THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2013
Vol. 66, No. 6, 1227–1240, http://dx.doi.org/10.1080/17470218.2012.736525

Tone of voice guides word learning in informative referential contexts

Eva Reinisch[1,2], Alexandra Jesse[3], and Lynne C. Nygaard[1]

[1]Department of Psychology, Emory University, Atlanta, GA, USA
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3]Department of Psychology, University of Massachusetts, Amherst, MA, USA

Listeners infer which object in a visual scene a speaker refers to from the systematic variation of the speaker's tone of voice (ToV). We examined whether ToV also guides word learning. During exposure, participants heard novel adjectives (e.g., "daxen") spoken with a ToV representing *hot, cold, strong, weak, big*, or *small* while viewing picture pairs representing the meaning of the adjective and its antonym (e.g., elephant–ant for *big–small*). Eye fixations were recorded to monitor referent detection and learning. During test, participants heard the adjectives spoken with a neutral ToV, while selecting referents from familiar and unfamiliar picture pairs. Participants were able to learn the adjectives' meanings, and, even in the absence of informative ToV, generalize them to new referents. A second experiment addressed whether ToV provides sufficient information to infer the adjectival meaning or needs to operate within a referential context providing information about the relevant semantic dimension. Participants who saw printed versions of the novel words during exposure performed at chance during test. ToV, in conjunction with the referential context, thus serves as a cue to word meaning. ToV establishes relations between labels and referents for listeners to exploit in word learning.

Keywords: Tone of voice; Prosody; Word learning; Word meaning; Speech perception.

When listening to a speaker describing a visual scene, listeners try to establish which referents in the scene the speaker may be referring to. To help listeners with this process, speakers can modulate their tone of voice (ToV) to provide additional information about the intended referent (e.g., their size; Nygaard, Herold, & Namy, 2009). A listener hearing a speaker say "Look at these!" in a low, slow, loud voice may correctly infer that the speaker refers to the big trees and not to the small flowers in the garden. ToV thus provides referent information through a modulation of the realization of suprasegmental speech features, such as pitch, speaking rate, and amplitude. This suprasegmental modulation is independent of the prosodic structure of the utterance and is not inherent to

Correspondence should be addressed to Eva Reinisch, Ludwig-Maximilian-Universität München, Institute of Phonetics and Speech Processing, Schellingstraße 3, 80799 Munich, Germany. E-mail: eva.reinisch@gmx.net

The project was funded by a grant of the German Research Foundation (DFG) JE510/2-1 awarded to the first and second authors. This research was also supported in part by National Institutes of Health Research Grant RO1 DC 008108 to Emory University. We thank Felicia Jackson, Felicia Long, and Hayley Heaton for helping with the preparation of materials, and Lauren Clepper for help with running the experiments. We also thank Patricia Bauer and the Memory at Emory laboratory for the use of the eye tracker.

the realization of the phonological word form. In the present study, we examined the role of ToV in word learning. Specifically, we asked whether listeners can use ToV information to learn the meaning of novel adjectives. We also tested what role the visual context plays in word learning from ToV.

Previous studies have shown that ToV can express information about the properties of referents (Kunihira, 1971; Nygaard, Herold et al., 2009; Shintel & Nusbaum, 2007; Shintel, Nusbaum, & Okrent, 2006). Speakers increase, for example, their speaking rate when describing a fast as opposed to a slowly moving object, even when the semantic content of the utterance refers to the direction rather than speed of the object ("It is going left/right"; Shintel et al., 2006). Speakers also modulate the pitch of their voice to express the direction of vertical movement of an object, with higher pitch indicating upward movement (Shintel et al., 2006). Importantly, listeners use ToV to resolve referential ambiguity (Kunihira, 1971; Nygaard, Herold et al., 2009; Shintel & Nusbaum, 2007; Shintel et al., 2006). In a two-alternative forced-choice task, listeners successfully inferred which of two objects the speaker described; for example, whether the fast or the slowly moving object was the original referent for the description "It is going left/right" (Shintel et al., 2006).

ToV can also be used to convey the meaning of novel words. Nygaard and colleagues (2009) found that the acoustic ToV signatures of novel adjectives were consistently related to the assigned meanings of the adjectives (i.e., *big–small*, *hot–cold*, *strong–weak*, *tall–short*, *happy–sad*, and *yummy–yucky*). For example, novel adjectives (e.g., "daxen") intended to mean *big* were consistently produced more loudly, more slowly, and with a lower pitch than when intended to mean *small*. Novel adjectives intended to mean *yummy* were consistently produced with a higher and more variable pitch than when intended to mean *yucky*. Although valence ratings for the adjective meanings correlated with some acoustic properties of the novel utterances, the production of novel adjectives differed across semantic domains. Each semantic domain was reflected in a unique acoustic profile across pitch level, pitch variation, amplitude, and duration. This finding suggests that ToV conveys information about referential properties and hence about the meaning of novel words.

Listeners are sensitive to ToV and use it to find the intended referent of novel adjectives, when presented with pictures related in meaning to the ToV (Herold, Nygaard, Chicos, & Namy, 2011; Nygaard, Herold et al., 2009). Picture pairs differed primarily along the semantic dimension of the assigned adjective meaning (e.g., a big and a small flower were shown for adjectives meaning *big* or *small*). Listeners were on average significantly better at selecting the correct picture if the ToV matched one of the pictures (e.g., the ToV for *big* cued the referent for a picture pair varying along the big–small dimension) than when a mismatching ToV was heard (e.g., the ToV for *big* was presented with the picture pair for the hot–cold contrast). This suggests that listeners take advantage of the unique acoustic signatures related to each adjectival meaning, at least when the semantic dimension of the adjectives could have been inferred from the contrast between the presented pictures.

Since listeners are sensitive to ToV as a cue to the intended referent, we asked here whether listeners can also use ToV to *learn* the meaning of novel adjectives. In the previous work, listeners used the association of ToV with a particular meaning to identify the intended referent. ToV thus established a momentary link between a novel auditory label and a visual referent. Here, we tested whether this momentary link can lead to the long-term learning of a word's meaning. If ToV enables listeners to learn the abstract meaning of a novel adjective, then listeners should subsequently be able to infer the intended referent by retrieving a novel adjective's acquired meaning without relying on ToV. That is, listeners should determine the intended referent from the adjective even when presented with a neutral, semantically uninformative ToV. In addition, listeners should then also infer the intended referent of an adjective when presented with a new visual scene. This generalization would show that

listeners did not simply learn associations between ToV and the referents presented during exposure, but indeed learned the abstract meaning of the novel adjectives.

In the present study we addressed this issue by using an exposure-test paradigm. During exposure, adults listened to sentences containing novel adjectives spoken in a ToV expressing one of six dimensional adjective meanings: *big*, *small*, *hot*, *cold*, *strong*, and *weak*. While listening to these sentences, listeners in Experiment 1 saw picture pairs differing, among other semantic features, along the critical adjectival dimensions (e.g., an elephant and an ant for *big–small*). As in previous work (e.g., Nygaard, Herold et al., 2009), these picture referents established a meaningful visual context that provided information about the critical semantic dimension (e.g., size, temperature, strength) of the novel adjectives. Picture pairs differed along more than just the critical dimension (e.g., ant–elephant are insect vs. mammal, they have different colours, etc.), but the critical dimension was nevertheless the most salient one, especially within the different picture pairs used for each dimension. Importantly, picture pairs alone did not disambiguate word meaning, as they did not provide information about to which of the endpoints within a dimension the novel adjectives were referring. For example, seeing an elephant and an ant could potentially inform listeners that the novel adjective refers to either *big* or *small*, but it cannot tell listeners whether the novel adjective means *big* or *small*. ToV information is thus necessary in order for listeners to infer the *appropriate* mapping within a semantic dimension. Listeners' eye fixations during exposure were tracked in order to monitor both ongoing learning and how listeners establish the momentary link between ToV and the intended referent. No explicit task was given. As listeners learn the meaning of a novel adjective, they should look more at the referent the adjective is referring to then at the referent depicting the adjective's antonym.

At a subsequent test, word learning was assessed by asking listeners to detect the visual referents of the novel adjectives, when presented in a neutral, uninformative ToV. Listeners were explicitly asked to click on the adjectives' visual referents while their eye fixations were recorded. Familiar and novel picture referents were used. If ToV can only be used momentarily during exposure to infer the intended referent, then no evidence of learning should be found at test in the absence of meaningful ToV. If listeners used ToV to learn the label–picture associations during exposure, then learning should be found for familiar referents, but should not transfer to unfamiliar referents. If listeners used ToV to infer and abstract word meaning, then they should be able to identify the intended referents during test, even in the absence of informative ToV and when presented with unfamiliar picture pairs.

In Experiment 2, we tested what role the visual context has in learning the meanings of words from ToV. Picture pairs used in previous ToV studies and in Experiment 1 differed primarily along the most relevant semantic dimension (e.g., elephant and ant). It could thus be the case that the visual context provides information about the semantic dimension of the novel adjective and ToV defines the mapping of the adjective to one of the endpoints within that dimension. Alternatively, previous studies suggest that ToV is expressed with different acoustic signatures for different adjectival meanings (Nygaard, Herold et al., 2009). ToV could thus provide sufficient information about both the semantic dimension and the exact mapping of the adjective. To address this issue, we presented the same auditory stimuli with written versions of the novel words (e.g., "seebow", "daxen") rather than pictures during exposure in Experiment 2. These referents were thus uninformative about the semantic dimension of the novel adjective. In order to learn the word meaning listeners would need to infer word meaning directly from ToV alone, without the constraining visual context.

Experiment 2 was otherwise identical to Experiment 1. During exposure listeners were presented again with the sentences in informative ToV. If ToV conveys information about the exact meaning of an adjective, then listeners should be able to establish a word–meaning relation even when presented with semantically uninformative

visual referents. At test listeners heard the words in sentences spoken with an uninformative ToV, and were asked to pick which pictures the novel adjectives referred to. If listeners had established word–meaning mappings through ToV during exposure, then they should be able to pick the intended referents at test, even in the absence of informative ToV. If, however, ToV constrains the word meaning within an informative semantic context, then learning may only occur when such a constraining visual or semantic context is present, as was the case in Experiment 1. ToV would thus guide word learning in conjunction with an informative referential context.

EXPERIMENT 1

In Experiment 1, we tested whether ToV can guide the learning of novel adjectives. If listeners use ToV to acquire word meaning then, during test, they should retrieve an adjective's meaning and identify the intended referent although adjectives were presented in a neutral ToV. Moreover, listeners should be able to do so even when encountering new potential referents.

We also tested whether listeners use ToV during exposure to establish a momentary link between an adjective label and its referent. Unlike in previous studies, listeners in the current study were not explicitly instructed to detect the matching picture referents. Rather, learning was implicitly monitored by recording listeners' eye fixations (cf. Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). As listeners recognize the adjectives' meanings, they should look more at the intended referent than at the referent of the antonym.

We also tested whether listeners can use ToV to infer the correct adjective–referent mapping during their first encounter with an adjective or only after multiple presentations. Previous work has not addressed the amount of exposure listeners need to establish a label–referent link. However, examining this issue is important, as it speaks to the question of how effective ToV is in resolving referential ambiguity.

Method

Participants

Twenty-four Emory University undergraduates participated for partial course credit or a small payment. All were monolingual native speakers of American English and reported normal hearing and normal or corrected-to-normal vision.

Materials

Six novel adjectives were created (see Table 1). Six English adjectives (*full*, *empty*, *old*, *new*, *hard*, *soft*) were used as fillers. All adjectives were recorded in the phrase *Can you find the [ADJECTIVE] one?* spoken by a female native speaker of American English. The same phrase had been used in previous studies (e.g., Nygaard, Herold et al., 2009). Novel adjectives were first recorded for the test

Table 1. *Acoustic measurements of novel adjective pairs (mean of the two recordings) with informative and neutral (uninformative) ToV*

Novel adjectives pairs	*ToV*	*Sentence duration (ms)*	*Adjective duration (ms)*	*Adjective mean pitch (Hz)*	*Adjective standard deviation pitch (Hz)*	*Adjective RMS amplitude (Pascal)*
foppick–riffel	*hot*	2126	989	285	79.6	0.024
	cold	2711	1224	247	24.7	0.033
	neutral	1768	469	206	23.0	0.027
blicket–tillen	*strong*	2642	862	198	24.9	0.030
	weak	2359	750	236	21.1	0.019
	neutral	1768	344	208	21.9	0.027
seebow–daxen	*big*	2760	969	206	12.7	0.029
	small	2207	838	389	29.4	0.015
	neutral	1825	494	201	35.8	0.026

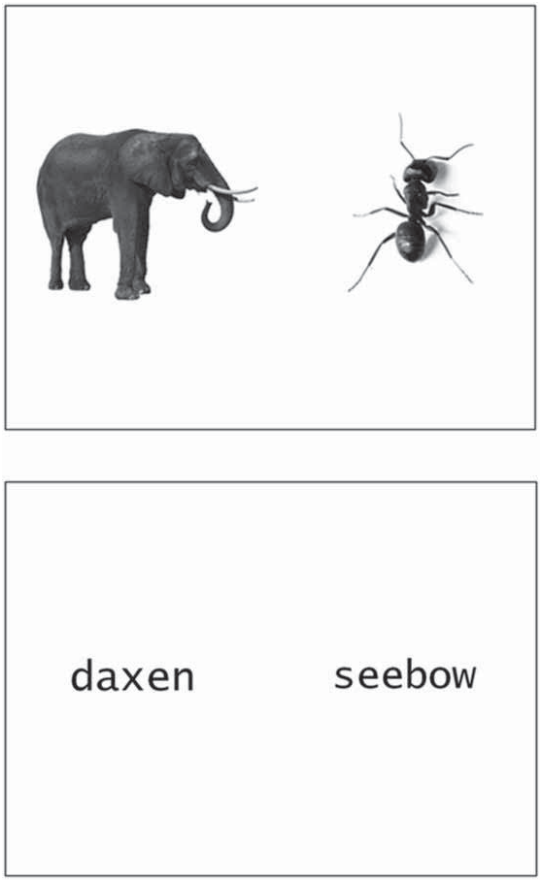

Figure 1. *Examples of screens from the exposure phase with informative referents in Experiment 1 (upper panel) and uninformative referents (i.e., printed versions of the novel adjectives) in Experiment 2 (lower panel). The pictures in Experiment 1 were of photo-like quality and were presented in full colour.*

phase in a neutral ToV without any assigned meaning. Filler and novel adjectives were then recorded in a meaningful ToV for use in the exposure phase. The meanings *hot*, *cold*, *strong*, *weak*, *big*, and *small*, were assigned to the novel adjectives. "Foppick"/"riffel" were selected for *hot–cold*, "blicket"/"tillen" for *strong–weak*, and "seebow"/"daxen" for *big–small*. The speaker was asked to produce the novel adjectives as if she was addressing a child that did not know the word's meaning. Each pair of novel words was recorded with both meanings within its assigned semantic dimension. The acoustic measures for the selected items (see Table 1) show that the ToV productions were clearly distinct within each pair but also different from the tokens spoken with a neutral ToV. In the experiment, word–ToV combinations were counterbalanced across participants. Word-inherent phonetic properties, such as their segmental make-up, could thus not contribute to word learning.

Eight picture pairs were selected to represent each of the meanings of each novel-adjective contrast (see Figure 1, upper panel, for an example). Four additional picture pairs were selected for each filler contrast. Pictures within each pair showed objects or scenes differing, among other properties, in the relevant contrast (e.g., elephant–ant for *big–small*, fire–snowman representing *hot–cold*). All pictures were rated in norming studies as described below to be good representations of the adjectives' meanings, and picture pairs were judged to be representative of the intended adjective contrasts.

Norming

Twenty-eight participants from the same population as in Experiment 1 rated between 9 and 14 picture pairs per adjective contrast (including filler contrasts) in a two-part norming study. In the first session, participants were asked to assign each picture pair to one of the six semantic contrasts. This determined whether a picture pair represented the intended adjective contrast. In the second session, each picture was shown with the adjective it was supposed to represent printed underneath (i.e., the picture of an elephant with the label *big*). Participants' task was to rate on a scale from 1 to 7 how good an example the picture was for the respective category. Assignment of labels ("bad example", "excellent example") to the scale's endpoints was counterbalanced across participants. Picture pairs were selected for a third round of norming if in the first norming session they were assigned to the intended semantic contrast more than 72% of the time, and had received ratings of 4.1 or higher in the second session (corrected for counterbalancing the endpoints of the scale to set "excellent match" to 7). In the third norming study, a new set of 23 participants was asked to rate how well these selected picture pairs represented the respective

adjective contrast on a scale from 1 to 7. Assignment of labels to scale endpoints was counterbalanced across participants. The best eight picture pairs for the critical adjective contrasts and the best four filler picture pairs were then selected for the main study. Mean ratings (when "excellent match" = 7) were: *big–small* 4.6, *hot–cold* 5.1, *strong–weak* 3.2, *empty–full* 5.3, *hard–soft* 4.2, *old–new* 4.3.

Design

On each of 96 exposure trials, participants saw two referents on a computer screen while listening to a sentence spoken with an informative ToV that contained a novel or filler adjective. Across trials, participants heard each adjective eight times, paired twice with each of four picture pairs. The participants thus saw half of the picture pairs selected for the novel adjectives. The picture-pair set shown during exposure was counterbalanced across participants. Assignment of meaning to novel adjectives within a contrast and the position of referents on the screen (left, right) were held constant within but counterbalanced across participants. Targets occurred equally often in each position (left, right) for a given participant. Order of presentation was pseudo-randomized. Participants first saw three filler trials and then all picture pairs once before items were repeated. The order of the rest of the trials was then fully randomized.

At test, participants saw picture pairs while listening to sentences with the novel adjectives spoken in an uninformative ToV. Each picture pair was presented four times, twice with each adjective. First, participants received trials with picture pairs that they had not seen during exposure, and then trials with familiar picture pairs. This order allowed us to assess the transfer of learning to new pictures before assessing the basic learning effect with familiar pictures.

Procedure

Participants were seated individually in a quiet room approximately 60 cm in front of a Tobii TT120 screen (Tobii Technology AB, 33.5 × 27 cm). The eye tracker, sampling at 120 Hz, was controlled using the Tobii Studio Software (version 1.7.2, Tobii Technology AB). The experiment was controlled by E-Prime 2.0 (version 2.0.8.73, Psychology Software Tools).

Each exposure and test trial started with participants looking at a centred fixation cross for 1000 ms before a picture pair appeared. A sentence was played over headphones at a comfortable listening level 500 ms after picture-pair onset. During exposure, participants were asked to listen to the sentences while looking at the screen. The sentences were presented in an informative ToV. At test, participants heard sentences in a neutral ToV and were also asked to click with the computer mouse on the picture the novel adjective was referring to. Pictures remained on the screen for 4500 ms during exposure and until participants responded during test. The inter-trial interval was 500 ms.

Results

Analyses used linear mixed-effect models (Baayen, Davidson, & Bates, 2008), implemented in the lme4 package (Bates & Sarkar, 2007) in R (version 2.10.0; the R Foundation for Statistical Computing). In the eye-tracking analyses, the dependent variable was the log-transformed preference for fixations on the target referent (i.e., the referent representing the meaning indicated by ToV during exposure) over the sum of fixations on the target and the antonym referent (preference = target/(target + antonym)). Fixation data were analysed in 20-ms time bins, from target onset until the average reaction times for clicks during test ($M = 1901$ ms). This time window was shifted by 200 ms, which is the common estimate of the average time needed to program and launch a saccade (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). *P*-values were based on Markov-chain Monte Carlo sampling. Click responses during test were analysed using a logistic linking function.

Best-fitting models were determined through step-wise model comparison using log-likelihood ratio tests. All models included participant as random factor. Initial models only contained an intercept term. A positive intercept significantly

different from zero indicated target preference. Whenever given such target preference, adjective contrast was evaluated as a fixed factor, mapping the effect of target preference for the hot–cold contrast onto the intercept. Regression weights significantly different from zero reflect the change in the intercept when accounting for performance for other contrast conditions (i.e., *strong–weak* and *big–small* compared to *hot–cold*). If contrast had an effect, separate intercept-only models were run for each adjective contrast. By further adding adjective as a fixed factor in these models, we tested whether both adjectives of a pair were recognized above chance. For the data collected during the exposure phase, we additionally analysed fixation behaviour during the first two presentations of each novel adjective. Each presentation was paired with previously unseen pictures. This analysis was intended to determine whether listeners immediately associated words and pictures by means of ToV or whether they needed multiple repetitions of words and referents in order to draw this relation. The number of trials needed to infer a ToV–picture relation is yet unknown, as previous studies only reported analyses pooled over the entire experiment (e.g., Nygaard, Herold et al., 2009). Here, we only compared the first part of the exposure with the complete exposure phase since only the first part was pseudo-randomized such that all picture pairs occurred once before they were repeated. The rest of the exposure phase was randomized so that listeners could not predict a word–meaning pairing when a picture pair was presented again. For the analyses of the test phase, picture familiarity was additionally evaluated as a fixed factor (familiar coded as –0.5, unfamiliar as 0.5).

Exposure phase

Target preference in eye fixations during exposure was analysed to track the learning process. Figure 2 shows the proportion of fixations over time by contrast. Intercept-only models indicate that listeners preferred looking at the target referent during exposure, $b_{(\text{intercept})} = 0.65$, $p_{(\text{MCMC})} < .001$.

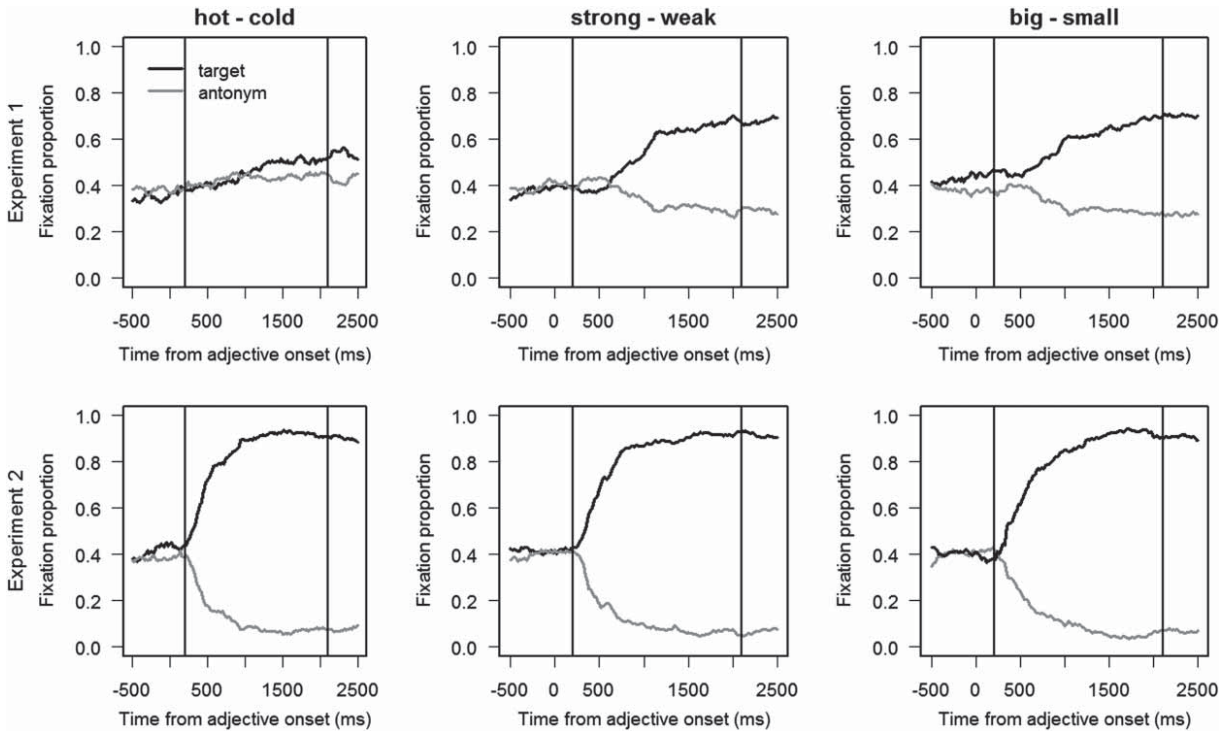

Figure 2. *Fixation proportion to the target referent and the referent of the antonym over time during exposure. Solid vertical lines indicate the time window of analysis.*

Table 2. *Results of statistical models for listeners' target preference during exposure (eye-fixation data for Experiment 1 only) and during test (click responses and eye-fixation data for both experiments)*

	Exposure phase		*Test phase*			
	Eye fixations		*Click responses*		*Eye fixations*	
	b	$p_{(MCMC)}$	*b*	*p*	*b*	$p_{(MCMC)}$
Experiment 1						
Intercept (hot–cold)	0.18	. 15	1.08	<.001	0.66	<.001
Adjustment for strong–weak	0.63	<.001	1.19	<.001	0.32	<.005
Adjustment for big–small	0.78	<.001	0.43	<.001	0.19	.09
Experiment 2						
Intercept (hot–cold)	n/a	n/a	–0.11	.55	–0.07	.59
Adjustment for strong–weak	n/a	n/a	0.51	<.001	0.29	<.01
Adjustment for big–small	n/a	n/a	0.50	<.001	0.18	.09

Table 3. *Target preference during exposure (eye fixations) and test (click responses and eye fixations) separately for each adjective contrast for both experiments*

	Exposure phase		*Test phase*			
	Eye fixations		*Click responses*		*Eye fixations*	
	b	$p_{(MCMC)}$	*b*	*p*	*b*	$p_{(MCMC)}$
Experiment 1						
Hot–cold	0.20	.25	1.54	<.05	0.66	<.005
Big–small	0.95	<.001	2.97	<.001	0.85	<.005
Strong–weak	0.80	<.001	2.41	<.001	0.99	<.001
Experiment 2						
Hot–cold	n/a	n/a	–0.23	.608	–0.08	.70
Big–small	n/a	n/a	0.76	.132	0.12	.53
Strong–weak	n/a	n/a	0.48	.083	0.21	.13

This target preference was significantly larger for the *big–small* and *strong–weak* contrast than for the *hot–cold* contrast (see Table 2). Follow-up analyses by contrast showed a significant target-preference effect only for *big–small* and *strong–weak* (see Table 3).

The analysis of data from the first part of the exposure phase, during which all picture pairs were presented only once, indicated no overall preference for the target over the competitor, $b_{(\text{intercept})} = 0.07$, $p = .57$. Listeners thus did not reliably relate words and referents through ToV on the first encounter, but eventually learned this association over the course of eight repetitions of each word.

Test phase

Click responses. Figure 3 depicts the mean number of correct click responses for each adjective meaning by participant. Intercept-only models show that participants learned the meaning of the novel adjectives, $M_{(\text{correct})} = 74.2\%$, $b_{(\text{intercept})} = 1.56$, $p < .001$. The performance was above chance for all three adjective contrasts (see Table 3), but better with the big–small and the strong–weak contrasts than with the hot–cold contrast (see

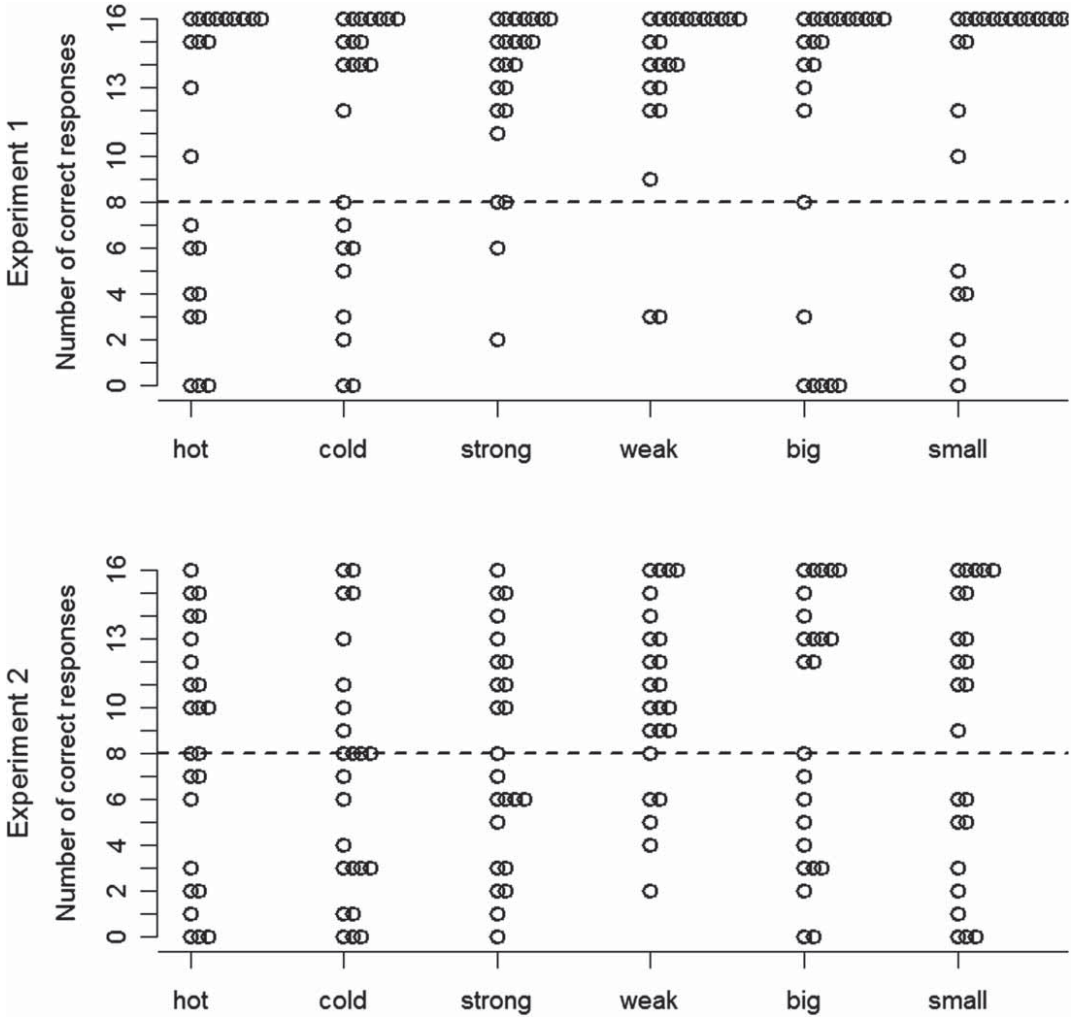

Figure 3. *Number of correct click responses for participants in Experiment 1 and Experiment 2 during test. Each circle represents the number of correct responses for one participant for the respective adjective meaning.*

Table 2). Picture familiarity did not affect performance, $\chi^2(1) = 0.027$, $p = .87$.

Eye-tracking data. Figure 4 shows the eye-fixation data during test. Participants fixated target pictures more frequently than antonym pictures during test (intercept-only model: $b_{(\text{intercept})} = 0.83$, $p_{(\text{MCMC})} < .001$). This target preference was significant for all contrasts (see Table 3) but smaller for the *hot–cold* contrast than for the other two contrasts (see Table 2). Analyses with adjective as additional factor showed that *hot* and *cold* differed significantly from each other. When *hot* was mapped onto the intercept, the target preference for this contrast was only marginally significant, $b_{(\text{intercept:hot})} = 0.45$, $p_{(\text{MCMC})} = .06$; $b_{(\text{cold})} = 0.42$, $p_{(\text{MCMC})} < .05$. The preference for *cold* as the target was, however, above chance, $b_{(\text{intercept:cold})} = 1.78$, $p_{(\text{MCMC})} < .01$; $b_{(\text{hot})} = -0.48$, $p_{(\text{MCMC})} = .05$. Adjectives of the *strong–weak* contrast were reliably recognized, $b_{(\text{intercept:strong})} = 1.04$, $p_{(\text{MCMC})} < .001$; $b_{(\text{weak})} = -0.09$, $p_{(\text{MCMC})} = .53$. The same was true for *big* and *small*, $b_{(\text{intercept:small})} = 0.97$, $p_{(\text{MCMC})} < .001$; $b_{(\text{big})} = -0.24$, $p_{(\text{MCMC})} = .10$. Picture familiarity did not affect performance in the model with contrast and familiarity as factors, $\chi^2(1) = 2.77$, $p = .09$.

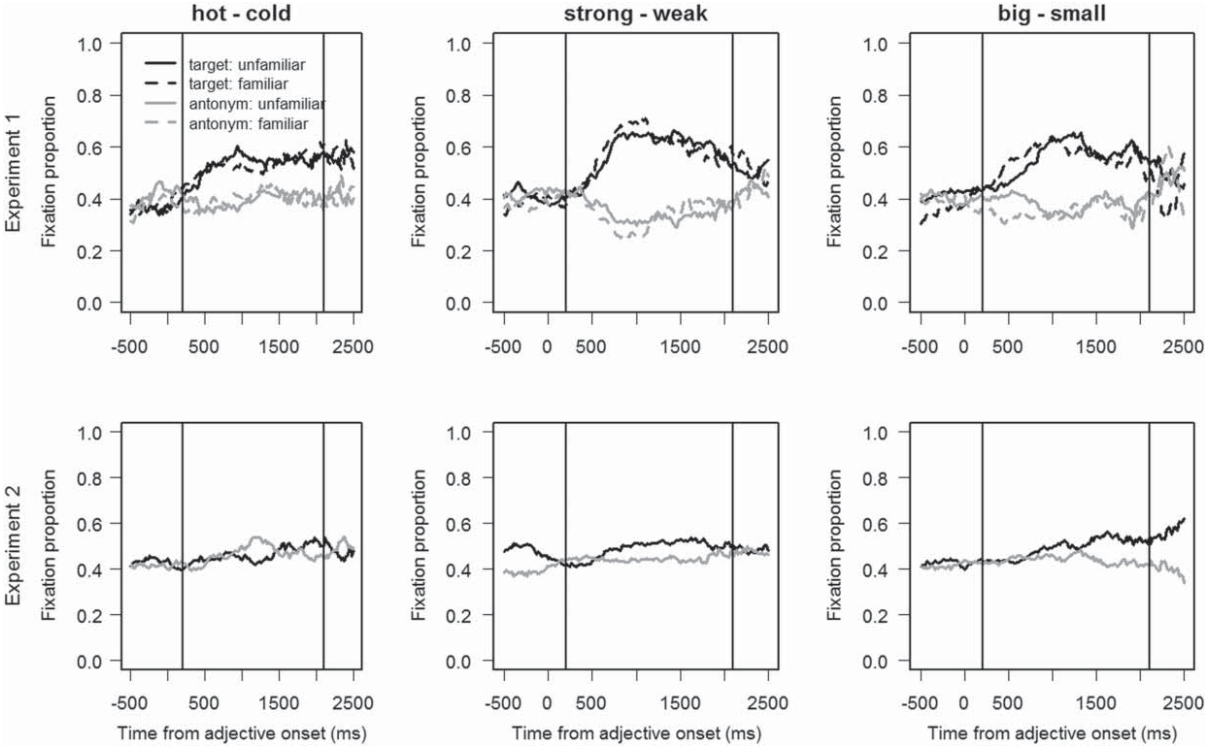

Figure 4. *Fixation proportion on target referents and antonym referents over time during test for the three adjective contrasts. Vertical lines indicate the time window of analysis.*

DISCUSSION

In Experiment 1, we showed that listeners were able to learn the meaning of novel adjectives through ToV. During exposure, listeners associated the words they heard spoken with informative ToV with the intended picture. Listeners did not show a preference for the intended referent on the first encounter, however, but eight repetitions of each word throughout the exposure phase were sufficient to trigger learning. During test, when listeners heard the same words in an uninformative ToV they were able to select the intended picture referents even when the referents had not been presented during exposure. Listeners performed above chance for all adjectives with the exception of the adjectives for *hot*. One possible explanation is that *hot* is associated with multiple meanings. Thus, although our pictures depicted only the temperature dimension of *hot*, ToV–referent mappings for *hot* may be more complex than for the other dimensions. Pooled over contrasts, however, all adjectives were learned. Listeners generalized the meaning of adjectives in the absence of informative ToV. Moreover, listeners were able to find the intended referents of the novel adjectives both in the absence of ToV and when presented with a set of new picture referents. This shows that listeners learned the meaning of the novel adjectives from ToV.

EXPERIMENT 2

In Experiment 1, we showed that listeners use ToV to learn the meaning of novel adjectives. In Experiment 2, we examined the role of the visual context in this process. In Experiment 1, just as in previous studies, the visual context consisted of pictures of objects. These objects differed in previous studies and in Experiment 1 most saliently, in the relevant semantic dimension (e.g., in size). The visual context thus was likely to provide the listener with information about the

relevant semantic dimension. That is, listeners could have determined from the visual context that the novel adjectives "seebow" and "daxen", for example, referred to either *big* or *small*. ToV could have subsequently provided the additional information necessary to allow the listener to map each adjective onto the correct endpoint of this semantic dimension (e.g., "daxen" means *small*). Note that although visual context might constrain the relevant semantic dimension, learning could not have occurred without ToV providing the necessary specific referent information. Thus, Experiment 1 clearly shows that listeners use ToV during word learning. Nevertheless, a critical question is whether word learning from ToV must operate within an informative referential context that provides information about the relevant semantic dimension. Prior work suggests that ToV is realized with a unique acoustic profile for a given meaning (Herold, Nygaard, & Namy, 2012; Nygaard, Herold et al., 2009) and hence, it is possible that ToV alone is sufficient to guide word learning. If this is the case and informative ToV is sufficiently specific, then word learning should also occur when the pictures during exposure are replaced with an uninformative visual context. We tested this in Experiment 2 by presenting listeners with printed versions of the novel adjectives during exposure. If ToV alone is sufficient to guide word learning, then we should replicate the learning effect found in Experiment 1 in Experiment 2, even when the visual context is not informative about the semantic dimension during exposure.

At the same time, the prior work on ToV is inconclusive with respect to how semantically specific ToV is and whether listeners can indeed extract specific meaning information from ToV, as listeners were always tested while being provided with a visual context that conveyed the semantic dimension of ToV (Nygaard, Herold et al., 2009). If visual context must provide information about the relevant semantic dimension for ToV to guide word learning, then no learning should be found in Experiment 2. In Experiment 2, we thus explored the mechanisms of learning the meaning of novel adjectives through ToV.

Methods

Participants

Twenty-four new undergraduates from the same population as for Experiment 1 took part for a small payment.

Materials, design, procedure

Materials, design, and procedure were identical to Experiment 1 with the exception that during exposure the picture pairs were replaced with written versions of the adjectives (see Figure 1, lower panel). The same sentences, spoken in an informative ToV, were presented as during exposure in Experiment 1. The test phase was identical to Experiment 1. Participants were presented again with the sentences spoken in a neutral ToV and with picture referents. The task was to click on the picture to which the adjective referred. Because printed words were shown during exposure, the picture pairs during test—which were the same as in Experiment 1—were all new to participants. The same presentation lists as in Experiment 1 were used, counterbalancing the side of the referent across participants and retaining the same trial order as in Experiment 1. The test phase was identical to Experiment 1. All analyses were conducted as described for Experiment 1.

Results

Exposure phase

Figure 2 shows the proportion of fixations over time by contrast. As expected, intercept-only models showed that listeners preferred looking at the target word, $b_{(\text{intercept})} = 2.43$, $p_{(\text{MCMC})} < .001$. This effect did not differ across contrasts, $\chi^2(2) = 1.16$, $p = .56$. This preference for looking at the printed target word cannot, however, inform about learning but rather reflects that while hearing words, listeners spontaneously direct their gaze to the corresponding printed words (e.g., McQueen & Viebahn, 2007). This result can thus be taken as evidence that listeners processed the sentences. However, unlike in Experiment 1 for the novel adjectives

and picture contexts, the link between heard word and printed-word "referent" was not a semantic one.

Test phase

Click responses. Figure 3 depicts the average number of correct click responses for each adjective meaning by participant. Intercept-only models show that participants did not learn the meaning of the novel adjectives, $M_{(\text{correct})} = 54.5\%$, $b_{(\text{intercept})} = 0.22$, $p = .21$. Participants did not perform better than chance for any of the three adjective contrasts (see Table 3). A numerically small preference for the non-target item in the *hot–cold* contrast (see Figure 3) led to a statistical difference in performance for that contrast compared to that for the other two contrasts (see Table 2).

Eye-tracking data. Figure 4 shows the eye-fixation data during test. Participants had no target preference overall, $b_{(\text{intercept})} = 0.09$, $p_{(\text{MCMC})} = .47$, or by contrast (see Table 3). The difference between the *strong–weak* condition and the *hot–cold* condition in click responses for this group (see Table 2) is hence not meaningful.

DISCUSSION

Experiment 2 tested the role of the referential context in listeners' use of ToV in word learning. By replacing the semantically informative picture referents from the exposure phase of Experiment 1 with semantically uninformative printed versions of the novel adjectives we showed that listeners do not learn the meaning of novel adjectives directly from the acoustic signatures of ToV alone. This suggests that the picture pairs in Experiment 1 provided essential additional information for establishing the word-to-meaning mapping, for example, by providing information about the semantic dimension of the adjectives. ToV then further constrained the exact adjective-to-referent mapping. Listeners used these mappings to extract and learn the meaning of the novel adjectives.

GENERAL DISCUSSION

Listeners can use ToV to learn the meaning of novel adjectives when presented with visual referents that contain information about the relevant semantic dimension. This learning was demonstrated by showing that, at test, listeners can infer the intended referent from hearing the adjective alone, in the absence of informative ToV. Importantly, listeners can still identify the intended referent even when presented with new picture pairs. ToV thus does not simply elicit the learning of particular adjective–referent pairings, but guides word learning. Listeners presented with printed versions of the novel adjectives as visual context during exposure did not learn the adjectives' meaning. The relationship between the visual scene and ToV thus mediates word learning. In summary, these results suggest that listeners use the momentary link that ToV establishes between a new auditory label and a visual referent to abstract and learn the meanings of novel adjectives.

The results of Experiment 1 extend previous findings by showing for the first time that language users use ToV to learn the meaning of novel words. Previous work has shown that listeners reliably associate informative ToV with picture referents (Nygaard, Herold et al., 2009), but the present study demonstrates that ToV can guide the learning of word meanings. Our study also provides information about how listeners use ToV to resolve referential ambiguity. We demonstrated that listeners draw links between ToV and visual picture referents during exposure without being explicitly asked to detect the intended referents. That is, listeners inferred by themselves that ToV provided referential information that relates to what they saw. Second, listeners readily remembered these relations after only eight repetitions presented during exposure. Although previous work (Nygaard, Herold et al., 2009) has shown that listeners reliably associate informative ToV with picture referents, this study demonstrates that ToV was used to learn word meanings, despite having only relatively few exposure trials and the absence of an explicit task.

Experiment 2 provided insight into the mechanisms underlying the use of ToV during word learning. Previous work (e.g., Nygaard, Herold et al., 2009) had not explored the role of the referential context in establishing the intended referent from ToV. Testing the role of ToV with different types of exposure allowed us to specify the role of referential context for the use of ToV in word learning. Listeners exposed to referents that were uninformative about referential properties were not able to learn the novel adjectives' meanings—at least not with the same amount of exposure. This suggests that ToV constrains word meaning in relation to possible referents present in the listening situation. However, hearing informative ToV during exposure, even in the presence of a constraining visual context, was necessary to guide learning. Seeing the picture pairs in Experiment 1 probably helped constrain the semantic dimension of the novel adjective, but could not have led to learning alone. Participants may have been sensitive to the fact that whenever they heard "seebow" or "daxen", for example, they saw two objects most saliently differing in size, and could have hence inferred that these adjectives must relate to the size dimension. Filler trials with known adjectives could also have helped listeners to realize that the pictures depicted an adjectival contrast. The pictures were, however, not informative about the mapping of the novel adjectives to the respective endpoints of the semantic dimension. If listeners did not use ToV during exposure, they may have started to map words consistently to pictures/meaning but the specific word–meaning associations would not have been systematic across participants. The consistent mapping we found here could have only been inferred from ToV. Our results show, therefore, that ToV operates within a referential context rather than as a stand-alone cue to word meaning.

These results form the basis for further research into the role of the referential context. Note that the visual context in Experiment 1 and in previous studies was highly informative about the critical semantic dimension. Future research is needed to establish the ubiquity of ToV as a tool for word learning by testing whether ToV can also guide learning in other, possibly less constraining, visual or semantic contexts, such as in more complex visual scenes, or without the presence of contrastive referents (e.g., when only one of the adjectives of a contrast is shown). Note that so far all previous studies on ToV–referent associations used contrastive referents.

Looking at word learning more generally, ToV surely is only one of many cues available to listeners to detect referents and establish word–meaning relations. Listeners, both children and adults, can, for example, sometimes use sound symbolism to relate the phonetic make-up of certain word forms to features of their referents (e.g., Maurer, Pathman, & Mondloch, 2006; Nygaard, Cook, & Namy, 2009a). Also, word learning can be established through cross-modal temporal alignment of referent motion and accompanying speech. Speakers align, for example, the motion they impose on a referent object to the prosodic structure of their speech (Jesse & Johnson, 2012). Listeners are sensitive to this prosodically mediated cross-modal alignment in detecting referents of novel noun labels (Jesse & Johnson, 2012). Toddlers use this cross-modal relationship to learn the meaning of novel nouns (Jesse & Johnson, 2008). In addition, children appear to use ToV to infer the meaning of novel contrastive adjectives (Herold et al. 2011) and mothers appear to use ToV spontaneously when reading storybooks to their children (Herold et al., 2012). ToV may thus operate jointly with other cues in referent detection and word learning. The relative importance of ToV and other linguistic and non-linguistic cues in learning the meaning of novel adjectives, and of other word types (e.g., nouns), thus needs to be determined across various referential contexts and at various stages of language learning.

The present study demonstrated that ToV can contribute to the learning of novel adjectives. Meaningful ToV in conjunction with a constraining visual referential context allows adult learners to infer word meaning quickly and to generalize these newly learned forms to novel referents. In line with studies on the role of prosody for establishing cross-modal label–referent mappings during referent detection for word learning (Jesse & Johnson, 2008, 2012),

the present study highlighted listeners' ability to exploit relations between auditory labels and visual objects in word learning. ToV facilitates referent detection and consequently word learning by establishing links between auditory labels and the visual properties of their referents. ToV thus reduces the arbitrariness between label and referent and may hence also be a potential candidate tool for word learning in both adults and children as lexical items are acquired.

Manuscript received 14 May 2012
Revised manuscript received 24 September 2012
First published online 8 November 2012

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.

Baayen, H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bates, D. M., & Sarkar, D. (2007). , *lme4: Linear mixed-effects models using S4 classes* (version 0.999375-27) [software application]. Retrieved June 28, 2011, from www.r-project.org

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

Herold, D. S., Nygaard, L. C., Chicos, K., & Namy, L. L. (2011). The developing role of prosody in novel word interpretation. *Journal of Experimental Child Psychology*, *108*, 229–241.

Herold, D. S., Nygaard, L. C., & Namy, L. L. (2012). Say it like you mean it: Mothers' use of prosody to convey word meaning. *Language and Speech*, *55*, 423–436.

Jesse, A., & Johnson, E. K. (2008). Audiovisual alignment in child-directed speech facilitates word learning. *Proceedings of the International Conference on Auditory–Visual Speech Processing* (pp. 101–106). Adelaide, Australia: Causal Productions.

Jesse, A., & Johnson, E. K. (2012). Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *Journal of Experimental Psychology: Human Perception and Performance*, Advance online publication. doi:10.1037/a0027921

Kunihira, S. (1971). Effects of the expressive voice on phonetic symbolism. *Journal of Verbal Learning & Verbal Behavior*, *10*, 427–429.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*, 316–322.

McQueen, J. M., & Viebahn, M. (2007). Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology*, *60*, 661–671.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*, 181–186.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, *33*, 127–146.

Shintel, H., & Nusbaum, H. C. (2007). The sound of motion in spoken language: Visual information conveyed by acoustic properties of speech. *Cognition*, *105*, 681–690.

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, *55*, 167–177.

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.