Abstract

Over the first year, infants begin to learn the words of their language. Previous work suggests that certain statistical regularities in speech could help infants segment the speech stream into words, thereby forming a proto-lexicon that could support learning of the eventual vocabulary. However, computational models of word segmentation have typically been tested using language input that is much less variable than actual speech is. We show that using actual, transcribed pronunciations rather than dictionary pronunciations of the same speech leads to worse segmentation performance across models. We also find that phonologically variable input poses serious problems for lexicon building, because even correctly segmented word forms exhibit a complex, many-to-many relationship with speakers' intended words. Many phonologically distinct word forms were actually the same intended word, and many identical transcriptions came from different intended words. The fact that previous models appear to have substantially overestimated the utility of simple statistical heuristics suggests a need to consider the formation of the lexicon in infancy differently.

Keywords: language acquisition, computational modeling, word segmentation, phonological variation

1. Introduction

Although infants are born knowing little about their native language, they quickly learn a great deal from the speech they hear. Within months, they become familiar with their native language's sound categories (Werker & Tees, 1984), as well as the relative frequency of different sequences of speech sounds (Archer et al., 2021; Jusczyk et al., 1993). Beyond learning about their language's phonology, infants also begin to learn words. Months before their first birthday, they recognize the meanings of some common words, including both concrete nouns (Bergelson & Swingley, 2012) and a little later, more abstract words (Bergelson & Swingley, 2013), and by the second half of the first year, they recognize the spoken form of a variety of words familiar from home experience or laboratory exposure (e.g., Hallé & Boysson-Bardies, 1994; Jusczyk & Aslin, 1995; Jusczyk & Hohne, 1997; Schreiner et al., 2016; Swingley, 2005a; Vihman et al., 2004).

An important step in the process of language learning is word segmentation, or pulling out words from the continuous stream of speech. It is easy to understand that this is a difficult problem—one only needs to listen to a parent speaking to an infant in an unfamiliar language to recognize that it is quite hard to infer where one word ends and the next begins. This problem is difficult for infants too, which is why infants learn words more easily when they are presented in one-word utterances than when they are embedded in longer utterances (Brent & Siskind, 2001; Keren-Portnoy et al., 2019; Swingley & Humphrey, 2018). Yet infants do manage to break utterances into parts. Laboratory studies demonstrate that infants can extract words from their phonetic contexts (e.g., Jusczyk and Aslin, 1995), and infants have some knowledge of grammatical words that never appear in isolation (e.g., Shi & Lepage, 2008).

Research into infants' early discovery of words has taken two forms: experiments that present continuous speech to infants and test which elements they retain, and computational models that evaluate what infants might learn were they to parse and retain speech sequences according to a particular set of computable heuristics. The present paper continues the latter line, but differs from most prior work in examining the consequences of normal phonological variability. When words are realized in more than one way, does the phonological structure of the lexicon still permit simple probabilistic heuristics to succeed in producing the foundation of the early vocabulary?

In principle, there are several cues that could be helpful in word segmentation, once the infant has some familiarity with phonological regularities present in the lexicon. For example, in English, strong syllables tend to coincide with word onsets, suggesting that English speakers could learn to use stress patterns or vowel-reduction patterns to detect where an unknown word begins (Cutler & Norris, 1988). Experiments have shown that infants do respond to such prosodic cues (Jusczyk et al., 1999; Nishibayashi et al., 2015; Seidl, 2007; Seidl & Johnson, 2006; Sundara & Mateu, 2018). Some consonantal sequences are much more common at word boundaries in English than within words, and infants respond to these phonotactic probabilities too (Mattys et al., 1999; Mattys & Jusczyk, 2001). These studies suggest that infants use the phonetic characteristics of a preliminary stock of words to form generalizations that they then apply in interpreting novel speech sequences.

Much of the laboratory research on word segmentation has focused on investigating how the initial stock of words is identified by infants, and what generalizations might follow. In principle, tabulating frequencies of occurrence, and relative frequencies of adjacent units, could be informative about word boundaries. Sequences of units (such as phones or syllables) *within*

words are expected to co-occur more often than sequences occurring *across* word boundaries (Harris, 1955). If infants could track this kind of information, they might be able to pull out candidate word forms from the speech stream (e.g., Aslin et al., 1998; Saffran et al., 1996).

Many laboratory experiments have confirmed that infants are sensitive to the statistical cohesiveness of sub-word units. Most of these studies involve familiarizing infants with an unknown, usually artificial, language, whose words are defined as the consistent phonological strings that were concatenated to create the listening sequence. Differences in infants' subsequent listening times to isolated words and nonwords show that infants must have computed, in some form, the probability differences among phone or syllable transitions between words and nonwords. The original findings by Saffran et al. (1996) have since been extended to other language learning populations (e.g., Mersad & Nazzi, 2012) and to infants as young as 5 months (E. K. Johnson & Tyler, 2010). In general, these experimental designs are well equipped to demonstrate which information sources are theoretically accessible to infants, and to reveal whether infants' strategies appear to have been shaped by the characteristics of the language they are learning.

However, laboratory experiments are not well equipped to show whether a particular cue is sufficient to support language acquisition given its actual availability in the language environment. For example, even if co-occurrence frequencies are sufficient to segment the small artificial languages that are typically used in experiments (though see Pelucchi et al., 2009), it is not necessarily the case that they can be used to successfully segment spontaneous natural language (e.g., Gambell & Yang, 2005; Swingley, 2005b; Yang, 2004). As a result, as a complement to experimental work, computational models can be deployed over language

corpora to test whether a proposed cognitive ability would be sufficient to account for the documented behavioral accomplishments of infants (e.g., Ludusan et al., 2022).

Computational models of word segmentation illustrate which word forms could be learned given different assumptions about the algorithm at work and the language input that the learner receives. Broadly speaking, a model is provided with a textual representation of speech without word boundaries, and returns as output the same text with the hypothesized word boundaries inserted, for example, in places where the transitional probability or conditional probability of two units (phones or syllables) occurring next to each other is relatively low (e.g., Saksida et al., 2017). This segmentation output can then be compared to the actual (gold-standard) words, to assess the model's performance. To make these comparisons, previous modeling work has typically focused on information retrieval metrics that compare the number of correct and incorrect segmentations, or, less commonly, on how psychologically plausible the errors seem (e.g., Daland & Pierrehumbert, 2011; Lignos, 2011). Relatively few studies have examined in detail what sort of language-learning foothold the output of a segmentation procedure would grant the infant.

Regardless of the particular model in question, most previous studies have made similar assumptions about the nature of the input to word segmentation. Specifically, the input to the model has typically been generated by taking an orthographic transcription of speech and replacing each word with its pronunciation according to a dictionary (e.g., Brent & Cartwright, 1996; see, for example, Börschinger et al., 2013; Elsner et al., 2013 for exceptions). This procedure tacitly assumes that any given orthographic word is always pronounced in the same canonical way. In real speech, however, this is not the case. Whole phones and even syllables can be dropped or added, or changed to incorporate features of nearby sounds (e.g., Ernestus &

Baayen, 2011; K. Johnson, 2004). It is well known that speech can vary in "non-contrastive" acoustic and phonetic dimensions (like pitch, amplitude, and creakiness, in English) where differences are usually not relevant to word identity. Here we highlight a different kind of variability—in the phones that are present or absent in a word form—which infants cannot reasonably disregard in trying to learn and recognize words. Providing computational models with dictionary pronunciations instead of a direct phonological transcription ignores this kind of variability and thus overestimates the clarity of the language input.

To address this potential limitation, the present study tested several existing models of word segmentation using two different phonological transcriptions of the same speech: a "dictionary pronunciations" version, derived using an orthographic transcription and a pronunciation dictionary, and a more realistic "transcribed pronunciations" version, or direct phonological transcription of the speech. This served two purposes. First, we wanted to assess how the performance of existing models would be affected by using more realistic input that incorporated phonological variation. If the models' previous successes relied on certain unrealistic features of dictionary-derived phonological transcriptions, then we would expect to see a substantial decrease in performance on the transcribed pronunciations version of the corpus. Similar performance on both versions of the corpus, on the other hand, would suggest that the models are robust to phonological variation, in line with infants' own learning, which proceeds despite the phonological variation present in actual speech. (We acknowledge that such a result would still leave open the question of whether infants' representation of spoken language resembles that of hand-transcribed corpora, a point we will return to later.)

Second, in addition to quantifying numerical differences in performance, we wanted to investigate what concrete effects more realistic input might have on the learner's developing

lexicon. Given input in which the same word can be pronounced in multiple ways, what kind of word knowledge could the learner achieve according to current proposed solutions to word segmentation? The typical view of word learning supposes that the output of word segmentation serves as the input to the process by which children map meanings to words (e.g., Graf Estes et al., 2007), but the present work highlights the fact that even when segmentation is successful, the resulting word forms can be difficult to link to word types.

We will begin by reviewing several recent models of statistically driven word segmentation. Next, we describe the corpus that we used as input to test how robust these models are to phonological variability, and present the performance results. Finally, we explore the nature of the segmented word forms under conditions of phonologically variable input and its broader implications for word learning.

1.1 Segmentation Algorithms

Building on experimental work using transitional-probability-based stimuli (Aslin et al., 1998; Saffran et al., 1996), several authors have implemented transitional-probability-based computational models of word segmentation (e.g., Gervain & Guevara Erra, 2012; Saksida et al., 2017; Yang, 2004). These models compute the transitional probability of each pair of units XY, which can be defined as the probability of XY divided by either the probability of X (forward probability), the probability of Y (backward probability), or their product (mutual information; in this case, the resulting fraction is also log-transformed). Then, the models insert word boundaries either wherever the transitional probability is lower than the transitional probability of the pairs around it (relative threshold) or wherever the transitional probability is lower than the corpus average (or some other absolute threshold). Work by Saksida et al. (2017) suggests that different variations of this transitional probabilities (TP) model may be more effective in different

languages, although overall performance was relatively high across languages and model variations.

The diphone-based segmentation model (DiBS) of Daland and Pierrehumbert (2011) also tracks the co-occurrence frequencies of sub-word units, in order to explicitly estimate the probability of a word boundary given a particular diphone. The key insight is that infants might detect through observation that some sounds are unusually common at the beginnings and ends of utterances (compared to their overall co-occurrence frequency). In the absence of any word boundary information, infants could use the utterance boundaries instead and treat these sounds as especially likely beginnings and ends of words. This is what DiBS does in its unsupervised instantiation. More specifically, DiBS estimates the probability of a word boundary occurring between two phones using the observed frequencies with which the first phone ends utterances and the second phone begins utterances, along with their co-occurrence frequency. If the estimated probability is greater than 0.5, a word boundary is deterministically inserted. Daland and Pierrehumbert (2011) found that DiBS was somewhat robust to phonological variation, although their focus was on the relative rates of different types of segmentation errors rather than absolute performance metrics.

As in the DiBS model, units that occur at the beginning and end of utterances play an important role in the PUDDLE (Phonotactics from Utterances Determine Distributional Lexical Elements) model of Monaghan and Christiansen (2010). PUDDLE is a subtractive algorithm that pulls out known chunks (previous utterances, to start) from new utterances, creating new chunks. However, this segmentation only occurs if the resulting new chunks start and end with n-grams (diphones by default) that the model has already learned as legal onsets and offsets. Thus, n-grams that occur next to utterance boundaries, which always get stored as legal onsets and

offsets, greatly inform the model's subsequent decisions. Monaghan and Christiansen (2010) tested the PUDDLE model on a phonological corpus derived by passing an orthographic corpus through a speech synthesizer. As they point out, this phonologization process is more realistic than most pronunciation dictionaries, since the speech synthesizer allows the same orthographic word to be pronounced differently in different part-of-speech contexts (e.g., "uses" as a verb versus "uses" as a noun). Still, this model does not incorporate all of the phonological variation present in actual speech.

These three models of word segmentation, along with a number of other models, continue to be used to investigate questions about children's early word learning. While such models are not necessarily seen as mechanistic explanations of what infants actually do, they are at least taken to demonstrate what information is potentially available to infants in different kinds of language input. Recent studies (Cristia et al., 2019; Fibla et al., 2021) highlight that a range of models can be used in parallel to better identify results that are stable across models. To facilitate this kind of multiple-model investigation, Bernard et al. (2020) developed the WordSeg software package, a coordinated collection of several different word segmentation algorithms, including TP, DiBS, and PUDDLE. So far, the WordSeg implementations have been used to compare the segmentability of adult-directed and child-directed speech (Cristia et al., 2019), to assess the segmentability of bilingual language input (Fibla et al., 2021), to test the value of prosodic breaks (Ludusan et al., 2022), and to measure the effects of morphological complexity on word segmentation (Loukatou et al., 2022). Given the importance of these and other questions to which these models can be applied, it seems especially crucial to investigate the consequences of the assumptions that such modeling efforts usually make about phonological variation.

2 Materials and Methods

2.1 Corpus

We used the Buckeye corpus (Pitt et al., 2007) because it is a large corpus that already has both a direct phonological transcription and also a phonological transcription derived via lookup of orthographic words in a pronunciation dictionary. The Buckeye corpus contains spontaneous speech from 40 American English-speaking adults living in Columbus, Ohio. Speech was recorded during one-on-one interviews about a variety of local issues, and then orthographically and phonologically transcribed. In the present study, we analyzed a subset of the Buckeye corpus composed of speech from four female talkers under 40 years of age. These speakers were selected so as to better approximate infant-directed speech, an issue taken up in more detail in the Discussion. In total, the smaller corpus used in this study included 30,910 words from 1,425 conversational turns. (Our rationale for collapsing across these four speakers when constructing the corpus can be found in the Supplementary Materials.)

During pre-processing, we removed all non-speech codes (e.g., VOCNOISE for non-speech vocalizations) and words containing non-speech codes from the corpus. We also replaced instances of syllabic consonants with a schwa vowel followed by that consonant. Before running the segmentation algorithms, we modified the corpus to include more frequent utterance boundary codes. The Buckeye corpus only marks conversational turn boundaries and not other utterance boundaries, so we probabilistically inserted additional utterance boundaries between words according to the rate observed in child-directed speech (the Brown (1973) files in the CHILDES database (MacWhinney, 2000)).

2.2 Syllabification

Some of the segmentation algorithms that we tested use syllables as the basic unit. To prepare the corpora for these algorithms, we used the program *tsylb2* developed at the National

Institute of Standards and Technology (Fisher, 1996). This program syllabifies words using information about which consonant clusters can begin and end words in English, in combination with the principle of maximal onset (intervocalic consonants are maximally assigned to syllable onsets).

2.3 Segmentation

We employed three proposed segmentation algorithms, TP, DiBS, and PUDDLE (described above), on both versions of the corpus using the WordSeg software package (Bernard & Cristia, 2018). (See the Supplementary Materials for a description of the parameters.) Since which unit, the phone or the syllable, is more appropriate to consider as the basic unit has been debated in the literature (e.g., Gambell & Yang, 2005; Swingley, 2005b), we tested both phone-based and syllable-based versions of each algorithm, with the exception of DiBS, for which we only tested the unsupervised phone-based version. Note that providing syllable boundaries rather than phone boundaries is much closer to providing the true word boundaries already, because English has many monosyllabic words. As a result, it is not meaningful to compare the performance of the phone-based algorithms to the performance of the syllable-based algorithms.

Instead, their performance can be compared to the performance of two different baseline algorithms. As the syllable-based baseline, we used the WordSeg (Bernard & Cristia, 2018) baseline algorithm, which identifies every syllable as a word. As the phone-based baseline, we coded an implementation of the Possible Word Constraint (Norris et al., 1997), according to which all segmentations should contain at least one vowel. To achieve this constraint, this baseline algorithm considers each pair of consecutive vowels in each utterance in the corpus and inserts a word boundary somewhere between them (with the location chosen at random)

according to the oracle probability of a word boundary occurring between two consecutive vowels within an utterance, across the corpus.

2.4 Performance Evaluation

To quantitatively evaluate the performance of each algorithm, we computed the standard information-theoretic measures of (token) precision, recall, and F-score. In the context of word segmentation, *precision* measures how many of the segmented words were correctly segmented (matched the gold text, i.e. [correct segmentations] / [all segmentations]), while *recall* measures how many of the words in the gold text were successfully extracted ([correct segmentations] / [all word tokens in the gold text]). In line with previous work (e.g., Cristia et al., 2019), we focus on token F-score, which is the harmonic mean of precision and recall. Results of analyses considering precision and recall separately are provided in the Supplementary Materials.

3 Results

3.1 Model Performance

Fig. 1 shows the relative performance of each algorithm on the dictionary pronunciations versus the transcribed pronunciations version of the corpus. The error bars represent two standard deviations across ten different runs of the procedure that probabilistically inserted additional utterance boundaries into the corpus. These pseudo-confidence intervals provide an estimate of the within-corpus noise introduced by different utterance boundary randomizations.

Across algorithms, performance on the transcribed pronunciations version of the corpus was lower than performance on the dictionary pronunciations version (as evidenced by the downward slope of the lines in Fig. 1), with the exception of the baseline algorithms. The average decrement in token F-score for the non-baseline algorithms was 12%, ranging from 3.65% to 22.5%. Using transcribed pronunciations also changed how performance varied

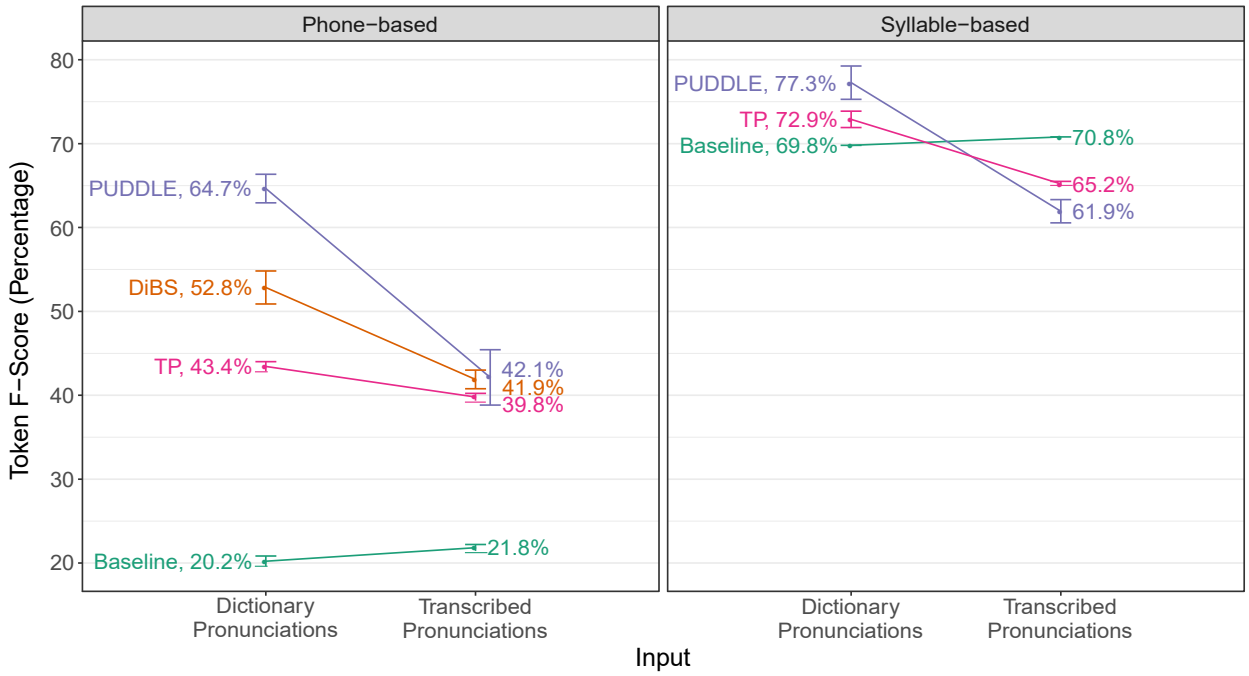


Figure 1. Segmentation performance on dictionary pronunciations versus transcribed pronunciations of the same speech. Each connected pair of points represents a particular word segmentation algorithm, and error bars show empirical 95% confidence intervals over different utterance boundary randomizations.

between algorithms, practically eliminating the differences observed on the dictionary pronunciations version of the corpus (phone-based algorithms) or even reversing the previous pattern (syllable-based algorithms).

The observed decrease in performance on the transcribed pronunciations can be explained by the underlying statistics of this version of the corpus. Let us consider the phone-based algorithms for simplicity. In the dictionary pronunciations version of the corpus, there are some pairs of phones that are extremely reliable cues to the presence or absence of word boundary. For example, when /h/ is followed by any other phone, the probability of a word boundary occurring between them is 0, since /h/ cannot end words in English. Conversely, any phone followed by /h/ is a fairly reliable cue to the presence of a word boundary (P(word boundary) = 0.887), since these can only belong to the same word if that word is multisyllabic (e.g., “clubhouse”). However, because /h/ is often deleted in conversational speech (e.g., “im”

instead of "him"), these helpful cues are less frequent in the transcribed pronunciations version of the corpus (1,358 instead of 1,635 occurrences). As another example, /ʊ/, which never starts words in the dictionary pronunciations version of the corpus and so always attaches to the phone before it, occurs 2,622 times in the dictionary pronunciations but less than half that often (1,196 times) in the transcribed pronunciations because of frequent vowel reduction.

In addition to cases like these, some phone pairs that are reliable cues in the dictionary pronunciations become less reliable in the transcribed pronunciations. For instance, /ɪ/, /ɛ/, and /ʌ/ never directly precede word boundaries in the dictionary pronunciations but do so about 10% of the time in the transcribed pronunciations due to final consonant deletion. On the whole then, the statistical landmarks that help in the dictionary pronunciations version of the corpus have been eroded in the transcribed pronunciations, leading to worse segmentation performance across models.

Despite the decrease we observed moving from dictionary pronunciations to transcribed pronunciations, the algorithms' absolute performance on the transcribed pronunciations version of the corpus was still relatively high in the case of the syllable-based algorithms and well above the relevant baseline for the phone-based algorithms (Fig. 1). This suggests that at least in terms of numerical performance, these models of word segmentation are somewhat robust to the phonological variation present in actual speech, though of course they leave open the question of how infants identify the phones or locate the syllable boundaries.

3.2 Proto-Lexicon of Word Forms

In addition to calculating the standard performance metrics, we also examined the nature of the word forms that appeared in the segmentation output when the models were given phonologically variable input. Since the typical view assumes that the output of word

segmentation gives rise to a proto-lexicon of word forms whose meanings are discovered during word learning, we wanted to assess the correspondence between segmented word forms and orthographic words, using orthographic words as a proxy for word meanings.

To visualize the relationship between correctly segmented phonological word forms and orthographic words, we can think of phonological word forms and orthographic words as the two kinds of *nodes* in a *bipartite* (or *bimodal*) *network*. In this network, an *edge* exists between two nodes A and B if that phonological word form A was ever correctly segmented when B was the speaker's intended orthographic word. With dictionary pronunciations as input, this network is guaranteed to consist of one-to-one links, or pairs of nodes that are only connected to each other (with the exception of homophones, where two phonological word forms would be linked to the same orthographic word). With phonologically variable input, however, such one-to-one correspondences are not guaranteed. Instead of one-to-one links where the meaning of each phonological word form is well defined, we could instead see a complex many-to-many relationship, where each orthographic word has several different pronunciations and these pronunciations overlap with the pronunciations of other orthographic words. In this case, learning which meaning to attach to a phonological word form would pose a problem with no clear solution.

For example, consider an English-learning child who has isolated [sɪd] ("sid") as a potential word, based on its statistical cohesiveness. The child might observe that this word's contexts of use are compatible with notions conveyed by "sit" and "said." Given this evidence, the child might suppose that these two meanings are, in fact, both members of some larger semantic category than previously hypothesized (e.g., LaTourrette & Waxman, 2020); or might guess that [sɪd] is a homophone. Similarly, a child who has isolated [kɔl] ("call") and [kɔ]

325 ("caw") as separate forms whose contexts of use (instances of the intended orthographic word
326 "call") seem identical might suppose that "call" has more than one pronunciation, or that the
327 phonological categories of [kɔl] and [kɔ] should actually be collapsed into one. Resolving these
328 possible errors, even if multiplied over many items in the lexicon, seems tractable. But there are
329 also many-to-many cases where any solutions would seem to be overwhelmed with ambiguity.
330 Imagine, for instance, a child who has isolated [kɪd] as a potential word. While canonically this is
331 simply the single pronunciation of the word "kid", in actual speech, "kid" and "could" are both
332 frequently pronounced as [kɪd]. Furthermore, "kid" can also be pronounced [kɪ], as can "could",
333 and other pronunciations of each word overlap with yet more orthographic words (e.g., "kit",
334 "good", "can", etc.). This scenario is clearly much less tractable, even given perfect knowledge
335 of each word form's context of use. (For a visual example of a small, many-to-many mapping,
336 see Fig. 2).

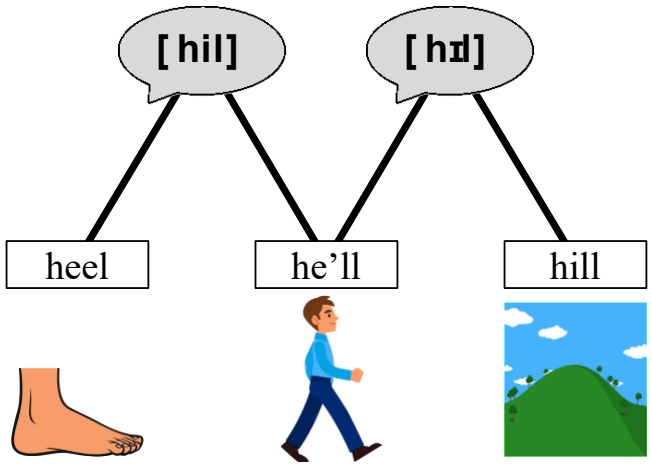


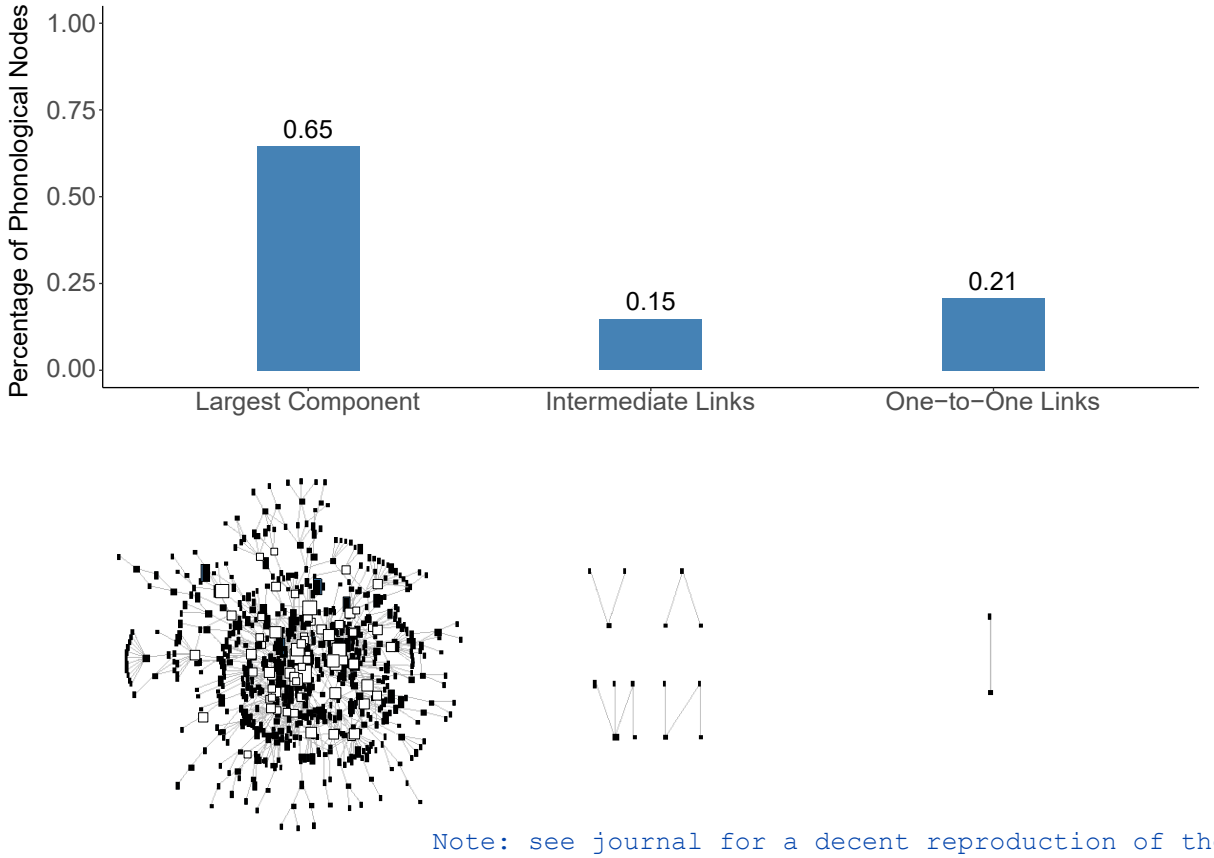
Figure 2. Example many-to-many mapping. Edges between phonological word forms (shaded) and orthographic words (white) represent attested pronunciations. In this network, the same orthographic word ("he'll") can have multiple pronunciations ([hil] and [hɪl]), and a single phonological word form can map onto multiple orthographic words and thus meanings (e.g., [hil] maps onto both "heel" and "he'll").

337 With these possibilities in mind, we turn to our results. We observed a variety of
338 outcomes, including both unambiguous one-to-one links and larger many-to-many components.

For simplicity, we focus here on the correct segmentations under the phone-based transitional probabilities model, but the overall pattern of results was similar using other models (see Supplementary Materials). As Fig. 3 shows, **most** of the phonological nodes or postulated word types (65%; 86% of tokens) ended up in a single giant component of phonological and lexical overlaps. By contrast, only 21% (7% of tokens) belonged to a one-to-one link.[1] Many of the one-to-one links were of extremely low frequency (just two or three occurrences), making it hard to say how many of these perfect correspondences would persist given a larger corpus.

This network of orthographic words and phonological word forms was very different from the equivalent network generated under the unrealistic simplifying assumption of no phonological variation from the dictionary pronunciations. The network generated from transcribed real pronunciations had a higher *density* (number of observed edges / total possible edges given the number of nodes). This increase in density was expected because when the dictionary pronunciations version of the corpus is used as input, the number of observed edges is bounded by the number of (correctly segmented) orthographic word types (i.e., each orthographic word has no more than one pronunciation). This is not true when the transcribed pronunciations are used. However, in addition to an increase in density, we also observed a giant component composed of overlaps, including a large number of many-to-many connections, and encompassing the majority of the postulated word types. In other words, it is not merely the case that each orthographic word had a few different pronunciations that would need to be grouped together by the learner. A given phonological word form was also ambiguous as to the intended orthographic word, indicating much more widespread homophony than is typically assumed.

[1] These estimates excluded segmentations that occurred only once (5% of the tokens). If these *hapax legomena* are included, the analogous proportions are 53% of types (84% of tokens) and 24% of types (6% of tokens).



Note: see journal for a decent reproduction of the network figure, which here didn't survive pdf/a conversion

Figure 3. Network visualization of the correctly segmented words under the TP model. Each shaded node represents a segmented word or phonological word form, which is linked to one or more intended orthographic words (white squares). Node size represents frequency, though any segmentations that occurred only once were excluded. For simplicity, only a few representative examples are plotted for the smaller components ("intermediate links" and "one-to-one links"). The bar graph shows the percentage of phonological word forms that ended up in each kind of component.

It is possible that our use of a binary edge condition, in which we ask whether the phonological word form A either was ever, or was never, an instance of the orthographic word B, overestimates the messiness of the input by weighting very infrequent pronunciation variants as strongly as frequent ones. If, for example, 9/10 instances of "rain" were segmented as [ɹeɪn] and 1/10 as [ɹeɪ] (which could also be "ray"), the child might be in a different position than if the proportions were 5/10 and 5/10. To incorporate frequency information, let us consider a *weighted network*, where each edge in the network has a weight representing how many times each orthographic word was realized as a particular phonological word form. Then, a

phonological word form and an orthographic word can be said to be in a close to one-to-one relationship if the weight of the edge between them is high relative to the *weighted degree* (total frequency = sum of weights of direct edges) of either node. Borrowing a measure from information theory, this is equivalent to saying that the (pointwise) mutual information (PMI) of the two nodes A and B (i.e., log[freq(A and B) / (freq(A)*freq(B)]) is high, or that the normalized pointwise mutual information (NPMI = PMI / -log[freq(A and B)]) is close to 1, where 1 indicates perfect correlation. In Fig. 4, we show what happens to the giant component from the original network when edges between nodes with NPMI close to 0 (where 0 indicates statistical independence) and edges with a weight of 1 (correspondences that occurred only once) have been pruned using oracle knowledge of the intended orthographic word. Taking frequency information into account in this way resolves some of the ambiguities, creating some one-to-one links and intermediate-size components where before there was only a single, densely connected component.

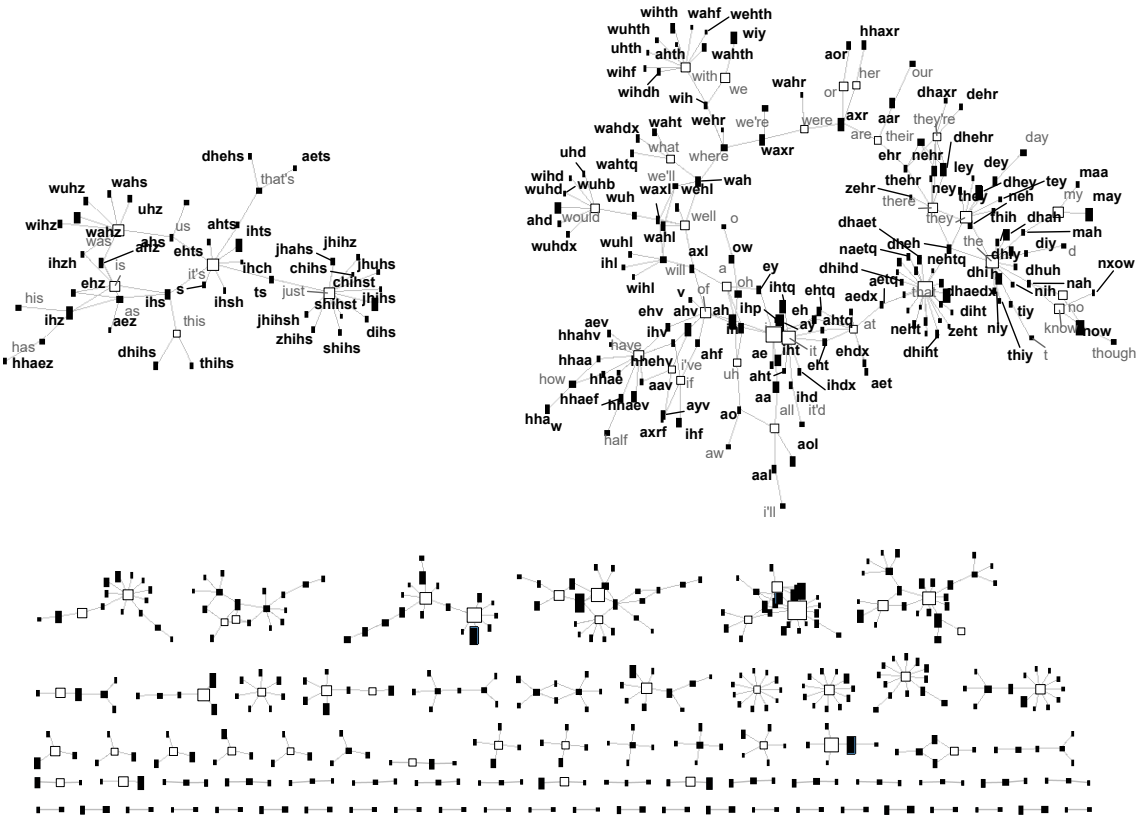


Figure 4. Giant component after pruning. Edges linking phonological word forms (shaded circles) and intended orthographic words (white squares) were removed if the two were only weakly informative about each other (NPMI ≤ 0.25) or if the correspondence occurred only once among the correct segmentations. For the components at the top of the figure, the orthographic nodes are labeled in gray (e.g., "his"), and the phonological nodes in black (e.g., **ihz** ([ɪz] in IPA)). For clarity, labels are not presented for the smaller components at the bottom of the figure.

The sparser network formed by pruning links with less tight an evidential connection shows that if children could identify and ignore infrequent phonological and lexical correspondences, or simply forget them, this would reduce the number of words involved in intractable many-to-many overlaps. Nevertheless, even after this pruning, some large connected components remained. For example, "as", "has", "his", "just", "is", "it's", "that's", "this", "was", and "us" all belonged to the same connected component. Such many-to-many relationships challenge the assumption that statistically driven word segmentation provides the learner with a strong lexicon-building foundation. In other words, considering word types and

phonological word forms separately reveals a hidden learning challenge that unsupervised statistics computed over a phonological transcription cannot untangle.

Pruning infrequent edges from the network provides a more nuanced characterization of the input but is probably unrealistic as a model of the learner, since infants do not have oracle knowledge of the intended orthographic word. Without this knowledge, the frequency of each correspondence is unlikely to be available. Thus, taking a more psychologically plausible approach, we modeled what would happen if phonological word forms below a certain overall frequency were excluded or forgotten. If we exclude word forms that happened only few times in the corpus, does this help reduce the number of many-to-many mappings? As Fig. 5 shows, thresholding by frequency does reduce the absolute number of orthographic words involved in many-to-many relationships, but their proportion in the proto-lexicon actually increases. In general, the phonological word forms involved in many-to-many relationships are high in overall frequency. As a result, excluding low-frequency word forms hurts the ideal, one-to-one links substantially more than it helps resolve the many-to-many entanglements.

So far, the networks we have considered have assumed that the learner tries to link up

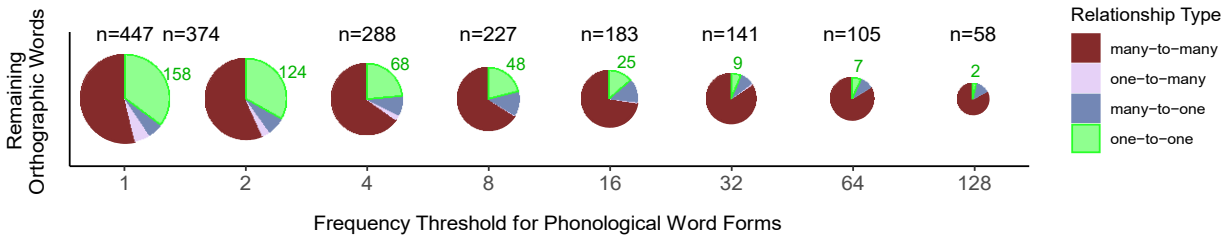


Figure 5. Composition of the proto-lexicon assuming low-frequency word forms are forgotten. For each frequency threshold, a pie chart shows which orthographic words remain when the phonological word forms at or below that frequency are excluded. (Note that because the corpus is a sample of the listener's experience, a frequency of 1 in the corpus does not necessarily correspond to a frequency of 1 in the listener's experience.) Including only higher frequency words forms reduces the absolute number of words involved in many-to-many entanglements (i.e., the absolute area of the dark red region decreases from left to right), but their proportion increases. By contrast, both the number (labeled) and the proportion of the ideal, one-to-one words (green, outlined) decrease dramatically as the frequency threshold increases.

every (correctly) statistically derived word form with a meaning. Even with perfect access to the word's semantic context (implemented here by identifying it with the corpus' orthographic transcription), our results demonstrate that this is an extremely hard problem to solve. As an alternative to this exhaustive processing of the input, infants might instead only consider as candidate word forms for their initial proto-lexicon those word forms that seem to have concrete referents. To investigate this possibility, we filtered the network by removing orthographic nodes with concreteness ratings (Brysbaert et al., 2014) below the median, and also removing any edges connected to them. Fig. 6 shows the result.[2] This exclusion of less concrete words significantly reduced the density of the network (0.0028) compared to removing the same number of orthographic nodes chosen at random (bootstrap 95% CI = [0.0033, 0.0043]). This happens because more concrete words tend to be phonologically heavier in their specification in the lexicon, and also less affected by reduction processes that would alter their transcription. Therefore, temporarily leaving aside words with less concrete meanings early in the learning process could make the mapping problem more tractable. This solution does not entirely rescue the learner, however, because some many-to-many components persisted after filtering, and because knowledge of the semantic context is not in fact given. In the real world, infants have to contend with variation both in the referential world and in words' pronunciations. Although some word use instances in parent-infant interaction are semantically transparent (Trueswell et al., 2016) and some demonstrate careful, relatively unambiguous phonetic presentations (Cychosz et al., 2021), these learning opportunities may not be the same ones (CITE BEECH & SWINGLEY 'gems' paper).

[2] As a complement to this simulation, we also used regression analyses to describe whether some kinds of words (e.g., content versus function words) were more likely to end up in one-to-one or close to one-to-one relationships in the original network. See the Supplementary Materials for details.

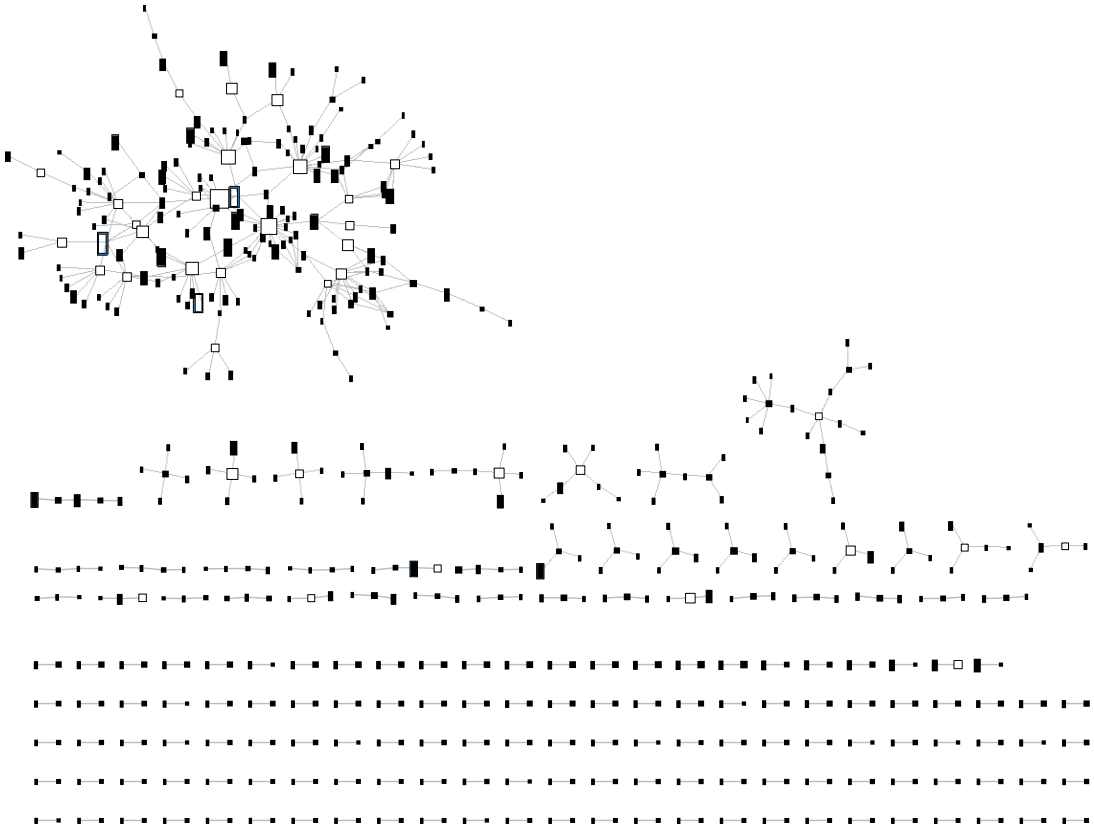


Figure 6. Entire network with less concrete words excluded. Concreteness norms were used to exclude orthographic words (white squares) of below average concreteness (2.64 out of 5) and any edges connecting them to phonological word forms (shaded circles). Despite the significant reduction in density compared to random orthographic node removal, some complex, many-to-many components persisted.

4. Discussion

In this study we investigated the effects of phonological variation on models of statistically driven word segmentation. We found that using transcribed pronunciations rather than canonical pronunciations led to consistently lower numerical performance across algorithms. In addition, we showed that phonological variability poses substantial problems for lexicon building. Most of the extracted phonological word forms ended up in a dense web of phonological and lexical overlaps, where phonological identities and differences were not consistent cues to word identity. This finding could explain why toddlers have sometimes failed to apply a mature phonological criterion in word learning experiments (Dautriche et al., 2015;

Stager & Werker, 1997; Swingley & Aslin, 2007). Perhaps they have learned that in real speech, a small phonological difference often does not imply a difference in meaning.

Traditionally, the fact that infants and toddlers often recognize words less well when the words are realized with phonologically deviant pronunciations has been interpreted as evidence that young children do use a phonological criterion for lexical differentiation, in line with textbook definitions of the function of a formal phonology. For example, two-year-old children learning Catalan, but not two-year-old children learning Spanish, find words harder to identify if their vowel /e/ is realized as /E/—phonetically the same change in the two stimulus sets, but phonologically quite different in the two languages (Ramon-Casas et al., 2009). This is consistent with the idea that "mispronunciation effects" derive from a mismatch signal triggered by the deviant phone, which is much stronger for phonological distinctions. Dietrich et al. (2007) obtained a similar result in a word teaching context, with younger children. But, does this strong mismatch signal, which we measure in experiments, prevail in children's interpretation of novel words? The answer, at least in the laboratory word-learning contexts we have tested, is no. Swingley and Aslin (2007) tried to teach 19-month-olds novel object labels that were phonologically distinct from familiar words (like "tog" for "dog"), and children failed again and again. Children of this age can hear the difference, but it does not make them posit a new lexical item. Swingley (2016) showed a similar result and eye-tracked the same children. From children's own eye movement data, it was clear that they noticed deviant realizations of words, but they did not interpret those differences as corresponding to novel words for unfamiliar objects. Why not? Why do children seem so skeptical, when they are confronted with a novel neighbor? Perhaps because their linguistic experience has been replete with tokens of words that often stray beyond their own phonological bounds. Children need to learn the canonical forms of

words, but they also need to learn the phonetic transform that connects reduced forms to one or more distinct canonical forms.

One limitation of the present work is the use of a corpus of adult-directed rather than infant-directed speech. Adult-directed and infant-directed speech clearly differ in their content, and a body of work suggests that infant-directed speech may be tailored to promote learning (e.g., Eaves et al., 2016; Kuhl et al., 1997; though see Ludusan et al., 2021). However, it is also clearly not the case that parents speak like dictionaries, producing only canonical forms, when conversing with their children (e.g., Bard & Anderson, 1983; Buckler et al., 2018; Lahey & Ernestus, 2014). In addition, infant-directed speech appears to have only a small and inconsistent advantage in segmentability, at least when the recording contexts for infant-directed and adult-directed speech are similar (Cristia et al., 2019).

Although this work introduced more realism in one way (by incorporating phonological variability), it still made simplifying assumptions. For instance, in considering the problem of attaching meanings to words, the orthographic word served both as the linguistic target and as a stand-in for the semantic context. Future investigations could model the semantic context separately, possibly by using a corpus with associated video data.

In addition, we have assumed that the statistical word segmentation algorithms operate over phonological categories. It is possible that the phonological units that infants use early in word learning are actually more continuous, in line with automatic speech recognition features derived directly from the acoustic signal. Although this changes the problem space, other computational modeling efforts have explored the feasibility of speech-based segmentation (Dunbar et al., 2020; Räsänen, 2011). It could be that some of the complexity we observed here in mapping variable segmented forms to distinct lexical items would be attenuated by avoiding

the imposition of phonological categories in the first place. For example, if the word “tree” were sometimes realized with an aspirated /t/ ([t^{h}ɹi]), and sometimes realized with frication ([tʃɹi]), these instances might be phonetically close but transcribed as phonemically distinct. At present, it is impossible to say whether a more “analog” and less “digital” conceptualization of infant speech processing would ameliorate the pervasive ambiguity problem we have identified, or exacerbate it. However, if infants do not adopt a categorical representation of speech one way or another, they also could not compute the same kinds of statistics that are widely presupposed to underlie performance in word segmentation experiments. This is a fruitful ground for further research efforts.

Despite these limitations, this study has important broader implications. Specifically, it suggests that the exhaustive parsing models that dominate current thinking about very early language development would, if true, place infants in a difficult position, by leading them to build extremely complex initial lexicons containing a strong proportion of unhelpful categorizations. This provides some impetus for thinking about the problem in another way. In particular, rather than conceptualize infants as trying to fully parse every sentence using rudimentary statistical segmentation heuristics, we might do better to suppose that infants begin language learning by attending primarily to salient islands of reliability in phonetic and semantic space, and building outward from there. If infants initially filter their input, homing in on moments where words are pronounced more clearly or canonically, or the intended meaning is more easily available, this could help them sidestep some of the problems that phonological variability poses for early word learning. Eventually though, young children must contend with speech on a larger scale. How children make this transition and learn to manage the phonological variability in speech remains an open question.

Acknowledgements

This work was supported by an NIH-funded predoctoral fellowship to CB (T32 DC016903) awarded to the University of Pennsylvania by the NIDCD, and by NSF grant 1917608 awarded to DS.

References

Archer, S. L., Czarnecki, N., & Curtin, S. (2021). Boosting the input: 9-month-olds' sensitivity to low-frequency phonotactic patterns in novel wordforms. *Infancy*, *26*(5), 745–755. https://doi.org/10.1111/infa.12423

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10*(2), 265–292. https://doi.org/10.1017/S0305000900007777

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

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397. https://doi.org/10.1016/j.cognition.2013.02.011

Bernard, M., & Cristia, A. (2018). *WordSeg* (0.7.1). https://doi.org/https://doi.org/10.5281/zenodo.1471532

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

Börschinger, B., Johnson, M., & Demuth, K. (2013). A joint model of word segmentation and phonological variation for English word-final/t/-deletion. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*, 1508–1516. http://web.science.mq.edu.au/~bborschi.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125. https://doi.org/10.1016/s0010-0277(96)00719-6

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. https://doi.org/10.1016/S0010-0277(01)00122-6

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Buckler, H., Goy, H., & Johnson, E. K. (2018). What infant-directed speech tells us about the development of compensation for assimilation. *Journal of Phonetics*, *66*, 45–62. https://doi.org/10.1016/j.wocn.2017.09.004

Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2019). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, *3*, 13–22. https://doi.org/10.1162/opmi_a_00022

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121. https://doi.org/10.1037/0096-1523.14.1.113

Cychosz, M., Edwards, J. R., Bernstein Ratner, N., Torrington Eaton, C., & Newman, R. S. (2021). Acoustic-lexical characteristics of child-directed speech between 7 and 24 months and their impact on toddlers' phonological processing. *Frontiers in Psychology*, *12*, 712647. https://doi.org/10.3389/fpsyg.2021.712647

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. https://doi.org/10.1111/j.1551-6709.2010.01160.x

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86. https://doi.org/10.1016/j.cognition.2015.06.003

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 United States of America*, *104*(41), 16027–16031. https://doi.org/10.1073/PNAS.0705270104

Dunbar, E., Karadayi, J., Bernard, M., Cao, X. N., Algayres, R., Ondel, L., Besacier, L., Sakti, S., & Dupoux, E. (2020). The zero resource speech challenge 2020: Discovering discrete subword and word units. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-Octob*, 4831–4835. https://doi.org/10.21437/Interspeech.2020-2743

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*(6), 758–771. https://doi.org/10.1037/rev0000031

Elsner, M., Goldwater, S., Feldman, N. H., & Wood, F. (2013). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 42–54.

Ernestus, M., & Baayen, R. H. (2011). Corpora and Exemplars in Phonology. In *The Handbook of Phonological Theory: Second Edition* (pp. 374–400). https://doi.org/10.1002/9781444343069.ch12

Fibla, L., Sebastian-Galles, N., & Cristia, A. (2021). Is there a bilingual disadvantage for word segmentation? A computational modeling approach. *Journal of Child Language*, 1–28. https://doi.org/10.1017/s0305000921000568

Fisher, W. (1996). *Tsylb2 syllabification software*. https://www.nist.gov/itl/iad/mig/tools

Gambell, T., & Yang, C. (2005). Word segmentation: Quick but not dirty. *Unpublished Manuscript, Yale University*, 1–36. http://www.ling.upenn.edu/~ycharles/papers/quick.pdf

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287. https://doi.org/10.1016/j.cognition.2012.06.010

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Hallé, P. A., & Boysson-Bardies, B. de. (1994). Emergence of an early receptive lexicon: Infants' recognition of words. *Infant Behavior and Development*, *17*(2), 119–129. https://doi.org/10.1016/0163-6383(94)90047-7

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x

Johnson, K. (2004). Massive reduction in conversational American English. *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*, 29–54. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=77FFE3C22602F55D083CAB7CA669DB0E?doi=10.1.1.142.5012&rep=rep1&type=pdf

599 Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent
600 speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010
601 Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993).
602 Infants′ sensitivity to the sound patterns of native language words. *Journal of Memory and*
603 *Language*, *32*(3), 402–420. https://doi.org/10.1006/jmla.1993.1022
604 Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, *277*(5334),
605 1984–1986. https://doi.org/10.1126/science.277.5334.1984
606 Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation
607 in English-learning infants. *Cognitive Psychology*, *39*(3–4), 159–207.
608 https://doi.org/10.1006/cogp.1999.0716
609 Keren-Portnoy, T., Vihman, M., & Fisher, R. L. (2019). Do infants learn from isolated words?
610 An ecological study. *Language Learning and Development*, *15*(1), 47–63.
611 https://doi.org/10.1080/15475441.2018.1503542
612 Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V, Ryskina,
613 V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of
614 phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.
615 https://doi.org/10.1126/science.277.5326.684
616 Lahey, M., & Ernestus, M. (2014). Pronunciation variation in infant-directed speech: Phonetic
617 reduction of two highly frequent words. *Language Learning and Development*, *10*(4), 308–
618 327. https://doi.org/10.1080/15475441.2013.860813
619 LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode
620 and remember objects. *Proceedings of the National Academy of Sciences of the United*
621 *States of America*, *117*(35), 21230–21234. https://doi.org/10.1073/pnas.2006608117
622 Lignos, C. (2011). Modeling infant word segmentation. *CoNLL 2011 - Fifteenth Conference on*
623 *Computational Natural Language Learning, Proceedings of the Conference*, 29–38.
624 Loukatou, G., Stoll, S., Blasi, D., & Cristia, A. (2022). Does morphological complexity affect
625 word segmentation? Evidence from computational modeling. *Cognition*, *220*, 104960.
626 https://doi.org/10.1016/j.cognition.2021.104960
627 Ludusan, B., Cristia, A., Mazuka, R., & Dupoux, E. (2022). How much does prosody help word
628 segmentation? A simulation study on infant-directed speech. *Cognition*, *219*, 104961.
629 https://doi.org/10.1016/j.cognition.2021.104961
630 Ludusan, B., Mazuka, R., & Dupoux, E. (2021). Does Infant-Directed Speech Help Phonetic
631 Learning? A Machine Learning Investigation. *Cognitive Science*, *45*(5), e12946.
632 https://doi.org/10.1111/cogs.12946
633 MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence
634 Erlbaum Associates.
635 Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by
636 infants. *Cognition*, *78*(2), 91–121. https://doi.org/10.1016/S0010-0277(00)00109-8
637 Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic
638 effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465–494.
639 https://doi.org/10.1006/cogp.1999.0721
640 Mersad, K., & Nazzi, T. (2012). When Mommy comes to the rescue of statistics: Infants
641 combine top-down and bottom-up cues to segment speech. *Language Learning and*
642 *Development*, *8*(3), 303–315. https://doi.org/10.1080/15475441.2011.609106
643 Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling
644 psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–

564. https://doi.org/10.1017/S0305000909990511
Nishibayashi, L. L., Goyet, L., & Nazzi, T. (2015). Early speech segmentation in French-learning infants: Monosyllabic words versus embedded syllables. *Language and Speech*, *58*(3), 334–350. https://doi.org/10.1177/0023830914551375
Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243. https://doi.org/10.1006/cogp.1997.0671
Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685. https://doi.org/10.1111/j.1467-8624.2009.01290.x
Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus.osu.edu]*. Department of Psychology, Ohio State University (Distributor).
Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*(1), 96–121. https://doi.org/10.1016/j.cogpsych.2009.02.002
Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, *120*(2), 149–176. https://doi.org/10.1016/j.cognition.2011.04.001
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926
Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, *20*(3), e12390. https://doi.org/10.1111/DESC.12390
Schreiner, M. S., Altvater-Mackensen, N., & Mani, N. (2016). Early word segmentation in naturalistic environments: Limited effects of speech register. *Infancy*, *21*(5), 625–647. https://doi.org/10.1111/infa.12133
Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, *57*(1), 24–48. https://doi.org/10.1016/j.jml.2006.10.004
Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565–573. https://doi.org/10.1111/j.1467-7687.2006.00534.x
Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, *11*(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x
Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102
Sundara, M., & Mateu, V. E. (2018). Lexical stress constrains English-learning infants' segmentation in a non-native language. *Cognition*, *181*, 105–116. https://doi.org/10.1016/j.cognition.2018.08.013
Swingley, D. (2005a). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432–443. https://doi.org/10.1111/j.1467-7687.2005.00432.x
Swingley, D. (2005b). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001
Swingley, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words.

691 *Developmental Psychology*, *52*(7), 1011–1023. https://doi.org/10.1037/dev0000114
692 Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning.
693 *Cognitive Psychology*, *54*(2), 99–132. https://doi.org/10.1016/j.cogpsych.2006.05.001
694 Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of
695 specific English words. *Child Development*, *89*(4), 1247–1267.
696 https://doi.org/10.1111/cdev.12731
697 Trueswell, J. C., Lin, Y., Armstrong, B. F., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L.
698 R. (2016). Perceiving referential intent: Dynamics of reference in natural parent-child
699 interactions. *Cognition*, *148*, 117–135. https://doi.org/10.1016/j.cognition.2015.11.002
700 Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in
701 early lexical representation. *Journal of Memory and Language*, *50*(3), 336–353.
702 https://doi.org/10.1016/j.jml.2003.11.004
703 Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual
704 reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.
705 https://doi.org/10.1016/S0163-6383(84)80022-3
706 Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10),
707 451–456. https://doi.org/10.1016/j.tics.2004.08.006