Journal of Memory and Language 63 (2010) 295–305

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task

Toni Cunillera [b,d,*], Matti Laine [d], Estela Càmara [b], Antoni Rodríguez-Fornells [a,c]

[a] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[b] *Department of Basic Psychology, Faculty of Psychology, University of Barcelona, 08035 Barcelona, Spain*
[c] *IDIBELL, University of Barcelona, 08907, L'Hospitalet de Llobregat, Barcelona, Spain*
[d] *Department of Psychology and Logopedics, Åbo Akademi University, FIN-20500 Åbo, Finland*

ARTICLE INFO

Article history:
Received 11 November 2009
revision received 27 May 2010
Available online 23 June 2010

Keywords:
Speech segmentation
Word-to-world mapping
Statistical learning
Language learning
Audio-visual speech

ABSTRACT

How are adult second language learners able to segment words and map them to referents in the new language? The present study explores this unresolved issue by using a new multimodal learning paradigm that tracks the first steps in learning new words and their mappings to visual referents. It encompasses a continuous audiovisual stream in which transitional probability of syllables is the only acoustic cue available to segment the stream into words, and a visual stream of object images that accompanies the novel words. The objects are systematically varied in terms of constancy of word-picture association and meaningfulness. The results indicated good word-referent mapping and word segmentation after short exposure to the audiovisual stream. Mapping words with pictures was more effective when the visual referents were meaningful objects. In word segmentation, the consistency of the word-picture association affected segmentation performance. The effect of associative strength on segmentation performance was most prominent with meaningful objects, albeit associative strength did not interact significantly with meaningfulness. The present results suggest that word segmentation and word-referent mapping are closely related processes: word segmentation is affected by the consistency of the mapping relationship and both segmentation and mapping can be accomplished under the same short exposure.

Introduction

To acquire words of a new language, learners need to face two difficult initial challenges. First, the continuous speech signal needs to be segmented into discrete words (the *segmentation problem*), and second, when these units are identified, they have to be mapped onto conceptual representations (the *word-to-world mapping problem*). The present study lies in the intersection of both processes: how adult learners assign word-to-world mappings while simultaneously segmenting words.

Speech segmentation is a considerable feat as the continuous and initially nonsense speech signal does not provide reliable acoustic cues signaling word boundaries. In order to solve the segmentation problem, the learner must use a number of cues available in the speech input such as prosody, stress patterns, allophonic variation, phonotactic regularities and the distributional properties of words (Jusczyk, 1999). Related to the last cue, it has been shown that language learners are able to detect word candidates in continuous speech by tracking the transitional probabilities of syllable combinations (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Low transitional probabilities (low likelihood of a syllable following another one) are found across

[*] Corresponding author at: Dept. Psicologia Bàsica, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, Barcelona 08035, Spain. Fax: +34 93 4021363.
E-mail address: tcunillera@ub.edu (T. Cunillera).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.05.003

words, whereas high transitional probabilities are found within words. This particular computational strategy is at the heart of statistical learning, a domain-general learning mechanism implicated in diverse sequential learning situations involving artificial syntax (Gomez & Gerken, 1999), tone sequences (Abla, Katahira, & Okanoya, 2008; Saffran, Johnson, Aslin, & Newport, 1999) and speech (Saffran, Aslin, et al., 1996; Saffran, Newport, et al. 1996).

Regarding the second challenge, the *word-to-world mapping problem*, learners need to correctly extract possible meanings from extralinguistic contexts and to link them to the phonological representations that have been identified and temporarily stored. This process is a very important step in language acquisition as it clearly lies at the core of language ability, i.e., the capacity to create conceptual representations that are linked to arbitrary sounds and symbols. Gleitman, Cassidy, Nappa, Papafragou, and Trueswell (2005) have recently argued that part of the bottleneck for young infants to learn a language resides in the mechanisms (innate or learned), constraints and tools available for solving the *word-to-world mapping problem*. Associationism has been the most popular mechanism to explain how children create a word-to-world mapping: when hearing a word, the infant associates it with the object that is simultaneously presented in the external world (Richards & Goldfarb, 1986). With repeated episodes, this sound-meaning association will be strengthened. This view is similar to the proposal that a general learning mechanism sensitive to statistical properties of the input (co-occurrences of syllables, words and objects, or words and other type of referents) might account for the discovery of linguistic structures, as in the case of speech segmentation (Saffran, 2003).

The associative perspective has encountered strong opposition from researchers advocating for the importance of other cognitive processes necessary for language learning, such as the understanding of others' intentions or syntactic knowledge (Bloom, 2000; Tomasello, 1992). In addition, the associative accounts cannot explain how the learner is able to choose the correct referent in a visual scene, a problem which was first introduced by Quine (1960), and which is known as the *reference uncertainty problem*: a new word could be heard in different contexts or visual scenes in which several possible referents are also available. In this regard, it is estimated that less than 7% of the speech directed to children consists of clear isolated words (Brent & Siskind, 2001; Weijer, 1998). Thus, the inherent variability of natural speech and natural scenes increases the difficulty of finding out clear spatial–temporal relations between specific speech sound combinations and referents.

In order to overcome this problem, a cross-situational learning mechanism has been proposed, which takes advantage of the distributional properties of both visual and auditory information (Gleitman, 1990; Pinker, 1989). For example, when a child hears a word, he/she can hypothesize a set of potential referents for that word. After hearing that word in several different utterances, each one produced in different contexts, the child can intersect the corresponding sets in order to discover a referent that is consistent across the different occurrences of that word. In a recent study, Yu and Smith demonstrated the validity of this account in human adults (Yu & Smith, 2007) and infants (Yu & Smith, 2008). In their study, participants were able to learn word-picture pairs in few consecutive trials through computations of cross-situational statistics, in a context where each time several words and pictures were simultaneously presented (for computational simulations of speech-visual cross-situational learning, see Roy and Pentland (2002), Siskind (1996), and Yu, Ballard, and Aslin (2005)).

As implied by the short review above, these two initial challenges of language learning, the *segmentation problem* and the *word-to-world mapping problem*, have been investigated successfully within two separate research programs: (i) the statistical learning approach (Aslin et al., 1998; Saffran, Aslin, et al., 1996) and (ii) the word-learning approach (Baldwin, 1993; Bloom, 2000; Carey & Barlett, 1978; Markman, 1990). Whereas the first line of research has focused mostly in studying the process by which infants can segment the continuous speech signal into isolated phonological units using statistical regularities, the second one has focused on how infants acquire the meaning of new words or how phonological representations and new labels are mapped onto external referents. Interestingly, both lines of research have neglected an intermediate learning process which should involve mapping the new segmented word into a possible conceptual representation (Saffran & Graf-Estes, 2006). Several interesting questions can be raised here: what is the nature of the output of the statistical learning mechanism and do the learning mechanisms underlying segmentation and mapping interact with each other?

The present study is devoted to these issues and explores the interface between speech segmentation and word learning. The first question concerns the nature of the output of the speech segmentation process based on statistical information. Saffran (2001) have proposed that the "representations emerging from statistical learning may serve as candidate lexical items for infants, available for integration into the native language" (p. 9). In that study, the author exposed infants to previously segmented words vs. non-words, presented at the end of meaningful and meaningless sentences using a standard familiarization task. The authors demonstrated that infants processed differently the newly segmented words vs. the non-words only when they were presented at the end of meaningful sentences. This result was further confirmed in a grammar learning paradigm in which 12-month-old infants were able to discover syntactic regularities during on-line segmentation of new words (Saffran & Wilson, 2003). Converging evidence is also found from recent neuroimaging studies (Abla et al., 2008; Cunillera, Toro, Sebastian-Galles, & Rodríguez-Fornells, 2006; Cunillera et al., 2009; De Diego-Balaguer, Toro, Rodríguez-Fornells, & Bachoud-Levi, 2007; Sanders, Newport, & Neville, 2002) demonstrating that during on-line segmentation, successfully isolated words showed a larger N400 component, an event-related brain potential that has been related to lexical and semantic processing (Kutas & Federmeier, 2000). Overall, these studies and the previous ones argue in favor of the existence of a special lexical or *proto-lexical* status acquired by the segmented new words and against a mere sound-string representation based on a pattern of high and low transitional probabilities (but see Endress & Mehler, 2009).

The second question deals with the possible interaction between segmentation and word-to-world mapping. It has recently been shown that meaning can be integrated more easily in recently segmented words when compared to other types of new words. Hollich (2006) observed that 23-month-olds learned better new word-object pairings if the new word was previously embedded in real English sentences. As children might have been able to segment a new word during the presentation of the two sentences containing it (Jusczyk & Aslin, 1995), this study seems to indicate a tendency of segmented new words to be linked with meaningful referents. In this vein, Graf-Estes, Evans, Alibali, and Saffran (2007) showed in 17-month-olds an advantage of segmented words vs. non-words in mapping to new meanings. After the speech segmentation task, infants had to learn object-word pairings using labels (words or part-words) heard during the segmentation task. The observed advantage of words over part-words suggests that the recently created phonological representations are available to be mapped with a visual referent. Moreover, in an analogous study to the previous one conducted with adults, Mirman, Magnuson, Graf-Estes, and Dixon (2008) showed that word-object pairings were learned faster when the auditory items were word-like units (defined by high string-internal transitional probabilities) or new words rather than part-words (low string-internal transitional probabilities).

To sum up, the earlier studies suggest that language learners might attach meaning to recently segmented words in two subsequent steps: first, they construct a meaningless lexicon based on the output of speech segmentation and second, conceptual information will subsequently be linked to the proto-lexical words already stored. In fact, the existence of this *proto-lexicon* might make the mapping-to-meaning process easier, as infants could focus exclusively on discovering the associated conceptual representation and not the phonological structure of the newly learned word. However, it is an open issue whether segmentation and conceptual mapping take place in a serial fashion as described above or whether they can interact in a more or less parallel process. Moreover, previous studies have not differentiated between two closely related but separate factors in word-to-world mappings, namely the meaningfulness of the referent and the consistency of the mapping.

In the present experiments, we studied the interplay of word segmentation and word-object mapping with a new audiovisual learning paradigm (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010). The auditory input consisted of artificial language streams (Graf-Estes et al., 2007; Saffran, Aslin, et al., 1996) composed of nonsense words and presented as a continuous syllabic stream (see Fig. 1). Along with the language stream, visual stimuli were added

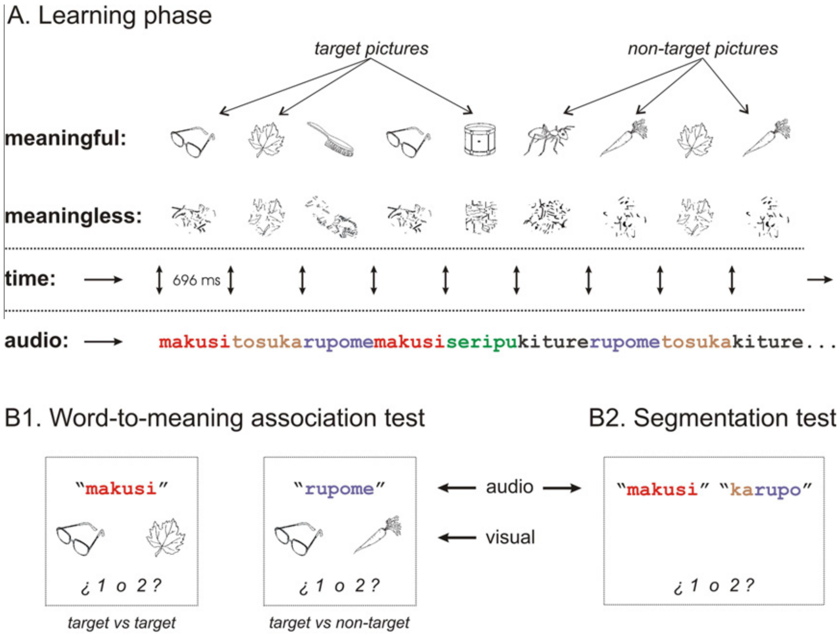

Fig. 1. Illustration of the procedure used at the learning phase and at the test phases. (A) Example of the audiovisual stream. Here *makusi*, *tosuka* and *seripu* are the associated novel words paired with three target pictures (*makusi*-glasses, *tosuka*-leaf, and *seripu*-drum). In the whole stimulus sequence, the association strength was about 92%. The other three novel words, *kiture*, *rupome*, and *seripu*, correspond to the non-associated words that are unsystematically paired with the non-target pictures. (B) Here B1 shows an example of the two types of word-picture matching trials presented in the meaningful condition. In the first type (left), two target pictures are presented simultaneously with an auditory word that has been consistently associated with one of the pictures at the learning phase (e.g., *makusi*-glasses). In the second type (right), a target and a non-target picture are presented together with an auditorily presented non-associated word (e.g., *rupome*). Here one should choose the non-associated picture that was weakly associated with the word, because the other picture had a consistent association with another word. B2 shows a visual example of the auditory speech segmentation test (2AFC test). One-color novel words correspond to "words" from the stream. Two-color novel words correspond to part-words from the stream. For both tests, the "¿ 1 o 2 ?" prompts the forced response.

and delivered in synchrony with the word onsets and offsets. Visual input consisted of meaningful (concrete objects) and meaningless (scrambled pictures) referents (see Fig. 1). At the same time, we manipulated the consistency of the word-referent associations, a prerequisite for creating word-to-world mappings (cross-situational co-occurrences). In order to have a more realistic setup, the association between the novel words and the objects was never fully consistent. Immediately after the exposure, participants performed either a word-learning test in which their ability to map the corresponding referent (visual object or scrambled picture) to the segmented words was evaluated (Experiment 1), or performed a standard word segmentation task (Experiment 2). Besides the important separation of word-referent association and its meaningfulness, our paradigm differs from that of Graf-Estes et al. (2007) in that in the present paradigm the detection of new words and the word-world mappings should take place more or less in parallel, as both words and visual referents were simultaneously presented.

Finally, it is worth noting that we use adults instead of infants as learners. Even though infant studies have been central in language acquisition research, studies on adult learners are also interesting in their own right because adults learning a second language are faced with the challenges of speech segmentation and word-to-world mapping. Because adults already master their native language, one could argue that they are in fact associating a new label to an already existing word-to-world association. However, our manipulation of meaningfulness forces the participants to create associations also with referents that are totally new (scrambled objects) and thus void of meaning.

Experiment 1

In the first experiment we exposed participants to an artificial language stream (Cunillera et al., 2009, 2010; Graf-Estes et al., 2007; Saffran, Aslin, et al., 1996) composed of six nonsense trisyllabic words. Visual stimuli consisting of either *meaningful* (common concrete objects) or *meaningless* (scrambled objects) pictures were presented synchronically to the word onsets/offsets (see Fig. 1A). Word-picture consistency was manipulated by presenting consistent word-picture associations (in 92% of the cases) for half of the items, while for the other half the association between the novel words and pictures was not consistent (31% of the cases). Right after the exposure to the audiovisual streams, a word-picture-learning test was presented in which participants were requested to map the corresponding referent (real or scrambled object picture) to a specific word.

Method

Forty-eight (12 males, mean age 20.4 ± 3.5 *SD*) undergraduate psychology students at the University of Barcelona served as subjects. All participants in the study received extra course credits for their participation. Twenty-four participants were randomly assigned to the meaningful condition and the other 24 participants to the meaningless one. For the participants in the meaningful condition, the visual stream consisted of real object pictures, while for the participants in the meaningless condition the visual stream was composed of scrambled object pictures.

Eighteen different consonant–vowel syllables were used to create two language streams which had the same structure as the ones created by Saffran, Aslin, et al. (1996). For each stream, six trisyllabic nonsense words were concatenated to form a nonstop speech stream by using the speech synthesiser MBROLA with a Spanish male diphone database at 16 kHz (Dutoit, Pagel, Pierret, Bataille, & van der Vreken, 1996). Afterwards the Cooledit software was used to equate the length of the different streams at the level of a millisecond. In all streams the transitional probability of the syllables forming a word was 1.0, while for syllables spanning word boundaries it was 0.2. Each language stream lasted for 1 min, 48 s and 576 ms, and it was concatenated three times to achieve the final stimulus stream with a duration of 5 min, 25 s and 728 ms. Moreover, for each language stream 60 part-words (trisyllabic sequences crossing word boundaries) were created by concatenating the last two syllables of a word and the first one of another word, or the last syllable of a word and the first two syllables of another word (30 part-words of each type).

For the meaningful condition, the visual stimuli consisted of 2 × 12 black-and-white drawings (Snodgrass & Vanderwart, 1980) comparable for word frequency in Spanish (13.94 per million words), name agreement (~99%), imageability (6.14), familiarity (6.04), and concreteness (5.92) (the last three variables are rated by a 1–7 scale where 7 denotes highest imageability, familiarity, or concreteness). For the meaningless condition, the visual stimuli consisted of scrambled versions of the object pictures from the meaningful condition (see Fig. 1A). All picture and word durations were equal (696 ms), and the pictures were presented on a computer screen in a perfect onset–offset synchrony with the words. All pictures had the same size (~7 × 7 cm), and were presented one after another at the center of a computer screen with an average viewing distance of 100 cm and subtending a visual angle of 4° × 4°.

In both the meaningful and meaningless conditions, each participant was instructed to listen carefully to the syllable stream and to identify novel words appearing in the stream, as well as to try to associate the novel words with the pictures. Two different pools of pictures were used in each audiovisual stream, and in all auditory streams, the assignment of the associated and non-associated words and picture pools were counterbalanced across participants and conditions.

The six nonsense words in the auditory streams were divided in three *associated* and three *non-associated words* (see Fig. 1A). The twelve pictures in the meaningful and meaningless conditions were divided in three *target* and three *non-target pictures* which were presented very frequently, and six infrequent *filler-pictures*. The three *associated words* were paired with the three *target pictures* approximately 92% of the time, with each associated word being paired with the same *target picture*. In 8% of the time, these words were paired with two of the other six pictures (equally often with each). In contrast, the two *non-associated words* were paired non-systematically with the three *non-target pictures* (~31% with each). The rest of the time the *non-associated words* were paired with two of the six *filler-pictures* (~7% with each). Picture assignment was randomized for each participant.

After the exposure to the audio-visual speech stream, the participants performed a word-picture-learning test. It included twelve trials in which they saw two pictures on a computer screen and simultaneously heard a word, either an associated or a non-associated one. They were asked to make a forced-choice decision as to which picture was coupled with the novel word. The *target* and *non-target pictures* were presented in two types of test trials (six items for each type), with a presentation order that was randomized for each participant. (i) In the first trial type, two *target pictures* were simultaneously presented together with an associated word. If the participants had acquired the consistent word-picture associations while exposed to the audiovisual stream, they should be able to choose the correct target picture significantly above chance. (ii) In the second trial type, participants heard a non-associated word and had to choose between a *target picture* (incorrect choice) and a *non-target picture* (correct choice). Even here an acquisition of the consistent word-picture associations should be of help, as it would enable the participants to rule out the incorrect target picture that has been consistently associated with another word during the learning phase. Thus following the idea of the mutual exclusivity constraint in language learning (Markman & Wachtel, 1988), we predicted that participants would tend to establish a new association by choosing the non-target picture for the non-associated word.

Results and discussion

In the first trial type of the word-picture-learning test, the two *target pictures* were presented simultaneously with each associated word. The mean percentages of correct word-picture matches for each experimental condition were as follows (see Fig. 2): meaningful: 74.3 ± 21.9%; meaningless: 61.8 ± 21.1%. One-sample *t*-tests revealed that in all conditions, the participants performed significantly above chance (chance level at 50%; meaningful: $t(23) = 5.42$; $p < .001$; meaningless: $t(23) = 2.74$; $p < .02$). A *t*-test for independent samples revealed that the meaningful group performed better than the meaningless one ($t(46) = 2.01$; $p = .050$), indicating that the participants acquired the consistent word-picture mappings more accurately when the picture represented a real object.

In the second trial type of the word-picture-learning test, a *target* and a *non-target picture* were presented together with a non-associated word. We hypothesized that the participants might be using a mutual exclusivity constraint ("the target picture has already another label"), and thus choosing the *non-target picture* as the correct alternative. The mean percentages of correct responses for each experimental condition were as follows (see Fig. 2): meaningful: 62.5 ± 23.7%; meaningless: 59.0 ± 19.6%. One-sample *t*-tests revealed that for both conditions the results were significantly above chance (i.e., 50%; meaningful: $t(23) = 2.58$; $p < .02$; meaningless: $t(23) = 2.25$; $p < .04$). A *t*-test for independent samples revealed no statistical differences between the two groups ($t(46) = 0.56$; $p > .5$), showing that regardless of the content of the picture, the participants favored the non-target picture on these trials.

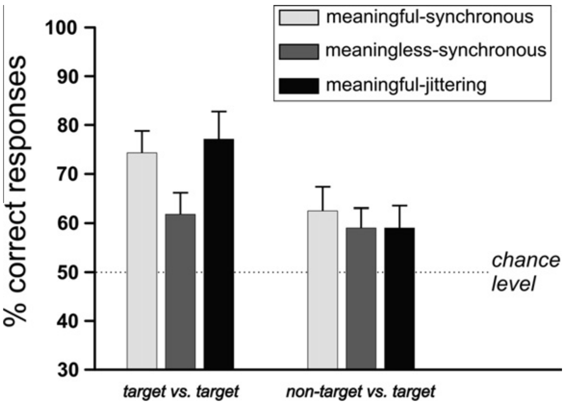

Fig. 2. Percentages of correct picture recognitions (±*sem*) in the picture-learning tests performed after the audiovisual streams (Experiment 1: meaningful-synchronous and meaningless-synchronous conditions; Experiment 3: meaningful-jittering condition). In the first test (target vs. target pictures), two target pictures were presented and an associated word was heard. In the second test (target vs. non-target pictures), a target and a non-target picture were presented and a non-associated word was heard.

The present results show that even after a short exposure, the participants were able to acquire the associations between those words and pictures that were systematically paired with each other. This result was evident in the first type of test trials (see Figs. 1B1 and 2) when participants had to choose between two target pictures. Besides, it also seems that the acquired picture-word association aided participants to choose a non-target picture as a referent of a word that was not consistently associated to any specific picture (see Figs. 1B1 and 2). This performance is predicted on the basis of the mutual exclusivity constraint (Markman & Wachtel, 1988).

Even though the participants learned word-picture associations with both meaningful concrete objects and scrambled pictures, they learned better the meaningful than the meaningless association (see Fig. 2). In other words, integration of information is easier when the association is created with an existing object representation than with novel visual information void of meaning. This difficulty may reflect higher demands on visual processing set by the scrambled objects that are unfamiliar and thus lack a mental representation. In two control experiments, we tested the visual discriminability of the real objects vs. their scrambled version using two different tasks. The results of these experiments confirmed this hypothesis and showed that the recognition and brief maintenance in working memory was more difficult for the meaningless than for the meaningful pictures.[1]

It should be noted that the facilitatory effect of meaningfulness was statistically significant in the first trial type only where an unambiguous correct response (a picture that is strongly associated with the target word) was present. In contrast, in the second trial type the best alternative was a picture that was only weakly associated (~31%) with the target word. Thus, the results from the second trial type most probably indicate that our participants were able to show their acquisition of the word-picture mappings through the application of a mutual exclusivity bias. Remarkably, the inherent difficulty of the task may have wiped away the meaningfulness effect on this trial type.

Experiment 2

After showing in the previous experiment that adults are able to acquire new word-referent mappings only after a limited number of exposures to consistent word-picture pairs that appear in synchrony in a continuous audiovisual stream, we aimed to evaluate directly the effect of this kind of exposure to word segmentation performance. We employed the same training phase as in the previous experiment, but this time the participants were asked to complete a speech segmentation test after being exposed to the audiovisual stream. This experiment should thus show whether participants are indeed able to segment the novel words during the short audiovisual exposure.

Method

Forty-eight (7 males, mean age 20.3 ± 3.2 *SD*) new undergraduate psychology students were recruited from the University of Barcelona, receiving extra course credits for participating in the experiment. Twenty-four participants were randomly assigned to the meaningful condition and the other 24 participants to the meaningless one. The procedure was otherwise identical to Experiment 1, but this time the audiovisual training phase was followed by a standard auditory two-alternative-forced-choice (2AFC) test. Test items consisted of the six words of each language stream and six part-words randomly selected from the pool of 60 part-words of the same stream (3 part-words with a 2-3-1 structure, and the other 3 with the 3-1-2 structure; see Fig. 1A and Appendix). Words and part-words were exhaustively combined, rendering a total of 36 pairs presented in random order in the 2AFC test. After hearing each pair of test items, the participants were asked to decide by pressing a button whether the first or the second item of the pair was a word of the new language. Presentation of the items of a pair was separated by a 400 ms pause. No visual information was provided during the test. The language streams, words, and part-words and the overall setup were the same as in Experiment 1.

Results and discussion

The overall percentage of segmented words for the meaningfulness contrast was as follows (see Fig. 3): meaningful: 61.2 ± 9.6%; meaningless: 63.7 ± 14.6%. These values were different from chance (chance level at 50%; meaningful condition: $t(23) = 5.73$; $p < .001$; meaningless condition: $t(23) = 4.62$; $p < .001$) but not different from each other ($t(46) = -.72$; $p > .4$).

We performed an overall analysis by subjecting the segmentation data to a two-way mixed-model ANOVA

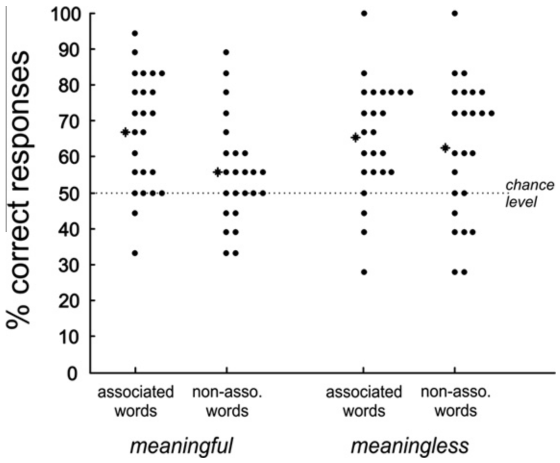

Fig. 3. Mean percentage of correctly segmented novel words (maximum possible score 16) in the auditory 2AFC test performed in Experiment 2 for the meaningful and meaningless conditions and associated and non-associated words. Each point corresponds to an individual participant score while stars denote the mean values for each condition.

[1] We tested the visual discriminability of our scrambled vs. non-scrambled pictures in two experiments employing different tasks. The first one was a 1-back working memory task in two blocks where 16 participants responded whether or not the displayed picture was the same as the one presented in the immediately preceding trial. This task should tap both the visual discriminability of the items and the easiness of keeping them in working memory for a brief moment of time. In one block, the stimuli were the 24 real object pictures employed in the meaningful condition of Experiment 1, whereas in the other block the 24 scrambled pictures from the meaningless condition were used. The timing of the task was identical to our word learning task, with each picture being displayed at the center of the screen for 696 ms, with a visual mask (visual noise grid of 90 x 65 cm) of the same duration presented in-between pictures. The proportion of targets (immediately repeated pictures calling for "yes" response) in the stimulus set was 25%. We calculated the reaction times for yes and no responses as well as the d-prime (d′) scores. The results showed that the mean reaction time was significantly slower for the scrambled than for the real object pictures (correct "yes" and "no" responses collapsed; meaningful pictures: 464 ± 66 ms; meaningless pictures: 492 ± 58 ms; within-subjects ANOVA: $F(1,15) = 11.2$; $p < .01$) and that the real object repetitions were detected significantly better than the scrambled object repetitions (mean d′ for the real object pictures 3.39; for the scrambled pictures 2.87; $t(15) = 2.62$; $p < .02$). The second control experiment consisted of a visual discrimination task composed of the same stimulus sets as in the previous control experiment (the rate of correct matches set at 50%). Sixteen participants were shown two pictures simultaneously, one on the left and one on the right side of a central fixation point. They were to respond whether or not the two pictures were the same. We calculated the reaction times for the correct responses. The results of this second experiment were similar than the previous one: the mean reaction time for the meaningless pictures was significantly slower when compared to meaningful pictures ("yes" and "no" responses collapsed; meaningful pictures: 513 ± 85 ms; meaningful pictures: 632 ± 111 ms; within-subjects ANOVA for the meaningfulness factor: $F(15) = 81.9$; p < .001).

with associativeness as the within-subjects factor (associated vs. non-associated) and meaningfulness (meaningful vs. meaningless) as the between-subjects factor. The results of the ANOVA revealed a main effect for associativeness ($F(1, 46) = 5.17$; $p < .03$), but not for meaningfulness ($F(1, 46) = 0.51$; $p > .4$), and the interaction term was non-significant ($F(1, 46) = 1.74$; $p > .1$). While the associativeness × meaningfulness interaction failed to reach significance, Experiment 1 revealed a facilitatory effect of meaning on word-referent mapping. Given this fact, we analyzed separately the associativeness effects on segmentation in the meaningful and the meaningless conditions.

For the meaningful condition, only those words that were consistently associated with a specific picture in the audiovisual stream were segmented above chance level (50%; associated words: 66.9 ± 16.1%, $t(23) = 5.15$; $p < .001$; non-associated words: 55.6 ± 14.5%, $t(23) = 1.88$; $p > .07$; see Fig. 3). A t-test comparing the two conditions revealed that the associated words were segmented more accurately than the non-associated words ($t(23) = 2.33$; $p < .03$). To confirm this finding, we also calculated the number of participants in each condition that performed better than chance as determined by a binomial test (with $p < .05$, in an 18-items test, chance level corresponds to scores lower than 66.67%). Thus, for the associated words, 14 out of 24 participants (58.3%) performed above chance level. In contrast, only five participants (20.8%) performed above chance with the non-associated words. The results of a chi-square test revealed that this difference between the associated and non-associated condition was significant ($\chi^2(1) = 7.06$; $p < .01$), in line with the results obtained with the parametric test.

The results obtained for the meaningless condition revealed that both associated and non-associated words were segmented above chance level (50%; associated words: 65.3 ± 15.9%, $t(23) = 4.71$; $p < .001$; non-associated words: 62.3 ± 19.2%, $t(23) = 3.14$; $p < .05$; see Fig. 3). In contrast with the significant associativeness effect encountered in the meaningful condition, no such effect was observed with the scrambled pictures ($t(23) = 0.75$; $p > .4$). We further studied the associativeness factor in the meaningless group by calculating the number of participants in each condition that performed better than chance as determined by a binomial test. For the associated words, 13 out of 24 participants (54.2%) performed above chance level, and half of the participants (12) performed above chance with the non-associated words. The results of a chi-square test confirmed the lack of an associativeness effect for the scrambled pictures ($\chi^2(1) = .08$; $p > .9$).

The present results confirmed that participants are able to segment the words from the audiovisual streams we used. For the meaningful word-referent pairs, only associated words were segmented above chance level, as one would expect. However, for the meaningless word-referent pairs, both associated and non-associated words were segmented above chance levels. One could speculate whether the presence of scrambled, semantically void pictures leads to attenuated attention to the visual stream, resulting in a more even segmentation performance for the associated vs. non-associated items. Taken together, the results from Experiment 1 and 2 indicate that during a short exposure to an audiovisual stream, it is possible to carry out both speech segmentation and word-world mapping. Finally, the systematic association between a word and a picture yielded a better segmentation performance when the visual referents were meaningful.

Experiment 3

In the present experiment we addressed the question of whether the results obtained in Experiments 1 and 2 hinged upon the fact that words and pictures were delivered in a perfect synchronous fashion in the audiovisual stream. It is obvious that in real-life learning situations, audiovisual information is not encountered at millisecond-level synchrony. Rather, one would expect to find a temporal window within which coinciding auditory and visual information can be integrated. In a previous study using similar audiovisual streams but with totally random picture-word pairings, we showed that speech segmentation performance for audiovisual conditions was boosted when compared to a pure auditory condition (Cunillera et al., 2010). More importantly for the present purposes, Cunillera et al. (2010) reported facilitation in segmentation performance even for asynchronous audiovisual streams where the picture onset was close to a low-probability syllable transition (i.e., a potential word boundary). Nevertheless, in order to explore directly the effect of synchrony with the present audiovisual streams, we conducted an additional experiment using asynchronous word-picture pairings in which we introduced a jitter in the visual stream that accompanies the auditory stream.

Method

Another 24 (8 males, mean age 22.6 ± 3.8 SD undergraduate psychology students were recruited from the University of Barcelona. The procedure was otherwise identical to Experiments 1 and 2. The audio-streams, words, part-words, and the overall setup were the same as in Experiment 1, but only meaningful pictures were used. Importantly, a jitter in the timing of the visual stream was introduced, taking as a reference the onset of the words in the auditory stream. More specifically, picture onset/offset deviated from word onset/offset by 0, ±100, ±150 or ±200 ms, keeping always picture durations in the range of 446–1046 ms, i.e. pictures switched always along the visual stream within the presentation of the first and the last syllable of the trisyllabic words of the auditory stream. In addition, a constraint was introduced in the setup so that two consecutive pictures coinciding with the same jitter never occurred in the visual sequence. In sum, this procedure resulted in an arrhythmic visual sequence in which picture switch did not point in any occasion to word onset/offset. After the training phase, the same word-picture-learning test was administered to participants as in Experiment 1.

Results and discussion

For the first trial type in the word-picture-learning task (two *target pictures*), the mean percentage of correct responses was 77.1 ± 27.7%, being significantly above chance (50%, $t(23) = 4.79$; $p < .001$; see Fig. 2, meaningful-jittering condition). We then compared this result with the one obtained in Experiment 1 for the meaningful word-picture pairs. The *t*-test revealed that the meaningful-synchronous group performed at the same level as the meaningful-jittering one ($t(46) = -0.39$; $p > .7$), showing that a perfect synchrony between words and pictures was not a prerequisite for learning the word-picture associations.

For the second trial type in the word-picture-learning task (*target* and a *non-target picture*), the mean percentage of correct responses was 59.0 ± 21.9%. This is only marginally better than chance (50%), ($t(23) = 2.01$; $p = .056$; see Fig. 2). A further analysis revealed no statistical differences between the synchronous (Experiment 1) and asynchronous conditions for this trial type ($t(46) = 0.53$; $p > .6$).

Given the present and previous results, we can conclude that the acquisition of word-picture mapping in the present paradigm is not an artifact of perfect timing with the picture-word onset and offset. This adds to the relevance of the audiovisual learning paradigm and indicates an integrative time window for cross-modal effects in speech segmentation. Cunillera et al. (2010) estimated that this time window spans at least 200 ms, corresponding roughly to one syllable in our language streams.

General discussion

In the present study, we explored for the first time whether adult learners are able to both segment new words and create word-to-world associations while being exposed to an audiovisual stream, and also whether constancy and meaningfulness of the word-world associations affects word learning and speech segmentation. Our participants' success in word-picture matching (Experiment 1) and speech segmentation (Experiment 2) indicate that both tasks can be accomplished in the same learning context. Even though these tasks were probed in separate experiments, it is reasonable to assume that word segmentation is a prerequisite for successful word-picture matching. The results from Experiment 1 revealed that acquisition of word-picture associations was better in the meaningful (common concrete objects) than in the meaningless condition (scrambled pictures), albeit word learning took place in both conditions. Experiment 2 showed that word segmentation performance is sensitive to the most basic mapping feature, namely the consistency of word-referent association, especially when the referents were meaningful. These results have important implications for the organization of speech segmentation and word learning processes at the initial stages of language acquisition during adulthood. The immediacy of the observed learning effects in both segmentation and in word-to-world mapping points to a learning process which allows more or less simultaneous discovery and isolation of new words and assignment of referents to them. In word-picture mapping, the associations that were consistent were acquired quite fast, after only 5 min continuous exposure to word-picture pairs.

Experiment 1 indicated that meaningful referents are more easily linked to newly segmented words. Establishing links with meaningful referents is a core aspect in word learning and it can be assumed that the familiarity of the meaningful objects also made the task easier in terms of processing demands. Control experiments mentioned in footnote 1 showed that the scrambled object pictures were visually harder to process. This finding is consistent with the notion that the knowledge of an object drives the search for a label (MacNamara, 1982). Assuming that forging a link between a label and a world referent is computationally demanding, the availability of the attentional resources for segmenting and associating the novel label to a visual referent might be compromised when the referent is unfamiliar to the learner. Accordingly, the degree of complexity of the word-referent mapping process is reduced when a recently segmented word has a potential association with a well-known referent. The effect of meaningfulness on word-referent mapping observed in the present study is reminiscent of the resource limitation hypothesis that postulates that depending on the cognitive complexity involved in the learning task, novice word learners may have difficulty in accessing the details of newly learned words (Fennell & Werker, 2003, 2004). In line with this hypothesis, it has been observed that infants fail in accessing phonetic details when the task calls for mapping of a novel label to a novel object but succeed when the mappings are to be made with known words and objects (Fennell & Werker, 2003, 2004). Infants at earliest age of 14 months, however, have been shown to have access to phonetic details when tested in a preferential looking task, indicating detailed phonological representation of consonants (Ballem & Plunkett, 2005) and vowels (Mani & Plunkett, 2008).

The most clear-cut effect in Experiment 2 was the fact that the consistency of picture-word associations in the audiovisual stream facilitated segmentation performance. Moreover, this effect appeared to be more prominent for the meaningful items, albeit the interaction between associativeness and meaningfulness failed to reach significance. It seems evident that the detection of consistent associative relationships must be the primary factor when establishing word-to-world mappings. In order to create a link between a novel word and a (meaningful) referent, they must co-occur.

An interesting property of the audiovisual paradigm used in the present study is that it simultaneously combines unimodal and cross-modal regularities in the input. The computation of transitional probabilities together with audiovisual synchrony informs the learner on word boundaries, while the detection of consistent word-picture mappings provides information on the referents of the to-be-segmented words. The effect of associativeness found in the speech segmentation test suggests that the detection of uni- and cross-modal regularities in a sequential input can be used in an interactive fashion during word learning, yielding an improvement in the detection of the regularities found in the input. In the same vein, a beneficial effect of learning word-to-meaning associations in audio-visual speech segmentation has been recently demonstrated by a computational model which acquires words from multimodal (audition and vision) sensory inputs (Roy & Pentland, 2002). Furthermore, Yu and Smith (2007) have demonstrated that word-picture mappings can be easily learned through audiovisual cross-trial statistical relations, whereas Vouloumanos (2008) has shown that learners are extremely sensitive to the frequency of the word-object combination in mappings. In this study (Vouloumanos, 2008), participants were able to differentiate between an object that appeared twice with a word and another object that occurred only once with the same word during the learning phase. Notably, both the two words and the two pictures were also paired with another set of six pictures and six words, respectively, during the training phase.

All in all, our results suggest that the earliest stages of adult word learning involve a learning mechanism that computes stimulus regularities both within and across modalities, and combines this information to arrive at candidates for words and word-referent pairs. Whether or not the referent corresponds to an existing concept does not prevent word learning at very early stages, although its meaningfulness clearly facilitates the mapping process (see results in Experiment 1). Further studies should address the issue of whether the low mapping performance observed in the meaningless condition would be overcome with longer exposures times that could compensate for the unfamiliarity of the scrambled pictures. Thus, familiarization with the scrambled pictures prior to the learning task might eliminate the advantage of meaningful word-object mappings (see e.g., Fennell, 2004).

The results of the current study are of particular interest because of the demonstration of the rapid achievement of the word-referent learning process. Thus the obtained results after a massive but short simultaneous exposure to words and their possible referents are convergent with several studies that have shown that novel words that have not been associated to any particular meaning are easily and rapidly integrated in the lexicon, showing as a consequence lexical competition effects when pitted against existing similar words (Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008). Nazzi and Bertoncini (2003) have proposed that infants' first learned words are acquired through an associationist mechanism and remain in a proto-word status (phonetically unspecified) until these words become genuine words, which takes place at the time that words turn into phonetically specified sound patterns and are linked to object categories. Together, all these studies point to the existence of a very plastic language learning system that can easily accommodate novel words associated with new referents. This is also reminiscent of the episodic theories of speech perception that suggest that perceptual details are stored in memory as episodes and afterwards integrated in perceptions (Goldinger, 1998). Similarly, language learners may posses a learning mechanism that flexibly creates and updates the information in the long-term memory associated to each word. Such a mechanism could permit learners to continuously adapt the representations associated to novel words heard in different contexts, with different speakers, dialects, etc. In the present experiment, the consistent association between the visual referent and the novel word may trigger the creation of an episodic trace that with further exposures in different contexts might ultimately become a more abstract lexical–semantic representation of the novel word.

Intuitively, the exploitation of multimodal statistical regularities seems necessary in order to reduce the ambiguity that exists when trying to link words with a visual referent, as words and their possible referents are susceptible to multiple sources of variation or distortion in real-life contexts. Thus, learners may rapidly learn word-referent associations when multiple co-occurrences are observed. A computational model created by Siskind (1996) also supports statistical cross-situational learning. The utilization of this combined uni- and cross-modal statistical information might explain why learners in our study were so fast to achieve a good performance on the word-picture mapping test even though the association strengths were set to a $\sim$90% of co-occurrences. It would be of interest to see whether the present multimodal word learning results would generalize even to infants.

The idea that speech segmentation is improved when multiple cues are exploited, either from the same or different sensory modalities, has been posited before (see e.g., Altmann, 2002; Plunkett, 1997). Indeed, our results indicate that visual referents may not only serve as the most important meaning attribute for concrete words, but also as a potential cue to identify novel words even when the visual properties are new and carry no meaning. It seems obvious that in natural learning contexts, different multimodal cues are available to segment real speech (Hollich, Newman, & Jusczyk, 2005; Hollich et al., 2000). This is supported by previous findings in which native language processing was observed to be facilitated with visible speech (Dodd, 1977; Reisberg, McLean, & Goldfield, 1987; Sanders & Goodrich, 1971; Thompson & Ogden, 1995). More important to the present study is the evidence that demonstrates that multimodal cues are exploited in foreign language learning as well (Davis & Kim, 2001; Reisberg et al., 1987). The coalition of attentional and associative processes (Plunkett, 1997; Smith, 2000) provides a plausible explanation on how multimodal word learning may work. For instance, Smith (2000) proposed that word learning in children is a process in which attention is initially captured by objects and actions that are the most salient ones in their environment. When attention is focused on a visual event, it may then aid in associating the visual object or action with the word spoken by the adult who is interacting with the infant. Finally, the associative process would rely on statistical learning of co-occurring linguistic sounds in extralinguistic contexts, i.e., in finding out the distributional features of the input.

The language learning paradigm that we have used in the present study (see Saffran, Aslin, et al., 1996) raises questions in regard to its ecological validity. It is obvious that in a real language learning situation learners are not exposed to a continuous language stream – composed of few words which are massively repeated in a brief period of time – and which is accompanied by visual referents that appear and disappear near word boundaries. Instead, it is more likely that the listener hears a set of distinct words within a given utterance while a visual referent may sometimes be present, appearing perhaps prior to the utterance and persisting beyond it. The present experimental set-up should be considered as an attempt to describe in an idealized and strictly controlled situation the limits of visual-auditory association in regard to language learning. Future experiments should benefit from the current results and study the visual-auditory association in the context of more natural language learning situations.

In sum, we provide evidence that after a brief exposure to an audiovisual stream, adult participants can both segment new words and associate referents to these words. These processes seem to be closely linked as the consistency of the word-referent association (a cross-modal feature), affects word segmentation performance. Moreover, the meaningfulness of the referents facilitates the mapping operation. These findings serve to bridge the gap between segmentation and word learning studies and highlight the efficiency of word learning mechanisms in adults.

Acknowledgments

We thank R. de Diego and S. Soto-Faraco for their comments for the earlier drafts of this manuscript. Special thanks are due to I. Nogué and V. Havas who helped in running previous pilot experiments related to the present ones. TC was supported by a fellowship of the Finnish Government Scholarship Pool (HH-05-3276). ML was supported by the Academy of Finland (Grants 200178 and 205204) and by a NOS-HS grant for a Nordic Centre of Excellence in Cognitive Control. ARF was supported by the Ramon y Cajal research program and a MEC Grant (SEJ2005-06067/PSIC).

Appendix: The artificial languages used in the different conditions

Language 1

Words: SERIPU, MAKUSI, PAMOTE, TOSUKA, RUPOME, KITURE

Part-words 3-1-2: RIPUMA, KUSIPA, MOTETO, SUKARU, POMEKI, TURESE, etc.

Part-words 2-3-1: PUMAKU, SIPAMO, TETOSU, KARUPO, MEKITU, RESERI, etc.

Language 2

Words: RIPUKU, SUKATE, TUMEPA, MOSIPO, KIMARE, TORUSE

Part-words 3-1-2: PUKUTU, KATEMO, MEPARI, SIPOSU, MARETO, RUSEKI, etc.

Part-words 2-3-1: KUTUME, TEMOSI, PARIPU, POSUKA, RETORU, SEKIMA, etc.

Snodgrass & Vanderwart (1980) pictures used in the study

Pool 1: ANT, BEAR, BOOT, BRUSH, CARROT, CHICKEN, DRUM, GLASSES, IRON, LEAF, MOON, and NECKLACE

Pool 2: ANCHOR, BOWL, CIGARETTE, COW, FLAG, ONION, ORANGE, POTATO, SEAL, SKIRT, SWEATER, and TRAIN

References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience, 20*, 952–964.

Altmann, G. T. (2002). Statistical learning in infants. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 15250–15251.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*, 395–418.

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language, 32*, 159–173.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*, B33–B44.

Carey, S., & Barlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology, 63*, 260–274.

Cunillera, T., Càmara, E., Toro, J. M., Marco-Pallares, J., Sebastian-Galles, N., Ortiz, H., et al. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage, 43*, 541–553.

Cunillera, T., Toro, J. M., Sebastian-Galles, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research, 1123*, 168–178.

Davis, C., & Kim, J. (2001). Repeating and remembering foreign language words: Implications for language teaching systems. *Artificial Intelligence Review, 16*, 37–47.

De Diego-Balaguer, R., Toro, J. M., Rodríguez-Fornells, A., & Bachoud-Levi, A. C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE, 2*, e1175.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception, 6*, 31–40.

Dutoit, T., Pagel, N., Pierret, F., Bataille, O., & van der Vreken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In: *Proc. ICSLP'96, Philadelphia* (Vol. 3, pp. 1393–1396).

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60*, 353–369.

Fennell, C. T. (2004). *Infants attention to phonemic details in word forms: Knowledge and familiarity effects*. Unpublished doctoral dissertation, University of British Columbia, Vancouver.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech, 46*, 245–264.

Fennell, C. T., & Werker, J. F. (2004). Infant attention to phonetic detail: Knowledge and familiarity effects. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceedings of the 27th annual Boston University conference on language development* (pp. 165–176). Somerville, MA: Cascadilla Press.

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

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1*, 3–55.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development, 1*, 23–64.

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

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.

Graf-Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*, 254–260.

Hollich, G. (2006). Combining techniques to reveal emergent effects in infants' segmentation, word learning, and grammar. *Language and Speech, 49*, 3–19.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., et al. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65*, i-123.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*, 598–613.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3*, 323–328.

Jusczyk, P. W., & Aslin, R. N. (1995). Infant's detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1–23.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*, 463–470.

MacNamara, J. (1982). *Names for things: A study of human learning*. Cambridge: MIT Press.

Mani, N., & Plunkett, K. (2008). Fourteen-month-olds pay attention to vowels in novel words. *Developmental Science, 11*, 53–59.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 57–77.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121–157.

Mirman, D., Magnuson, J. S., Graf-Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition, 108*, 271–280.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science, 6*, 136–142.

Pinker, S. (1989). *Learnability and Cognition*. Cambridge, MA: MIT Press.

Plunkett, K. (1997). Theories of early language acquisition. *Trends in Cognitive Sciences, 1*, 146–153.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–114). Hillsdale, NJ: Lawrence Erlbaum Associates Inc..

Richards, D., & Goldfarb, J. (1986). The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development, 1*, 183–219.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*, 113–146.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81*, 149–169.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J. R., & Graf-Estes, K. M. (2006). Mapping sound to meaning: Connections between learning about sounds and learning about words. In R. Kail (Ed.), *Advances in child development and behavior* (pp. 1–38). New York: Elsevier.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy, 4*, 273–284.

Sanders, D. A., & Goodrich, S. J. (1971). The relative contribution of visual and auditory components of speech to speech intelligibility as a function of three conditions of frequency distortion. *Journal of Speech, Language, and Hearing Research, 14*, 154–159.

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience, 5*, 700–703.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39–91.

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–81). Oxford, England: Oxford University 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.

Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *Quarterly Journal of Experimental Psychology, 61*, 361–371.

Thompson, L. A., & Ogden, W. C. (1995). Visible speech improves human language understanding: Implications for speech processing systems. *Artificial Intelligence Review, 9*, 347–358.

Tomasello, M. (1992). The social bases of language acquisition. *Social Development, 1*, 67–87.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729–742.

Weijer, J. van de (1998). Language input for word discovery. MPI Series in Psycholinguistics 9. Nijmegen: Max Planck Institute of Psycholinguistics.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29*, 961–1005.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414–420.

Yu, C., & Smith, L. B. (2008). Infants rapidly learn word-referent mapping via cross-situational statistics. *Cognition, 106*, 1558–1568.