COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12731

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12731

# Cross-Situational Learning of Phonologically Overlapping Words Across Degrees of Ambiguity

Karen E. Mulak,[a,b,c] Haley A. Vlach,[b,d] Paola Escudero[a,b]

[a]*The MARCS Institute for Brain, Behaviour and Development, Western Sydney University*
[b]*Australian Research Council Centre of Excellence for the Dynamics of Language, Western Sydney University*
[c]*Department of Hearing and Speech Sciences, University of Maryland, College Park*
[d]*Department of Educational Psychology, University of Wisconsin-Madison*

Received 19 December 2017; received in revised form 1 February 2019; accepted 20 March 2019

**Abstract**

Cross-situational word learning (XSWL) tasks present multiple words and candidate referents within a learning trial such that word–referent pairings can be inferred only across trials. Adults encode fine phonological detail when two words and candidate referents are presented in each learning trial (2 × 2 scenario; Escudero, Mulak, & Vlach, 2016a). To test the relationship between XSWL task difficulty and phonological encoding, we examined XSWL of words differing by one vowel or consonant across degrees of within-learning trial ambiguity (1 × 1 to 4 × 4). Word identification was assessed alongside three distractors. Adults finely encoded words via XSWL: Learning occurred in all conditions, though accuracy decreased across the 1 × 1 to 3 × 3 conditions. Accuracy was highest for the 1 × 1 condition, suggesting fast-mapping is a stronger learning strategy here. Accuracy was higher for consonant than vowel set targets, and having more distractors from the same set mitigated identification of vowel set targets only, suggesting possible stronger encoding of consonants than vowels.

*Keywords:* Cross-situational learning; Statistical learning; Word learning; Phonological encoding; Consonants vs. vowels

## 1. Introduction

Explicit mappings between words and their referents are not typically part of day-to-day conversations, making word learning a difficult task. In addition, there are often a wide range of possible referents for one linguistic label, and linguistic labels are

Correspondence should be sent to Karen Mulak, Department of Hearing and Speech Sciences, University of Maryland, 0100 LeFrak Hall, College Park, MD 20740. E-mail: kmulak@umd.edu

often presented in running speech, rather than in isolation. To understand how learners determine word–referent mappings, researchers have sought to identify the information used to resolve referential ambiguity, revealing that learners are able to track co-occurring words and referents across moments in time. This behavior is often termed *cross-situational word learning* (XSWL) or *statistical word learning* (Benitez, Yurovsky, & Smith, 2016; Blythe, Smith, & Smith, 2016; Fitneva & Christiansen, 2011; Kachergis, Yu, & Shiffrin, 2016; Vlach & DeBrock, 2017; Vlach & Sandhofer, 2014; Yurovsky & Frank, 2015).

In a typical XSWL task, participants are presented with a series of ambiguous learning trials that consist of several words and several objects within a single trial. After the learning trials, participants' word–object mappings are assessed at an immediate forced-choice test. Studies using this paradigm have shown that learners can infer word–object mappings at test, suggesting they can make use of the co-occurrence statistics between words and objects presented during the learning phase. Moreover, learners can infer word–object mappings across varying degrees of referential ambiguity (e.g., two words and two objects per trial [2 × 2] vs. four words and four objects per trial [4 × 4]; Yu & Smith, 2007) and retain these mappings over time (Vlach & Sandhofer, 2014).

Outside the laboratory, encoding fine phonetic detail in novel words is necessary due to the vast amount of phonological overlap that occurs across the lexicon, the most extreme case being minimal pair words, where two words are differentiated by only a single phonological segment (e.g., BET vs. DEBT or BET vs. BIT). Phonological encoding ability also directly affects XSWL efficiency. Reliably encoding a word in fine phonological detail across multiple occurrences would strengthen word–referent pairings over a shorter number of occurrences compared to encoding a word unreliably or with less specification, which would lead to more erroneous or weaker connections between target words and referents across encounters. But XSWL tasks have typically used words that contain minimal phonological overlap, such as FEP and DAX (e.g., Fitneva & Christiansen, 2011; Vlach & Sandhofer, 2014). While fine phonetic encoding may have taken place in these studies, these studies did not afford a measure of phonetic detail encoding.

Recent research has shown that learners can encode fine phonetic detail while tracking word–referent co-occurrence probabilities in a XSWL task. Escudero et al. (2016a) presented participants with a 2 × 2 XSWL task in which each learning trial consisted of two words and two objects. At test, participants were asked to identify the object corresponding to the auditory word in the context of one visual distractor. The learning and test trials formed three different pair types based on the word associated with the target and distractor image: (a) non-minimal pairs, in which two or all three segments in each word differed (e.g., BON–DEET); (b) consonant minimal pairs, in which the initial consonant differed, but vowel and final consonant were shared (e.g., BON–TON); and (c) vowel minimal pairs, in which the vowel differed but the initial and final consonants were shared (e.g., DEET–DIT). Accuracy was above chance for all pairs, but it was lower for vowel minimal pairs than consonant minimal pairs or non-minimal pairs.

Although this work shows that learners can encode fine phonetic detail during XSWL, learning trials involved only two words and two objects, which is likely a very small amount of ambiguity compared to situations outside the laboratory. Previous studies have suggested that learners can determine word mappings in more ambiguous learning situations; researchers have examined XSWL across varying amounts of ambiguity during learning (such as three words and three objects per trial; for example, Yu & Smith, 2007). However, these studies have presented learners with very phonologically distinct words. Consequently, it is unknown whether learners can still encode the phonetic detail of words across greater levels of referential ambiguity. While it is likely that as referential ambiguity increases, fewer cognitive resources are available for fine phonological encoding, the details of this relationship are unknown. Understanding this relationship is important for understanding XSWL's efficacy in natural implicit word learning. For instance, if participants' ability to encode fine phonological detail decreases too rapidly as referential ambiguity increases, this may lead accuracy to fall to chance, perhaps implying that XSWL is not a widely used or primary learning strategy in real-world scenarios. If, however, performance drops only marginally and/or the rate of decline decreases as ambiguity increases, this would suggest fine phonological encoding can successfully occur in scenarios of greater ambiguity than previously tested (Escudero et al., 2016a), supporting XSWL as a viable word learning strategy in such scenarios. Thus, we tested participants' ability to learn minimally different words via XSWL under differing degrees of ambiguous learning scenarios in which two ($2 \times 2$), three ($3 \times 3$), or four ($4 \times 4$) images and auditory labels occurred in each learning trial. The primary goals of this study were to (a) understand whether fine phonological encoding occurs during XSWL under greater degrees of referential ambiguity and (b) characterize the nature of the relationship between fine phonological encoding ability and amount of referential ambiguity.

We also included a $1 \times 1$ learning scenario where only one word and one image were presented per learning trial, making the word–object association non-ambiguous. Thus, this learning scenario did not test cross-situational word learning, but fast-mapping. This afforded, to our knowledge, the first direct comparison of phonological encoding ability between these two word learning strategies, comparing explicit and ambiguous word learning scenarios. Performance in the $1 \times 1$ learning condition therefore provides an important baseline for understanding overall effects of introducing ambiguity into word learning. We hypothesized that the addition of ambiguity would increase the task difficulty, taxing cognitive resources available for fine phonological encoding. While the primary motivation of this experiment was to test the strength of phonological encoding in increasingly ambiguous word learning scenarios rather than directly inform the mechanisms underlying fast-mapping and XSWL, one proposal is that XSWL shares an underlying mechanism with fast-mapping (e.g., Trueswell, Medina, Hafri, & Gleitman, 2013). In this view, XSWL is not achieved through automatic statistical tracking of object-label co-occurrences (e.g., Yu & Smith, 2007), but begins with the same explicit association that underlies fast-mapping. That is, participants may hypothesize a word–object pairing, and stick with this association until presented with conflicting information in a trial (e.g., hearing a label but not seeing its hypothesized referent in a trial), at which point another hypothesis is made. Because both types of word learning (ambiguous versus explicit) are proposed to have the same underlying mechanism, it is possible that the ability to encode phonological detail will not differ between the explicit and ambiguous word learning conditions. Regardless, as fast-mapping is typically used in formal language instruction, understanding differences between word learning efficacy between these two word learning strategies has implications for formal language teaching.

We also compared encoding of consonants and vowels across varying degrees of ambiguity. As mentioned above, Escudero et al. (2016a) found that participants had lower accuracy when the words associated with the target and distractor formed a vowel minimal pair, compared to when they formed a consonant minimal pair or non-minimal pair. The authors proposed this signalled poorer encoding of vowels than consonants during the learning phase, suggesting a consonant bias. While it is increasingly clear that whether participants demonstrate a consonant or vowel bias (or any bias at all) likely depends on language-specific factors (e.g., Højen & Nazzi, 2015; Wiener & Turnbull, 2016), research suggests that adult native listeners of English perceive consonants and vowels in qualitatively different ways, perceiving consonants more categorically than vowels (Beddor & Strange, 1982; Fry, Abramson, Eimas, & Liberman, 1962; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This may reflect differing roles of consonants and vowels in speech perception. For instance, in Dutch, which is closely related to English, consonants are proposed to have a more prominent role in lexical access than vowels (Cutler, Sebastián-Gallés, Soler-Vilageliu, & van Ooijen, 2000; see also Nazzi & Cutler, 2019). These factors may more readily support consonant encoding in word learning tasks, leading to the prediction of stronger encoding of consonant information over vowel information even under increased ambiguity. However, native English listeners' short-term memory for vowels appears to be stronger than for consonants (Crowder, 1971), which may enhance vowel encoding in a word learning task under increasing task load. By manipulating the task ambiguity during word learning (by varying the number of words and objects presented in each learning trial) and on test (by varying the number of phonologically overlapping distractors), the relative effects of these sources of ambiguity on consonant and vowel encoding in a naturalistic word learning paradigm can be examined.[1]

We presented adults with an XSWL task that consisted of eight novel words spoken by a female native speaker of Australian English (AusE). Four words differed by their initial consonant (BON, DON, PON, TON) and four by their vowel (DEET, DIT, DOOT, DUT). Words and objects occurred across four relative levels of referential ambiguity, classified by the number of auditory words and visual referents presented in each trial: no ambiguity (1 × 1, that is, one word and one referent per learning trial), low ambiguity (2 × 2), medium ambiguity (3 × 3), and high ambiguity (4 × 4). At test, participants were asked to identify the referent corresponding to an auditory word in the context of three visual distractors, which were referents associated with the other auditory words. We predicted that participants who were taught word–referent pairings in the unambiguous context (1 × 1) would outperform those who were taught in an ambiguous context, and that as referential ambiguity increased, word learning performance would decrease, as shown in previous research (Vlach & Sandhofer, 2014; Yu & Smith, 2007).

Because Escudero et al. (2016a) found that target identification accuracy was lower when the target word differed in only one vowel from the distractor (e.g., DEET vs. DIT) compared to when they differed in a consonant (PON vs. BON) or more than one phonological segment (e.g., DEET vs. PON), we also predicted that participants would again show poorer learning of the set of words differing in a vowel than those differing in a consonant and that subsequently, performance for vowel-differing words (and not those differing in consonants) would be influenced by within-trial ambiguity and number of distractors.

## 2. Method

### 2.1. Participants

Participants were 86 students aged 17.1 to 54.4 years ($M = 23.5$, $SD = 7.5$, 65 females) who participated for course credit at Western Sydney University. Fifty-two participants reported via a language background form that they were raised in a household where at least one parent spoke a language other than English. Incorporating participants' language background in our analysis did not improve the predictive power of our model. This was similarly found in Escudero et al. (2016a) whose participants came from the same population, and is likely due to the heterogeneity of the non-monolinguals. Thus, we do not discuss language background further. All participants provided informed consent in accordance with the Western Sydney University Human Research Ethics Committee. Data from another nine participants were excluded from the final sample due to participant-reported speech or language difficulties ($N = 3$) or experimenter error ($N = 6$).

### 2.2. Materials

#### 2.2.1. Novel words

Eight monosyllabic nonsense words were produced by a female native speaker of AusE. These stimuli were also used in an experiment with infants (Escudero, Mulak, & Vlach, 2016b), and so were recorded in child-directed speech. As shown in Fig. 1, the words followed a CVC structure, adhered to English phonotactics, and have been used in previous research on minimal pair learning (Curtin, Fennell, & Escudero, 2009; Fikkert, 2010) and XSWL (Escudero et al., 2016a,b). Two tokens of each were selected and prosodic contours were matched impressionistically across all words. Each word belonged to one of two sets: Four differed minimally in their first consonant (consonant set) and four in their vowel (vowel set).

#### 2.2.2. Novel visual referents

Each word was randomly paired with a visual referent (see Fig. 1). All participants viewed the same word–referent pairings. The visual referents for the words were colourful pictures of novel items used in previous studies of XSWL (Escudero, Mulak, Fu, &

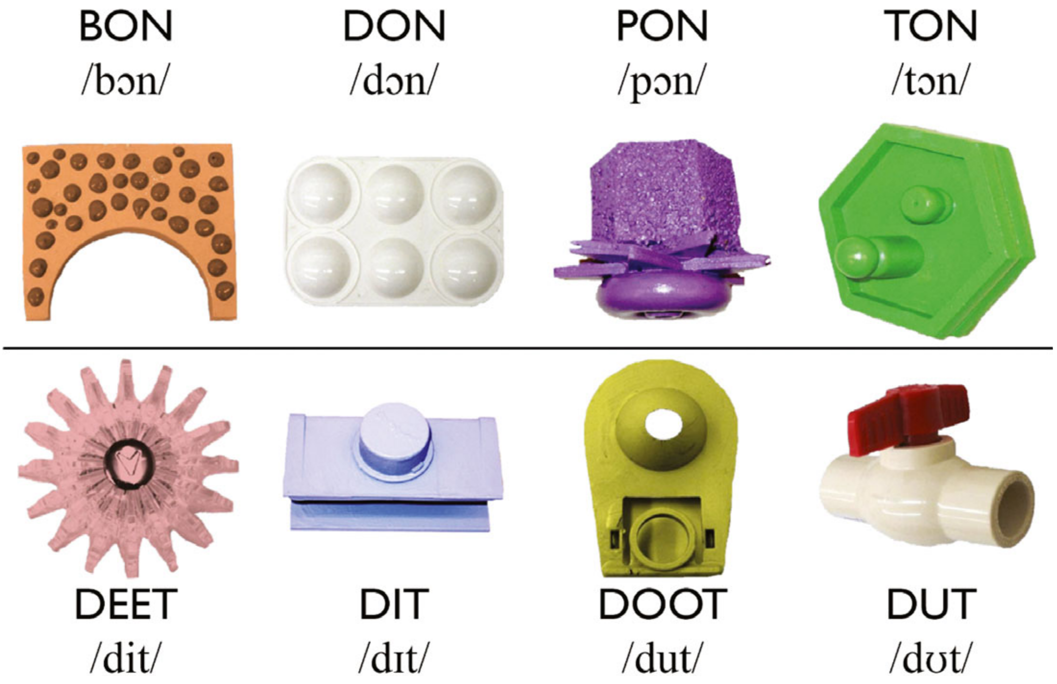


Fig. 1. The eight novel words and their visual referents. The four words in the top row are minimally different in their initial consonant and comprise the consonant target set. The words on the bottom row are minimally different in their vowel, and make up the vowel target set. The vowel used for the consonant set in the top row is /ɔ/ as in POT. The vowels used in the vowel minimal pairs are /i/ as in BEAT, /ɪ/ as in BIT, /u/ as in BOOT, and /ʊ/ as in PUT.

Singh, 2016; Escudero et al., 2016a,b; Vlach & Sandhofer, 2014) and measured 336 × 330 pixels.

### *2.3. Setup and procedure*

Participants sat 50 cm in front of a 24-in. monitor displaying at 1680 × 1050. The experiment consisted of a learning phase and testing phase. Examples of learning and testing phase trials can be seen in Fig. 2. E-Prime (version 2.0, Psychology Software Tools, Inc., Sharpsburg, Pennsylvania, United States) was used to present stimuli and record responses.

#### *2.3.1. Learning phase*

Learning conditions differed in the number of auditory words and visual referents presented in each learning trial, numbering from one word and one visual referent (1 × 1 condition) to four words and four referents (4 × 4 condition). Thus, the 1 × 1 condition directly taught word–object pairings, whereas pairings in the 2 × 2, 3 × 3 and 4 × 4 conditions were presented in a XSWL paradigm with differing degrees of within-trial

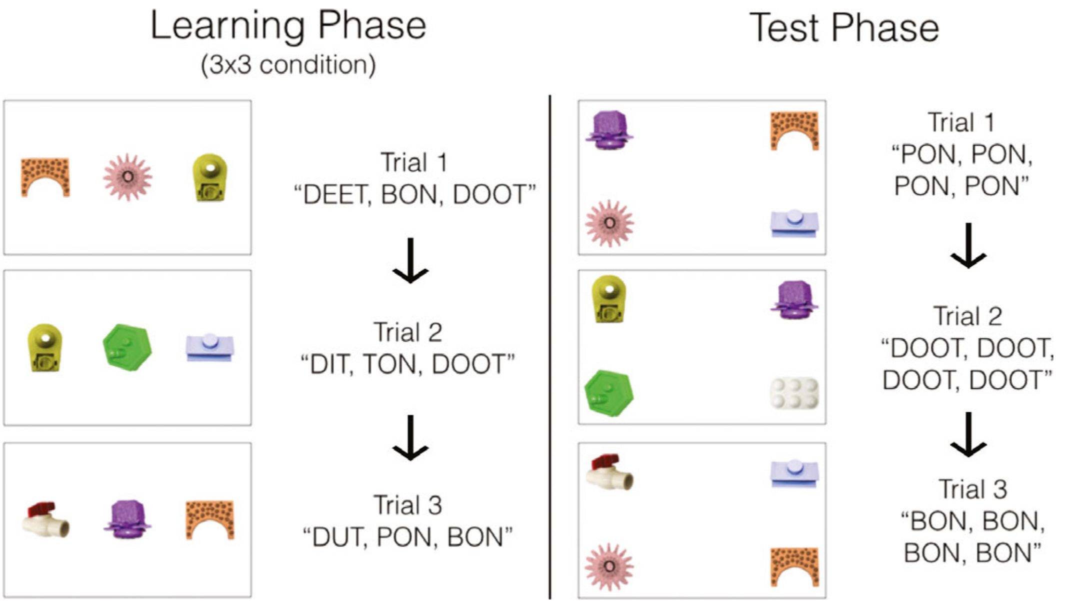


Fig. 2. Examples of learning trials presented to participants in the 3 × 3 condition, and test trials presented to all participants. Participants were instructed to simply view and listen to the stimuli during the learning phase. For the test phase, they were asked to pick the image they believed was associated with the spoken word in each trial.

ambiguity. Participants were assigned to one of the learning conditions ($N_{1\times1} = 23$, $N_{2\times2} = 19$, $N_{3\times3} = 21$, $N_{4\times4} = 23$). In the 1 × 1 condition, the image was centered on the *x*-and *y*-axis. In the other conditions, images were centered on the *y*-axis and were symmetrically arrayed about the *x*-axis. So that there would be no explicit indication of which word referred to which referent, we created one randomisation of the temporal presentation of the words and physical location of the objects for each trial in the 2 × 2–4 × 4 learning conditions. The order of the learning trials in all learning conditions was randomized across participants. Thus, each participant in a given learning condition received the same trials, but in a random order.

As can be seen in Table 1, the number of trials across the four learning conditions was controlled such that each participant was exposed to each word–object pairing six times, hearing the two tokens of each word three times. Thus, the 1 × 1 learning condition contained 48 trials, the 2 × 2 condition contained 24 trials, the 3 × 3 condition contained 16 trials, and the 4 × 4 condition contained 12 trials. Each trial began after participants fixated on a central cross for 1 s. This was followed by presentation of the visual referents for 0.5 s before onset of the first word. In the 1 × 1 learning condition, a trial would end 3 s after word onset. In the other conditions, 3 s marked onset of the second word, and so on until all words had been presented. Thus, trial lengths after the fixation criterion was met were 3.5 s in the 1 × 1 condition, 6.5 s in the 2 × 2 condition, 9.5 s in the 3 × 3 condition, and 12.5 s in the 4 × 4 condition.

Table 1
Trial composition for the four learning conditions

| Learning Condition | Word–Object Pairings | Presentations of Each Pairing | Trials | Time Per Trial (s) | Total Learning Time (s) |
|---|---|---|---|---|---|
| 1 × 1 | 8 | 6 | 48 | 3.5 | 168 |
| 2 × 2 | 8 | 6 | 24 | 6.5 | 156 |
| 3 × 3 | 8 | 6 | 16 | 9.5 | 152 |
| 4 × 4 | 8 | 6 | 12 | 12.5 | 150 |

#### 2.3.2. *Testing phase*

The testing phase immediately followed the learning phase. All participants received the same test regardless of which learning condition they completed. In each trial, participants heard two alternating repetitions of the two tokens of the target word (i.e., four repetitions of one word; starting token counterbalanced) in the context of four possible referents presented across the four corners of the screen. Each of the 8 words served as the target 7 times, resulting in 56 total trials presented in random order. Each of the three distractors belonged to either the same or opposite set (vowel or consonant) as the target. That is, if the target word belonged to the set of words that differed from one another by one vowel (the vowel set), each distractor either also belonged to the vowel set or belonged to the consonant set. This created four target-distractor quartet relationship types, as shown in Table 2. Thus, the target could occur in the context of (a) no distractors from the same set, and three from the other set, (b) one distractor from the same set, and two from the other, (c) two distractors from the same set and one from the other, or (d) three distractors from the same set, and no distractors from the other. Across trials, the combination of distractor images paired with each target was never repeated so that each trial was novel in this way. Each target word occurred twice with zero to two same-set distractors (a–c), and once with the sole combination of three same-set distractors (d), except for DIT and DUT, which occurred as targets once with zero same-set distractors (a) and three times with two same-set distractors (c) due to a coding error.

As in the learning phase, each trial began after participants fixated on a central cross for 1 s, at which point the four visual referents appeared on the screen. After 0.5 s, the first word repetition began, with the next token beginning 1.5 s after onset of the previous token, and so on for the four total repetitions. Thus, each test trial was 6.5 s. Participants were instructed to indicate via keyboard press to which item they thought the spoken word referred, and to answer as quickly and accurately as possible. Participants could make their selection at any point after onset of the first word repetition, at which point the trial ended. The total test duration was approximately 7 minutes.

## 3. Results

We were interested in whether the degree of within-trial ambiguity during training affected subsequent identification of words which differed from others in only one

Table 2
Different target–distractor relationship types for test trials, depending on the number of distractors belonging to the same or opposing minimal set as the target

| | | Distractors | | Example | |
|---|---|---|---|---|---|
| | Target Set | Number of Same-Set | Number of Opposite-Set | Target | Distractors |
| a | C | 0 | 3 | BON | DEET (V)—DIT (V)—DOOT (V) |
| | V | 0 | 3 | DEET | BON (C)—PON (C)—TON (C) |
| b | C | 1 | 2 | BON | PON (C)—DEET (V)—DIT (V) |
| | V | 1 | 2 | DEET | DIT (V)—BON (C)—PON (C) |
| c | C | 2 | 1 | BON | PON (C)—TON (C)—DEET (V) |
| | V | 2 | 1 | DEET | DIT (V)—DOOT (V)—BON (C) |
| d | C | 3 | 0 | BON | PON (C)—TON (C)—DON (C) |
| | V | 3 | 0 | DEET | DIT (V)—DOOT (V)—DUT (V) |

*Note* C, consonant set; V, vowel set.

consonant or vowel across test trials with differing amounts of same-set distractors. Prior to analysis, we removed test trials in which the reaction time was lower than 601 ms or greater than 6,500 ms. The lower limit of 601 ms corresponded to the onset of the first auditory token occurring at 500 ms plus 100 ms for auditory processing (e.g., Salthouse & Ellis, 1980). The maximum time of 6,500 ms corresponded to the onset of the final auditory token plus 1,500 ms, which was the inter-stimulus interval used in the previous token presentations. This upper limit was only 6 ms greater than another commonly used criterion, the median value (3,894 ms) plus two times the standard deviation ($SD$ = 1,947 ms). These criteria removed one sample that did not meet the minimum and 251 samples that exceeded the maximum time, in total removing 5.23% of the samples, which is within recommended guidelines for dealing with reaction time data (Ratcliff, 1993). Means and standard deviations of participants' accuracy and reaction times for correct responses across all within- and between-subjects factors are in Table 3.

We then fitted a mixed-effects binomial logistic model to participants' correct and incorrect responses to test trials in R (version 3.3.2; R Core Team, 2016) using the glmer function from the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). We included random intercepts[2] for participant and introduced our between-subjects factor of learning condition (1 × 1, 2 × 2, 3 × 3, 4 × 4) and within-subjects factors of minimal set (consonant vs. vowel) and same-set distractors (0–3) as factors in a factorial design. This revealed main effects of learning condition ($F[3, 85] = 7.56$, $p < .001$) and set ($F[1, 4452] = 8.03$, $p = .005$), but not same-set distractors ($F[3, 4452] = 1.87$, $p = .133$). While interactions involving learning condition were not significant (learning condition × set: $F[3, 4452] = 0.95$, $p = .414$; learning condition × same-set distractors: $F[9, 4452] = 0.97$, $p = .464$; learning condition × set × same-set distractors: $F[9, 4451] = 0.84$, $p = .576$), there was a two-way interaction of set and same-set distractors ($F[3, 4451] = 5.79$, $p < .001$).

Table 3
Means and standard deviations of participants' accuracy and reaction time (RT) for correct responses across all within- and between-subjects variables

| Learning Condition | Target Set | Same-Set Distractors | Accuracy (%) | | RT (ms) | |
|---|---|---|---|---|---|---|
| | | | Mean | *SD* | Mean | *SD* |
| 1 × 1 | Consonant | 0 | 63.98 | 28.90 | 2,784.04 | 1,036.10 |
| | | 1 | 57.76 | 29.20 | 2,702.44 | 1,238.72 |
| | | 2 | 57.78 | 25.83 | 2,586.66 | 1,089.85 |
| | | 3 | 68.84 | 26.85 | 2,978.99 | 1,222.21 |
| | Vowel | 0 | 63.48 | 31.24 | 3,081.13 | 1,122.30 |
| | | 1 | 55.67 | 17.01 | 2,981.26 | 974.04 |
| | | 2 | 50.48 | 20.06 | 2,886.22 | 1,098.79 |
| | | 3 | 43.48 | 29.83 | 2,772.34 | 1,131.54 |
| 2 × 2 | Consonant | 0 | 53.82 | 29.23 | 2,337.09 | 924.35 |
| | | 1 | 50.00 | 27.64 | 2,297.66 | 1,016.14 |
| | | 2 | 54.23 | 28.74 | 2,633.63 | 1,168.22 |
| | | 3 | 49.12 | 32.50 | 2,380.37 | 1,214.77 |
| | Vowel | 0 | 61.05 | 24.85 | 2,467.73 | 953.09 |
| | | 1 | 49.53 | 25.33 | 2,310.01 | 1,080.12 |
| | | 2 | 47.95 | 25.76 | 2,674.50 | 1,096.35 |
| | | 3 | 52.19 | 23.71 | 2,445.71 | 1,088.87 |
| 3 × 3 | Consonant | 0 | 32.37 | 20.99 | 2,552.73 | 1,416.15 |
| | | 1 | 42.45 | 15.46 | 2,828.06 | 1,569.61 |
| | | 2 | 42.29 | 22.73 | 2,939.15 | 1,423.57 |
| | | 3 | 45.24 | 23.36 | 2,780.38 | 1,589.79 |
| | Vowel | 0 | 40.56 | 24.60 | 3,315.46 | 1,574.62 |
| | | 1 | 33.65 | 18.66 | 2,818.29 | 1,422.48 |
| | | 2 | 35.53 | 12.37 | 3,205.27 | 1,594.42 |
| | | 3 | 30.56 | 26.00 | 2,876.81 | 1,519.93 |
| 4 × 4 | Consonant | 0 | 37.76 | 23.23 | 3,126.32 | 1,074.53 |
| | | 1 | 36.78 | 25.38 | 3,260.02 | 1,365.69 |
| | | 2 | 44.02 | 23.34 | 3,139.96 | 1,337.13 |
| | | 3 | 36.59 | 28.62 | 2,905.04 | 1,305.86 |
| | Vowel | 0 | 42.32 | 20.16 | 3,056.08 | 1,141.56 |
| | | 1 | 34.55 | 18.91 | 2,977.48 | 1,151.15 |
| | | 2 | 31.93 | 22.29 | 2,945.72 | 1,314.49 |
| | | 3 | 28.03 | 22.79 | 2,504.65 | 1,140.56 |

Participants in all learning conditions performed above chance (25%) at test (1 × 1: $M = 58\%$, $t[22] = 7.81$, $p < .001$, 95% CI [49, 66]; 2 × 2: $M = 52\%$, $t[18] = 5.32$, $p < .001$, [41, 63]; 3 × 3: $M = 38\%$, $t[20] = 5.54$, $p < .001$, [33, 42]; 4 × 4: $M = 37\%$, $t[22] = 4.04$, $p < .001$, [31, 43]). Planned orthogonal linear comparisons using asymptotic $z$-tests revealed that participants in the 1 × 1 unambiguous, fast-mapping learning condition outperformed participants who received training in a cross-situational learning paradigm (2 × 2 to 4 × 4; $z = -3.42$, $p = 0.002$; Fig. 3). Participants in the 2 × 2 learning condition outperformed those in the 3 × 3 and 4 × 4 conditions ($z = -2.68$, $p = .022$), but performance did not differ between participants in the 3 × 3 and 4 × 4 conditions ($z$ = 0.69, $p$ = 0.869).

As mentioned above, the model revealed a main effect of set, such that overall accuracy was greater for targets in the consonant than in the vowel set. With respect to the interaction of set and same-set distractors, there was a difference in the slope between the consonant and vowel set in terms of the relationship between number of same-set distractors and accuracy ($z$ = 3.60, $p$ < .001; see Fig. 4). As the number of same-set distractors increased, performance for vowel set targets showed a linear decrease ($z$ = −3.76, $p$ < .001), while there was no such decrease in accuracy for consonant set targets ($z$ = 1.27, $p$ = .206). Nevertheless, performance in each case was above chance (all $p$s < .001).

We next analyzed participants' reaction times to correct responses in a mixed-effects linear model using the lmer function from the lme4 package (Bates et al., 2015). We included random intercepts for participant and introduced our between-subjects factor of learning condition (1 × 1, 2 × 2, 3 × 3, 4 × 4) and within-subjects factors of set (consonant vs. vowel) and same-set distractors (0–3) in a factorial design. This revealed a main effect of set ($F$[1, 2155.80] = 13.26, $p$ < .005) whereby reaction time was faster for consonant set targets ($M$ = 2,745 ms, $SD$ = 1,410 ms) than vowel set targets ($M$ = 2,957 ms, $SD$ = 1,395 ms). There was also an interaction of learning condition and the number of same-set distractors ($F$[9, 2153.40] = 1.98, $p$ = .034), but comparing the pattern across 0–3 same-set distractors in linear contrasts for each learning condition

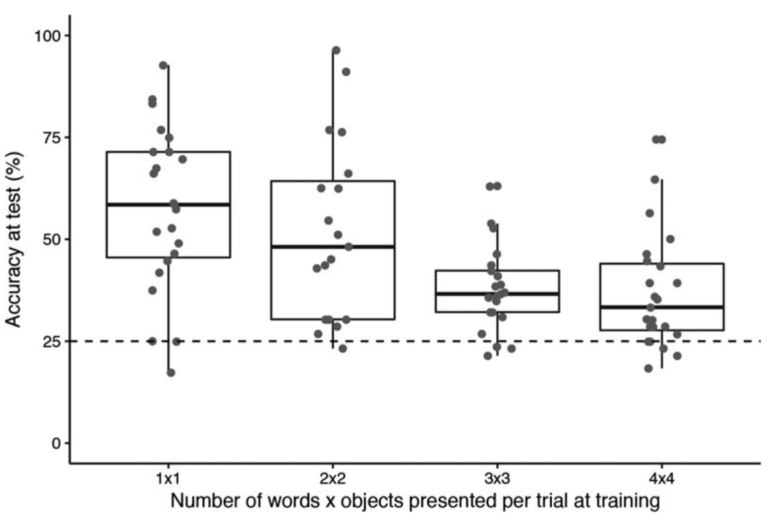


Fig. 3. Percent accurate identification of target word–object pairing during test across participants who received differing degrees of within-trial ambiguity at training (see Fig. 2). Performance was above chance (25%) in each instance. Participants who received training in the 1 × 1 fast-mapping condition outperformed those in the XSWL conditions. Participants in the 2 × 2 condition outperformed those in the 3 × 3 and 4 × 4 conditions, but performance did not differ between participants in the 3 × 3 and 4 × 4 conditions. Data are jittered horizontally and vertically.

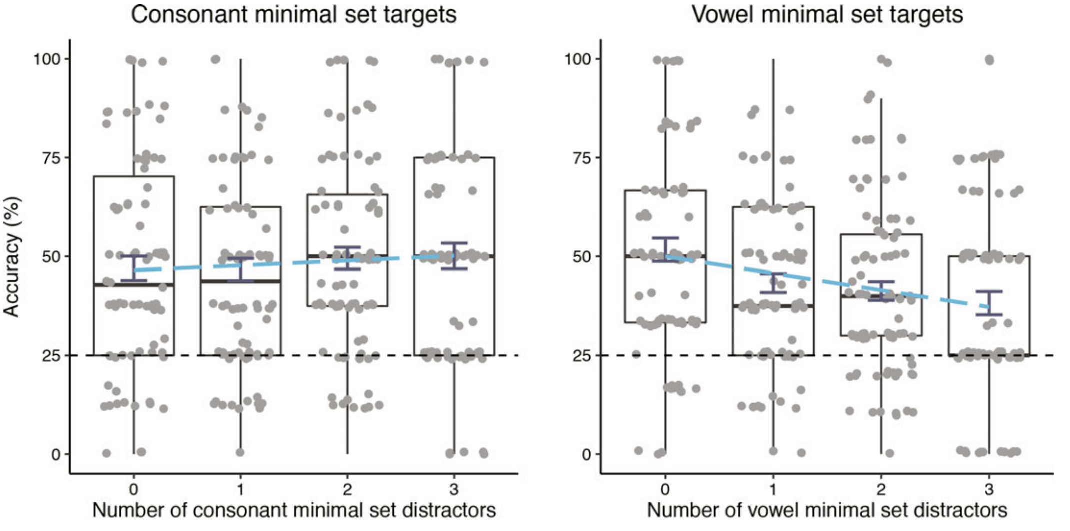


Fig. 4. Percent accurate identification of target word–object pairings at test across test trials with differing numbers of same-set distractors. Performance was above chance (25%) in each instance, though overall performance for minimal consonant targets was greater than for minimal vowel targets. As the number of same-set distractors increased, performance decreased for minimal vowel targets, but not consonant targets. Data are jittered horizontally and vertically.

revealed no effects, making it difficult to interpret the interaction in a meaningful way. There were no main effects of condition ($F$[3, 81.93] = 1.13, $p$ = .342) nor same-set distractors ($F$[3, 2153.68] = .069, $p$ = .560), and interactions involving set were not significant (condition × set: $F$[3, 2155.45] = 1.80, $p$ = 0.146; set × same-set distractors: $F$[3, 2154.15] = 1.37, $p$ = 0.250; condition × set × same-set distractors: $F$[9, 2153.75] = 0.431, $p$ = 0.919).

## 4. Discussion

We directly compared word learning between fast-mapping and XSWL, examined whether more ambiguous XSWL situations impact learners' ability to encode phonetic detail of consonants and vowels, and tested the strength of that encoding. We compared learning of two sets of four phonologically overlapping words that differed from other words in the set only in the initial consonant or vowel, following a learning condition in which participants were presented with one to four words and candidate referents in each trial. Learning was assessed via participants' speed and accuracy in identifying the visual referent corresponding to a target auditory word in the context of three visual distractors. To assess the strength of encoding, we manipulated the number of distractors whose associated word came from the same set as the target (differing in only one segment) or the other set.

Participants in all four learning conditions (1 × 1 through 4 × 4) learned and encoded words with sufficient phonetic detail to allow above-chance performance. As predicted, participants in the unambiguous 1 × 1 learning condition were more accurate than those in the 2 × 2 to 4 × 4 ambiguous XSWL conditions, demonstrating for the first time that word learning via fast-mapping is more robust (at least initially) than XSWL. This finding also suggests fast-mapping is a fundamentally simpler task than XSWL, and it allows us to estimate the overall word encoding difficulty prior to the effects of having to track occurrences of targets and distractors between trials. The absence of interaction effects between learning condition and the number of phonologically overlapping distractors on test suggests similar phonological encoding processes between fast mapping and cross-situational word learning. Indeed, Trueswell et al. (2013) have suggested an alternative account of cross-situational word learning in which the initial underlying process is shared between fast-mapping and XSWL, which may explain similar encoding abilities. In this view, XSWL begins as fast mapping of one candidate referent with subsequent hypothesis testing. These data are not inconsistent with that view—by this interpretation, fast-mapping of a single target word would remain the easiest because the initial word-object pairing hypothesis is always correct and no conflicting evidence has to be accounted for.

Past studies of XSWL of very phonologically dissimilar words show that as referential ambiguity in learning trials increases from 2 to 4 items (2 × 2 to 4 × 4 conditions), word learning decreases or appears to decrease near-linearly (Vlach & Sandhofer, 2014; Yu & Smith, 2007). Our data show a broadly similar pattern, with participants in the 2 × 2 learning condition more accurate than those in the 3 × 3 and 4 × 4 learning conditions. However, while accuracy has been shown to significantly decrease between participants in the 3 × 3 and 4 × 4 learning conditions when tested on phonologically dissimilar words (Vlach & Sandhofer, 2014), here, performance between participants in the 3 × 3 and 4 × 4 learning conditions did not differ. This suggests that using phonologically overlapping words, which require encoding of greater phonological detail, added demand to the task that effectively removed the scope for further decrease in accuracy between the 3 × 3 and 4 × 4 conditions. It is possible that this added demand taxed memory resources, which were already constrained in the 3 × 3 and 4 × 4 conditions, as visual working memory is generally limited to three to four items (Luck & Vogel, 1997; Sperling, 1960) and memory abilities are critical to XSWL performance (Vlach & DeBrock, 2017).

The added demand imposed by phonetic encoding may also be reflected in the apparent overall lower accuracy across learning conditions relative to prior work. Accuracy in our 2 × 2 condition was 52%, whereas accuracy exceeded 80% in studies that tested learning of phonologically dissimilar words also using four-option test trials (Vlach & Sandhofer, 2014, immediate test; Yu & Smith, 2007, experiment 1). Notably, the accuracy reported in these previous XSWL studies appears to exceed even the accuracy in our 1 × 1 unambiguous condition (58%). Likewise, accuracy in our 4 × 4 condition (37%) appears considerably lower than that recorded across a variety of 4 × 4 learning conditions of phonologically dissimilar words (>50%; Yu & Smith, 2007, experiment 2). This suggests an absolute effect of both phonological similarity and XSWL on increasing task difficulty, and it warrants future investigation via direct comparison between phonologically similar and dissimilar words across degrees of within-trial ambiguity during training. Reduced accuracy may have arisen through other factors such as test fatigue as our test phase included many more trials (56 vs. 4: Vlach & Sandhofer, 2014). However, this appears unlikely, as inspection of the accuracy for the first four test trials by each participant appears to result in similar or lower values across learning conditions ($1 \times 1$: 52%; $2 \times 2$: 47%; $3 \times 3$: 36%; $4 \times 4$: 27%).

The lack of a decrease between the $3 \times 3$ and $4 \times 4$ conditions following decreases across the less ambiguous learning conditions is suggestive of a non-linear decrease in accuracy as ambiguity increases, though additional degrees of ambiguity are certainly required to adequately assess the true shape of this relationship. This is in contrast to the linear appearance of results from earlier studies that have compared $2 \times 2$ through $4 \times 4$ learning scenarios using phonologically dissimilar words (Vlach & Sandhofer, 2014; Yu & Smith, 2007). This may suggest the pattern here may be attributed to the increased task difficulty incurred by presenting participants with phonologically overlapping words. However, more than four degrees of ambiguity are required to adequately assess the true shape of this curve, and thus a future avenue of research that presents even greater amounts of ambiguity and compares performance between phonologically overlapping and non-overlapping sets of words would further detail the processing cost of fine phonological encoding and the limits of XSWL.

This study includes fewer target words than Vlach and Sandhofer (2014) and Yu and Smith (2007; 8 vs. 18 words), but the same number of presentations of each word–object pairing in the learning phase. Modeling of XSWL suggests that the number of trials needed until word–object pairings can be statistically derived increases logarithmically as ambiguity increases (Blythe et al., 2016). If this reflects natural XSWL, one would expect that participants would perform better at learning 8 words compared to 18 words. The fact that we appear to find similar or lower accuracy scores rather higher accuracy scores across our three XSWL learning conditions compared to studies requiring participants to learn over twice as many (non-minimal) words (Vlach & Sandhofer, 2014; Yu & Smith, 2007) implies a dramatic cost of phonological overlap to working memory and points to the importance of mastery of native categories in making word learning possible.

Regarding the strength of encoding of consonants and vowels in our task, our results suggest more robust phonetic encoding for consonants than vowels: Accuracy was higher for words differing in a consonant than for those differing in a vowel. This mirrors the finding by Escudero et al. (2016a). In the first demonstration of a possible consonant bias in a naturalistic XSWL scenario, they found that adults who learned the same words and referents here in a $2 \times 2$ scenario were more accurate at identifying the target referent at test in a two-alternative forced choice task when the word associated with the distractor referent formed a consonant minimal pair with the target word compared to when it formed a vowel minimal pair. This study extends this finding of a possible consonant encoding bias to what is in theory an even more naturalistic word learning situation, as there is a greater amount of ambiguity at learning, and learning was assessed in the context of a greater number of candidate referents. Both factors likely more closely reflect the ambiguity or range of ambiguity encountered in natural implicit word learning settings.

More important, however, this study allowed, for the first time, examination of the *strength* of encoding of consonants and vowels in a naturalistic word learning scenario, by comparing accuracy across trials containing differing numbers of distractor images whose referent words formed a minimal pair with the target word. Not only was overall accuracy greater for consonant set targets than vowel set targets, but accuracy for consonant set targets was seemingly *unaffected* by the number of same-set distractors at test; only words differing in a vowel showed a linear decrease in accuracy as same-set distractors increased. That is, even if consonants were encoded more strongly than vowels, one would reasonably still expect performance to decrease as the amount of more highly confusable distractors increased. Thus, this finding uncovers a remarkably high degree of strength for consonant encoding in this task, and it suggests powerful encoding of consonant information in natural word learning.

As mentioned in the Introduction, these results may be driven by the tendency for native adult listeners of English to perceive vowels less categorically (Beddor & Strange, 1982; Fry et al., 1962; Polka, 1995; Stevens, Liberman, Studdert-Kennedy, & Öhman, 1969) than consonants (Liberman et al., 1967). This less categorical perception may make it more difficult to encode vowels in fine details, and/or may also make it more difficult to match the vowel properties of the speech input to encoded lexical entries. A consonant bias is well attested in lexical tasks in English, and as such, it has been reasoned that consonants are more critical to a word's identity, driving lexical access and processing (see Nazzi & Cutler, 2019 for a review). But stimulus and task can influence which information is most critical or attended to. For instance, when listening to running speech, vowel information is more critical to sentence comprehension (Fogerty, Kewley-Port, & Humes, 2012; Kewley-Port, Burkle, & Lee, 2007), evidenced by participants' higher sentence repetition accuracy when the vowel, but not consonant, information was intact, compared to the inverse (Kewley-Port et al., 2007). Our results suggest that in a cross-situational word learning task in which words are presented in citation style, consonant information is more strongly encoded than vowel information. While this suggests a consonant bias in a naturalistic word learning task, there is scope to further the naturalness of the task by testing XSWL when words are presented within sentences—in which case it is possible that vowel information may be more strongly encoded (see Kewley-Port et al., 2007)—to further clarify how consonant and vowel attention and encoding is driven by task, stimulus, and the particular linguistic skill being tested.

It is important to note that the position of the critical consonant and vowel segment may have also contributed to the differences in performance (see also Escudero et al., 2016a). While the consonant differentiating the consonant set occurred as the first segment, the critical vowel in the vowel set occurred as the second segment, with /d/ as the initial segment. Indeed, our results demonstrate that reaction time was faster for consonant set targets than vowel set targets, which likely reflects this difference in location of the critical segment between sets. Given that initial segments may have a more prominent role than later segments in lexical access and identification (Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson & Zwitserlood, 1989), participants may have paid more attention to initial segments, masking any intrinsic differences in the processing of consonants versus vowels. Future research should compare processing of consonants and vowels while controlling for position effects. For instance, comparing word learning across four minimal sets of words that follow the structure (with the critical segment underlined) CVCV, CVCV, VCVC, VCVC, which would allow examination of the role of syllable structure, segment type, and position of the critical segment. Notably, the first research to address this issue suggests that position effects do not account for consonant-vowel asymmetries. Twenty-month-old Canadian English–French bilinguals failed to learn VCVC words that differed in their initial vowel, which also occurred as the initial segment, but learned words that differed in their initial consonant, which occurred as the second segment (Nazzi & Polka, 2018). Together with previous work showing that French-learning 20-month-olds show a consonant bias when learning new words—differentiating between CVCV words that differ in either consonant—but struggle to differentiate between words that differ in their vowel (Nazzi, 2005), this supports poorer encoding of vowel information regardless of segment position. At this point, more research is required to see if this is also the case in adults, and how the language background of the listener may contribute to this relationship.

There is scope for future work to build off the current experiment by testing a wider variety of contrasts to test the generalizability of the consonant bias in XSWL. Both the consonant and vowel contrasts were selected to differ from one another within each minimal set by one or two features. Vowels differed in tenseness (/i-ɪ/,[3] /u-ʊ/; see Fig. 1 for examples of English words containing these vowels), backness (/i-u/, /ɪ-ʊ/), or both (/i-ʊ/, /ɪ-u/). Consonants differed in place (/b-d/, /p-t/), voicing (/b-p/, /d-t/), or both (/d-p/, /b-t/). Manner of consonant articulation did not differ, however; all consonants were plosives. It is possible that the consonant bias attested here may be specific to these consonant contrasts, since plosives are most distinct from vowels on the sonority scale (Clements, 1990). Using more vowel-like consonants (e.g., liquids or glides) in this task may mitigate a consonant bias since they may be perceived more like vowels, that is, less categorically (Beddor & Strange, 1982; Fry et al., 1962; Liberman et al., 1967) than the plosive consonants tested here. Thus, testing a wider range of contrasts would determine whether the consonant bias holds categorically for all consonants relative to all vowels, or if the bias is driven by certain acoustic-phonetic features characteristic of only a subset of contrasts.

In summary, XSWL provides a plausible approach to understanding the problem of multiple referents. While previous research shows us that adults can use this method to learn words implicitly among many candidate referents, our research demonstrates how phonetic encoding of both consonants and vowels takes place in this learning scenario, which is crucial for successful perception and learning of words in the real world. While participants' word learning accuracy via fast-mapping was higher than via XSWL, we have shown that adults can learn new words when presented with up to four candidate referents simultaneously, and they can encode those words with fine-grained phonetic detail. However, learning is impaired as the number of referents during learning increases, it appears to be dependent in part on segmental properties. These findings lay groundwork for future exploration into the nature and limits of XSWL as a word learning strategy. Testing a wider range of segments and varying the position of critical segments will better inform the effects of these properties on phonetic encoding, and testing speakers of different languages in these properties will help to categorize the language-specific and universal mechanisms at play. Going forward, the manipulation of task difficulty by varying the number of competitors may be key to uncovering these differences.

## Acknowledgments

This research was funded by the ARC (Australian Research Council) Centre of Excellence for the Dynamics of Language (project number CE140100041) where Paola Escudero is Chief Investigator and in part by R03 HD081153-01 awarded to Haley Vlach.

## Notes

1. Of course, as with the majority of such studies, it is also possible that any differences between consonant and vowel encoding are due to the specific consonants and vowels tested rather than to consonants and vowels generally, as the stimuli comprise a limited set of consonant and vowel contrasts which have features that cannot be easily compared. This possibility is explored further in the Discussion.
2. Because participants were of a wide age range (17.1–54.4 years), and because of the coding error described in the testing phase description, we also introduced random intercepts for participant age and whether a target word was one of the two subjected to the error. This did not improve the model ($\chi^2[2, 35] = 1.42$, $p = .491$) and resulted in a 2.5-point increase in AIC. Thus, they were not included in the final model.
3. These vowels are also differentiated by their dynamic properties in AusE (Elvin, Williams, & Escudero, 2016; Escudero, Mulak, Elvin, & Traynor, 2018; Williams, Escudero, & Gafos, 2018).

## 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*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beddor, P. S., & Strange, W. (1982). Cross-language study of perception of the oral-nasal distinction. *Journal of the Acoustical Society of America*, *71*(6), 1551–1561. https://doi.org/10.1121/1.387809.

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *Journal of Memory and Language*, *90*, 31–48. https://doi.org/10.1016/j.jml.2016.03.004.

Blythe, R. A., Smith, A. D. M., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18–27. https://doi.org/10.1016/j.cognition.2016.02.017.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 283–333). Cambridge, UK: Cambridge University Press.

Crowder, R. G. (1971). The sound of vowels and consonants in immediate memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(6), 587–596. https://doi.org/10.1016/S0022-5371(71)80063-4.

Curtin, S. A., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science*, *12*(5), 725–731. https://doi.org/10.1111/j.1467-7687.2009.00814.x.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, *28*(5), 746–755. https://doi.org/10.3758/BF03198409.

Elvin, J., Williams, D., & Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the Acoustical Society of America*, *140*(1), 576–581. https://doi.org/10.1121/1.4952387.

Escudero, P., Mulak, K. E., Elvin, J., & Traynor, N. M. (2018). “Mummy, keep it steady”: Phonetic variation shapes word learning at 15 and 17 months. *Developmental Science*, *21*(5), e12640. https://doi.org/10.1111/desc.12640.

Escudero, P., Mulak, K. E., Fu, C. S., & Singh, L. (2016). More limitations to monolingualism: Bilinguals outperform monolinguals in implicit word learning. *Frontiers in Psychology*, *7*, 1218. https://doi.org/10.3389/fpsyg.2016.01218.

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016a). Cross-situational learning of minimal word pairs. *Cognitive Science*, *40*(2), 455–465. https://doi.org/10.1111/cogs.12243.

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016b). Infants encode phonetic detail during cross-situational word learning. *Frontiers in Psychology*, *7*, 1419. https://doi.org/10.3389/fpsyg.2016.01419.

Fikkert, P. (2010). Developing representations and the emergence of phonology: Evidence from perception and production. In C. Fougeron, B. Kühnert, & M. D’Imperio (Eds.), *Laboratory phonology 10: Variation, phonetic detail and phonological representation* (Vol. 10, pp. 227–258). Berlin: De Gruyter Mouton.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science*, *35*, 367–380. https://doi.org/10.1111/j.1551-6709.2010.01156.x.

Fogerty, D., Kewley-Port, D., & Humes, L. E. (2012). The relative importance of consonant and vowel segments to the recognition of words and sentences: Effects of age and hearing loss. *The Journal of the Acoustical Society of America*, *132*(3), 1667–1678. https://doi.org/10.1121/1.4739463.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language & Speech*, *5*(4), 171–189.

Højen, A., & Nazzi, T. (2015). Vowel bias in Danish word-learning: Processing biases are language-specific. *Developmental Science*, *19*(1), 41–49. https://doi.org/10.1111/desc.12286.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*, *41*(3), 590–622. https://doi.org/10.1111/cogs.12353.

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *122*(4), 2365–2375. https://doi.org/10.1121/1.2773986.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 576–585. https://doi.org/10.1037/0096-1523.15.3.576.

Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition*, *98*(1), 13–30. https://doi.org/10.1016/j.cognition.2004.10.005.

Nazzi, T., & Cutler, A. (2019). How consonants and vowels shape spoken-language recognition. *Annual Review of Linguistics*, *5*(1), 25–47. https://doi.org/10.1146/annurev-linguistics-011718-011919.

Nazzi, T., & Polka, L. (2018). The consonant bias in word learning is not determined by position within the word: Evidence from vowel-initial words. *Journal of Experimental Child Psychology*, *174*, 103–111. https://doi.org/10.1016/j.jecp.2018.05.011.

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America*, *97*(2), 1286–1296. https://doi.org/10.1121/1.412170.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *The American Journal of Psychology*, *93*(2), 207. https://doi.org/10.2307/1422228.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759.

Stevens, K. N., Liberman, A. M., Studdert-Kennedy, M., & Öhman, S. (1969). Cross-language study of vowel perception. *Language and Speech*, *12*, 1–23. https://doi.org/10.1177/002383096901200101.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001.

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217–230. https://doi.org/10.1016/j.jml.2016.10.001.

Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science*, *38*(4), 757–774. https://doi.org/10.1111/cogs.12092.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*(1), 59–82. https://doi.org/10.1177/0023830915578000.

Williams, D., Escudero, P., & Gafos, A. (2018). Spectral change and duration as cues in Australian English listeners' front vowel categorization. *The Journal of the Acoustical Society of America*, *144*(3), EL215–EL221. https://doi.org/10.1121/1.5055019.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62. https://doi.org/10.1016/j.cognition.2015.07.013.