COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1618–1641
© 2018 Cognitive Science Society, Inc. All rights reserved.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.12620

Lexical Learning May Contribute to Phonetic Learning in Infants: A Corpus Analysis of Maternal Spanish

Daniel Swingley, Claudia Alarcon

Department of Psychology, University of Pennsylvania

Received 5 June 2017; received in revised form 6 February 2018; accepted 9 February 2018

Abstract

In their first year, infants begin to learn the speech sounds of their language. This process is typically modeled as an unsupervised clustering problem in which phonetically similar speech-sound tokens are grouped into phonetic categories by infants using their domain-general inference abilities. We argue here that maternal speech is too phonetically variable for this account to be plausible, and we provide phonetic evidence from Spanish showing that infant-directed Spanish vowels are more readily clustered over word types than over vowel tokens. The results suggest that infants' early adaptation to native-language phonetics depends on their word-form lexicon, implicating a much wider range of potential sources of influence on infants' developmental trajectories in language learning.

Keywords: Language; Phonetics; Categorization; Perception; Lexicon; Infancy

1. Introduction

Language acquisition is often said to begin with phonological learning. Accounts of this process share the premise that infants are born with the potential to learn any human language, a potential that includes the ability to tell apart clear instances of any language's phonologically contrastive speech sounds (e.g., Kuhl et al., 2008; Werker & Hensch, 2015; although see Narayan, Werker, & Beddor, 2010). By the second half of the first year, infants have learned enough of their native language to adapt to its particular sound structure, notably by becoming less capable of distinguishing foreign sounds and better at distinguishing sounds that are linguistically relevant in the native language (e.g., Kuhl et al., 2008; Yeung, Chen, & Werker, 2013). The precocious development of this adaptation—as early as 6 months of age, in Polka and Werker (1994)—seemed to rule

Correspondence should be sent to Daniel Swingley, Department of Psychology, University of Pennsylvania, 425 S. University Avenue, Philadelphia, PA 19104. E-mail: swingley@psych.upenn.edu

out the possibility that infants learn their language's sounds using the lexicon, by attending to the phonetic differences between phonological "minimal pairs." If words like *wheel* and *whale* were to signal to English learners that [i] and [e^{I}] are distinct, children would need a basis for differentiating the two words, and using the words' meanings to do this is the obvious path. But infants have widely been viewed as not knowing any words' meanings before 9 months of age (e.g., Bloom, 2000; Tomasello, 2001), and certainly not in numbers adequate to provide minimal pairs that would clarify the language's phonetic categories starting around 6 months.

The apparent implausibility of the minimal-pair proposal led to the conclusion that the lexicon *per se* is not involved in infants' early phonetic category learning. Instead, researchers turned to the notion that infants learn phones via an unsupervised distributional clustering process (e.g., Guenther & Gjaja, 1996; Kuhl, 1992; Lacerda, 1995). The idea behind unsupervised clustering is that categories are identifiable as relatively tightly packed instances in a perceptual space. These instances form statistical modes that can be discovered by an observer without any labeled training examples (Duda & Hart, 1973). For example, if a sample of sounds varies bimodally in a dimension like duration, the learner may discover that there are many short sounds and many long sounds but few medium sounds, and infer that there are two categories that differ in their typical duration.

That such a mechanism might account for infants' phonetic category learning has two preconditions: first, that infants have the cognitive wherewithal to perform this clustering (which requires skills of perception and memory and the clustering itself), and second, that the speech infants hear offers distributionally distinct categories to be discovered. The literature provides more reason to be confident about the first than the second. Experimental studies show that with well-separated categories, both infant and adult listeners can quickly modify their interpretation of sounds in a manner consistent with category learning without being given feedback (e.g., Cristia, McGuire, Seidl, & Francis, 2011; Francis, Kaganovich, & Driscoll-Huber, 2008; Goudbeek, Swingley, & Smits, 2009; Liu & Holt, 2015; Maye, Werker, & Gerken, 2002; Yoshida, Pons, Maye, & Werker, 2010). Thus, at least in principle, distributional clustering is a learning mechanism available to infants as a way of discovering their language's phonetic categories.

Are the speech sounds in infant-directed speech sufficiently distinct for unsupervised clustering over tokens to be feasible? Perhaps not. The infant-directed vowels of Russian, English, and Swedish measured by Kuhl et al. (1997), figs. 1, 2) showed substantial spread and overlap even though the analysis only included the point vowels [i,a,u] which are arguably the most distinct possible human monophthongs; the same was true of the Norwegian infant-directed speech vowels measured by Englund and Behne (2005). The mid-front vowels studied by Cristia and Seidl (2014) also showed substantial overlap, even looking within speakers. Jones, Meakins, and Muawiyath (2012) looked at a child-directed speech corpus in the northern Australian language of Gurindji Kriol and found that k-means clustering was unable to separate the five vowels. Adriaans and Swingley (2017) analyzed English vowels from the Brent corpus of infant-directed speech (Brent &

Siskind, 2001) and found that vowels were not, in general, separable using a range of clustering techniques.

Vallabha, McClelland, Pons, Werker, and Amano (2007) showed that Gaussian mixture models could recover the mid-front vowels of English or Japanese from data simulating first and second formant distributions. These models were usually quite successful, but the training data were not measured directly from speech tokens. They were sampled from Gaussian distributions estimated from speech tokens, and those speech tokens themselves came from a book-reading task in which parents taught their children phonologically similar nonce words. It is likely that these speech samples significantly underestimated the variability present in infant-directed speech as a whole (see also McMurray, Aslin, & Toscano, 2009). Thus, in our view, there is room for doubt that infant-directed speech presents infants with distinct distributions of speech sounds that map onto phonetic categories (see also Burnham, Wieland et al., 2015; Martin et al., 2015; Narayan, 2013; Sundberg & Lacerda, 1999). Indeed, whether infant-directed speech should be considered a better teaching signal than adult-directed speech has been questioned, with much of the current evidence showing that infant-directed vowel categories are not particularly separated in phonetic space (e.g., Martin et al., 2015; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013; Miyazawa, Shinya, Martin, Kikuchi, & Mazuka, 2017) even if the average formant values of the point vowels [i,a,u] are more spread out (e.g., Kalashnikova, Carignan, & Burnham, 2017).

Yet infants do learn their language's speech sounds. How? The hypothesis we evaluate here is this: If two phonologically distinct categories overlap phonetically, they will be difficult to learn using token distributions alone; but if the different phones tend to appear in different words, the uneven distribution of sounds over words could signal to the learner that the sounds are distinct (Swingley, 2009). For example, if every time the child hears *please*, the [i] vowel has a high second formant, and every time the child hears *give*, the vowel has a lower second formant, this could indicate that the [i] and [ɪ] vowels are distinct in the language, even if phonetically the [i] and [ɪ] tokens do not form two distinct modes. Note that this could work, in principle, whether children knew the meanings of the words or not. Even if infants only know "protowords" made of phonological content and minimal or no semantics, the lexical contexts could still sketch out the boundaries of the speech sounds they contain.

The idea that lexical context might guide infants' phonetic category learning was raised by Swingley (2006), who proposed that word or syllable contexts might motivate Dutch children's intuition that vowel duration can be contrastive—a conclusion English learners do not reach even though vowel duration distributions are quite similar in the two languages (Dietrich, Swingley, & Werker, 2007). The idea that lexical context might fine-tune existing categories (Norris, McQueen, & Cutler, 2003) or make existing categories more salient for word learning (Thiessen, 2007) has received substantial empirical support, so the argument that words might help children set up phonetic categories in the first place is perhaps not a great leap. Yet it could be an important one, given the implication that the determinants of vocabulary learning could also be determinants of phonetic learning.

The best experimental evidence for the influence of the lexicon on infants' learning of phonetic categories comes from a study by Feldman, Myers, White, Griffiths, and Morgan (2013), using methods developed by Thiessen (2007). Feldman et al. created a continuum of resynthesized vowels from English [a] to [ɔ]. These vowels were embedded in the word contexts *gutah, gutaw, litah,* and *litaw*. The *...tah* words ended with the first four tokens from the continuum (i.e., [a]s), and the *...taw* words with one last four tokens from the continuum (i.e., [ɔ]s), in equal numbers. Thus, in terms of phonetic distributions, the [a] and [ɔ] did not form categories but were presented as elements from a uniform distribution. Infants in the control condition heard all of the vowels from the continuum equally in all four words; for example, they heard [guta] and [gutɔ] the same amount. Infants in the experimental condition heard only words sounding like [guta] and [litɔ], or (counterbalanced) [gutɔ] and [lita]. Thus, for infants in the experimental condition, the situation mimicked the learning of, for example, the English [i] and [ɪ] sketched above, where lexical contexts were aligned with limited portions of the phonetic space.

Following exposure to 128 instances of the words (total), the infants were probed for their preference for repeated instances of one of the two vowels in a monosyllabic context ([ta, ta, ta...]) or alternations of the two ([tɔ, ta, tɔ...), a test of categorization (Best & Jones, 1998). Feldman et al. found that only infants who had heard vowel exemplars in distinct lexical contexts then responded differently to alternating and non-alternating trials. This result showed that at least in a laboratory context, 8-month-olds keep track of the words that phonetically similar vowels appear in, and this information can affect their differentiation of the vowels.

Thus, infants appear to be up to the task. The next question is whether parental speech to children actually fulfills the necessary preconditions, among which two predominate. First, if words are to delineate speech-sound categories, the words must be sufficiently distinct phonologically. If the English learner hears not only *please* and *give* (which are very distinct), but also *beet* and *bit*, *seat* and *sit*, and *feet* and *fit*, the distinctiveness of the different-sounding words could be swamped by these similar-sounding pairs that could signal to the child that in fact [i] and [ɪ] should be treated as equivalent. Of course, adults know that *seat* and *sit* are different words, because adults know what the words mean; but researchers have long supposed that 8-month-olds are more advanced in their extraction of word-forms than they are at discovering word meanings (e.g., Jusczyk & Hohne, 1997; Martin, Peperkamp, & Dupoux, 2013; Swingley, 2005), and in such a state, minimal pairs hinder more than they help. Thus, the first precondition is that sufficiently distinctive words compose the early lexicon.

The second precondition is that the word-forms actually do illustrate their component sounds consistently from word to word. If a given sound is phonetically variable as a function of which word it appears in, the average or prototypical pronunciations of words would not collectively yield a good estimate of each sound's phonetics. (Of course, this would be a problem for a clusterer operating over individual tokens too.) Sounds might vary substantially as a function of the word they are in because of phonetic context effects. To take a few familiar examples, /u/ tends to be fronted when adjacent to sounds like /t/; vowels near nasals tend to be nasalized themselves; and voicing in coda consonants is linked to lengthening in the preceding vowel. Such context effects are numerous and pervasive. Although infants might "undo" some of these effects perceptually (e.g., Eimas & Miller, 1980), from an acoustic perspective, we should expect that to some degree each consonant or vowel in a given word will have a particular realization that depends on the context. The average [i] in *please* might not resemble the average [i] in *mean* or *feet*. If the prototypical realizations of each word's sounds do not line up with the prototypical realizations in other words, then clustering over word types rather than tokens could be unhelpful. (A fuller analysis of this problem is given in the Supporting Online Materials.)

Previous analyses have demonstrated the in-principle utility of the lexicon for deriving phonetic categories but have not evaluated whether infant-directed speech ever meets these preconditions. Feldman, Griffiths, Goldwater, and Morgan (2013) provided the most thorough treatment of this problem. Feldman et al. presented a Bayesian approach to simultaneous optimization of the lexicon and the vowel system. The input to their analysis (Simulations 3 and 4) was a corpus of English words sampled from the CHILDES database (MacWhinney, 2000). Consonant identities were estimated from the canonical pronunciations given in a dictionary and were given to the model as discrete symbols. Vowels' formants were estimated as samples from Hillenbrand, Getty, Clark, and Wheeler (1995). The Hillenbrand dataset is a collection of formant and duration measurements made from many different talkers' utterances of isolated, clear instances of the syllable *hVd*, where V stands for each of the English vowels. (Given that the syllables were hyperarticulated relative to ordinary conversational speech [or child-directed speech], the substantial overlap among measured vowels in this dataset probably derived in large part from individual differences among the speakers.) To estimate vowel formant data as an input to the learning model, Feldman et al. computed mean and covariance values from the Hillenbrand vowels and used these parameters to generate random samples that simulated vowel productions' first and second formant values. Models were judged on how well they correctly assigned vowels with the same (gold standard) identity together into the same category.

Considering a range of parameter values and two different versions of the vowel corpus data, the authors consistently found that models jointly estimating word identity and vowel identity were superior to models without lexical information. This result suggests that minimal-pair words in English infant-directed speech are not frequent enough to sabotage the general utility of the lexicon. However, as Feldman et al. acknowledge, (a) laboratory isolated-syllable *hVd* vowels may not resemble day-to-day child-directed vowels; (b) spoken words are not always uttered with their canonical set of consonants; and (c) by assigning simulated vowels' phonetics to words independently of their context, the model did not evaluate the second of the preconditions described above, namely that for words to be helpful, vowels need to be sufficiently similar across word types, in spite of phonetic context effects and other word-specific phonetic effects.

The goal of the present research is to evaluate the utility of words for vowel category discovery, using phonetic data measured from a sample of unscripted infant-directed speech. We chose to examine a Spanish sample because Spanish has a number of useful properties for this purpose. Spanish has only five phonological vowels, which makes the problem somewhat simpler, and balances to some degree the dominance of English-language analyses in the literature. Also, Spanish vowels are essentially monophthongal and are therefore reasonably well represented using formant values at vowel midpoint. Spanish is widely spoken and researched, so the acquisition trajectory of the Spanish vowels has been documented experimentally to some degree (e.g., Sebastian-Galles & Bosch, 2009). The present authors are a native speaker (C.A.) and a once-fluent speaker (D.S.) of Spanish. Finally, an annotated corpus of Spanish child-directed speech was available (as described below).

Our analysis is fundamentally a descriptive one, based on a case study. First, we test whether a general-purpose classifier succeeds in discovering the vowels of Spanish in this sample. Then, we compare that model's performance with other models that are provided with lexical information. We find that using words as anchors of the category-finding process is more successful than working without words. To the extent that this sample is representative of Spanish infant-directed speech, this result suggests that infants might be able to learn the Spanish vowels because they are learning words at the same time.

2. Methods

2.1. Preparation of the corpus

The analyses used a portion of the Ornat corpus (López Ornat, 1994), which includes video recordings taken during bathing, play, or feeding interactions between a girl of 1;7–4;0 and her parents. The conversation is in Castilian Spanish. Only the recording corresponding to the youngest age was used, namely recording session 01, age 1;7, corresponding to a total of just over 1 h of interaction time. Ideally, we would have used recordings from a parent speaking Spanish to an infant of 6–12 months, but no such recordings were available. Child-directed speech evolves to some degree as infants grow into young children, so it is possible that results with a different corpus would be different. There is no consensus about whether speech-sound categories are more readily identifiable in infant-directed speech or adult-directed speech, or in different varieties of infant-directed speech (e.g., Kalashnikova et al., 2017; Miyazawa et al., 2017, Stern, Spieker, Barnett, & MacKain, 1983). This issue merits further consideration with additional corpora.

Corpus materials were downloaded from the CHILDES repository (MacWhinney, 2000) in 2009. The Ornat corpus is annotated with an orthographic rendering of each sentence, a morphological parse of the utterance, a syntactic parse, and the temporal boundaries of each utterance within the session tape. These temporal boundaries (plus a slight pad at each end) were used to extract each utterance of the mother into its own soundfile. Various text processing steps were done over the maternal utterances to regularize the corpus: removing untranscribed utterances, removing sounds or words that were marked as unspoken, removing punctuation, removing codes like "@f" (words particular to a given family), “@c” (words invented by the child), and so on. Soundfiles that were very noisy, sung, or uninterpretable were discarded. The final analysis corpus consisted of 402 utterances.

An initial phonetic transcription of the corpus was generated using the LDC Spanish lexicon (Garrett, Morton, & McLemore, 1997). To mimic Iberian Spanish, the (Mexican) LDC dictionary was modified to replace some /s/ sounds with /θ/ according to context. Out-of-dictionary words were added by hand. To generate basic word and phone alignments for starting the annotation process, the corpus and soundfiles were run through the Penn Phonetics Lab Forced Aligner Toolkit (Yuan & Liberman, 2008). This system is based on the HTK Speech Recognition Toolkit (Young et al., 2006), and it uses an acoustic model derived from English to estimate phonetic boundaries. (Of course, a Spanish acoustic model would work better, but none was available; our goal was really just to set up the structure of the alignments to speed the hand-annotation process.) To fit the assumptions of the aligner, the Spanish dictionary was converted to approximate English equivalents. These alignments were then converted to the Praat TextGrid format for hand-realignment in Praat (Boersma & Weenink, 2001). Three fluent speakers of Spanish (two native) went through each sentence, realigning the boundaries of phones and words, correcting pronunciations that deviated from the dictionary pronunciation, and adding or (more often) removing phones as needed to match the speech signal.

2.2. Formant measurement

Once word and phone boundaries were established, the first and second vowel formants (vocal tract resonances) were measured using a Praat script. Formants provide an imperfect characterization of vowels, but they are more readily interpretable than, for example, the MFCC representation used in speech recognition (Davis & Mermelstein, 1980), and are the standard basic metric for characterizing vowels in the phonetics literature (e.g., Labov, Ash, & Boberg, 2005). Unlike English vowels, Spanish vowels are not heavily diphthongized and are not differentiated by duration, and so they may be reasonably characterized by their midpoint first and second formants (F1 and F2). That said, measuring formants is not trivial, so a number of checks were undertaken to ensure the accuracy of our measurements and the integrity of the data.

First, every token was extracted from its context (plus 60 ms on each side), sorted by transcribed vowel, and presented to a native speaker listener (C. Alarcon) who judged (a) whether the vowel was, indeed, an instance of the transcribed category; and (b) whether the vowel was even interpretable as speech at all when taken out of context. Of the 2,706 tokens, 466 (17%) were judged uninterpretable, typically because of background noise, poor recording quality, or extreme hypoarticulation or elision. These tokens were removed from the analysis. About 8.9% were judged to not match the transcribed label; however, we let these instances stand (prioritizing transcriptions in context). Forty-three vowels with very short durations (<25 ms) were excluded because formants are difficult to estimate in very short tokens. The exclusions yielded a sample of 2,217 vowels.

Second, mean F1 and F2 values were computed for each of the five Spanish vowel categories. Instances greater than 1.5 standard deviations from the mean were checked by visual inspection of the waveform and spectrogram, as were a randomly selected 10% of tokens. Formants were corrected if (for example) setting a different number of formants for Praat to locate provided clearer formant tracks with more reasonable values (a frequent error was for the first-pass algorithm to label one formant as another one); otherwise the measurements were left as is. During this process, 27 vowel tokens were rejected because hand-inspection revealed no clear, reliable formant values, usually because the voice signal was quiet relative to background noise. The total sample was thus 2,190 vowels, distributed as follows: /a/, 797; /e/, 572; /i/, 292; /o/, 471; /u/, 48.

Fig. 1 shows the dataset, with each point representing the first and second formants for each token. Substantial overlap among categories is visible.

To what extent does collapsing token measurements onto word-type averages reduce variability? Fig. 2 shows the dataset again, this time with each token represented not as a point, but as the endpoint of a line segment that leads from that token's values to the mean values of the token's "host" word, with words appearing five times or more (and their tokens) displayed. The plot reveals visibly distinct clusters for [i] and [e] and perhaps [a], though there is also substantial spread in the types for [o]. This suggests that clustering over words could be more successful than clustering over tokens, for at least some vowels.

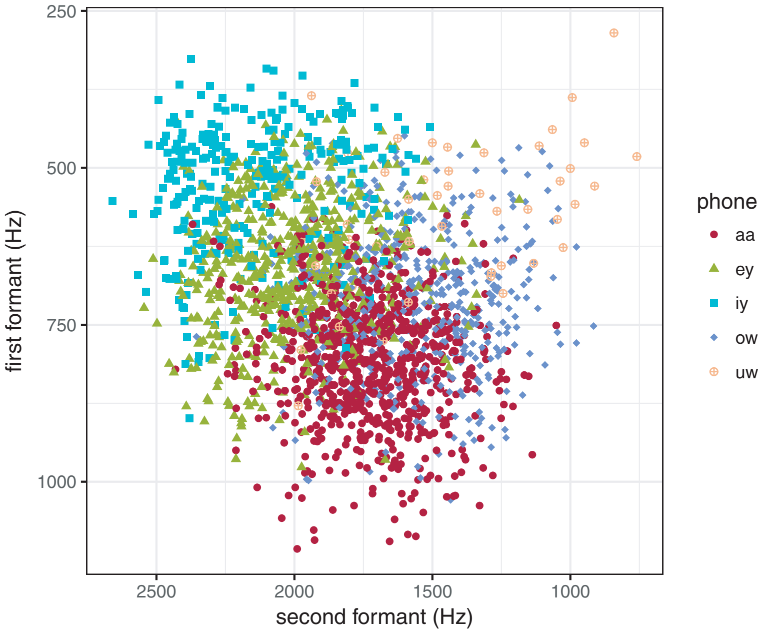

Fig. 1. First and second formants of 2,200 Spanish infant-directed vowels.

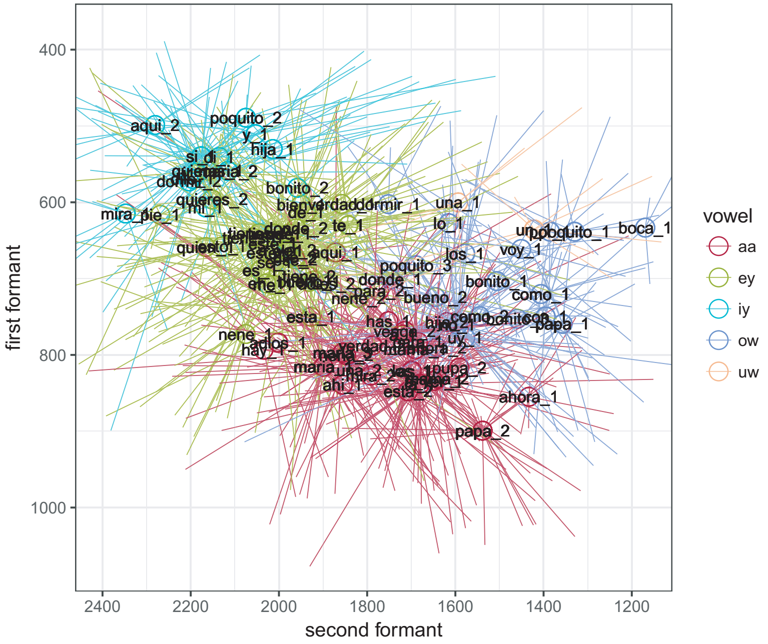

Fig. 2. First and second formants of vowel tokens appearing in words with a corpus frequency of five or more, each connected by a line to the mean formant values of the word type (indicated by the orthographic spelling and the number of the relevant syllable).

3. Results

3.1. Analysis strategy

There are several ways one might evaluate an unsupervised category learner. Ideally, the model (or the child) would learn the number of categories the language distinguishes and each category's normal range in phonetic space, and could therefore identify instances of each sound by choosing the same label a native speaker would. However, given the variability present in our speech sample (and natural conversational samples in general), we were not optimistic about achieving this ideal level of performance in any of our models, and indeed, we did not achieve it. For the purpose of evaluating the utility of words for identifying categories, then, our strategy was as follows.

First, we applied statistical clustering methods to our sample of vowel tokens, comparing the categories proposed by the analysis to their transcribed category labels. Next, we used the same methods clustering not over tokens, but over average formant values for each of the frequent words in the sample. Thus, for example, if the word "que" occurred 40 times, rather than including the 40 tokens' first and second formant values as input to the clusterer, we input the mean of these first and second formants as an estimate of the infant's representation of that word. Words were weighted by frequency, so that both the tokens analysis and the types analysis included “que” the same number of times. If wordforms could be helpful in guiding infants to the language’s vowel categories, the clusters found over types should match the transcribed category more often than clusters found over tokens do.

There are many clustering algorithms. Nothing is known about which might match infants’ mental processes the best. We did not formally explore the space of possible algorithms, but informal testing resulted in similar improvements in performance from token-based to type-based analyses among various algorithms. The results we present here come from hierarchical cluster analysis using the average-linkage method (UPGMA) implemented in R (R Core Team, 2015).

In our first analysis, we test the lexical hypothesis using words defined as they were transcribed in the corpus. It is not realistic to suppose that infants can always identify spoken words as accurately as the corpus’ transcribers, but this analysis provides an in-principle upper bound on the utility of words. In a second analysis, we test the hypothesis that words still help with vowel identification under the assumption that infants can identify consonants correctly but initially treat all vowels as the same. In a third analysis, we relax the assumption that infants can correctly identify consonants.

3.2. Analysis 1: Orthographic words

Clustering over frequent words requires setting a criterion for “frequent.” Entering all words that occur once or more uses the greatest amount of the sample, but unique words do not test the hypothesis (because each type is a token); on the other hand, entering only words with very high frequencies results in a dataset that is too small. We (more or less arbitrarily) chose five as the minimum word frequency to be included in the types analysis, but we present summary results for other frequencies as well. Another parameter that must be set is the number of categories to find, that is, where to section the cluster analysis’ dendrogram. Hierarchical cluster analysis yields a hierarchy, not one flat level of partitions, so to describe its results it is necessary to cut the tree at a certain level. The infant, of course, does not know in advance how many vowels there are. Here we display in-depth analyses for a partitioning at six splits, which gives a good feel for the model’s structure, but we also show summary results for a range of other splits. To keep the types and tokens analyses on fair footing, the tokens analyses only included tokens from the words that were frequent enough to be included in the types analysis. Analyses extending from these to the full dataset are provided in the supporting online materials.

An ideal classifier would assign all instances of a given vowel into one category and would discover categories that each only contain one vowel. To visualize the degree to which partitioning by types and by tokens met these goals, Fig. 3 presents a grid with each vowel as a column and each split (category derived from the analysis) as a row. Circles’ areas indicate the number of vowels in each cell of the vowel:split matrix; counts are printed when nonzero. The analysis by types is on the left, by tokens on the right. For example, looking at the lower left of the left panel, we see that 325 [a] vowels were assigned to split 1. The circle is shaded fairly dark because 325 of (325 + 10 + 105) represents a strong majority of the vowels in split 1. Perfect classification would be shown by all circles falling on the diagonal, and a 1:1 assignment of vowels to categories. It is clear from inspection of the figure that the analysis by word types approaches this ideal much more closely than the analysis by word tokens. The tokens analysis, for example, placed 167 tokens into its fifth category (second row from top), including a diverse range of vowel types which were also placed frequently in other splits.

To see more directly which words were clustered together and to characterize the errors, Fig. 4 shows the individual words in formant space, and the categories to which they were assigned (left panel) and the tokens-only categorizations (right panel). In the types-based analysis, a few outliers (the first syllables of *boca* and *ahora*, and the second syllable of *papa*) were assigned to two outlier categories, and the remainder were placed fairly accurately into four groups, an exception being [u], a vowel appearing in only two frequent words. The greatest source of error was a number of [o] words (e.g., *donde, bueno, como*) that infiltrated the [a] category. The tokens-based analysis showed relatively little alignment with the true vowel categories. The tokens analysis was also relatively unstable, because it drew boundaries between categories that, statistically speaking, hardly exist. This instability is made evident by re-performing the clustering analyses over random subsamples of the dataset; even different random samples of 99.5% of the data result in token-based categorizations with widely varying morphologies (see Supporting Online Materials).

The preceding types-based analyses required fixing two parameters: the minimum word frequency threshold for including a word type in the analysis and the number of splits that the clustered similarity space was divided into. The superiority of type-based

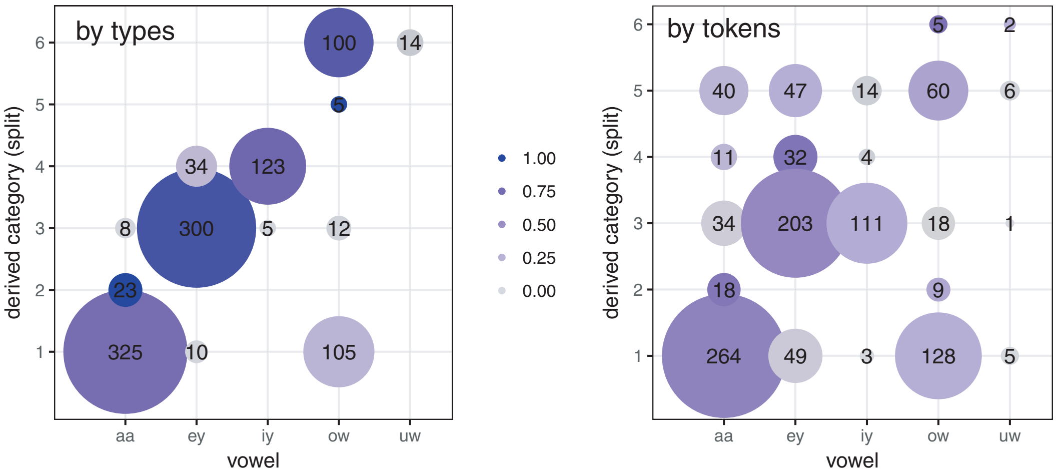

Fig. 3. Assignment of vowels to categories based on orthographic words (left panel) and based on tokens (right panel). Circle area represents the number of instances of a given Spanish vowel that were assigned into a given category; this number is also printed in the circle. Circle darkness indicates how “pure” the category split is, where a derived category containing instances of only one Spanish vowel would be the darkest.

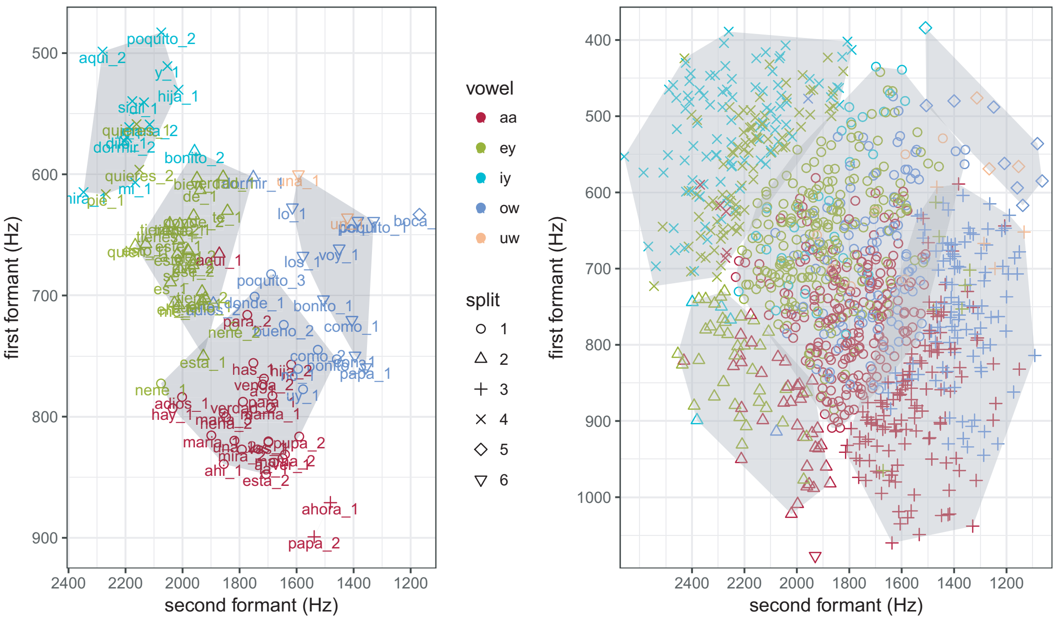

Fig. 4. Assignment of vowels to categories based on orthographic words (left panel) and based on tokens (right panel). Gray shading indicates the boundaries of the derived categories. Colored points indicate the data entering the categorization, color-coded by true (transcribed) vowel. Point shape indicates the derived category.

clustering does not depend on the particular values selected. This is shown in Fig. 5, which summarizes results of analyses varying the word frequency threshold from 4 to 8 and the number of categories from 2 to 14. The left panel shows how homogenous each split is for a given analysis, with homogeneity defined as the proportion of instances in a given split that correspond to the most common Spanish vowel that was assigned to that split (e.g., if a category has mostly [a] in it, what is its proportion of [a]s?). This figure is high if each discovered category tends to have only one kind of vowel in it. The right panel shows, on average, the proportion of vowel types (such as [a]) that were placed into a single category. This statistic is high if each Spanish vowel tends to go into a category dedicated to that vowel, as opposed to being distributed over several categories.

In sum, although unsupervised clustering over orthographically defined words did not solve the problem of discovering the Spanish vowel categories, it appears that the failure of Spanish vowel tokens to present as clear statistical modes in formant space is ameliorated significantly by first aggregating formant values into means word by word, and clustering over these lexcial prototypes.

However, it is optimistic to imagine that infants can identify words with the precision implied by the orthographic labeling of our dataset. One obvious problem is that if infants do not know how vowels differentiate words, the only way they could know that (e.g.) *boot* and *boat* are different words to cluster over would be to use their meanings (and to

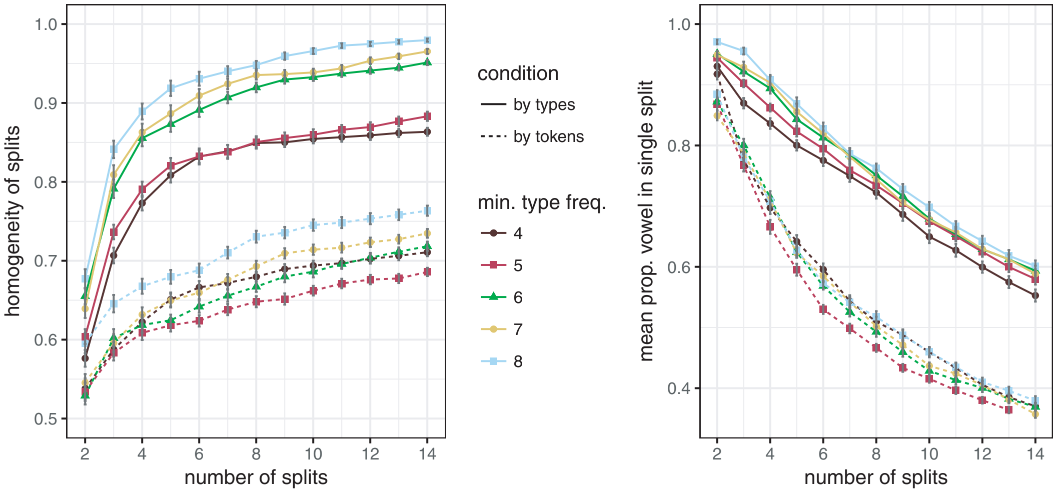

Fig. 5. Robustness of the type-analysis advantage over the tokens analysis, for orthographically defined lexical types. Curves represent the predominance of the majority vowel in each derived category (left panel) and the proportion of each vowel assigned to a single category (right panel), while varying how common a word needed to be to enter the analysis (plot colors) and how many derived categories were stipulated (x axis). Dotted lines (the lower set in each plot) show the tokens analysis; full lines show the types analysis. Error bars show standard errors over 50 model runs each including a random 90% of the corpus.

have some resistance to homophony). Infants likely know at least some words during the time that they learn their language's vowel categories (e.g., Bergelson & Swingley, 2012, 2015; Tincoff & Jusczyk, 1999, 2012), but we should not presuppose that they know most or all of them. Thus, next we present an analysis that backs off on this assumption, identifying words using the transcribers' phonetic labels, and reducing every vowel to a generic symbol covering all vowels. Thus, vowel-based minimal pairs like *hay* (there is/are) and *hoy* (today) are collapsed into a single "word," thereby becoming a data point for the clusterer. Intuitively, the more vowel-differentiated minimal pairs are present in the corpus, the greater the potential for this analysis to be catastrophically flooded with instances which, being conflations of two different vowels, fail to correspond to an actual Spanish vowel.

3.3. Analysis 2: Neutralized-vowel lexical types

Phonetic transcriptions resulting from hand annotation of the input corpus were modified by replacing each vowel specification with a generic character representing any vowel. These phonologically defined lexical types were used as an input to clustering models that were otherwise identical to those examined in Analysis 1.

Assuming a minimum occurrence frequency of five, there were 82 distinct vowel environments over which the clustering algorithms of the types analysis operated, amounting to 1,105 tokens (thus, the vowel dataset was slightly different from the dataset of the prior analysis). As Fig. 6 shows, the derived categories lined up much more closely with Spanish vowel categories in the type-based analysis (left panel) than the token-based analysis (right panel).

The main cost to the change from orthographic words to vowel-neutralized, transcribed words was in the greater confusion of [a] and [o] instances. This is easily explained: the Spanish gender distinction is marked on many words that come in pairs and that differ only in including a word-final *a* or *o*, such as *esta* (this [f.]) and *esto* (this [m.]), or *hija* (daughter) and *hijo* (son). Averaging formant values over these pairs produces some data points that are in between the true [a] and [o] categories. This can be seen in the large category whose first formant ranges from about 650 to 850 in the left panel of Fig. 7.

Although the conflation of vowels (and therefore words) rendered category-finding less accurate, there was still a marked advantage to clustering over types rather than tokens, and this held over all parameter settings evaluated, as shown in Fig. 8.

3.4. Analysis 3: Ambiguous consonants

The utility of lexical contexts for grouping vowels into categories depends partly on the accuracy with which the lexical contexts may be identified. Analysis 2 showed that even if vowels are rendered completely ambiguous as contexts, their consonantally defined lexical environments still yield superior categorization performance. But it is unlikely that infants can identify words with complete fidelity. To model the noise that this uncertainty might introduce, one option would be to add random variability to the

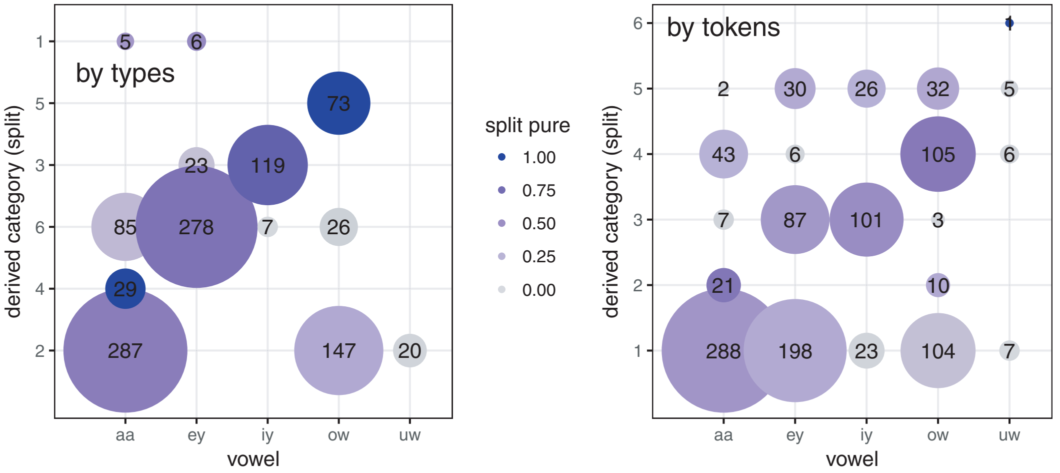

Fig. 6. Assignment of vowels to categories based on phonologically transcribed words (with the vowel neutralized; see text) and based on tokens. Circle area represents the number of instances of a given Spanish vowel that were assigned into a given category; this number is also printed in the circle. Circle darkness indicates how "pure" the category split is, where a derived category containing instances of only one Spanish vowel would be the darkest. The analysis by words is in the left panel; by tokens, the right.

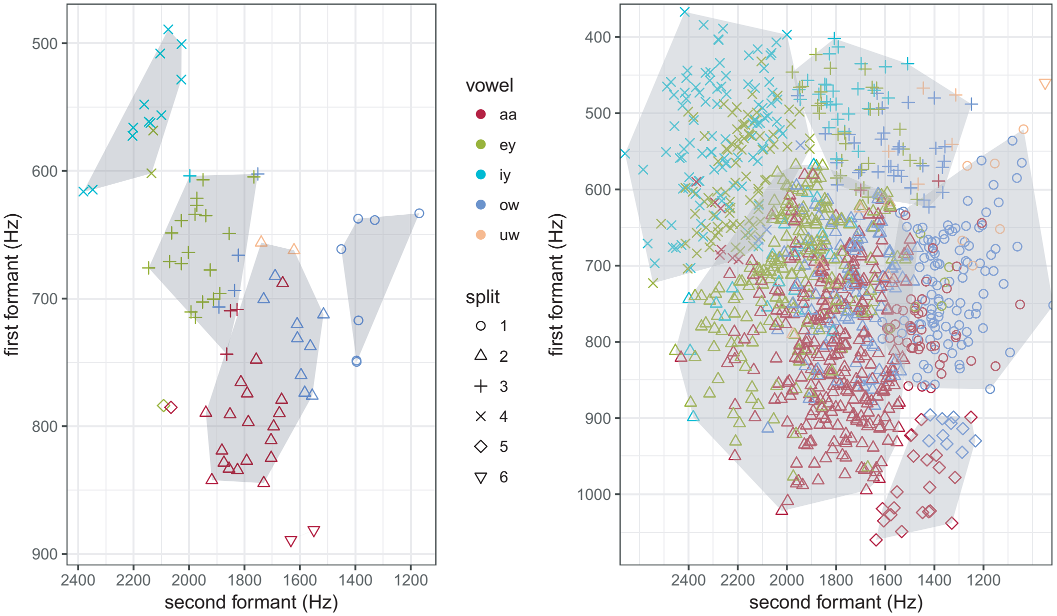

Fig. 7. Assignment of vowels to categories based on transcribed, vowel-neutralized words (left panel) and based on tokens (right panel). Gray shading indicates the boundaries of the derived categories. Colored points indicate the data entering the categorization, color-coded by true (transcribed) vowel. Point shape indicates the derived category.

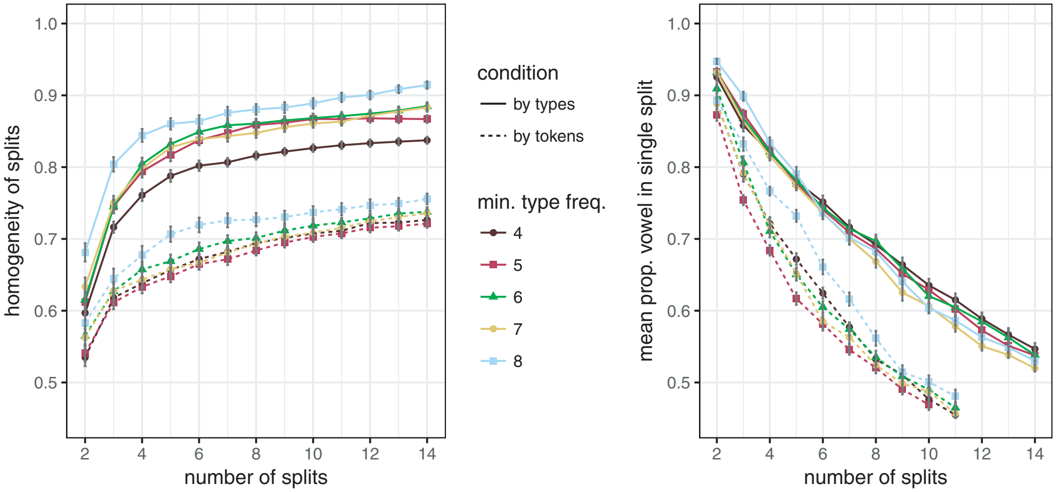

Fig. 8. Robustness of the type-analysis advantage over the tokens analysis, for transcribed, vowel-neutralized lexical types. Plotting conventions are as in Fig. 5.

consonants and examine the resulting decrement in performance. However, random variability tends to create many low-frequency events that can be filtered out using occurrence-frequency criteria, and might ultimately simulate little more than a reduction in effective corpus size. For this reason, we instead explored collapsing phonological features, creating systematic forms of ambiguity. To this end, we created two new versions of the corpus: one in which all consonantal *voicing* distinctions were collapsed, as were all vowel distinctions, and another in which all consonantal *place of articulation* distinctions were collapsed, as were the vowels. These manipulations were not meant as serious proposals that infants fail to encode these features in words; rather, they present "worst-case scenarios," or at least "quite-bad-case scenarios" of low-fidelity lexical representation. Thus, in the first, [t] and [d] were combined into a single category, and also [ð] and [θ], [k] and [g], and so on. Words like *beso* (kiss) and *pasa* (happen), for example, were therefore taken as identical lexical contexts in the no-voicing analysis. The no-place analysis collapsed [b,d,g]; [m,n], [w,y], and so on, so that (for example) *donde* (where) and *venga* (come) were identical contexts. Otherwise, the analyses were the conducted in the same way as prior analyses. Type frequencies were computed over these new "collapsed" stand-ins for word forms.

Once again, the categories derived by each of these analyses lined up more closely with the true Spanish vowels in the analysis over types than the analysis over tokens. Fig. 9 presents the overall alignment of derived splits and true vowels for each of the lexical-types analyses. Accuracy did suffer, although not catastrophically, relative to the prior analyses.

As was the case for the previous analyses, the superiority of the types-based clustering over token-based clustering was clear regardless of the number of categories specified or the frequency criterion stipulated for entering types into the analysis. This is shown in Fig. 10.

As is apparent from the figures, the assumption that infants might fail to use consonant voicing or place of articulation in differentiating words for category labeling had fairly

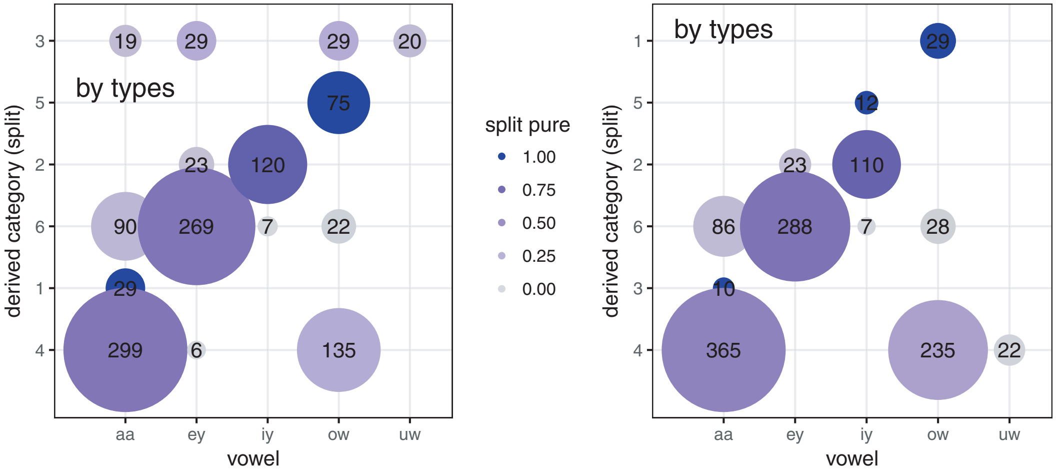

Fig. 9. Assignment of vowels to categories based on phonologically transcribed words, with all vowels neutralized and with consonant voicing distinctions ignored (left panel) or consonant place of articulation ignored (right panel). Plotting conventions are otherwise as in Fig. 3. By-types analyses only.

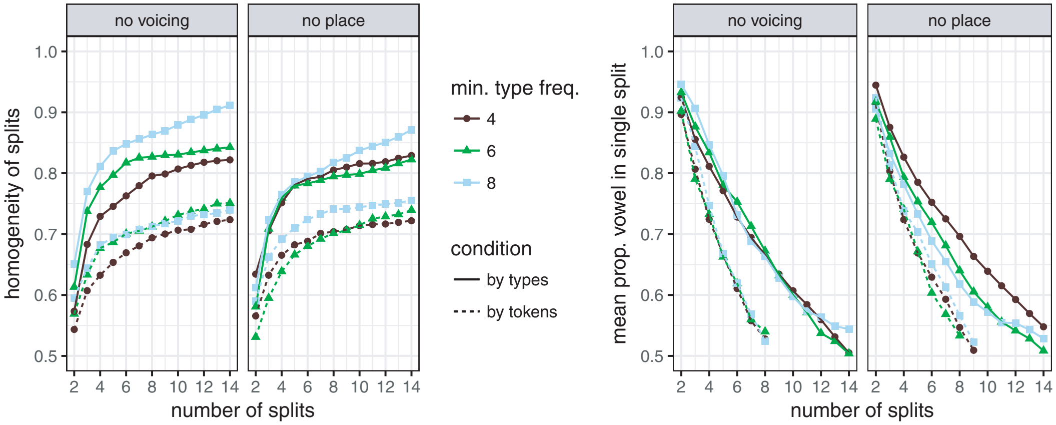

Fig. 10. Robustness of the type-analysis advantage over the tokens analysis, for transcribed, vowel-neutralized lexical types. Plotting conventions are as in Fig. 5, with the left panel indicating how homogenous the derived categories were, and the right indicating how divided the true vowels were. For clarity, fewer frequency criteria are shown.

little impact on the types analysis. An implication of this outcome is that Spanish child-directed speech contains a sufficient proportion of phonologically distinct words to render ignorance of fine details of consonantal phonetics only a minor barrier to the potential utility of word contexts for vowel category clustering.

4. General discussion

Experiments probing infants' categorization of vowels appear to show that over the course of the first year, infants come to interpret vowels sounds in a manner consistent with the distinctions made by their language or languages. These results are usually interpreted as indicating that infants learn phonetic categories that correspond to statistical modes in the distribution of speech sounds in their environment. However, several attempts to measure the existence of these modes, including the present effort, have failed to find clear separations between categories (e.g., Cristia & Seidl, 2014). Infants appear to be better at finding them than we are.

We have suggested that part of the solution is to supplement the phonetic data with information from the developing lexicon, an approach that has been explored previously in other ways (e.g., Feldman et al., 2013). We know that infants in the relevant age range do learn the sound-forms of words (e.g., Jusczyk & Hohne, 1997), and there is evidence that infants in constrained experimental contexts can use word-form contexts to guide category differentiation (Feldman et al., 2013). The contribution of the present work is to show that a simple clustering scheme with very little adaptive fine-tuning is substantially aided in finding Spanish vowel categories by clustering over words rather than over every experienced token. In this Spanish sample, phonetic contexts do not obscure the similarity shared by a given vowel as it appears in different words, nor does the existence of minimal pairs in the language interfere enough to prevent words from helping.

Phonetic context effects do make the vowel category learning problem harder, as shown in the analysis in which vowels of a given category were randomly assigned to words, and then averaged over these “words” (Supporting Materials, Analysis 3). In this simulation, phonetic context effects (and indeed any other acoustic consequence of a vowel’s being in a particular word) were averaged away, and categorization over these (theoretical) words was nearly perfect. This suggests that vowel category learning would be easier if infants could model the sources of this variation. For example, if infants were capable of attributing nasalization in a vowel to its nasal context, they might submit to their vowel-categorization mechanism a derived, hypothetical token with the nasalization abstracted away, and likewise for any other feature that might be consistently affected by context. There is not a good basis now for estimating infants’ abilities in this regard (see Seidl & Cristia, 2012, for relevant discussion).

Our study has some important limitations. First, we have not provided a model of infant category learning that succeeds in identifying the number and distribution of the Spanish vowel categories. Clustering algorithms generally, and ours as well, are better considered as laying out a similarity space than as yielding a discrete set of categories. Such clustering models have a distinguished history in computational studies of language acquisition (e.g., Redington, Chater, & Finch, 1998) but from a descriptive perspective might not have the right outputs. That said, it is not clear that the infant perception experiments force the conclusion that infants have discrete phonetic categories rather than language-specific similarity spaces. Along similar lines, Dillon, Dunbar, and Idsardi (2013) have pointed out that acoustically defined phonetic categories may not be the appropriate targets of acquisition models anyway, because children ultimately need phonological categories that are defined on linguistic terms and not purely phonetic ones. These distinctions merit further exploration.

A second limitation of the present study is that our characterization of the vowels themselves is limited, in our use of first and second formants to specify them. Although Spanish vowels are generally monophthongal and do not contrast importantly in duration, there are undoubtedly other speech features that play into vowel representation in Spanish speakers. Indeed, it is possible that formants are not merely insufficient, but simply inaccurate as proxies for children’s phonetic representations. There are other representational methods (such as the narrow spectral slices more common to automatic speech recognition systems), and whether using other representational methods makes the category-discovery process easier or harder remains to be seen. Any model that is not an infant auditory system will be inaccurate to some unspecified degree, and we cannot know a priori whether these inaccuracies interact with the utility of lexical contexts (or the need for them). This being said, at present, there is no evidence contradicting our claim that the phonetic distributions of vowels in conversational child-directed speech are insufficient for vowel learning.

A third limitation is that we have only tested one simple model of word learning, namely a frequency-criterion model in which common word-forms are assumed to be represented by the infant. Candidate phonological sequences were drawn from the corpus' true word-forms, correctly segmented from their utterances. Although these sequences were then degraded and collapsed in various ways in Analyses 2 and 3, the starting points were words. In current work, we are testing alternative models of word-form learning in which infants are not assumed to make correct segmentation estimates.

It also bears noting that our corpus of about 2,000 vowels is very small compared to the total learning experience of a 6- or 8-month-old infant, who may have experienced more than a thousand times more speech than we measured. It does not seem particularly likely that simply increasing the size of the dataset would lead vowel categories to emerge or would expose a preponderance of frequent words whose typical vowel realizations cloud the language's canonical categories. But these are possibilities we cannot exclude. Hand-labeling orders of magnitude more phonetic data than we have done here might not be practicable, but converging evidence from automated measurement might reinforce our conclusions here, and it could extend beyond just one talker (given that we have no information about how typical this particular talker was).

If infants use words to find phonetic categories, what does this imply about the course of language acquisition? The most significant practical implication is that to the extent that individual differences in infants' phonetic categorization are environmentally determined, they are attributable not only to the phonetic properties of parental speech, but also the myriad sources of parental influence on vocabulary acquisition. To take one example, Cristia (2011) found that infants whose mothers produced more distinct [s] sounds were better at perceptually differentiating [s] and [ʃ] than infants whose mothers produced less distinct [s]s. If phonetic category learning is primarily a matter of perceptual learning over acoustic distributions, one would expect that the infants who are better at it would be those whose interlocutors speak to them more distinctly. That is, an important source of phonetic learning performance would be the clarity of the acoustic data provided to the infant by his or her parents. But if phonetic category learning is also dependent upon the infant's acquisition of word-forms, a range of other features of parent–child interaction become relevant as well. These include some that seem to support extraction of word-forms, such as using short or one-word utterances (e.g., Johnson, Seidl, & Tyler, 2014; Lew-Williams, Pelucchi, & Saffran, 2011) or employing exaggerated "infant-directed" prosody (e.g., Singh, Nestor, Parikh, & Yull, 2009), as well as other features that facilitate word learning more generally, such as repeated presentation of words over short intervals (e.g., Horst, 2013; Swingley & Humphrey, 2017), overall use of a varied vocabulary (Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010), follow-in labeling and contingent, referential language (e.g., Tamis-LeMonda, Kuchirko, & Song, 2014), or provision of clear contextual support (e.g., Cartmill et al., 2013). At this point, we cannot say whether word-forms are more effective in guiding phonetic category formation if children know what the words mean, except in the sense that semantic knowledge could render minimal pairs unambiguous and prevent the misleading averaging of vowels from distinct categories.

The idea that word learning and phonological learning are mutually supportive suggests a refinement of the typical interpretation of correlations between speech perception and vocabulary development. To take one important example, Kuhl, Conboy, Padden, Nelson, and Pruitt (2005) found that infants' performance in distinguishing native-language consonants was positively correlated with later vocabulary size, whereas performance on nonnative consonant contrasts was negatively correlated with later vocabulary. A natural explanation of this relationship is that skilled phonetic categorization makes word learning more successful. But if vocabulary development itself contributes to the learning of native phonetic categories, such correlations may reflect the opposite causal path, at least in part (Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014).

It might seem simpler for infants to "solve" the problems posed by their native language by cleaving the language into linguistic domains and attacking each in sequence: first sounds, then word-forms, then word meanings, then syntax. More and more, though, it seems that children take their first tentative steps in every domain at once. It may be that this is not only descriptively true, but also necessary.

Acknowledgments

This work was supported by NIH grant R01-HD049681 to D. Swingley. Portions of the research reported here were presented at the Workshop on Infant Speech Perception (Bilbao) in 2017. Some of the initial annotations were done by Madeline Schnur and Juan Cabrera when they were undergraduates at Penn. We thank Susana López Ornat for making her corpus available via the CHILDES database.

References

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *Journal of the Acoustical Society of America*, *141*, 3070–3078. https://doi.org/10.1121/1.4982246

Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the USA*, *109*, 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, *11*(4), 369–380. https://doi.org/10.1080/15475441.2014.979387

Best, C. T., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development*, *21*, 295.

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

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44. https://doi.org/10.1016/s0010-0277(01)00122-6

Burnham, E. B., Wieland, E. A., Kondaurova, M. V., McAuley, J. D., Bergeson, T. R., & Dilley, L. C. (2015). Phonetic modification of vowel space in storybook speech to infants up to 2 years of age. *Journal of Speech, Language, and Hearing Research, 58*(2), 241–253. https://doi.org/10.1044/2015_JSLHR-S-13-0205

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*, 11278–11283. https://doi.org/10.1073/pnas.1309518110

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/category. *The Journal of the Acoustical Society of America*, *129*, 3271–3280.

Cristia, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, *39*(3), 388–402.

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*, 913–934. https://doi.org/10.1017/S0305000912000669

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345. https://doi.org/10.1111/cdev.12193

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences of the USA*, *104*, 16027–16031. https://doi.org/10.1073/pnas.0705270104

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, *37*(2), 344–377. https://doi.org/10.1111/cogs.12008

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science*, *209*, 1140–1141.

Englund, K. T., & Behne, D. M. (2005). Infant directed speech in natural interaction: Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research*, *34*(3), 259–280. https://doi.org/10.1007/s10936-005-3640-7

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*, 751–778. https://doi.org/10.1037/a0034245

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. J. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *Journal of the Acoustical Society of America*, *124*, 1234–1251.

Garrett, S., Morton, T., & McLemore, C. (1997). *Callhome Spanish lexicon LDC96L16*. Philadelphia: Linguistic Data Consortium.

Goudbeek, M., Smits, R., & Swingley, D. (2009). Supervised and unsupervised learning of multidimensional auditory categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1913–1933.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*, 1111–1121. https://doi.org/10.1121/1.416296

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Horst, J. S. (2013). Context and repetition in word learning. *Frontiers in Psychology*, *4*, 149. https://doi.org/10.3389/fpsyg.2013.00149

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*, 343–365.

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS One*, *9*(1), e83546.

Jones, C., Meakins, F., & Muawiyath, S. (2012). Learning vowel categories from maternal speech in Gurindji Kriol. *Language Learning*, *62*(4), 1052–1078. https://doi.org/10.1111/j.1467-9922.2012.00725.x

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, *277*, 1984–1986.

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: Smiling, teaching or social convergence? *Royal Society Open Science*, *4*(8), 170306. https://doi.org/10.1098/rsos.170306

Kuhl, P. K. (1992). Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 293–332). Washington, DC: American Psychological Association. https://doi.org/10.1037/10119-012

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T.(2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London, B*, *363*, 979–1000.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the "critical period." *Language Learning and Development*, *1*, 237–264.

Labov, W., Ash, S., & Boberg, C. (2005). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. Berlin: Mouton.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the XIIIth international congress of phonetic sciences* (Vol. *2*, pp. 140–147). Stockholm, Sweden.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329. https://doi.org/10.1111/j.1467-7687.2011.01079.x

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Hillsdale, NJ: Erlbaum.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103–124. https://doi.org/10.1111/j.1551-6709.2012.01267.x

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347. https://doi.org/10.1177/0956797614562453

Maye, J., Gerken, L. A., & Werker, J. F. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129* (2), 362–378. https://doi.org/10.1016/j.cognition.2013.07.015

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84–93. https://doi.org/10.1016/j.cognition.2017.05.003

Narayan, C. (2013). Developmental perspectives on phonological typology and sound change. In A. C. L. Yu (Ed.), *Origins of sound change: Approaches to phonologization* (pp. 128–146). Oxford: Oxford University Press.

Narayan, C. N., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, *13*, 407–420. https://doi.org/10.1111/j.1467-7687.2009.00898.x

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Ornat, S. L. (1994). *La adquisicion de la lengua Española*. Madrid: Siglo XXI.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 421–435.

R Core Team. (2015). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria. Retrieved from http://www.R-project.org/ (R version 3.2.0)

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.

Sebastian-Galles, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: Is the distributional account all there is to it? *Developmental Science*, *12*(6), 874–887. https://doi.org/10.1111/j.1467-7687.2009.00829.x

Seidl, A., & Cristia, A. (2012). InfantsâĂŹ learning of phonological status. *Frontiers in Psychology*, *3*, 448. https://doi.org/10.3389/fpsyg.2012.00448

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654–666. https://doi.org/10.1080/15250000903263973

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: infant age and context related changes. *Journal of Child Language*, *10*, 1–15.

Sundberg, U., & Lacerda, F. (1999). Voice onset time in speech to infants and adults. *Phonetica*, *56*(3-4), 186–199.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132.

Swingley, D. (2006, June). *Data-driven phonological distinction without phonetic opposition*. Paper presented at the XVth Biennial Conference on Infant Studies. Kyoto.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*, 3617–3622. Retrieved from https://doi.org/10.1098/rstb.2009–0107https://doi.org/10.1098/rstb.2009.0107

Swingley, D., & Humphrey, C. (2017). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development.* https://doi.org/10.1111/cdev.12731

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*, 16–34.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*, 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*, 432–444. https://doi.org/10.1111/j.1532-7078.2011.00084.x

Tomasello, M. (2001). Could we please lose the mapping metaphor, please? *Behavioral and Brain Sciences, 24*. 1119–1120.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the USA*, *104*, 16027–16031.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual Review of Psychology*, *66*, 173–196.

Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language*, *68*(2), 123–139. https://doi.org/10.1016/j.jml.2012.09.004

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, *15*(4), 420–433. https://doi.org/10.1111/j.1532-7078.2009.00024.x

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D.., &Woodland, P. C. (2006). *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.

Yuan, J., & Liberman, M. Y. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*, 3878. https://doi.org/10.1121/1.2935783

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Additional analyses.