Research Report

Prosodic Knowledge Affects the Recognition of Newly Acquired Words

Keren B. Shatzman and James M. McQueen

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT—An eye-tracking study examined the involvement of prosodic knowledge—specifically, the knowledge 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, its duration was longer than, shorter than, or equal to its learning-phase duration. Listeners' fixations indicated that short syllables tended to be interpreted as the first syllables of the bisyllables, whereas long syllables generated more monosyllabic-word interpretations. Recognition of newly acquired words is influenced by prior prosodic knowledge and is therefore not determined solely on the basis of stored episodes of those words.

Are newly acquired words processed in the same way as wellknown words? Throughout their lives, people continue learning new words. This ability is fundamental to the human 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?

Current models of speech comprehension agree that during word recognition, multiple lexical candidates consistent with the acoustic-phonetic information in the speech signal become activated and compete with 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 for all sentences. In the study we report here, we examined ambiguity resolution in the recognition of novel words.

Onset-embedded words, such as ham in hamster, are a critical case of temporary ambiguity. The phonemic overlap between the embedded word and the longer word suggests that definitive recognition of the short word could occur only after its offset. Fine-grained acoustic information can, however, bias the lexical competition in favor 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 proportion of transitory fixations to pictures representing that monosyllabic embedded word. Salverda et al. manipulated the duration of the initial syllable of words containing onset-embedded words and found that longer durations generated more monosyllabic-word interpretations and shorter durations generated more polysyllabic-word interpretations (see also Salverda, 2005). They argued that these durational differences reflect the prosodic structure of the utterances and that listeners compute this prosodic structure during the word recognition process.

In the present study, we examined 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 sequences (e.g., *bap*) was held constant. In the test phase, we manipulated the duration of these sequences to see if this would modulate the frequency of monosyllabic- and polysyllabic-word interpretations, as it does with existing words (Salverda et al., 2003).

This issue is critical to 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

Address correspondence to Keren Shatzman, Utrecht Institute of Linguistics OTS, Trans 10, 3512 JK Utrecht, The Netherlands, e-mail: keren.shatzman@let.uu.nl.

phonologically abstract form (e.g., Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; Norris, 1994). Other authors have suggested instead 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, 1997b; 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, whether it is based on the stored episodes of those words, or whether recognition of newly acquired words is also determined by prior experience with similar-sounding real words. We employed the eye-tracking paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which was recently used in an artificial-lexicon study (Magnuson, Tanenhaus, Aslin, & Dahan, 2003). In that study, participants learned to associate novel shapes with novel words, some of which were cohort and rhyme competitors 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' frequency of occurrence 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 information that was not present in the learning phase. If recognition is based purely on the episodes that listeners are exposed to, then newly acquired words should be recognized fastest when they are spoken exactly the same as during the learning phase. If, however, there is transfer of knowledge about the relative duration of syllables in existing words (i.e., knowledge that monosyllabic words tend to have longer durations than the initial syllables of polysyllabic words), then listeners should be more likely to interpret newly learned ambiguous sequences (e.g., *bap*) as monosyllabic words when these sequences have a long duration and as polysyllabic words when they have a shorter duration.

METHOD

Participants

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

Materials

Twenty line drawings of nonsense objects were randomly selected from a database of nonobjects (see Fig. 1 for examples). Ten consonant-vowel-consonant (CVC) Dutch nonwords (e.g., *bap*) were selected as monosyllabic novel words. Ten bisyllabic novel words were constructed by adding a second syllable to





Fig. 1. Examples of stimulus displays presented to participants in a two-alternative forced-choice trial (above) and a four-alternative forced-choice trial (below).

each of these monosyllables (e.g., *baptoe*). The items are listed in Table 1. The nonsense-object pictures were randomly assigned to the novel words.

The auditory stimuli for the learning and test phases were spoken instructions to use a computer mouse to click on the picture of the novel word and then on one of four geometric forms appearing on the screen (see Fig. 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").

TABLE 1

Mean Duration of the Monosyllabic Words, the First Syllables of the Bisyllabic Words, and the Training Versions

	Phonetic	Duration (ms)			
Item pair	transcription	Monosyllabic	Bisyllabic	Training ^a	
bap-baptoe	bap-baptu	316	266	292	
fiem-fiemser	firm-firmser	358	252	305	
jom-jomtie	jəm-jəmtir	320	237	278	
kes-keste	kes-kestə	322	257	289	
kuin-kuinwes	kœyn-kœynwɛs	312	288	300	
nim-nimsel	nım-nımsəl	329	221	275	
soer-soerket	surr-surrket	386	281	333	
taaf-taafpag	ta:f-ta:fpax	335	273	304	
tuik-tuikfom	tæyk-tæykfom	331	228	280	
zaf-zafkes	zaf-zafkəs	362	277	313	
Average		337	258		

^aThis duration was the average of the syllable's duration in the monosyllabic and bisyllabic words.

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 a 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 the recordings of the four instruction 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 (Boersma, 2001). For the training version of each monosyllabic word, 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 original monosyllabic and bisyllabic words. For the other two versions, the same token of the syllable was resynthesized such that its duration was either the average for that syllable in the monosyllabic word (the long version) or the average for that syllable in the bisyllabic word (the short version).

The manipulated 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 sentences with monosyllabic targets, the manipulated syllables were followed by the phrase *en dan op de* and the name of each of the four geometric forms. In the sentences with bisyllabic targets, the same manipulated syllables were followed by the second syllable of the word and the same four continuations. 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 eyetracker (an SMI Eyelink system; SensoMotoric Instruments GmbH, Teltow, Germany), 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 (Wittenburg, Nagengast, & Baumann, 1998).

Each trial was structured as follows. First, a central fixation dot appeared on the screen for 500 ms. Then a spoken sentence was presented, and simultaneously a 5×5 grid with the pictures of the novel objects and the geometric forms appeared (see Fig. 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 the participant clicked on the geometric form, the distractor pictures disappeared, leaving only the correct referent on display. At the same time, a sentence indicating whether the response was correct or incorrect was played (*Dat was goed/fout*, "That was right/wrong"). The feedback sentence followed. The trials in the test phase were identical to those in the last three training blocks, except that no feedback was given.

In each of the six blocks of the learning phase, each word was presented three times (in total, 360 trials). During this phase, participants heard only the training version of each novel word (i.e., the version with intermediate duration). For each trial, one or three distractor items were selected randomly from the set of nonsense objects. 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 the presentation of paired items (e.g., *bap* and *baptoe*). Order of presentation in the training phase was identical for all participants.

The test phase included 60 experimental trials in which the target picture was presented with the picture associated with the target word's competitor (e.g., the target *bap* with the *baptoe* object) and two unrelated distractors. Each novel word was presented three times, once each in its training, long, and short 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 of the novel words. The unrelated distractors for all trials were randomly selected. A random order was created for the test block, with the constraints that at least one filler intervened between two experimental trials and at least five items intervened between trials involving paired words (e.g., *bap* and *baptoe*). We created six lists that varied in the order in which the training, long, and short versions

of each novel word were presented. Participants were randomly assigned to one list.

RESULTS

Listeners successfully learned the new words. Error rates dropped from 34.9% in the first training block to 4.3% in the last block. 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, each fixation in the test phase was 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 (target, competitor, unrelated distractor) was computed in each 10-ms slice, in each condition, by summing the number of fixations to each type of picture and dividing it by the total number of fixations in the same time interval. Blinks and saccades were not included in this calculation.

The average proportions of fixations to the target and competitor pictures were computed for the long, training, and short versions of the novel words. Our analyses focused on fixations during the time interval when the proportion of fixations to the competitor was higher than the proportion of fixations to the distractors. This time interval extended until approximately 1,000 ms after target onset in the monosyllabic condition (i.e., when the target was monosyllabic) and until 1,400 ms after target onset in the bisyllabic condition (i.e., when the target was bisyllabic). Because of this difference, analyses were conducted separately for the bisyllabic and monosyllabic conditions. The difference in time intervals is due to the inherent asymmetry between the competitors in the two conditions: The evidence in favor of an embedded monosyllabic competitor given a bisyllabic target word is stronger than the evidence in favor of a bisyllabic competitor given a monosyllabic target word. Consequently, the ambiguity takes longer to resolve when the input is a bisyllable. In both conditions, we did not analyze fixations prior to 200 ms after target onset because that is approximately the earliest time point at which fixation proportions reflect 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).

Analyses of variance (ANOVAs) were computed by subjects (F_1) and by items (F_2) on the average fixation proportions. The average fixation proportions and the results of the ANOVAs are shown in Table 2.

In the bisyllabic condition, participants fixated the target picture most often when the short version was presented, less often when the training version was presented, and least often when the long version was presented. Planned comparisons (two-tailed matched-pairs t tests; see Table 3) indicated a statistically significant difference in fixation proportions between the long and short versions, given a Type I error rate of .05. The analysis of the proportion of fixations to the competitor pictures showed that participants fixated these pictures 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.

In the monosyllabic condition, fixation proportions to the target pictures 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 .05 level. The average proportion of fixations to the competitor pictures was higher when listeners heard the short version, compared with when they heard the long or training version. The

TABLE 2

	Version of syllable		ble		
Condition and picture	Long	Training	Short	Analyses of variance	
Bisyllabic condition					
Target (e.g., <i>baptoe</i>)	.44	.46	.50	$F_1(2, 46) = 3.45, p < .05, p_{rep} = .89, \eta^2 = .13$	
				$F_2(2, 18) = 3.53, p = .08, p_{rep} = .84, \eta^2 = .28$	
Competitor (e.g., <i>bap</i>)	.24	.21	.19	$F_1(2, 46) = 6.06, p < .01, p_{rep} = .97, \eta^2 = .34$	
				$F_2(2, 18) = 5.64, p < .05, p_{rep} = .95, \eta^2 = .38$	
Monosyllabic condition					
Target (e.g., <i>bap</i>)	.35	.40	.31	$F_1(2, 46) = 3.10, p = .055, p_{rep} = .87, \eta^2 = .12$	
				$F_2(2, 18) = 2.98, p = .09, p_{rep} = .83, \eta^2 = .25$	
Competitor (e.g., <i>baptoe</i>)	.21	.21	.27	$F_1(2, 46) = 3.70, p < .01, p_{rep} = .90, \eta^2 = .14$	
				$F_2(2, 18) = 2.91, p = .08, p_{rep} = .84, \eta^2 = .24$	

Average Proportion of Fixations to the Target and Competitor Pictures and Results of Analyses of Variance Conducted on These Means

Note. The proportions shown are averages for the time interval 200 through 1,400 ms after target onset and 200 through 1,000 ms after target onset in the bisyllabic and monosyllabic conditions, respectively.

	Versions compared					
Condition and picture	Long vs. training	Training vs. short	Short vs. long			
Bisyllabic condition						
Target (e.g., <i>baptoe</i>)	$t_1 < 1, t_2 < 1$	$t_1 = -1.77, p = .08, p_{rep} = .82,$ d = 0.28	$t_1 = -2.61, p < .05, p_{rep} = .94,$ d = 0.44			
		$t_2 = -3.37, p < .01, p_{rep} = .96,$ d = 0.33	$t_2 = -2.36, p < .05, p_{rep} = .89,$ d = 0.55			
Competitor (e.g., <i>bap</i>)	$t_1 = 1.94, p = .06, p_{rep} = .86, d = 0.32$	$t_1 = 1.55, p = .13, p_{rep} = .78, d = 0.27$	$t_1 = 3.43, p < .01, p_{rep} = .98,$ d = 0.59			
	$t_2 = 1.91, p = .09, p_{rep} = .83, d = 0.35$	$t_2 = 1.50, p = .17, p_{rep} = .74, d = 0.30$	$t_2 = 3.18, p < .05, p_{rep} = .95, d = 0.60$			
Monosyllabic condition						
Target (e.g., <i>bap</i>)	$t_1 = 3.18, p < .05, p_{rep} = .76,$ d = -0.26	$t_1 = 2.35, p < .05, p_{rep} = .92, d = 0.53$	$t_1 < 1, t_2 < 1$			
	$t_2 = 1.72, p = .12, p_{rep} = .79,$ d = -0.77	$t_2 = 2.02, p = .07, p_{rep} = .84,$ d = 0.48				
Competitor (e.g., <i>baptoe</i>)	$t_1 < 1, t_2 < 1$	$t_1 = 2.21, p < .05, p_{rep} = .90, d = 0.48$	$t_1 = 2.16, p < .05, p_{rep} = .89,$ d = 0.59			
		$t_2 = 1.94, p = .08, p_{\text{rep}} = .83, d = 0.52$	$t_2 = 2.21, p = .055, p_{rep} = .87, d = 0.39$			

TABLE 3

Two-Tailed Matched-Pairs t Tests Comparing the Proportion of Fixations to the Target and the Competitor Pictures

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

planned comparisons indicated that the difference in fixation proportions between the short version and the long version was the statistically strongest difference. Perhaps the effects are clearer for the bisyllabic targets because of the inherent asymmetry between the two conditions: The ambiguity takes longer to resolve when the input contains a bisyllable.

DISCUSSION

The results demonstrate that with very little exposure, listeners were able to make fine-grained distinctions between newly learned word forms. Subtle variations in syllable duration influenced the pattern of fixations that participants made to the target and competitor pictures: Short syllables tended to be interpreted as the first syllable of a bisyllabic word, and long syllables generated more monosyllabic-word interpretations. Additional analyses reported elsewhere (Shatzman, 2006) indicate that the real-word neighborhoods of newly acquired words modulate the influence of prosodic information on the recognition of those words.

The overall pattern of results is incompatible with a simple episodic model of word recognition in which episodes of newly learned words are retained in memory, subsequent speech is compared only with these exemplars, and the best-matching exemplars are identified as the words in the speech input. Such a model predicts that a word will be recognized best when the information in the acoustic signal perfectly matches the stored episodes. Our findings indicate that this is not the case. Rather, recognition of newly acquired words is guided by prosodic knowledge about the relative duration of syllables in existing words (i.e., the knowledge that monosyllabic words tend to have longer durations than the initial syllables of polysyllabic words). 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.

REFERENCES

- Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–439.
- 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.
- Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. Glot International, 5, 341–345.
- 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.
- 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.
- 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.
- 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.

- Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 251–279.
- Hallett, P.E. (1986). Eye movements. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 10:1–10:112). New York: Wiley.
- Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. In *Ohio State University Working Papers in Linguistics*: *Vol. 50* (pp. 101–113). Columbus: Ohio State University, Department of Linguistics.
- Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.
- 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.D. (1987). Functional parallelism in spoken wordrecognition. Cognition, 25, 71–102.
- 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., 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.
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189–234.

- 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, Germany: Mouton de Gruyter.
- Salverda, A.P. (2005). Prosodically-conditioned detail in the recognition of spoken-words (Max Planck Series in Psycholinguistics Vol. 33). Wageningen, The Netherlands: 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.
- Shatzman, K.B. (2006). Sensitivity to detailed acoustic information in word recognition (Max Planck Series in Psycholinguistics Vol. 37). Wageningen, The Netherlands: Ponsen & Looijen.
- 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.
- Wittenburg, P., Nagengast, J., & Baumann, H. (1998). NESU—the Nijmegen experiment setup. In A. Trapp, N. Hammond, & C. Manning (Eds.), *CIP98 conference proceedings* (pp. 92–93). York, England: CTI Centre for Psychology.
- Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32, 25–64.

(Received 5/27/05; Revision accepted 10/31/05; Final materials received 12/6/05) This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.