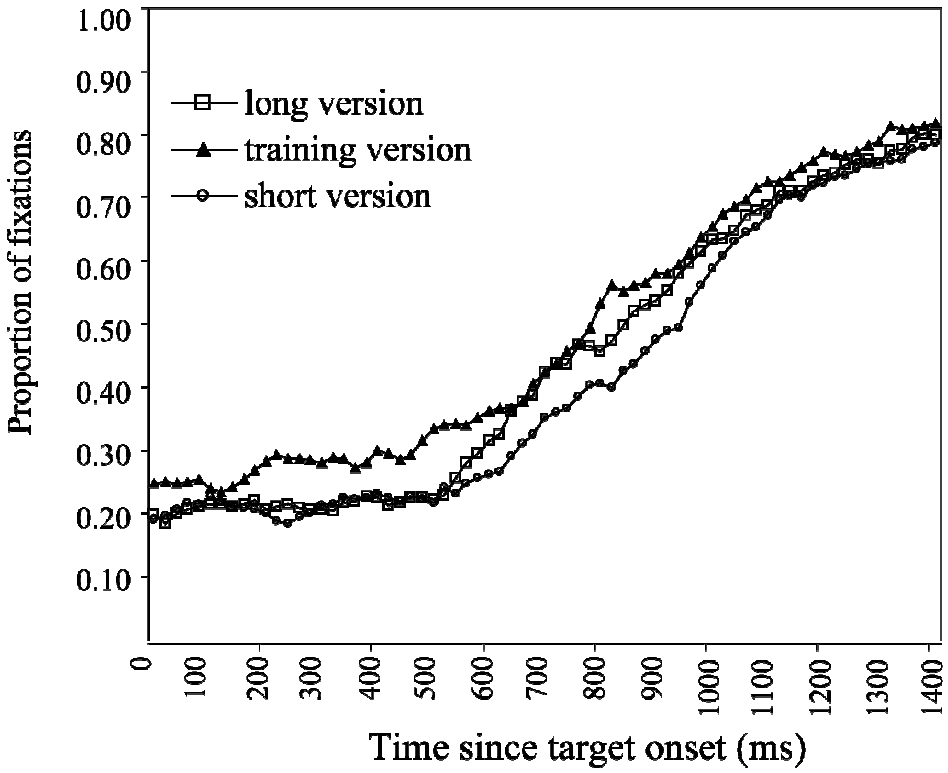
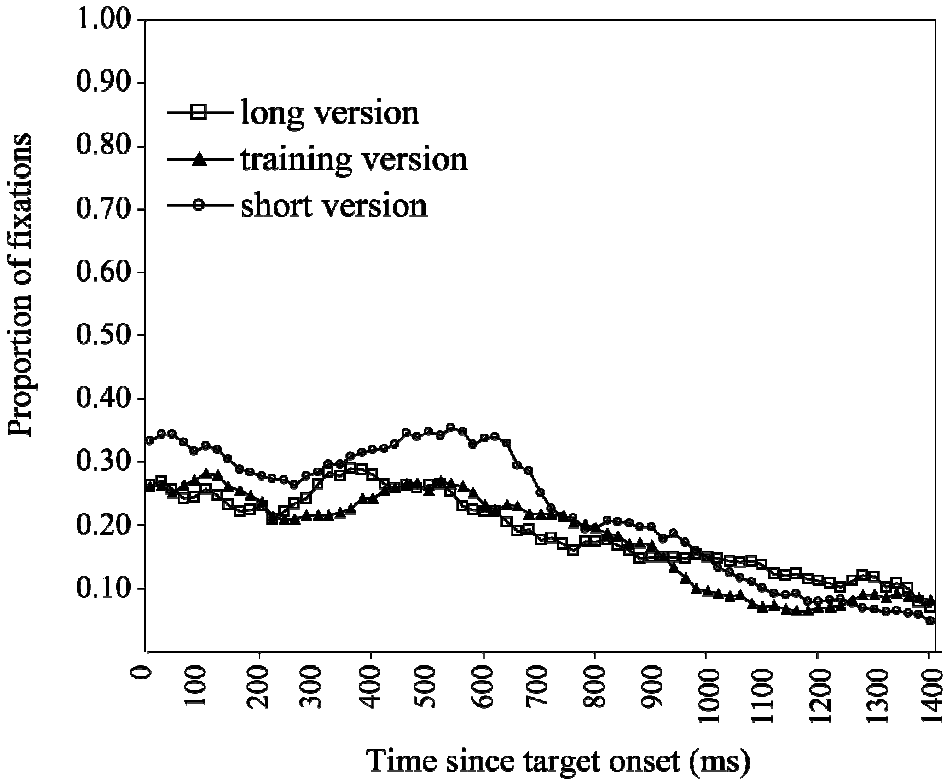


Figure 3. *Fixation proportions to the target and competitor pictures in response to the long, training and short versions: (A) bisyllabic condition, targets (e.g., baptoe); (B) bisyllabic condition, monosyllabic competitors (e.g., bap); (C) monosyllabic condition, targets (e.g., bap); and (D) monosyllabic condition, bisyllabic competitors (e.g., baptoe).*

C


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
Proportion of fixations
long version
training version
short version
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Time since target onset (ms)

D


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
Proportion of fixations
long version
training version
short version
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Time since target onset (ms)

Analyses of variance were computed by subjects (F_1) and by items (F_2) on the average fixation proportions to targets and competitors, in the time interval during which fixation proportions to the competitor were higher than fixation proportions to the distractors: in the monosyllabic condition, until approximately 1000 ms after target onset; in the bisyllabic condition, until 1400 ms. This difference is due to the inherent asymmetry between the competitors in the two conditions: the evidence in favour of an embedded monosyllabic competitor upon hearing a bisyllabic target word is stronger than the evidence in favour of a bisyllabic competitor given a monosyllabic target word. Analyses were therefore carried out separately for each condition. In both cases, analysis intervals began 200 ms after target onset because that is approximately the earliest time point at which fixation proportions start reflecting significant events in the speech stream (Fischer, 1992; Hallett, 1986; Matin, Shao, & Boff, 1993; Saslow, 1967; but see Altmann & Kamide, 2004, for a discussion of the time-locking lag between speech and fixations). The results of these analyses are shown in Table 2. The results of two-tailed matched-pairs t-tests are presented in Table 3.

In the bisyllabic condition, participants were most likely to fixate the target given the short version, less given the training version and least given the long version. Planned comparisons indicated a statistically significant difference in fixation proportions between the long and short versions, given a Type I error rate of 0.05. The analysis of the proportion of fixations to the competitor showed that participants fixated the competitor the most given the long version, less given the training version and least given the short version. Planned comparisons showed that the difference in fixation proportions between the long and short versions was again statistically significant.

Table 2. *Average fixation proportions (in percentages) to the target pictures and the competitor pictures, in the bisyllabic condition and the monosyllabic condition, separately for each version, and the ANOVA results related to these means.*

	Version						
	Long	Training	Short	ANOVAs			
Bisyllabic condition				*F*	*p*	p_{rep}	η^2
Target (e.g., baptoe)	44%	46%	50%	F_1 = 3.45	< .05	.89	.13
				F_2 = 3.53	.08	.84	.28
Competitor (e.g., bap)	24%	21%	19%	F_1 = 6.06	< .01	.97	.34
				F_2 = 5.64	< .05	.95	.38
Monosyllabic condition							
Target (e.g., bap)	35%	40%	31%	F_1 = 3.10	= .055	.87	.12
				F_2 = 2.98	= .09	.83	.25
Competitor (e.g., baptoe)	21%	21%	27%	F_1 = 3.70	< .01	.90	.14
				F_2 = 2.91	= .08	.84	.24

Note. Fixation proportions were averaged in the time intervals 200-1400 and 200-1000, in the bisyllabic and monosyllabic conditions respectively. Degrees of freedom are 2,46 and 2,18 for F_1 and F_2, respectively.

Table 2 indicates that fixation proportions to monosyllabic targets were lowest when listeners heard the short version, higher when they heard the long version, and highest when they heard the training version. In the planned comparisons, no differences were significant by both subjects and items at the 0.05 level. The average proportion of fixations to the (bisyllabic) competitor was higher when listeners heard the short version, compared to when they heard the long or training version. The planned comparisons indicated that the difference in fixation proportions between the short version and the long version was statistically significant.

Table 3. *Two-tailed matched-pairs t tests comparing fixation proportions to the target and the competitor pictures in the long, training and short versions, in the bisyllabic condition and the monosyllabic condition.*

	t_1	p	p_{rep}	d	t_2	p	p_{rep}	d
Bisyllabic Condition								
Target (e.g., baptoe)								
Long vs. Training	< 1	n.s			< 1	n.s		
Training vs. Short	-1.77	= .08	.82	.28	-3.37	< .01	.96	.33
Short vs. Long	-2.61	< .05	.94	.44	-2.36	< .05	.89	.55
Competitor (e.g., bap)								
Long vs. Training	1.94	.06	.86	.32	1.91	.09	.83	.35
Training vs. Short	1.55	= .13	.78	.27	1.50	= .17	.74	.30
Short vs. Long	3.43	< .01	.98	.59	3.18	< .05	.95	.60
Monosyllabic Condition								
Target (e.g., bap)								
Long vs. Training	3.18	< .05	.76	-.26	1.72	= .12	.79	.77
Training vs. Short	2.35	< .05	.92	.53	2.02	= .07	.84	.48
Short vs. Long	<1	n.s			<1	n.s		
Competitor (e.g., baptoe)								
Long vs. Training	<1	n.s			<1	n.s		
Training vs. Short	2.21	< .05	.90	.48	1.94	= .08	.83	.52
Short vs. Long	2.16	< .05	.89	.59	2.21	= .055	.87	.39

Note. Degrees of freedom are 23 and 9 for t_1 and t_2, respectively.

Listeners' fixations thus indicate that syllable duration modulates the activation of new words, such that short syllables tended to be interpreted as the first syllable of a bisyllabic word, while long syllables generated more monosyllabic word interpretations. The effects are clearer for the bisyllabic targets perhaps because of the inherent asymmetry between the two conditions: The ambiguity takes longer to resolve when the input contains a bisyllable.

Influence of the native lexicon. We also explored whether words from the listeners' native lexicon that are phonologically related to the experimental items influenced the results. The item bap, for example, has existing cohort words such as bad [bath], bal [ball], bank [bank], but also longer words like badjas [bathrobe] and balpen [ball pen], as well as rhymes such as hap [bite] and sap [juice].

For each monosyllabic item in the experiment, the following measures were calculated: (1) Number of neighbours and frequency-weighted neighbourhood density ("neighbourhood density"); (2) Number of cohort words that were monosyllabic and frequency-weighted monosyllabic cohort ("MS cohort density"); (3) Number of cohort words that were polysyllabic and frequency-weighted polysyllabic cohort ("PS cohort density"); (4) Number of rhymes and frequency-weighted rhyme set ("rhyme density").

Densities were computed by comparing the monosyllables to real words, using the CELEX lexical database (Baayen, Piepenbrock & Gulikers, 1995). Following Newman, Sawusch and Luce (1997), an item's neighbour was defined as every real word that differed from the item by a one-phoneme addition, substitution or deletion. A cohort word was defined as beginning with the same consonant and vowel. A rhyme was defined as differing only in onset. The frequency counts used were raw frequencies from CELEX, which are based on a corpus of approximately 42.4 million words. For each word in the sets (i.e., neighbourhood, cohort or rhyme set), the raw frequency of the word was multiplied by 10. The log-transformations of the product of all words in the set were summed to yield the frequency-weighted density. Words which had a frequency of 0 were excluded. The results of these calculations are displayed, per item, in Table 4.

Table 4. *Frequency-weighted neighbourhood, monosyllabic cohort, polysyllabic cohort and rhyme densities, per experimental item and, in brackets, the number of words in each set.*

Item	Neighbourhood Density		Monosyllabic Cohort Density		Polysyllabic Cohort Density		Rhyme Density	
bap	63	(22)	88	(32)	695	(337)	84	(27)
fiem	18	(6)	14	(50)	139	(65)	17	(6)
jom	52	(16)	32	(13)	170	(79)	61	(18)
kes	74	(27)	54	(23)	917	(438)	60	(20)
kuin	65	(19)	16	(5)	82	(41)	27	(8)
nim	33	(11)	20	(8)	64	(28)	23	(10)
soer	53	(19)	14	(6)	73	(37)	55	(17)
taaf	33	(12)	21	(7)	364	(1780)	27	(9)
tuik	60	(20)	21	(8)	281	(1390)	33	(11)
zaf	44	(13)	49	(13)	358	(177)	45	(14)

We then examined whether any of these lexical measures correlated with the eye-tracking results. For each experimental item, the perceptual effect was defined as the difference in average fixation proportions between the two most diverging versions. In the bisyllabic condition, fixation proportions differed most between the long and short versions, both for target and competitor fixations. In the monosyllabic condition, target fixations diverged most between the training and short versions, while in the competitor fixations the biggest difference was between the long and short versions. The correlations between the perceptual effects and the lexical measures are presented in Table 5. An additional lexical measure was examined: the ratio between the frequency-weighted monosyllabic cohort and the frequency-weighted polysyllabic cohort (“cohort ratio”).

Table 5. *Correlations of native-lexicon lexical density measures with the perceptual effect in the fixations to the targets and competitors in the bisyllabic and monosyllabic conditions.*

Measurement	Bisyllabic Condition		Monosyllabic Condition	
	Target *baptoe*	Competitor *bap*	Target *bap*	Competitor *baptoe)*
No. of neighbours	-.433	-.076	.050	.091
Neighbourhood density	-.336	-.122	-0.19	.040
No. of monosyllabic cohorts	-.547	.125	-0.45	.284
MS cohort density	-.484	.076	-.073	.275
No. of polysyllabic cohorts	-.463	.368	.342	.250
PS cohort density	-.466	.374	.339	.259
No. of rhymes	-.538	-.033	-.220	.332
Rhyme density	-.493	-.002	-.197	.308
Cohort ratio	-.027	-.655*	-.909**	.147

*Note. * = p<0.05; ** = p<0.001. For all correlations, N=10.*

There was a strong correlation between the perceptual effect in the monosyllabic target condition and the cohort ratio, such that the perceptual effect increased as the cohort ratio decreased (due to the density of the MS cohort decreasing and/or the density of the PS cohort increasing). Although cohort densities on their own did not correlate with the perceptual effect, a step-wise regression analysis showed that MS cohort density did add significant explanatory value, compared to a model with only the cohort ratio as a predictor (Model 1, cohort ratio: $\underline{R^2}$ = .83; Model 2, cohort ratio and MS cohort density: $\underline{R^2}$ = .92). The cohort ratio also proved to be a significant predictor for the perceptual effect in the competitor fixations in the bisyllabic condition (i.e., fixations to the monosyllabic pictures), although it did not explain as much of the variance as with the monosyllabic targets. None of the lexical measurements correlated with the perceptual effect on the bisyllabic items (either in the bisyllabic targets condition or in the monosyllabic competitor condition). It is not clear why these effects were stronger for the monosyllabic items. Listeners' fixations were nevertheless

sometimes modulated by the relative densities of the monosyllabic and polysyllabic real-word cohorts of the novel words.

Conclusions

The results demonstrate that with very little exposure, listeners were able to make fine-grained distinctions between newly-learned forms. Subtle variations in syllable duration influenced the pattern of fixations that participants made to the targets and the competitor. The overall pattern of results is incompatible with a simple episodic model, in which episodes of the newly-learned words are retained in memory and subsequent speech is compared only to these exemplars, with the best matching exemplar being identified. Such a model would predict that a word will be recognized best when the information in the acoustic signal perfectly matches the stored episodes, compared to when the signal deviates from the stored episodes. Our findings indicate that this is not the case. Rather, recognition of the newly-acquired words is guided by prosodic knowledge about the relative duration of syllables in existing words (i.e., that monosyllabic words tend to have longer durations than the initial syllables of polysyllabic words). Moreover, our analyses show that the perceptual effect, as reflected in listeners' fixations to pictures of the monosyllabic words, is modulated by the relative density of monosyllabic and polysyllabic cohorts of those words.

The contribution of the present study is thus twofold. First, it demonstrates that recognition of newly-acquired words is influenced by prosodic knowledge, despite the absence of such prosodic information in the exposure that listeners received. Second, it shows that the lexical neighbourhood can modulate the influence that prosodic knowledge has on the recognition of newly-acquired words.

These results, together with previous findings, indicate that purely episodic or abstractionist models are insufficient. The fact that recognition of newly-learned words is not

determined solely on the basis of episodes, and other results (the lack of a same-speaker advantage on repetition priming in lexical decision, Luce & Lyons, 1998; generalization of perceptual learning about speech sounds to new words, McQueen, Cutler & Norris, submitted) suggest that the mental lexicon is not episodic. If one wanted to salvage this class of models, a more sophisticated version would need to be developed. Such a model might be able to accommodate our results by taking into account the influence that temporary activation of phonologically-related words has on the activation and subsequent recognition of newly-acquired words.

Purely abstractionist accounts of spoken-word recognition are equally untenable, given the growing evidence concerning talker-specific effects (see Goldinger, 1998 for a review). Furthermore, current abstractionist models do not adequately account for the influence that prosodic information has on word recognition, as shown here and elsewhere (e.g., Salverda, 2005; Salverda et al., 2003). Modeling this effect will additionally be constrained by the present study's finding that lexical neighbourhood can modulate the influence that prosodic information has on word recognition.

Our results add to recent findings regarding the integration of newly-acquired words into the mental lexicon. Magnuson et al. (2003) found that the frequency of occurrence of novel words in the exposure phase elicited subsequent frequency effects similar to those found with existing words. In contrast to our study, however, they found no evidence for an interaction between artificial lexicon frequency and native-language lexical density, perhaps because they only computed neighbourhood density measures. Neighbourhood density also did not correlate with the perceptual effect reported here – but the cohort ratio did. This might indicate that for newly-learned words, existing words that overlap with their onsets are more important than other neighbours. Bilingual studies have already shown that native lexicon cohorts are active in second-language listening.

In Gaskell and Dumay's (2003) study, participants were repeatedly exposed to novel words which began like existing words (e.g., the pair cathedruke-cathedral). The existing words were then presented in a lexical decision task, either immediately after exposure or after a delay. There was an immediate facilitatory effect implying that the novel words had activated the existing words, and a late-emerging inhibitory effect, like that observed for words with real-word competitors, suggesting that fully integrating a novel word into the lexicon takes some time.

Magnuson et al. (2003) and Gaskell and Dumay (2003) examined primarily if and how new words are integrated into the lexicon. In contrast, we have shown how phonological knowledge influences the on-line recognition of a newly-acquired word. This interaction between new experience and prior knowledge suggests that storage of lexical episodes is not a sufficient account of word acquisition. Abstract knowledge of fine-grained phonetic signatures of prosodic structure in the listener's native language modulates the interpretation of new words as they are heard.

Summary and conclusions

CHAPTER 7

Understanding spoken language entails recognizing the words in spoken utterances. The speech stream, however, is a continuous signal and does not typically have clear breaks between words. In order to understand what is said, the information in the speech signal needs to be analyzed and compared to stored knowledge about what words sound like. This thesis investigated the acoustic information and the stored knowledge that are involved in this process. While word boundaries are only occasionally marked explicitly in the speech signal, there is a plethora of detailed acoustic information that correlates with word boundaries. That is to say, properties such as duration, amplitude and pitch of segments and syllables can vary depending on the location of word boundaries. However, these acoustic correlates of word boundaries are effective cues only to the extent that listeners can perceive and use them during the word recognition process. This thesis focused therefore on two issues: what fine-grained acoustic information receives the listener's attention, and how the listener then makes use of it.

The experiments in Chapters 2 and 3 investigated the degree to which listeners use different acoustic correlates of word boundaries in fully ambiguous Dutch sentences (e.g., *ze heeft wel eens pot gezegd* [she said once jar]). In these sentences, the stop-initial target words (e.g., *pot*) were preceded by *eens*; the sentences could also refer to cluster-initial words (e.g., *een spot* [a spotlight]). Listeners' eye movements were monitored as they heard recordings of such sentences and saw four pictured objects. Their task was to click on the object mentioned in the sentence. The recordings of the sentences were manipulated by splicing. In one experiment (Chapter 2, Experiment 1), the target word (e.g., *pot*) and the preceding [s] were replaced either by a recording of the cluster-initial word (e.g., *spot*) or by another recording of the target and the preceding [s]. In another experiment (Chapter 3), only the stop and the preceding [s] were cross-spliced. In both experiments, participants were slower to fixate the

target picture (e.g., a jar) when the cross-spliced portion was taken from the cluster-initial word context than when it was taken from the stop-initial word context. This indicates that listeners were attending to the fine-grained acoustic differences between the two spliced versions of the sentences.

The acoustic analyses revealed several differences between the realization of the ambiguous phrase in the two contexts. In other words, there were potentially several acoustic cues to word boundaries that listeners could use. However, only one of these differences – the duration of the [s] – correlated with listeners' performance in the eye-tracking task, indicating that, in this context, the [s] duration information is an important factor guiding word recognition. This was further confirmed in an experiment in which the duration of the [s] preceding the target word was manipulated (Chapter 2, Experiment 2). The only difference between the stimuli in the two conditions was whether the [s] was lengthened or shortened. Listeners were slower to identify the stop-initial target when the duration of the [s] in the spoken signal was lengthened.

Chapter 4 describes an experiment in which [s] duration was manipulated in sentences containing a temporary ambiguity. For example, in the sentence *ik zou ooit eens pijp willen roken* [I would like to smoke a pipe some time] temporary ambiguity arises because before hearing all of *pijp* the sentence could also refer to *een spijker* [a nail]. In temporarily ambiguous sentences, in contrast to fully ambiguous sentences, subsequent information resolves the ambiguity. Again, participants' eye-movements were tracked as they listened to sentences in which a stop-initial target word (e.g., *pijp*) was preceded by an [s]. Participants made more fixations to pictures of cluster-initial words (e.g., *spijker*) when the duration of the [s] was lengthened, compared to when it was shortened. At the same time, participants made more fixations to pictures of the stop-initial words when they heard a short [s], compared to when they heard a long [s]. The results of this study show that the duration of the [s] differentially favours lexical candidates. Segment duration appears to modulate the lexical

competition process by winnowing down the set of competing words, thus affecting how the competition is resolved.

These findings demonstrate that the word recognition process is sensitive to the acoustic correlates of word boundaries. However, the process does not appear to be influenced by all the cues in the speech signal: Listeners' segmentation of the ambiguous sequences could be predicted only from the duration of the [s]. This does not mean that the other acoustic measurements can not influence segmentation. It is quite likely that if the duration of the [s] is kept constant, other acoustic correlates of word boundaries would be used for segmentation. But given the normal variation in natural speech, listeners seem to rely on the duration of the [s] to segment these ambiguous sequences.

It might seem strange or even contradictory that the speech recognition system, which has been shown to be very sensitive to subtle acoustic cues marking word boundaries, would ignore potentially useful information. Why would only the duration of the [s] be used? One possible explanation is that in these sequences, [s] was the first segment containing disambiguating information that the listeners heard. That is to say, the duration of the [s] might have had more influence because at that point the lexical competition had not yet been biased towards one interpretation. It is possible that other potentially disambiguating information has less impact because, by the time that information has been heard, the competition has already shifted in favour of one of the interpretations. This suggests that the influence of fine-grained acoustic information on lexical competition is in itself a function of the state of the competition and how that state changes over time.

Chapter 5 examined the influence of fine-grained acoustic information on the recognition of words containing offset-embedded words. Participants heard sentences mentioning such carrier words (e.g., in the sentence z*e kon de grotere pipet niet vinden* [she could not find the bigger pipette] the word *pipet* contains the word *pet* [cap] at its offset). In one version of the sentence, the embedded sequence was replaced by another token of the sequence from the

carrier-word sentence (e.g., by splicing in the second syllable of the word *pipet*). In the other version the embedded sequence originated from a matched sentence mentioning the embedded word (e.g., *ze kon de grote hippie pet niet vinden* [she could not find the big hippy cap]). In the first series of experiments, participants' eye-movements were monitored as they heard these sentences. The results showed that the recognition of the carrier words was not influenced by the acoustic differences between the two versions (i.e., between the realization of the sequence embedded in the carrier word and that of the monosyllabic word). Furthermore, participants looked at pictures of the embedded words more than at distractors only when they heard carrier words in which the embedding was aligned with the internal syllable boundary and was unstressed. But even with these items there was no effect of the splicing manipulation. When participants heard carrier words with aligned stressed embeddings or with misaligned embeddings, they fixated pictures of the embedded words as much as they did pictures of phonologically unrelated words, even when the embedded sequence originated from the monosyllabic word. In the second series of experiments, listeners performed a cross-modal identity priming task, with either the carrier words or the embedded words as visual targets. Similar to what was found with eye-tracking, the recognition of the carrier words was not influenced by the acoustic differences between the two versions of the prime. When the embedded words were presented as the visual targets, lexical decision latencies were slower after related primes (i.e., the carrier words) than after unrelated primes. This inhibitory effect was taken as evidence that the embedded word had been suppressed after losing the lexical competition with the carrier word prime. The splicing manipulation did not appear to have any effect on the amount of inhibition of the embedded words.

The findings reported in Chapter 5 thus reveal a dissociation between the results obtained with eye-tracking and those obtained with cross-modal identity priming. While the eye-tracking data showed evidence for the activation of the offset-embedded words only with

carrier words with aligned unstressed embeddings, the inhibitory effect in the cross-modal priming experiment was equally large for all types of carrier words. The inhibitory effect indicates that the embedded words must have been activated at some point by the carrier word prime (or else responses to them would not be different than those to targets after an unrelated prime). Taken together, the results from the two experiments suggest that spoken language comprehension involves the activation of a word's phonological form (which can be detected by cross-modal identity priming) and the activation of the word's conceptual representation (which can be measured with eye-tracking). The activation of an offset-embedded word's phonological form (reflected by the inhibitory priming effect) appears to be obligatory, but does not automatically lead to activation of its conceptual representation (no effect in the eye-tracking results for the carrier words with aligned stressed and misaligned embeddings).

The results in Chapter 5 indicate that the word recognition process is insensitive to the differences between the recordings of offset-embedded sequences and monosyllabic words. This finding contrasts with evidence from eye-tracking studies investigating word-initial embeddings (e.g., Salverda, Dahan & McQueen, 2003; Salverda, 2005), which showed that the duration of the embedded sequence (e.g., *ham* in *hamster*) influenced the interpretation of the sequence as either a monosyllabic word or the first syllable of a longer word. The results in Chapter 5 are therefore congruent with the idea that the modulation of the lexical competition process by detailed acoustic information can only be detected if the competition is in a relatively balanced state. Offset-embedded words are always at a disadvantage relative to the carrier word, because the carrier word matches more of the signal. The "head start" that the carrier word enjoys effectively eliminates any serious competition from the embedded words, making the usage of fine-grained information redundant.

It is important to note that all the stop-initial misaligned embeddings in Chapter 5 (e.g., *peen* [carrot] in *speen* [pacifier]) were used as targets in the experiments reported in Chapters 2 and 3. Listeners were not influenced by the splicing manipulation in Chapter 5, in contrast

to what was found in Chapters 2 and 3. The splicing manipulations in these experiments were similar, but not identical: In Chapters 2 and 3 the cluster-initial word replaced the stop-initial target word and the preceding [s]; in Chapter 5 the stop-initial embedded word was spliced into the cluster-initial target. It is not likely, however, that this is the reason for the difference in findings. Rather, placing the stop-initial word in an ambiguous context (i.e., with a preceding [s]) makes the competition between the stop-initial and cluster-initial candidates more equal so that the modulation of the lexical competition by detailed acoustic information can be observed. Detailed acoustic information favouring one candidate is of little use, however, if it arrives at a time at which another candidate is already winning the lexical competition, as in the situation examined in Chapter 5. Again, this suggests that the influence of fine-grained acoustic detail is a function of how the state of the lexical competition changes over time. The effect that detailed acoustic information has on the lexical competition process appears to decrease progressively as the lexical ambiguity is resolved. Where there is little ambiguity (as is the case with offset-embedded words), fine-grained information appears to have no effect. Where there is a high degree of lexical ambiguity (as in the ambiguous sequences in Chapters 2 and 3), incoming acoustic cues modulate the competition, but their impact also varies as a function of time and competitor environment. That is to say, the effect of fine-grained acoustic information on the competition diminishes as the competition between candidates gradually approaches its conclusion.

The experiment reported in Chapter 6 investigated the stored knowledge that underlies listeners' ability to use subtle acoustic information to resolve lexical ambiguity. Is it word-specific knowledge (i.e., that a certain word is typically produced with a certain duration) or more abstract knowledge? The ambiguity under investigation was that which arises with onset-embedded words. In the experiment, participants learned to associate novel shapes with novel words. During the learning phase, participants heard a spoken instruction to click on one of two nonsense objects displayed on a computer screen (e.g., "click on the *bap*"). After

they clicked on one of the objects they received feedback as to whether they had selected the correct item. Later on, the display was altered to include four objects. The new words were matched pairs of bisyllables (e.g., *baptoe*) and onset-embedded monosyllabic words (e.g., *bap*). In the learning phase, the duration of the ambiguous sequence (e.g., *bap*) was held constant. In the test phase, it was longer, shorter or equal to its duration in learning phase. In the test phase, participants' eye-movements were monitored as they heard the spoken instruction. In the critical trials, pictures of both members of the pair were displayed on the screen.

The results showed that when the target was bisyllabic (e.g., *baptoe*) participants looked more at the picture of the monosyllabic competitor (e.g., *bap*) when they heard the long version than when they heard the short version. When the target was monosyllabic, participants looked more at the bisyllabic competitor when they heard the short version than when they heard the long version. In other words, short syllables tended to be interpreted as the first syllable of a bisyllabic word, while long syllables generated more monosyllabic word interpretations. This is also the pattern that has been found in a previous study, using real words (Salverda et al., 2003). Thus, the recognition of newly-acquired words is influenced by stored knowledge about the relative duration of syllables in existing words (i.e., that monosyllabic words tend to have longer durations than the initial syllables of polysyllabic words), despite the fact that the newly-acquired words did not exhibit these durational patterns during either the learning or the test phase.

These results are incompatible with a simple episodic model, in which episodes of the newly-learned words are retained in memory and subsequent speech is compared only to these exemplars, with the best-matching exemplar being identified. Such a model would predict that a word will be recognized best when the information in the acoustic signal perfectly matches the stored episodes, compared to when the signal deviates from the stored episodes. Our findings indicate that this is not the case. Such an episodic account is therefore insufficient.

The shortcoming of the sketched episodic model is due to the fact that recognition is assumed to be based only on the episodes of newly-learned words. However, several eye-tracking studies with bilinguals have shown that the native lexicon is active in second-language listening, that is, even when it is irrelevant to performing the task (Spivey & Marian, 1999; Weber & Cutler, 2004). It is therefore reasonable to assume that in the present experiment, words in the listener's native lexicon are temporarily activated. That is to say, while the only exemplars that participants have of the newly-acquired items are the ones they have been exposed to during the experiment, they do have knowledge about phonologically related words. For example, for the item *bap*, participants will have knowledge about phonologically related existing Dutch words such as *bal* [ball], *bad* [bath], *bank* [bank], but also longer words like *balsem* [balm] and *banjo* [banjo]. If, as the speech signal unfolds, these items are temporarily activated, it is plausible that the observed pattern of results is partly due to the competition of the newly-acquired word with these phonologically related words.

To investigate the influence of the listener's native lexicon, several aspects of the real-word neighbourhood of each newly-learned word were calculated. Subsequently, the correlation of these measures with the observed pattern of results was computed. The results showed that the perceptual effect, as reflected in listeners' fixations to pictures of the monosyllabic words (i.e., the difference in fixations to these pictures between two versions), was modulated by the real-word neighbourhoods of the newly-acquired words. More specifically, the relative density of monosyllabic and polysyllabic cohorts of those words correlated with the observed pattern of results. Cohort density is the frequency-weighted number of words beginning with the same consonant and vowel as the novel word. It was computed separately for monosyllabic continuations of the onsets of the novel words, and for polysyllabic continuations. The ratio of the monosyllabic and polysyllabic cohort densities was a highly significant predictor of the perceptual effect. This indicates that the degree to

which durational differences influence the interpretation of the ambiguous sequence is related to the distribution of similar-sounding monosyllabic and polysyllabic words.

Why would this be so? One plausible explanation is that this is, again, due to the dynamics of the lexical competition process. As speech unfolds over time, words that are fully or partially consistent with the input become activated and compete among one another. The state of the competition is determined by the number of active lexical candidates and their goodness of fit with the signal. The number of active lexical candidates and their activation levels depends, in part, on cohort density. Goodness of fit is determined, amongst other things, by fine-grained acoustic information. This information biases the competition in favour of certain lexical candidates. The number of candidates that benefit from that biasing information affects therefore the impact of that information on the competition. In the present experiment, the resolution of lexical embedding of novel words displays the same sensitivity to detailed acoustic information that real words do, apparently because the real words also participate in the process; the number of participating words modulates the impact of the detailed information.

The findings of this thesis demonstrate that detailed acoustic information helps in the resolution of lexical ambiguity. The degree to which this information affects the lexical competition process varies during the course of the competition. Chapters 2 and 3 provide evidence that listeners use segment duration to resolve lexical ambiguity. Chapter 4 indicates that segment duration modulates the lexical competition. The results showed, however, that listeners do not use all the acoustic correlates of word boundaries in their segmentation of the ambiguous sequences. Detailed acoustic information which arrives when the competition is in an unresolved state appears to have more influence on the competition than information which arrives after the competition has shifted in favour of one competitor. The results in Chapter 5 support this conclusion by indicating that when there is little lexical ambiguity, detailed acoustic information does not have an effect on the competition. Finally, Chapter 6 provided

evidence that fine-grained durational information is used to resolve lexical ambiguity with newly-learned words in the same way that it does with real words, despite the fact that the newly-learned words did not exhibit the same durational patterns as that found with real words. Stored lexical and phonological knowledge is therefore involved in the recognition of newly-learned words.

The influence of detailed acoustic information on the lexical competition process therefore varies according to the degree of ambiguity that is involved, that is, with the number of competing lexical candidates and their activation levels. Because sensitivity to detailed acoustic information operates at the level of lexical competition it is inherently related to stored knowledge about words and their sound forms. As speech unfolds over time, a certain degree of ambiguity arises, at least temporarily, in almost all utterances. Sensitivity to detailed acoustic information reduces this ambiguity. This enables the word recognition process to proceed with such seemingly effortless speed and efficiency, so that we, as listeners, hardly ever have to be aware of its operation.

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

Alphen, P. M. van, & McQueen, J. M. (2006). The effect of Voice Onset Time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance.*

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.Henderson & F. Ferreira (Eds.), *The integration of language, vision and action* (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 [Computer software]. (1995). Calabasas, CA: Nova Development Corporation.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database. Philadelphia, PA: Linguistics Data Consortium, University of Pennsylvania.

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, MA: 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 and 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.

Content, A., Kearns, R. K., & Frauenfelder, U. H. (2001). Boundaries versus onsets in syllabic segmentation. *Journal of Memory and Language*, **45**, 177-199.

Cutler, A., McQueen, J. M., Norris, D. G., & Somejuan, A. (2001). The roll of the silly ball. In E.Dupoux (Ed.), *Language, Brain and Cognitive Development: Essays in honor of Jacques Mehler* (pp. 181-194). Cambridge, MA: MIT Press.

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. (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. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review,* **12**, 453-459.

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 and Performance,* **28**, 218-244.

Dumay, N., Frauenfelder, U. H., & Content, A. (2002). The role of the syllable in lexical segmentation French: Word- spotting data. *Brain and Language,* **81**, 144-161.

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 Verlag.

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., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition,* **89**, 105-132.

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

Glucksberg, S., Kreuz, R. J., & Rho, S. H. (1986). Context can constrain lexical access: Implications for models of language comprehension. *Journal of Experimental Psychology-Learning Memory and Cognition,* **12**, 323-335.

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 and 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 and 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.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition,* **96**, B23-B32.

Isel, F., & Bacri, N. (1999). Spoken-word recognition: The access to embedded words. *Brain and Language,* **68**, 61-67.

Johnson, K. (1997a). 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, CA: Academic Press.

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

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 and Hearing Research,* **17**, 51-63.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition*, **9**, 587-598.

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

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America,* **51**, 2018-2024.

Lindblom, B., Lyberg, B., & Holmgren, K. (1981). *Durational patterns of Swedish phonology: do they reflect short-term motor memory processes?* Bloomington, Indiana: Indiana University Linguistics Club.

Luce, P. A., & Cluff, M. S. (1998). Delayed commitment in spoken word recognition: Evidence from cross-modal priming. *Perception & Psychophysics,* **60**, 484-490.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition,* **26**, 708-715.

Luce, P. A., & Lyons, E. A. (1999). Processing lexically embedded spoken words. *Journal of Experimental Psychology-Human Perception and Performance*, **25**, 174-183.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology-General,* **132**, 202-227.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review,* **101**, 3-33.

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

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

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

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. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language,* **39**, 21-46.

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 255-275). London: Sage Publications.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes,* **10**, 309-331.

McQueen, J. M., Cutler, A., & Norris, D. (submitted). Phonological abstraction in the mental lexicon.

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. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology-Learning Memory and Cognition,* **20**, 621-638.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology-Human Perception and 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.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology-Human Perception and Performance,* **23**, 873-889.

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

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (submitted). Phonological and conceptual activation in speech comprehension.

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.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137-157). Amsterdam: John Benjamins.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101-140). Berlin: Mouton de Gruyter.

Pitt, M. (1994, November). Lexical competition: The case of embedded words. *Poster presented at the 34th annual meeting of the Psychonomic Society*, Washington, D.C.

Prather, P. A., & Swinney, D. (1988). Lexical processing and ambiguity resolution:An autonomous process in an interactive box. In S. L. Small, G. W. Cottrell & T. M.K. (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology and artificial intelligence.* San Mateo, CA: Morgan Kaufmann Publishers.

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

Salverda, A. P. (2005). *Prosodically-conditioned detail in the recognition of spoken-words.* Doctoral dissertation, Radboud University Nijmegen (MPI Series in Psycholinguistics, Vol. 33). Wageningen: Ponsen & Looijen.

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.

Schreuder, R., & Baayen, R. H. (1994). Prefix stripping re-revisited. *Journal of Memory and Language*, **33**, 357-375.

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. *Proceedings of the Eighth International Conference on Spoken Language Processing*, Jeju Island, Korea (pp. 329-332).

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, **68**, 1-16.

Shatzman, K. B., & McQueen, J. M. (in press). Prosodic knowledge affects the recognition of newly-acquired words. *Psychological Science*.

Shatzman, K. B., & McQueen, J. M. (submitted). The modulation of lexical competition by segment duration. *Psychonomic Bulletin & Review*.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming. *Journal of Experimental Psychology-Learning Memory and Cognition*, **18**, 1191-1210.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 24-49). Cambridge, MA: MIT Press.

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 and Memory*, **6**, 174-215.

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

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, **10**, 281-284.

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 and Performance*, **26**, 758-775.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language and 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.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology-Human Perception and Performance,* **23**, 710-720.

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

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language,* **50**, 1-25.

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America,* **119**, 597-607.

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

Samenvatting

Het verstaan van gesproken taal vereist dat men de woorden in de spraakstroom herkent. Deze spraakstroom is echter een continu signaal – normaliter zitten er geen duidelijke pauzes tussen woorden. Om te kunnen verstaan wat er gezegd werd, moet men de informatie in het spraaksignaal analyseren en vergelijken met de kennis die in ons hoofd ligt opgeslagen over hoe een bepaald woord klinkt. Deze dissertatie onderzocht de akoestische informatie en de opgeslagen kennis die bij dit proces betrokken zijn. Hoewel woordgrenzen maar zelden expliciet in het spraaksignaal gemarkeerd zijn, is er een rijkdom aan gedetailleerd akoestische informatie die met de woordgrenzen correleert. Dat wil zeggen, eigenschappen zoals de duur, amplitude en pitch van klinkers en lettergrepen kunnen variëren afhankelijk van de locatie van de woordgrenzen. Het onderzoek beschreven in dit proefschrift heeft zich op twee kwesties gericht: welke gedetailleerd akoestische informatie wordt opgemerkt door de luisteraar, en hoe maakt de luisteraar gebruik van deze informatie?

In de experimenten beschreven in Hoofdstuk 2 en 3 heb ik onderzocht, in welke mate luisteraars gebruik maken van verschillende akoestische variabelen die met woordgrenzen correleren in ambigue zinnen zoals *ze heeft wel eens pot gezegd*. In al deze zinnen zat een woord met een initiële plofklank (bijv. *pot*), dat werd voorafgegaan door het woord *eens*. Daardoor kon het voor de luisteraar lijken alsof er een woord werd uitgesproken dat met een cluster begon (bijv. *spot*). Terwijl luisteraars naar deze zinnen luisterden, werden hun oogbewegingen geregistreerd. Tegelijkertijd zagen zij vier objecten op een computerscherm. Van tevoren kregen ze de instructie om met de muis op het object te klikken dat in de zin werd genoemd. De opnames van de gesproken zinnen waren gemanipuleerd door middel van splicing (uitknippen en vervangen van stukjes spraak). In één experiment (Hoofdstuk 2, Experiment 1), werd het kritieke woord (bijv. *pot*) en de voorafgaande [s] vervangen hetzij door een opname van het cluster-initiële woord (bijv. *spot*), of door een andere opname van

het kritieke woord en de [s]. In een ander experiment (Hoofdstuk 3) werd alleen de plofklank en de [s] ervoor uitgeknipt en vervangen. In beide experimenten deden proefpersonen er langer over om naar het plaatje van het kritieke woord (bijv. een pot) te kijken wanneer ze een zin hoorden waarin het ingeplakte deel van een cluster-initieel woord kwam, dan wanneer het van het kritieke woord kwam. Dit laat zien dat luisteraars gevoelig zijn voor de subtiele akoestische verschillen tussen de twee gemanipuleerde zinnen.

Akoestische analyses onthulden dat er een aantal verschillen waren tussen de twee realisaties van het ambigue zinsdeel (*eens pot/een spot*). Er waren dus meerdere akoestische cues die gebruikt konden worden door de luisteraar om te bepalen waar de woorgrens lag. Slechts één van deze aanwezige cues, namelijk de duur van de [s], correleerde met het gedrag van de proefpersonen in het oogbewegingsexperiment. Dit laat zien dat in deze context de duur van de [s] bepalend is voor het woordherkenningsproces. Deze conclusie werd bevestigd door de uitkomsten van een tweede experiment waarin de duur van de [s] werd verlengd of verkort (Hoofdstuk 2, Experiment 2). Uit de resultaten bleek dat luisteraars er langer over deden om het kritieke woord beginnend met een plofklank te herkennen als deze werd voorafgegaan door een lange [s].

Hoofdstuk 4 beschrijft een experiment waarin de duur van de [s] werd gemanipuleerd in zinnen met een tijdelijke ambiguïteit. Bijvoorbeeld, de zin *ik zou ooit eens pijp willen roken* is tijdelijk ambigu omdat in deze zin tot en met de klinker van *pijp* ook nog verwezen kan worden naar een spijker. In tegenstelling tot de ambigue zinnen in de hiervoor beschreven experimenten, wordt de ambiguïteit in de tijdelijk ambigue zinnen opgelost door de rest van de zin. Net als in de andere experimenten werden de oogbewegingen van de proefpersonen geregistreerd terwijl ze naar zinnen luisterden met plofklank-initiële woorden (bijv. *pijp*) met een [s] ervoor. Proefpersonen keken vaker naar het plaatje van het cluster-initiële woord (bijv. *spijker*), wanneer de [s] langer was gemaakt. Daarnaast keken ze vaker naar het plaatje van het plofklank-initiële woord wanneer ze een korte [s] hoorden. De resultaten van dit

experiment laten zien dat de duur van de [s] bepaalt welke lexicale kandidaat meer steun krijgt. De duur van een klank lijkt invloed te hebben op het lexicale competitieproces door de groep strijdende lexicale kandidaten in te perken. Op deze manier bepaalt de klankduur op welke manier en wanneer de competitie afloopt.

Deze bevindingen bewijzen dat het woordherkenningsproces gevoelig is voor subtiele akoestische informatie, die aangeeft waar woordgrenzen liggen. Het proces lijkt echter niet door alle aanwezige cues in het spraaksignaal beïnvloed te worden; alleen de duur van de [s] kon voorspellen hoe luisteraars de ambigue zinnen interpreteerden. Dit betekent niet dat de andere akoestische cues geen invloed kunnen hebben op het segmentatieproces. Het is zeer aannemelijk dat deze akoestische cues wel invloed zouden hebben als de duur van de [s] constant wordt gehouden. Maar gegeven de variaties in natuurlijke spraak lijken luisteraars met name gevoelig te zijn voor de duur van de [s].

Het lijkt misschien vreemd, of zelfs tegenstrijdig, dat het spraakherkenningsysteem dat gevoelig blijkt te zijn voor subtiele akoestische informatie, potentieel nuttige informatie zou negeren. Waarom zou men alleen gebruik maken van de duur van de [s]? Een mogelijke verklaring is dat disambiguerende informatie in de [s] eerder kwam dan de andere potentieel nuttige informatie. In andere woorden, de duur van de [s] kwam binnen op een moment dat het competitieproces nog lang niet beslist was. De overige potentiële cues zouden minder impact hebben, simpelweg omdat ze op een moment binnenkwamen dat de strijd al bijna gestreden was. Dit suggereert dat de mate waarin subtiele akoestische informatie het competitieproces beïnvloed, wordt bepaald door de staat waarin het competitieproces zich op dat moment bevindt en de manier waarop deze in de tijd verandert.

In hoofdstuk 5 werd onderzocht wat de invloed was van akoestisch detail op de herkenning van woorden, waar aan het eind een korter woord zat ingebed (bijv. *pet* aan het eind van *pipet*). Proefpersonen kregen zinnen te horen zoals z*e kon de grotere pipet niet vinden*. Van

iedere zin waren weer twee verschillende versies gemaakt met behulp van splicing. In de ene versie kwam het ingebedde woord van de tweede lettergreep van het woord *pipet*; in de andere versie kwam het van de zin z*e kon de grote hippie-pet niet vinden*. Uit de resultaten van een serie oogbewegingsexperimenten bleek dat de akoestische verschillen tussen de twee versies geen effect hadden op de herkenning van de langere kritieke woorden (bijv. *pipet*). Daarnaast bleek dat proefpersonen vaker naar het plaatje van het ingebedde woord keken dan naar een ongerelateerd object enkel en alleen als het ingebedde woord geen klemtoon had en op de lettergreepgrens begon (bijv. het woord *zon* in *bizon*). Maar zelfs in deze conditie was er geen effect van de akoestische manipulatie. Wanneer de ingebedde woorden niet op de lettergreepgrens begonnen (bijv. *peen* in *speen*) of klemtoon hadden (bijv. *pet* in *pipet*) keken de luisteraars even vaak naar het plaatje van het ingebedde woord als naar een ongerelateerd object – ook wanneer ze de versie hoorden waarin het ingebedde woord oorspronkelijk uit de zin met het korter woord kwam.

Dezelfde gemanipuleerde zinnen fungeerden als stimuli in een tweede serie experimenten waarin de cross-modal identity priming taak werd gebruikt. In deze taak krijgen proefpersonen een zin te horen (de prime) en een reeks letters op het scherm (de target) te zien, die al dan niet een bestaand woord vormden. De taak van de proefpersoon is om zo snel mogelijk te beslissen of de visuele target wel of geen bestaand Nederlands woord is. In deze serie experimenten kwam de target overeen met of het langere kritieke woord (bijv. *pipet*) of het ingebedde woord (bijv. *pet*). Opnieuw bleek uit de resultaten dat de akoestische verschillen tussen de twee versies geen effect hadden op de herkenning van de langere woorden. Wanneer de ingebedde woorden als target werden aangeboden, waren de lexicale beslissingen langzamer als de luisteraar net de zin met het langere woord had gehoord, dan na een zin met een ander ongerelateerd woord. Dit inhibitie- effect wordt beschouwd als bewijs dat het ingebedde woord de competitie heeft verloren van het langere woord. De akoestische manipulatie leek geen invloed te hebben op de mate van inhibitie.

De bevindingen in Hoofdstuk 5 laten dus een dissociatie zien tussen de resultaten van de oogbewegingsexperimenten en de primingexperimenten. Terwijl de oogbewegingsdata er op wezen dat alleen ingebedde woorden zonder klemtoon en beginnend op de lettergreepgrens geactiveerd worden, lieten de primingdata een inhibitie-effect zien, dat even groot was voor alle type ingebedde woorden. De resultaten van deze twee series experimenten suggereren dus dat het verstaan van gesproken taal twee soorten activatie met zich meebrengt: dat van de fonologische woordvorm (welke met priming kan worden gemeten) en dat van de conceptuele representatie (welke met een oogbewegingstaak kan worden gemeten). De activatie van de fonologische representatie van een finaal-ingebed woord (gereflecteerd door het inhibitie-effect) lijkt verplicht te zijn, maar leidt niet automatisch tot activatie van het concept (gereflecteerd door een oogbewegingseffect voor slechts één type ingebedde woorden).

De resultaten van Hoofdstuk 5 geven aan dat het woordherkenningsproces ongevoelig is voor de akoestische verschillen tussen finaal-ingebedde woorden en monosyllabische woorden. Dit verschilt van oogbewegingsexperimenten waarin de activatie van initieel-ingebedde woorden (bijv. *ham* in *hamster*) werd onderzocht (bijv. Salverda, Dahan & McQueen, 2003; Salverda, 2005). Uit deze studies bleek dat de interpretatie van de uiting als monosyllabisch dan wel als het begin van een bisyllabisch woord werd bepaald door de duur van de kritieke lettergreep. De afwezigheid van een dergelijk effect in Hoofdstuk 5 suggereert opnieuw dat akoestisch detail slechts invloed kan hebben op het competitieproces als de activatie van de competerende kandidaten nog dicht bij elkaar ligt. Finaal-ingebedde woorden hebben qua activatie altijd een achterstand op het langere woord. Het langere woord heeft immers al steun gekregen van het spraaksignaal op het moment dat het ingebedde woord nog moet beginnen.

In Hoofdstuk 6 werd de opgeslagen kennis die het gebruik van gedetailleerde akoestische informatie mogelijk maakt onderzocht. Is deze kennis woordspecifiek (dat zou betekenen dat een woord meestal met één bepaalde duur wordt uitgesproken) of is deze kennis op een

abstractere manier opgeslagen? Dit werd onderzocht door proefpersonen nieuwe woorden te leren. Tijdens de leerfase hoorden de proefpersonen instructies zoals *klik op de bap*, en zagen ze twee plaatjes van niet-bestaande objecten op het scherm. Nadat ze op één van de plaatjes hadden geklikt, kregen ze te horen of hun keuze correct was. Later in de leerfase moesten ze in plaats van uit twee plaatjes, uit vier plaatjes kiezen. De nieuwe woorden bestonden uit paren van een bisyllabisch woord (bijv. *baptoe*) en het initieel-ingebedde monosyllabisch woord (bijv. *bap*). In de leerfase werd de duur van het ambigue deel (bijv. *bap*) in beide woorden constant gehouden. In de testfase, waarin proefpersonen dezelfde taak uitvoerden als in de leerfase, was de duur van het ambigue deel variabel; het was langer dan wel korter dan wel even lang als in de leerfase. Bovendien werden in de testfase de plaatjes van zowel het bisyllabisch als het monosyllabisch woord (bijv. van *bap* en van *baptoe*) tegelijkertijd op het scherm getoond. Tijdens de testfase werden de oogbewegingen van de proefpersonen geregistreerd.

Uit de resultaten bleek dat als proefpersonen een bisyllabisch woord te horen kregen (bijv. *baptoe*), ze vaker naar het plaatje van het monosyllabische woord keken (bijv. *bap*) wanneer de duur van het ambigue deel lang was, dan wanneer het kort was. Als ze een monosyllabisch woord kregen te horen (bijv. *bap*), keken ze vaker naar het plaatje van het bisyllabische woord wanneer ze de korte versie van het woord hoorden dan wanneer ze de lange versie hoorden. Met andere woorden, korte lettergrepen werden vaker geïnterpreteerd als het begin van een bisyllabisch woord, en lange lettergrepen werden vaker als een monosyllabisch woord geïnterpreteerd. Dit is precies het patroon dat eerder werd gevonden met bestaande woorden (Salverda et al., 2003). Hieruit kunnen we concluderen dat het herkennen van *nieuwe* woorden bepaald wordt door opgeslagen kennis over de typische duur van lettergrepen in *bestaande* woorden (dat wil zeggen dat een monosyllabisch woord meestal met een langere duur wordt uitgesproken dan de eerste lettergreep van een langer woord). Dit is opmerkelijk, omdat de nieuwe woorden in het experiment niet zo een dergelijk systematisch duurpatroon

hadden; in het experiment waren de monosyllabische woorden en de eerste lettergreep van de bisyllabische woorden even lang.

Hoe komt het dat bij de herkenning van nieuwe woorden kennis wordt gebruikt over de duur van lettergrepen in bestaande woorden, terwijl de nieuwe woorden een ander duurpatroon hebben? Een mogelijke verklaring hiervoor, ligt in het feit dat de nieuwe woorden een zekere mate van overlap vertonen met bestaande woorden. Bijvoorbeeld, voor het woord *bap*, kennen luisteraars fonologisch gerelateerde woorden zoals *bal*, *bad*, *bank*, maar ook langere woorden zoals *balsem* en *banjo*. Terwijl luisteraars het woord *bap* horen worden deze bestaande gerelateerde woorden tijdelijk geactiveerd. De lexicale competitie tussen het nieuw woord en deze gerelateerde bestaande woorden zou dus verantwoordelijk kunnen zijn voor het resultatenpatroon in Hoofdstuk 6. Steun voor deze bewering komt uit verdere analyses die laten zien dat het resultatenpatroon ook beïnvloed wordt door het aantal fonologische gerelateerde bestaande woorden en hun frequentie. Kortom, luisteraars zijn net zo gevoelig voor subtiele akoestische informatie tijdens het herkennen van nieuwe woorden als tijdens het herkennen van bestaande woorden, omdat ook bestaande woorden tijdelijk worden meegenomen in dit proces. De gevoeligheid voor akoestische informatie in het herkennen van de nieuwe woorden hangt samen met het aantal en de frequentie waarmee de bestaande woorden voorkomen.

Dit proefschrift beschrijft hoe gedetailleerde akoestische informatie luisteraars helpt bij het woordherkenningsproces. De mate waarin deze informatie effect heeft varieert gedurende het lexicale competitieproces. Als de competitie nog lang niet gestreden is, heeft akoestische detail meer invloed op het competitieproces, dan wanneer één lexicale kandidaat veel actiever is dan alle anderen. Dit zorgt ervoor dat het woordherkenningsproces bijna altijd snel en moeiteloos verloopt, waardoor wij als luisteraars maar zelden merken hoe complex het eigenlijk is.

Curriculum Vitae

Keren Shatzman was born in Jerusalem, Israel, in 1972. She studied psychology, biology and cognitive sciences at the Hebrew University in Jerusalem and received her B.A. *magna cum laude* in 1997. In 1998 she moved to the Netherlands and started studying cognitive psychology at the University of Nijmegen, graduating *cum laude* in 2001. During this period, she worked as a research assistant in the Comprehension Group at the Max Planck Institute for Psycholinguistics. From 2002 to 2005 she carried out her dissertation research at the same group, the results of which are described in this thesis. Since September 2005 she is employed as a postdoctoral fellow in the project *Phonotactic constraints for speech segmentation: The case of second language acquisition*, in the Linguistics Department at the University of Utrecht.

MPI SERIES IN PSYCHOLINGUISTICS

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsaggital articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*