COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 49 (2025) e70047

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70047

Modeling How Suffixes Are Learned in Infancy

Canaan M. Breiss,[a] Bruce P. Hayes,[b] Megha Sundara,[b] Mark E. Johnson[c]

[a]*Department of Linguistics, University of Southern California*
[b]*Department of Linguistics, University of California, Los Angeles*
[c]*School of Computing, Macquarie University*

Received 7 October 2024; received in revised form 31 January 2025; accepted 6 February 2025

Abstract

Recent experimental work offers evidence that infants become aware of suffixes at a remarkably early age, as early as 6 months for the English suffix *-s*. Here, we seek to understand this ability though the strategy of computational modeling. We evaluate a set of distributional learning models for their ability to mimic the observed acquisition order for various suffixes when trained on a corpus of child-directed speech. Our best-performing model first segments utterances of the corpus into candidate words, thus populating a proto-lexicon. It then searches the proto-lexicon to discover affixes, making use of two distributional heuristics that we call Terminus Frequency and Parse Reliability. With suitable parameter settings, this model is able to mimic the order of acquisition of several suffixes, as established in experimental work. In contrast, models that attempt to spot affixes within utterances, without reference to words, consistently fail. Specifically, they fail to match acquisition order, and they extract implausible pseudo-affixes from single words of high token frequency, as in [pi-] from *peekaboo*. Our modeling results thus suggest that affix learning proceeds hierarchically, with word discovery providing the essential basis for affix discovery.

Keywords: Infant language acquisition; Morphology; Suffixes; Computational modeling of language acquisition; Distributional learning; Morpheme discovery

1. Introduction

Acquiring a language involves discovering its smallest meaningful units, that is, morphemes. Morphemes can be stems (as in *mute* in *unmuting*), or they can be affixes, which in English include prefixes (like *un-*) and suffixes (like *-ing*). In many languages, affixes can

The Supplementary Materials for this article may be obtained at https://osf.io/ybvxu/?view_only=6b3cd50b9d2047e8b19f65fef92eb39f.

Correspondence should be sent to Canaan M. Breiss, Department of Linguistics, University of Southern California, 3601 Watt Way, Los Angeles, CA 90089, USA. E-mail: cbreiss@usc.edu

act as building blocks of syntax, and learning them is likely one of the first steps that infants take in acquiring syntactic knowledge. Affixes have long been studied by cognitive scientists, and their role in language comprehension and production is now well-established in the literature; for overviews, see Amenta and Crepaldi (2012) and Cohen-Goldberg (2013).

Our research is part of the general effort by linguists and other cognitive scientists to understand how infants discover these basic building blocks. In this article, we use computational modeling to help identify the mechanisms by which infants could discover the morphemes of their language. The models we consider operate on unsegmented infant-directed speech, can be fitted to existing experimental data, and generate novel testable predictions about infant behavior. Our data all involve affixes of English.

Classical work on the acquisition of affixes has typically examined production, with data coming from young children rather than infants. This research (e.g., Brown, 1973; Davies, Rattanasone, & Demuth, 2020; Golinkoff, Hirsh-Pasek, & Schweisguth, 2001; Soderstrom, Wexler, & Jusczyk, 2002; Davies, Rattanasone, & Demuth, 2017; van Heugten & Johnson, 2011) has shown that English-learning children begin to produce affixes in the second year of life. However, subsequent experiments targeting the *perception* of affixes have shown that infants have at least some knowledge of affixes long before they can produce them.

Using perception experiments, Willits, Seidenberg, and Saffran (2014) found that English-learning 7.5-month-olds are better able to notice the familiar verbs *kiss, give, drink*, and *walk* when they have been familiarized with the *-ing* form of these verbs (*kissing*, etc.). By 15 months, infants show a preference for lists of nonce words suffixed with *-ing*, but not with a pseudo-morpheme *-ot* (Mintz, 2013). Similar findings have been reported for infants learning French (Marquis & Shi, 2012).

More recent findings show that infants can be aware of affixes even earlier than this. With data from ∼500 infants, Kim and Sundara (2021), using a method similar to that of Willits et al., showed that English-learning 6-month-olds become more likely to detect the novel stem *bab* when they have been trained earlier on sentences that include the affixed form *babs*; that is, *bab-s*. At the same age, they do not detect *bab* when trained on novel words like *babsh* (*bab*, plus nonsuffix *-sh*), indicating that they are aware of the status of *-s* as a suffix. Further experiments showed that while 6-month-old infants are not aware of the suffixes *-ing* and *-ed*, they shortly do become aware of them: *-ing* at 8 months (roughly matching Willits et al.), and *-ed* somewhat later.

Given that the stems tested by Kim and Sundara are not words of English, it is unlikely that infants are detecting a suffix with reference to meaning. Instead, it is likely that infants obtain this information via some form of *distributional learning*. Since the appearance of a seminal paper by Saffran, Newport, and Aslin (1996), the application of distributional learning models to language has been a major research topic; see, for instance, the surveys in Aslin and Newport (2014) and Finley (2018). Among these learning models are models capable of detecting affixes. In this article, we assess such models according to their ability to mimic the infant developmental timeline for affix learning as attested in the experimental literature.

The research we report here blends influences often thought to be mutually opposing. In particular, we follow an increasing trend within theoretical linguistics to adopt distributional learning as an account of language acquisition.[1] In additional to working well, distributional

learning models generate highly detailed predictions that lend themselves to a traditional research ethos within linguistics, namely, close scrutiny of the details of linguistic patterns. Indeed, below we will be paying much attention to our models' treatment of specific affixes and affix candidates. From a larger perspective, we are followers of a point of view put forth by Dupoux (2018) and others, who advocate computational modeling as a way of reverse-engineering the process of language acquisition.

1.1. Evaluating the models against infant data

To achieve our purpose, we must evaluate the computational models in a nonstandard way. A typical approach in computational modeling, say, of detecting affixes, is to seek models that are very accurate at the final state, finding all and only the actual affixes of the target language. For this purpose, model evaluation is typically based on metrics such as precision, recall, and F-score. This method of model comparison tells us which models best align with (the researcher's intuitions about) adult speech segmentation, but not about which models best approximate the course of acquisition in infants. Our own interest, in contrast, is in identifying the mechanisms by which infants might discover suffixes, so we evaluate model success based on whether or not their performance successfully matches the human developmental timeline.

The models we examine are unsupervised, in the sense that they work directly from corpora; they are not given (for example) singular-plural pairs or any similar form of supervised instruction. To train our models, we chose a standard corpus of phonemically transcribed infant-directed speech and extracted from it a series of subcorpora of increasing size. We presented the series of subcorpora to each of our models, and assessed the behavior of each at each point. This procedure is meant to simulate the gradual increase in language experience as an infant gets older. As a result, we can derive developmental predictions from the models: a successful model should capture the fact that infants, receiving an ever-growing sample of data and processing it with principles assumed in the model, should become aware of the affixes in a particular order. Models are evaluated, in part, by how their own learning paths match the order in which English-learning infants are observed to discover suffixes.

With this method of evaluation, we can also make explicit predictions about acquisition order for affixes not yet experimentally studied by examining model outputs. This is necessary from both a scientific and a practical point of view. Concerning the latter, we note that experimentation with infants is very expensive in terms of both money and time; a model that can reasonably approximate the infants' developing morphological knowledge can help advance the research program by pinpointing the most potentially fruitful areas of inquiry.

1.2. Varieties of models

A key focus of this article is a comparison of models embodying two alternative architectures, instantiating distinct hypotheses about how infants learn affixes. In one architecture, affixes are treated as if they were small words, discovered on a par with the other words of the language. Since all morphemes, both words and affixes, are discovered in a single layer, we will call this the **flat** architecture. In this approach, the spoken sentence *Sammy wants out*

would be parsed as shown in (1), with the affixes [-i][2] (diminutive, spelled *-y*) and [-s] (third singular present) put on a par with full words like *out*.

(1) *Parsing* Sammy wants out *in a flat model*

a. sæmiwɔntsaʊt	*unsegmented training data*
b. sæm i wɔnt s aʊt	*output of parse*

Here, the output parse is not annotated in any way for what parts of the string actually are affixes; this task would have to be carried out, in some unspecified way, with subsequent processing.

In the other architecture we consider, affixes are treated as parts of words. The model must include a mechanism of word discovery, which establishes a **proto-lexicon**, that is, a list of possible word candidates whose meaning is not necessarily yet known. The affixes are learned by processing the entries of the proto-lexicon, using a mechanism which may be distinct from the one employed to discover proto-lexical entries from running speech. We will call this architecture the **hierarchical** approach. In the hierarchical approach, *Sammy wants out* would be parsed as in (2).

(2) *Parsing* Sammy wants out *in a hierarchical model*

a. sæmiwɔntsaʊt	*unsegmented training data*
b. sæmi wɔnts aʊt	*parsed into words*
c. sæm-i wɔnt-s aʊt	*affixes located within words*

To preview our results, we find that at least one hierarchical model (Section 8) provides a fair match for the time course of acquisition, whereas all of the flat models we were able to examine proved defective. In particular, the flat models learned the affixes in an order that fails to match that of English-learning infants (Section 7), and the models are prone to extract "pseudo-affixes" (like the hypothetical prefix *pee-* [pi-] in *peekaboo*) from frequent words (Section 7.3). In the final sections, we note some yet-unchecked predictions made by our favored model (Section 11), and address ways in which our model might be improved (Section 12).

2. The affixes under study

First, it will be helpful to be more precise about the affixes that the infants are acquiring. Following well-studied principles of English phonology (see, e.g., Pinker & Prince, 1988:Section 4.2), two of these affixes take on distinct forms in different contexts, often called *allomorphs*, which we describe below.

The affix spelled *-s* is used for plurals (*dogs*), possessives (*Ernie's*), the third singular present tense of verbs (*wags*), and various other uses. In distributional learning, where the infant is unlikely to be aware of these distinct usages, we treat all the various cases as one single class. The three allomorphs of *-s* are phonetic [-s], appearing after voiceless sounds (*cats* [kæts]); [-əz], appearing after sibilants (*kisses* [kɪsəz]); and [-z], appearing elsewhere

(*dogs* [dɔgz]). Since the allomorph [-əz] is rare, it will not be treated here; for the distinction between [-s] and [-z], see Section 9.2. We will refer to this affix (taken as an allomorph group) as [-z/-s].

The affix spelled *-ed* is used for regular past tenses (*Ernie jumped*) and past participles (*Ernie has jumped*). It, too, has three allomorphs: phonetic [-t], appearing after voiceless sounds (*kissed* [kɪst]); [-əd], appearing after [t] and [d] (*patted* [pætəd]); and [-d], appearing elsewhere (*hugged* [hʌgd]). Like [-əz], the allomorph [-əd] is rare and will not be treated here; see Section 9.2 for more on the distinction between [-t] and [-d]. We will refer to this affix as [-d/-t].

The affix spelled *-ing* ([-ɪŋ]) is used for present participles (*Ernie is jumping*) and gerunds (*Jumping is fun*). For present purposes, we will assume just one allomorph. We will refer to this affix as [-ɪŋ].

3. The experimental data

We will begin the presentation of our modeling efforts with just a subset of the data, namely, the three suffixes studied by Kim and Sundara (2021). This work collapsed the two allomorphs of [-z/-s] and [-d/-t] into single categories, testing them together in the same experiment. The goal at this modeling stage is to determine, insofar as is possible, the order in which infants acquire the three target affixes [-z/-s], [-d/-t], and [-ɪŋ]. With this accomplished, we will confront our model with additional data, involving the separate allomorphs [-z] and [-s] (Section 9.2) as well as the suffix [-i] (Section 8.6). This procedure, which reflects the order in which the research was carried out, in essence, enables us to test the model on unseen data. We will find that the model as originally set up (same parameter values) accommodates the additional data without difficulty.

The experiments we discuss all employed the Headturn Preference Procedure (Kemler-Nelson et al., 1995). In the training phase, infants were played a brief monologue that included multiple instances of a nonce (made-up) word that included the target affix, such as *babs* [bæbz]. In the test trials, the infants were played a series of repetitions of the nonce form alone (*bab* [bæb]). The amount of time an infant attended to the nonce form in the test trials served as a measure to indicate whether the infant had detected the affix.

Kim and Sundara (2021) found that 6-month-old infants familiarized with the nonwords *babs* and *dops* listened longer to *bab* and *dop* than they did for other nonwords to which they were not exposed (*kell* [kɛl] and *teep* [tip]). Others were exposed to *kells* [kɛlz] and *teeps* [tips], and listened to *kell* and *teep* longer than they did for *bab* and *dop*, which in this case served as the controls. This suggests that 6-month-olds are able to isolate the [bæb] of [bæbz] in the presence of the following [z], and likewise [dɑp] when [s] follows.

A further control was to perform a similar experiment with a nonsuffix, namely, *-sh* ([-ʃ]). This obtained a null result, suggesting that the ability of infants to detect *bab* within *babs* is not simply due to their phonological overlap (as in [bæbʃ]), but depends on the infant's knowledge that [-z/-s] is a suffix.

This basic procedure was used to study the three affixes given above ([-z/-s], [-d/-t], [-ɪŋ]). Infants were tested in a cross-sectional design at the approximate ages of 6, 8, and 10 months, and the following results were found.

(3) *The course of learning for three English suffixes*

a. 6-months: [-z/-s] is detected, [-d/-t] and [-ɪŋ] are not detected (Kim and Sundara 2021). The non-affix [-ʃ] is not detected.

b. 8-months: [-ɪŋ] is detected (Willits et al., 2014; Kim & Sundara, 2021).[4] Follow-up work by Sundara and Johnson (2024), carried out with identical methods, shows that [-d/-t] is still not detected at this age.

c. 10-months: The results of Sundara and Johnson (2024) indicate that [-d/-t] is detected by this age.

These results include gaps that can be covered only by extrapolation. In particular, we assume that infants do not regress in their knowledge, so further work would show the ability to detect [-z/-s] at 8 months and older, and [-ɪŋ] at 10 months or older. We also assume that the negative result obtained for the nonaffix [-ʃ] would continue to hold good at all subsequent ages. While we think these assumptions are reasonable, obviously, it would be helpful for the relevant experiments to be conducted in the future.

To restate our initial modeling goal: we seek to model the ordering just described, [-z/-s] > [-ɪŋ] > [-d/-t]. This ordering should emerge as the combined consequence of the model and the characteristic distribution of these suffixes in utterances accessible to English-learning infants. Later on, when we have a model in place that is able to fit this basic order, we will test the model to see if it generalizes to other suffixes and suffix allomorphs.

4. Data for training the models

For our purpose, we needed a set of English-language utterances similar to what might be heard, or overheard, by English-learning infants. The corpus must be represented in phonetic transcription, not English orthography, to better reflect young infants' perceptual experience.[4] Here, we employed the **Pearl-Brent derived corpus** (hereafter, "Pearl-Brent"), from Pearl, Goldwater, and Steyvers (2010). According to the authors, it "contains child-directed speech to children between 8 and 9 months old, consisting of 28,391 utterances (96,920 word tokens, 3,213 word types, average words per utterance: 3.4, average phonemes per word: 3.6)."[5] The original corpus was compiled by Brent and Siskind (2001), with the phonemic transcription added by Pearl et al., and was distributed on CHILDES (MacWhinney, 2000). To this corpus, we added our own annotation for which suffixes were present. We also examined the Bernstein-Ratner Corpus (Bernstein-Ratner, 1987) employed by Goldwater, Griffiths, and Johnson (2009) and other researchers; the pattern of results was similar, so we report just the Pearl-Brent corpus results here.

When training our models, we followed standard practice in letting the model have access only to full utterances, with no spaces or hyphens; hence, the model is responsible for dis-

covering the correct segmentation below the utterance level. For example, the first three utterances in the Pearl-Brent corpus are given below, both as the model sees them and in an orthographic key.

(4) *Sample sentences from the Pearl-Brent corpus*

a.	[hirðəkɪtimərgi]	*hear the kitty Morgie*
b.	[sæmiwɔntsaʊt]	*Sammy wants out*
c.	[okeðəkɪtiɪzaʊt]	*okay the kitty is out*

5. Model architectures

In choosing a model approach, we need to consider the specific situation of very young infants, who know very little about the vocabulary and structure of the language they encounter. Hence, models that obtain important results for later ages, notably Vitevich and Storkel (2012) and Jones, Cabiddu, Andrews, and Rowland (2021), are not what is needed for present purposes (for instance, these models assume input forms already segmented into words, which is something that we want the model itself to accomplish). The models that follow assume very little prestructuring of the input—it is treated solely as the phoneme streams of spoken utterances.

5.1. Two modeling strategies: Flat versus hierarchical

In this section, we specify more explicitly the contrast between flat and hierarchical models laid out in the introduction.

5.1.1. Flat models

A flat model is fed whole utterances, as in (4) above, and discovers all linguistic units at once, at the finest-grained level; hence, these units can be simple words, stems, or affixes. The question of how these undifferentiated units get arranged into appropriate hierarchical structures (affixes and stems form words, words form phrases, etc.) is a task left for later learning. To give an example, for the utterances in (5), a successful "flat" parse would be as shown; affixes such as [-i] (*-y*) and [-s] (*-s*) are placed on the same footing as words like [aʊt] *out* and [dɔg] *dog*.

(5) *Successful "flat" parses for three sentences*

a.	sæm i wɔnt s aʊt	*Sammy wants out*
b.	o no wʌn wɔnt s to gɛt drɛs t	*oh no one wants to get dressed*
c.	no it ŋ dɔg fud	*no eating dog food!*

Where can models that find such structure be obtained? We have not found computational models in the literature that deliberately seek flat structure, yet in fact, close approximations can be found if we look among models that are designed to discover words. It is a widespread

behavior of such models that alongside the words that they discover, there are also a fair number of affixes.[6] For example, the model of Johnson, Pater, Staubs, and Dupoux (2015), discussed below in Section 5.3, parses the sentence (5a) as [sæmi wɔnt s aʊt], thus finding an instance of [-z/-s]. Modelers differ as to whether they regard such parses as pernicious (Goldwater et al., 2009:39) or innocuous (Pearl & Phillips, 2018:17). From our point of view, however, it is an actual advantage that models intended to discover only words often return affixes in abundance; it is this trait that makes them usable here as flat models.

The appeal of word-discovery models for use in affix discovery is increased by the fact that some of them include parameters that govern how fine-grained a parse they return. Thus, the Johnson et al. (2015) model returns the following three parses of a representative sentence when its word-length penalty (*d*) is set to ever higher values:

(6) *Changing the word-length penalty* d *of the Johnson et al. model*

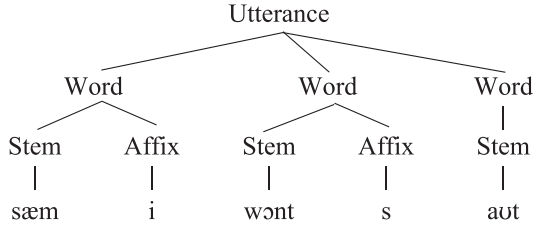
a.	Input:	aɹjufɪnɪʃtɪtɔkɪŋɔnðəfonmoɹgɪn 'Are you finished talking on the phone Morgan'
b.	$d = 1.44$:	aɹju fɪnɪʃ**t** tɔk**ɪŋ** ɔnðəfon moɹgɪn 'Areyou finish**ed** talk**ing** onthephone Morgan'
c.	$d = 1.55$:	aɹ ju fɪnɪʃ**t** tɔk **ɪŋ** ɔn ðə fon moɹgɪn 'Are you finish**ed** talk **ing** on the phone Morgan'
d.	$d = 1.64$:	aɹ ju fɪn ɪʃ **t** tɔk **ɪŋ** ɔn ðə fon moɹ gɪn 'Are you fin ish **ed** talk **ing** on the phone Mor gan'

It can be seen that at $d = 1.44$, the model detects neither [-t] nor [-ɪŋ]; at $d = 1.55$, it detects [-ɪŋ] but not [-t], and at $d = 1.64$, it detects both. Thus, for purposes of a "flat" model, as desired here, it would be sensible to pick $d = 1.64$.

5.1.2. Hierarchical models

We turn next to what we are calling hierarchical models. Such models aspire to find the full structure depicted in (7), with both words and their internal structure.

(7) *Target structure for a simple hierarchical model: "Sammy wants out"*



The process of learning in a hierarchical model is mediated by a **proto-lexicon**, defined here as the tentative list of word forms stored by the infant at any particular point of time during learning. The entries in the proto-lexicon do not necessarily have meanings attributed

to them; they may simply be distributionally detected units.[7] The idea of a proto-lexicon has come to play a major role in the acquisition and learnability literatures. For instance, Ngon et al. (2013) find that French-learning infants respond preferentially to sequences (like [tule]) that are frequent in French but are *not* words—these plausibly are erroneous proto-lexical entries that will be expunged with further learning. Oh, Todd, Beckner, Hay, and King (2023) show that non-Māori-speaking New Zealanders, by the time they reach adulthood, have internalized a large Māori proto-lexicon; they "can readily identify many more Māori words than they can define, and … the number of words they can reliably define is quite small." Feldman, Griffiths, Goldwater, and Morgan (2013) demonstrate through simulation that the discovery of phonetic categories is assisted by the compilation of a proto-lexicon, rather than attempting the same discovery using the raw speech data. The same is shown, for the discovery of classical phonemes, by Martin, Peperkamp, and Dupoux (2013). For a review of the proto-lexicon literature, see Todd, Youssef, and Vásquez-Aguilar (2023).

The proto-lexicon is at the core of our hierarchical models; it plays the same mediating role in affix discovery that it does for Feldman et al. and Martin et al. for phoneme discovery. For us, the proto-lexicon is obtained by segmenting utterances into words, and affixes are obtained by processing the proto-lexicon.[8]

For the hierarchical strategy to be effective, the segmentation of utterances into words must be conservative, discovering words per se and refraining from splitting off affixes as if they were words (since the latter task is delegated to a different portion of the model). For example, as applied to *Sammy wants out*, an optimal procedure for word discovery would aspire to output not the "flat" parse given in (5), but rather the output shown in (8).

(8) *Optimal output for the word-discovery component of a hierarchical model*

sæmi wɔnts aʊt	*Sammy wants out*
o no wʌn wɔnts tu gɛt drɛst	*oh no one wants to get dressed*
no itɪŋ dɔg fud	*no eating dog food!*

In other words, for this purpose, we should choose from among available models the ones that *least* tend to parse out affixes as separate units. For example, if were to choose among the three parameter values in (6), we here would opt for $d = 1.44$.

The second step of the hierarchical strategy is to use the candidate words from these segmentations to compile a proto-lexicon. A miniature proto-lexicon is given in (9). To illustrate the fact that words of the proto-lexicon occasionally have meanings, we include an informally stated meaning in the entry for *Sammy*. The words are listed by type, hence, there is only one

entry for [wɔnts]. Observe also that, to the extent that the word-discovery model is able to avoid the error [wɔnt s], the proto-lexicon will include inflected forms.

(9) *A partial proto-lexicon based on (8)*

{ [sæmi] 'nearby animal', [wɔnts], [aʊt] … [tu], [drɛst], [itɪŋ] … }.

In the third and final stage of a hierarchical model, the entries of the proto-lexicon are processed to discover the affixes; for example, [wɔnts] is discovered to be [wɔnt-s]. Once this is done, it is possible to combine the results of word discovery and affix discovery to provide something closer to the full structure of utterances. This is shown in (10), which assumes a fully successful parse.

(10) *Sample result from the hierarchical strategy*

a. Proto-lexicon

{ [sæm-i] 'nearby animal', [wɔnt-s], [aʊt] … [tu], [drɛs-t], [it-ɪŋ] … }.

b. Learned segmentation of the sample sentences

sæm-i wɔnt-s aʊt	*Sammy wants out*
o no wʌn wɔnt-s tu gɛt drɛs-t	*oh no one wants to get dressed*
no it-ɪŋ dɔg fud	*no eating dog food!*

What processes might work best at finding affixes within words? One possibility is to proceed recursively, reapplying the same model that discovered words to discover affixes. However, our own preliminary checks indicated that this strategy tends to perform poorly, and for this reason, we opt here for employing a distinct model (Section 8) to find affixes within words.

In evaluating the flat and hierarchical strategies, we employ standard metrics such as precision, recall, and F-score.[9] However, these metrics must be adapted to the goals of a particular model type. For flat models, we would want to determine the extent to which the boundaries posited in the model match with *any* linguistic boundary, as in (5) above. For a model using the hierarchical strategy, we are more stringent, counting its guess as correct only if it matches word boundaries to word boundaries in the real language, and morpheme boundaries to morpheme boundaries, as in (10b).

5.2. Roadmap

At this point, we have outlined our scheme to the point that we can intelligibly lay out the organization of the remainder of this article. Section 5.3 gives a description of the flat models we have employed. After the training data have been presented in Section 6, Section 7 provides an initial evaluation of the flat models. Section 8 in turn describes and evaluates hierarchical models. Section 9 tests our models against additional experimental data. In Section 10, we summarize all of the evidence, concluding that the hierarchical approach is far

more promising. In Section 11, we offer additional predictions that the hierarchical approach makes. Section 12 addresses issues for future research.

5.3. Some candidate flat models

Recall (Section 5.1) that we are readapting available models of word detection to the purpose of detecting affixes. We also want to use these models to serve as the first (word-finding) stage of a hierarchical model. To find appropriate choices, we examined as many models as we could that included downloadable software.[10] We were particularly aided in this task by the WordSeg software created by Bernard et al. (2020).

For purposes of discovering *words*, the models given in (11) performed the best, as determined by F-score.[11]

(11) *Five best models for word-discovery*

Model and source	*Word F-score*
JPSD MaxEnt (Johnson et al. 2015), $d = 1.55$	0.860
Adaptor Grammar, Phonotactic	0.780
JPSD MaxEnt (Johnson et al. 2015), $d = 1.64$	0.760
Adaptor Grammar, U-T-Phon	0.751
PUDDLE (Monaghan et al. 2012)	0.724

All models are described in the references cited, except for the Adaptor Grammars, for which see Johnson, Griffiths, and Goldwater (2006) and below.

For purposes of discovering *affixes*, the models given in (12) performed the best. Here, we calculated F-score on the basis of the 12 most frequent affixes in the corpus. Note that three of the models are also among the best for word discovery, and so appear on both lists.

(12) *Five best models for affix discovery*

Model and source	*Affix F-score*
JPSD MaxEnt (Johnson et al. 2015), $d = 1.64$	0.577
Adaptor Grammar, U-X-T-X-Phon	0.473
Adaptor Grammar, U-X-T-X-X-Phon	0.405
Adaptor Grammar, U-T-Phon	0.400
JPSD MaxEnt (Johnson et al. 2015), $d = 1.55$	0.376

The chosen models fall into three families, described below.

The model of Johnson et al. (2015), henceforth "JPSD MaxEnt" employs Maximum Entropy Optimality Theory (Goldwater & Johnson, 2003), with features referring to various phonological properties of words. As a rough approximation, it seeks the segmentation for the corpus that has the maximum likelihood. As noted earlier, the model includes a word length penalty d, which penalizes segmentations that create long words (else it would return the null segmentation).[12] A higher value for d creates segmentations that are more fragmentary,

as was shown above in (6). The value of d that yields the best F-score for word-finding in the Pearl-Brent corpus is $d = 1.55$; the value that yields the best F-score for affix-finding is $d = 1.64$.

The PUDDLE model (Monaghan & Christiansen, 2010) depends on phonotactics, the principles of legal segment sequencing. The model compiles the segment sequences present at utterance edges, and uses this information as a guide to the (presumed) phonotactics of word edges, which in turn are employed as the basis for placing word breaks. This model stands out because in our experience it never makes the error of segmenting an affix separately; thus, it would be terrible for use as a flat model, but such restraint gives it potential for use as the word-finding component of a hierarchical model.

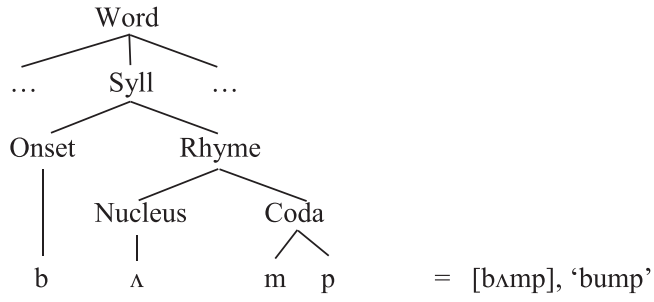
The remaining four models are Adaptor Grammars, as originally proposed by Johnson et al. (2006). Adaptor Grammars permit the analyst to prespecify a hierarchy of linguistic levels using a context-free grammar. This grammar is then augmented with probabilities attached to each production rule; at the same time, the system employs a Pitman-Yor process to accumulate a set of memorized entities at various levels, which typically embody the more frequent linguistic units.

What varies in the Adaptor Grammars we deploy is the hierarchy of prespecified levels. The top level of structure is always Utterance and the bottom level is always the sequence of phonemes. One of the intermediate levels is taken to be the Target level, which may be assessed for finding morphemes in the flat strategy or for finding words in the hierarchical strategy. We experimented with adding intermediate levels both above and below the Target level. These additional levels tend not to be linguistically meaningful (see Section 12.3), but they are included in the hope of facilitating the discovery of Target-level units. In our nomenclature for Adaptor Grammars, we specify the entire hierarchy, using "U" for Utterance, "T" for the Target level, "Phon" for the level of phonemes, and "X" for additional intermediate levels. The three models we cover here are U-T-Phon, U-X-T-X-Phon, and U-X-T-X-X-Phon.

We also tried a different Adaptor Grammar, AG Phonotactic, that was more specifically phonological in content, following Johnson (2008) and, more distantly, Coleman and Pierrehumbert (1997). It specifies a framework for phonotactic knowledge; namely that Utterances are made up of one or more Words; Words are made of one of more Syllables; and Syllables

are represented with a widely adopted internal structure (see, e.g., Zsiga, 2013:336), with Onsets, Rhymes, Nuclei, and Codas.

(13) *Syllable structure used in the AG Phonotactic model*



The system is told which segments are consonants and vowels, that all and only vowels occur in Nuclei, and that consonants must occur outside Nuclei. Segments in word-initial and word-final positions are given distinct treatment.

In Section 7, we turn to the task of empirically evaluating these five flat models.

5.4. Hierarchical models

The word-discovery capacity of the models just discussed forms one component of our hierarchical strategy, which also needs a basis for discovering affixes within words. Our discussion of hierarchical models is resumed in Section 8.

6. Modeling the course of acquisition

To model the course of suffix learning across time, we presented our models with a series of data samples of ever-increasing size, each one a subset of the full Pearl-Brent corpus, and scrutinized the results obtained by each model for each sample size. Our assumption is that the developmental bottleneck is not the need to carry out the learning calculations on known words (a task that we suspect can be accomplished swiftly), but rather the slowness with which the relevant data can be accumulated from ambient speech. We chose a set of data samples that varied in size a great deal: the smallest samples were only 1/512th of the full corpus (55 utterances), and we moved upward by powers of two: 1/256, 1/128, and so on up to 1/4, 1/2 and the full corpus (28,391 utterances). Each sample is meant as the basis for learning by the infant at a given point in time. Note that we cannot expect to equate a given sample size with a particular age of the infant; yet the models still make testable predictions, for they indicate the *relative* timing of acquisition of the various suffixes, for which we have data (Section 3).

Unsurprisingly, for the smaller sample sizes, there is a great deal of variation from sample to sample, so for each of these, we made 32 different subcorpora and averaged the modeling results. At the 1/16 sample size and larger, we simply used as many nonoverlapping slices as

could be made (16, 8, etc.). We ran our models and calculated our evaluation metrics (e.g., F-score) for each sample size, then averaged the results, thus tracking each model's behavior across simulated time.

7. Evaluating flat models

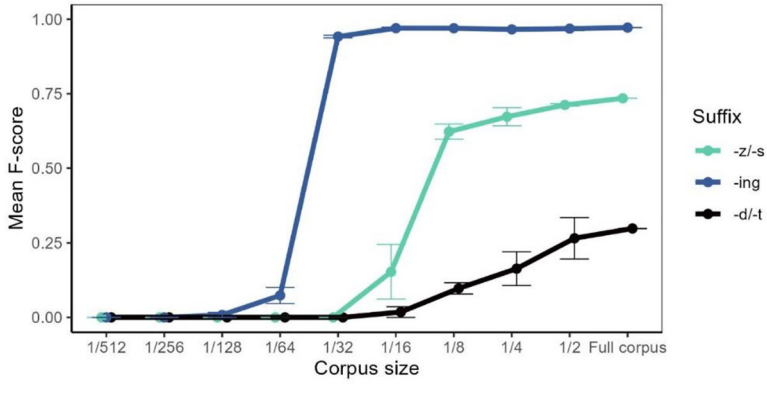
We fed each of the five models from table (12), with the series of datasets just described. A model was deemed successful, in general terms, if it learned the affixes in the same order that infants do. Recall that under the flat strategy, a model is required to learn a suffix as if it were a word. For instance, in the sentence from (5b), *Sammy wants out*, a parse of the form [sæm i wɔnt s aʊt] contributes positively to the evaluation of a model's performance (F-score) for each of the affixes [-i] and [-s].

Because of software limitations, our assessments are of the models' ability to assign affixed representations to the real words in the corpus. We use this as a proxy for what the models would do when confronted by a novel stem like [bæbz], as in the Kim−Sundara experiments.

7.1. JPSD$_{1.64}$ as a flat model

We begin with the JPSD$_{1.64}$ model, which in our preliminary survey yielded the highest full-corpus F-score for affixes. In Figure (14), the vertical axis represents the F-scores obtained for each of the three suffixes [-z/-s], [-ɪŋ], and [-d/t]. The horizontal axis gives virtual time; that is, expanding corpus size. The *x*-axis labels denote corpus size, each one a 1/*n*-sized portion of the full corpus. The plotted values represent a mean over from 1 to 32 subcorpora, as described above; error bars (in all figures) represent one standard error, pooled across allomorphs in the cases of [-z/-s] or [-d/-t].

(14) *Learning of three suffixes by the JPSD$_{1.64}$ MaxEnt flat model: F-scores*



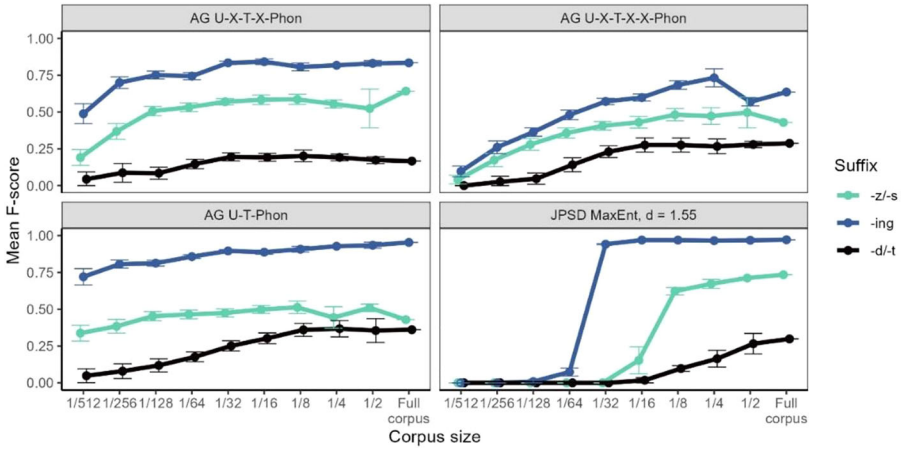
We evaluate these predicted acquisition trajectories against the developmental timeline of English affix discovery summarized in Section 3: a successful model should discover [-z/-s] first, then [-ɪŋ], and lastly [-d/-t]. In the present case, the model does indeed learn [-d/-t] last, but it fails in that it learns [-ɪŋ] before [-z/-s]. Thus, we consider it inadequate as an account of how English-learning infants learn suffixes.

We suggest that the error of learning [-ɪŋ] too early is not accidental, but is a direct consequence of the model's structure: because $JPSD_{1.64}$ includes a length penalty, it is (all else being equal) less likely to parse out a shorter string like [z] or [s] than a longer string like [ɪŋ]; in particular, shortening a candidate word by two phonemes reduces the word length penalty by more than shortening it by one phoneme.

7.2. *Other flat models*

The failure of $JPSD_{1.64}$ to learn the suffixes in the correct order is shared by the other four flat models we chose for scrutiny. Figure (15) gives time series plots analogous to Figure (14).

(15) *Learning of three suffixes by four flat models: F-score values*



Of the four models shown, AG U-X-T-X-X-Phon does the least harm in favoring [-ɪŋ] too much; perhaps this arises because its target level is further from the bottom of the a priori hierarchy than its sister Adaptor Grammar models. But even this model falls short, failing to favor [-z/-s] at any point.

In sum, among these five flat models, as well as others we have inspected but do not report here, we have found none that can account for the observed developmental timeline. A pattern widely seen among the failed models is that they learn [-ɪŋ] first, rather than the observed [-z/-s]. We think this finding makes sense, since any model that is specifically designed to discover words is likely to incorporate a bias against words that are very short, very short words being unusual. So, [-ɪŋ] is benefiting here, wrongly, from its greater length.

7.3. *The affixes that flat approaches discover*

An adequate model should not just learn the affixes that really exist, but it should also *refrain* from learning affixes that, linguistically speaking, are random nonsense. From a methodological point of view, any model that "discovered" the correct suffixes by positing a host of nonsensical pseudo-affixes would hardly be convincing.

As a way of testing the suffixes that our flat models discovered, we developed a **peripheral-position test**: intuitively, when a flat model isolates a very short sequence at the end of an

utterance, it is reasonable to treat this as the discovery of a suffix. For example, if a model divides *Uh-oh, that's Mommy's* into [ʌʔo ðæts mɑmi **z**], it is reasonable to infer that it is treating [z] as a suffix. The actual test includes a number of refinements; see Supplementary Materials B.

The test revealed that our flat models were learning quite a few putative "affixes" that, from a linguistic point of view, are bizarre. All of them were extracted from highly frequent words in the corpus, for instance, [pi-], extracted largely from the frequent word *peekaboo*—parsed by the models as [pi[kəbu]]. Here are similar cases (in all of them the frequent word is a family member): [-ri], obtained from *Henry* [[hɛn]ri]; [-di] from *Mandy* [[mæn]di]; [-ən] from *Dillon* [[dɪl]ən]; [æl-] from *Alexander* [æl[əgzændɹ]; and (from a different model) [-dɹ] from *Alexander* [[æləgzæn]dɹ]. No flat model was exempt from this kind of error, which moreover was not made by the hierarchical models discussed below.

Why do flat models discover these quirky pseudo-affixes? We cannot offer a full explanation (which doubtless would vary from model to model), but a likely cause is that these models, rather than forming a proto-lexicon, work directly with utterances—and, therefore, are based on token counts rather than type counts. We presume that this is the only mechanism that could give words like *peekaboo* or *Henry* such outsized influence.

7.4. Summary of results on flat models

The critique of flat models that emerges from our testing is twofold. First, these models consistently fail to match the acquisition order found in English-learning infants, in particular, learning bisegmental [-ɪŋ] before they learn [-z/-s], probably because they favor the discovery of relatively longer morphemes. Second, they tend to learn peripheral substrings of single frequent words (like *peekaboo*) as outlandish affixes, probably because they are based on token frequency. Together, these findings suggest a tentative conclusion: that flat models, by virtue of their very structure, are not a good approach for explaining how infants discover suffixes.

8. Hierarchical models

We turn next to the second of the two general alternatives we are exploring, hierarchical models. To review: we assume that in such a model, the sequence of utterances is initially parsed into (something approximating) its component words. These words are stored as entries in a proto-lexicon, where they are represented as types; that is, even a word of great frequency will still have just one entry in the proto-lexicon. The proto-lexicon serves as the training set for an affix-discovery module.

As noted earlier, to obtain a good proto-lexicon, it is best to pick a slightly different set of models to serve as the word-discovery front end. In Section 6, where we evaluated flat models, we made use of word-discovery models that performed well at discovering affixes. Here, we want to use a model that is specifically good at discovering only words (its original intended purpose). A list of good candidate models was given above under (11).

8.1. Models for learning affixes within words

The other task in a hierarchical model is to take the entries of the proto-lexicon and use them to learn words. For the task of finding affixes within words, we might in principle have sought out a model from the large literature (Hammarström & Borin, 2011). In this case, however, we have opted for a different strategy, namely, using only very simple models. We feel that any loss of accuracy is likely to be compensated by a clearer understanding of what data patterns matter and how the model deals with them.

The model we ultimately opt for is based on the work of Baroni (2000, 2003), where the concepts we call Terminus Frequency and Parse Reliability below are explicitly defended and implemented under a Minimum Description Length approach. Here, we will start by exploring some simple options, each of which is unsatisfactory on its own but offers some help in finding a more adequate model.

The simplest possible model we can imagine is one that says that affixes can be found because they are frequent terminal strings: if a particular phoneme string *y* occurs frequently in word-initial position, it is likely to be a prefix ([*y* [*x*]]) and if *y* frequently occurs word-finally, it is likely to be a suffix ([[*x*] *y*]). A more precise characterization of the intended criterion is given in (16):

(16) **Terminus Frequency**

Let *y* be some string of phonemes and *C* be a list of words. The fraction

$$\frac{\text{Number of entries in } C \text{ of the form } xy}{\text{Total entries in } C}$$

defines the Terminus Frequency of *y* as a candidate suffix with respect to *C*. The analogous definition, with *yx*, is used for prefixes.

We assume here that list C consists simply of the set of distinct word types in a given corpus. For the alternative of having C consist of tokens, see Sections 7.3 and 8.6.

Of course, not all instances of a candidate affix actually are affixes; some of them are "false friends." [-z] is indeed an affix in *sees* [si-z], but in *sneeze* [sniz] it is not ([sni-z], implying **snee*). Despite this, we will show that high Terminus Frequency does have considerable heuristic value in detecting affixes. In the full Pearl-Brent corpus, fully 20.2% of the word types end in [z] or [s]; 17.6% in [d] or [t]; and 8.7% in [ɪŋ].[13]

However, taken alone, Terminus Frequency is unlikely to provide the basis for a successful acquisition model. A simple argument to this effect is based on the fact that the terminal string [-ŋ], plainly not a suffix, has a higher Terminus Frequency than the real suffix [-ɪŋ], since it is present in all of the words ending in [-ɪŋ] as well as many other words such as *song* [sɔŋ]. In light of this, it is perhaps unsurprising that the Terminus Frequency values for the three affixes on which we focus, calculated across simulated time, do not match acquisition order: [-d/-t], [-z/-s], [-ɪŋ], rather than correct [-z/-s], [-ɪŋ], [-d/-t]. For calculations and time series, see Supplementary Materials C.

Consider next a different metric. At least for English, if [[*x*] *y*] is a suffixed form, then for any data corpus *C*, *C* is likely also to include *x*, the base from which [[*x*] *y*] is derived. For example, the Pearl-Brent corpus includes the suffixed forms *falling*, *feeding*, and *barking*, and we are not surprised to find that it also includes *fall*, *feed*, and *bark*. This suggests that a metric that tracks such correspondences might be useful for affix discovery.

To operationalize this idea, we will make the same assessment across all word types that end in *y*: the more often *x* is in corpus C whenever *xy* is in C, the more likely *y* is to be a suffix (and analogously for prefixes). This assessment can be expressed as a ratio, the total cases of *xy* where *x* also occurs, divided by the total cases of *xy*. We call this ratio **Parse Reliability**, giving a more careful definition in (17), which presupposes a list of word types:

(17) **Parse Reliability**

a. *Definition applicable to potential suffixes*

Let *y* be a string of phonemes.
Let *n* be the number of word types of the form *xy*.
Let *m* be calculated thus: for all word types of the form *xy*, if *x* exists, augment *m* by one.
The **Parse Reliability** of the potential suffix *y* is defined as *m*/*n*.

b. Parse reliability for potential prefixes is defined analogously.

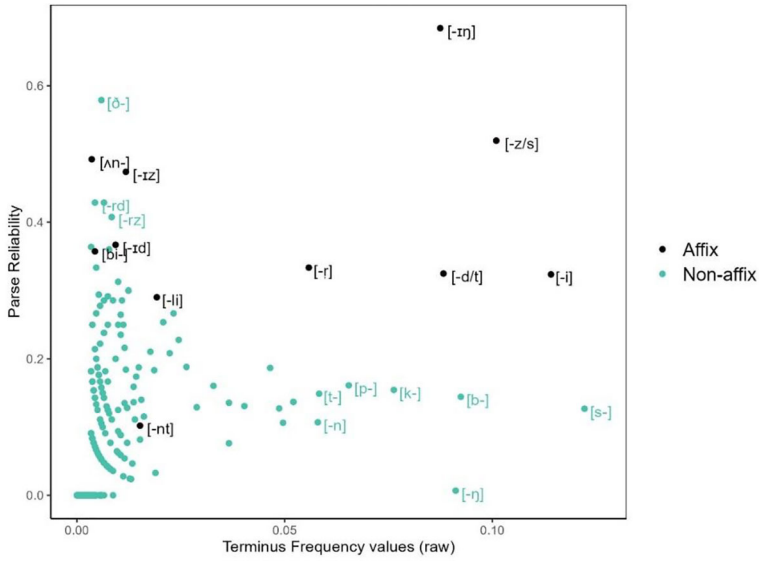
Here are some examples of how Parse Reliability works. An English word ending in [-ɪŋ] is quite likely to be suffixed, reflecting the fact that suffixed words like *jumping* greatly outnumber monomorphemic words like *lemming*. Thus, even assessed in a small corpus like the Pearl-Brent, [-ɪŋ] has a rather high Parse Reliability value, 0.684. In contrast, if *y* is a randomly chosen nonsuffix, we will find *x* within *xy* only if *xy* is a "false friend" in the sense just given. For example, the Parse Reliability of the nonsuffix [-k] is 0.136; it is above zero only because English has a modest number of false-friend pairs like *thing* [θɪŋ]—*think* [θɪŋk].

However, as before, it turns out that Parse Reliability alone cannot serve as a good basis for affix learning; it, too, learns our three affixes in the wrong order, in this case [-ɪŋ], [-z/-s], [-d/-t], rather than correct [-z/-s], [-ɪŋ], [-d/-t]; details are given in Supplementary Materials D.

Although working alone, Terminus Frequency and Parse Reliability fail to provide adequate models, we show here that they can be *combined* to provide an effective model. We first demonstrate this in qualitative terms, and then provide the full model. In Figure (18), we consider not just the primary affixes of interest here but a large set of affix candidates: we inspect every one of the 700 or so word-peripheral phoneme sequences of length 1 or 2 that occur in the words of the Pearl-Brent corpus. Considering just the full corpus (the earlier

subcorpora are considered shortly), we plot Terminus Frequency and Parse Reliability values together in the scattergram.[14]

(18) *Joint distribution of Terminus Frequency and Parse Reliability values (700 terminal phoneme sequences)*



We see that the three focus suffixes all occur in the upper right region of the figure, accompanied by two other important suffixes, adjectival/diminutive [-i] (*-y*, *-ie*) and agentive/comparative [-ɚ] (*-er*); the latter are discussed in Section 8.6. The remaining real affixes of English are much rarer, and we expect that research would show that they are not learned early. We also note that the troublesome nonsuffix [-ŋ], which has very high Terminus Frequency, also has very low Parse Reliability. The general point that emerges from Figure (18) is that *the nonaffixes occupy an L-shaped region*. Hence, neither Parse Reliability alone nor Terminus Frequency alone can single out the sequences of interest as real affixes, but in an appropriate combination, they should do fairly well.

Terminus Frequency and Parse Reliability not only single out the right affixes in the final learning state, but we also find that they provide the key to a model that predicts acquisition order correctly. In particular, when we plot the *trajectories* taken by our target affixes over simulated time, we find that they gradually emerge from the L-shaped default region into the privileged upper-right corner (see Supplementary Materials E and graph (24) below). Under a suitable choice of model parameters, we find that the target suffixes enter the upper right region in an order that matches the experimental findings. We turn to formulating a model that accomplishes this task.

8.2. An affix-finding model combining Terminus Frequency and Parse Reliability

Our model is reminiscent of MaxEnt Optimality Theory in linguistics (Goldwater & Johnson, 2003): we enumerate a set of candidate morphological parses and employ the mathemat-

ics of multinomial logistic regression to assign a probability to each candidate, based on a set of features.

For us, the set of candidate parses for any given word has up to 11 members. One candidate treats the word as unaffixed, and the remaining candidates posit a single prefix or suffix, whose length ranges from one to five segments. To give a simplified example, for the past tense form *shared*, phonetically [ʃerd], the model would allot probability to the following set of nine candidates.

(19) *The nine candidates for "shared"* [ʃerd]

Candidate	*Affix if any*	*Description*
[ʃerd]	*none*	Monomorphemic (no affix)
[[ʃer]d]	-d	correct candidate
[[ʃe]rd]	-rd	logically possible but wrong
[[ʃ]erd]	-erd	"
[[]ʃerd]	-ʃerd (null stem)	"
[ʃ[erd]]	ʃ-	"
[ʃe[rd]]	ʃe-	"
[ʃer[d]]	ʃer-	"
[ʃerd[]]	ʃerd- (null stem)	"

A longer word like *spilled* [spɪld] would have the full 11 candidates.

The output of the model is a probability distribution over each candidate set; a typical output of our system for (19) is given in (20) (for the weights employed, see (22) below).

(20) *Characteristic probabilities assigned to the 9 candidates in (19)*

[ʃerd]	0.274
[[ʃer]d]	0.726
[[ʃe]rd]	0.000
[[ʃ]erd]	0.000
[[]ʃerd]	0.000
[ʃ[erd]]	0.000
[ʃe[rd]]	0.000
[ʃer[d]]	0.000
[ʃerd[]]	0.000

It can be seen that in this case virtually all the probability is assigned to the two intuitively reasonable candidates; that is, unparsed and suffixed with [-d].

The model employs multinomial logistic regression (see, e.g., Jurafsky & Martin, in preparation, ch. 5). There are three features, of which two are the model's representations of Terminus Frequency and Parse Reliability. In addition, it is helpful also to include a "Prefer Monomorphemic" feature that serves as an intercept term, reflecting the overall eagerness or reluctance of the model to detect affixes in the data.

For any given candidate, each feature bears a *value*, calculated in the ways described below. The features also bear real-number *weights* (*w*) that express the degree to which the feature

values raise or lower the probability of a candidate. In our notation, features with positive weights lower the probability of candidates that match their description, and features with negative weights raise the probability.

The three features of our model are essentially as given above, defined more explicitly. We will use small capitals to distinguish features from the metrics they are based on.

TERMINUS FREQUENCY. The value of TERMINUS FREQUENCY for any affixed candidate is the natural log of its Terminus Frequency (i.e., type frequency), as defined in (16). It is natural to employ log frequency, since log frequencies often do better for purposes of modeling experimental data involving words, for instance, in reading time (White, Drieghe, Liversedge, & Staub, 2018), word recognition (Magnuson, Mirman, & Harris, 2012), naming (Howes, 1979), and morphological productivity (Hay & Baayen, 2003).

Here is a representative calculation. Of the 3225 word types in the full Pearl-Brent Corpus, 228, or 7.1%, end in [d], so the raw Terminus Frequency metric for the [-d] suffixed candidate is 0.071. The TERMINUS FREQUENCY value used in our model for any [-d]-suffixed candidate (i.e., of the form [[*x*]d]) is the log of 0.071, −2.65.

We define the TERMINUS FREQUENCY value for unaffixed candidates as 1. This is an arbitrary choice, since whatever value we choose will be counterbalanced by the weight of PREFER MONOMORPHEMIC.

PARSE RELIABILITY. To assign values to PARSE RELIABILITY, we use the metric defined in (17). Like TERMINUS FREQUENCY, PARSE RELIABILITY acts as a reward, and to be effective must bear a negative weight. As with TERMINUS FREQUENCY, we assign the arbitrary value 1 to unaffixed candidates.

PREFER MONOMORPHEMIC. This feature bears the value 1 for all affixed candidates and 0 for the monomorphemic candidate. In fitted models, it normally bears a negative weight, boosting the probability of the affixed candidates.

In what follows, we will frequently refer to a simple particular model based on these three features, given in (21).

(21) *Features and weights for a simple hierarchical model*

a. PREFER MONOMORPHEMIC	−28
b. TERMINUS FREQUENCY	−5
c. PARSE RELIABILITY	−13

These weights are ad hoc; they were chosen to fit the modeled data patterns of (3) when we assume all-correct real words as the output of the word discovery stage. Our intent at present is simply to provide an "existence proof"; that is, with appropriate weights, it is possible to use Terminus Frequency and Parse Reliability to derive the observed outcomes with a hierarchical model. In this respect, we claim that the hierarchical approach differs from the flat approach discussed above, in which the prospects for there being *any* feasible model appear to be bleak. For exploration of weights different from (21), and of more realistic word discovery procedures, see Supplementary Materials F and Section 8.5.

Although our findings are primarily an existence proof, we will see in Section 9.3 that model (21) does extend beyond the data to which it was fitted; it turns out to predict further experimental data that had not been gathered when we proposed the model.

We next cover how model (21) is employed to calculate probabilities for the various morpheme parses. The logistic regression calculations applied to the candidates in (20) are given in (22). Candidates 4–9 are assigned probabilities similar to [[ʃe]rd] and are omitted.

(22) *Generating the probabilities of (20) (first three candidates) using model (21)*

	PREFER MONOMORPHEMIC $w = -28$	TERMINUS FREQUENCY $w = -5$	PARSE RELIABILITY $w = -13$	linear predictor	p
1. [[ʃerd]]	0	1	1	−18	.274
2. [[ʃer]d]	1	−2.65	0.32	−18.97	.726
3. [[ʃe]rd]	1	−5.43	0.43	−6.37	∼ 0

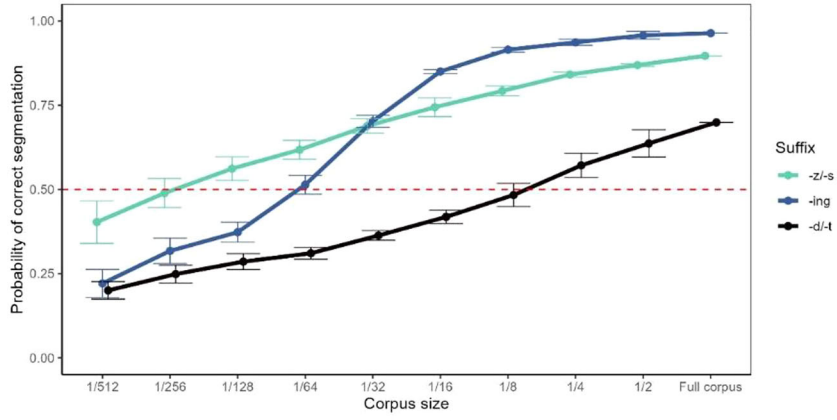
8.3. Addressing the data with the hierarchical model

As before, we fed the training data to our hierarchical model in batches, employing the same sequence of subcorpora (Section 6). As an initial demonstration of feasibility, we first use correct words as the training data instead of the output of the model's word discovery stage, turning to the latter, more realistic task in Section 8.5.

To see whether the model can mimic the behavior of infants studied experimentally, we calculate the probability it assigns to the suffixed candidate for each of the nonce stems used in the experiments, for all three suffixes (e.g., *babs*, *babbed*, and *babbing*). We suggest that it is reasonable for the model, and the infant, to conclude that *babs* is suffixed if it is the [[bæb]z] parse that receives the greatest probability. In the normal case, this arises when the probability of [[bæb]z] exceeds 50%, since usually just one other candidate, namely, unsuffixed [bæbz], is viable. In an adequate model, [-z/-s] should cross the 50% threshold first, followed by [-ɪŋ], followed by [-d/-t].

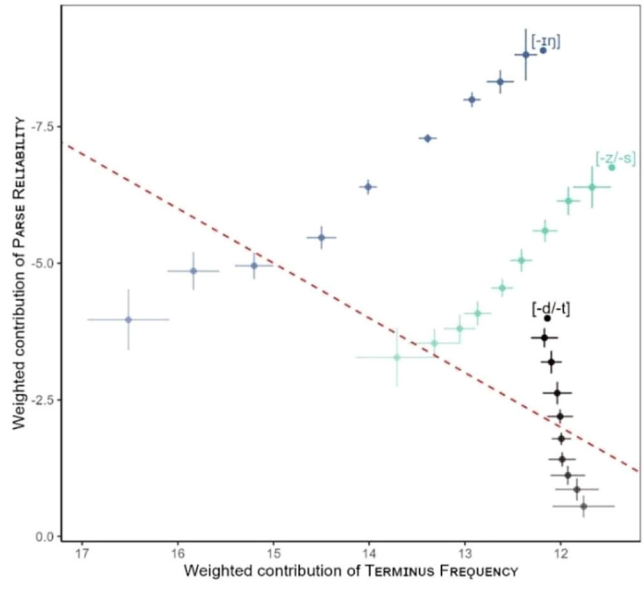
This is in fact the case for model (21); in (23), we give the model probabilities for the suffixed candidate for all three suffixes and all 10 time stages. The key point is that the three lines cross the 50% threshold at clearly distinct stages, in the correct order.

(23) *Probability of model (21) finding correct parse over simulated time (trained on real words)*



We can understand how the model succeeds more closely by examining not the output probabilities themselves, but rather the individual contributions to the linear predictor made by the features TERMINUS FREQUENCY and PARSE RELIABILITY. These are plotted in two-dimensional space, scaled by the feature weights, in (24). Each series of datapoints shows how an affix moves through this space, crossing the "50% line" at which the probability of the suffixed candidate comes to exceed that of the monomorphemic candidate.

(24) *Weighted contributions over time of* TERMINUS FREQUENCY and PARSE RELIABILITY *to the linear predictor*



Over simulated time, the three paths cross the line in the correct order [-z/-s], [-ɪŋ], [-d/-t]. In sum, the formal model (21) helps to validate our original suggestion that the properties of Terminus Frequency and Parse Reliability can be used to predict acquisition order.

8.4. Further issues

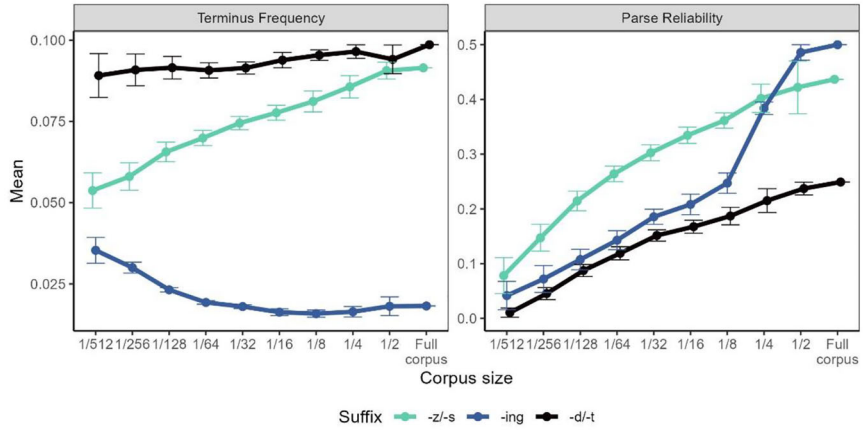
Having demonstrated the potential of the model to match the experimental data, we use the Supplementary Materials to shore up our proposal in various ways. Section G demonstrates that the Terminus Frequency and Parse Reliability values seen in the Pearl-Brent corpus are not accidental to that corpus, but are stable across a range of English corpora. Section H offers explanations for why Terminus Frequency and Parse Reliability change over time in the observed way. Section I shows that our model, in its final state, gives parses that approximately match those of a gold-standard parse of the Pearl-Brent corpus. Section J explores whether adding further predictors to the model would improve its accuracy.

8.5. A complete hierarchical model

We next present a more realistic simulation, in which the affix-finding component is fed the results of an actual word-finding model, rather than an artificially correct diet of known English words. As the word-finding model, we chose AG Phonotactic, described in Section 5.3. Among the models examined there, this model ranked second at the task of discovering words, and it was the most effective when serving as the word-finding component of a hierarchical model.

We first give the trajectories of the values for Terminus Frequency and Parse Reliability when the proto-lexicons on which they are calculated are not real English, as above, but are obtained from the AG Phonotactic model. In the Supplementary Materials (C, D), we have given the trajectories for our hierarchical model as calculated from real-word data, and they are unproblematic, providing the basis for successful learning. In contrast, the trajectories obtained using the AG Phonotactic model are a source of trouble (Figure (25)).

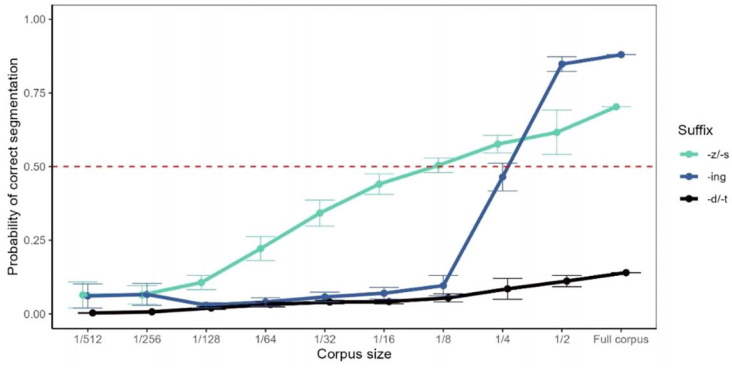
(25) *How Terminus Frequency and Parse Reliability change over simulated time, when fed the output of AG Phonotactic*



In particular, we see that [-ɪŋ] actually declines somewhat in Terminus Frequency, rather than rising to catch up with the others as it does for real words. The probable cause is that AG Phonotactic, once it has seen enough data, has a strong tendency to parse [-ɪŋ] as a separate word (about half of all [ɪŋ] sequences, at the final stage), thus reducing its frequency as a suffix and eliminating cases that could have boosted Terminus Frequency. We had chosen AG Phonotactic because, among our word-finding models, it is the most conservative about parsing [ɪŋ] as a word, but nevertheless, learning was impacted.

In spite of the misparsing problem, it emerges that this hierarchical model can be made to work, after a fashion: we dispense altogether with TERMINUS FREQUENCY (no longer useful, in light of the distortion induced by the word-discovery module) and employ only PREFER MONOMORPHEMIC (weight −9) and PARSE RELIABILITY (weight −13 as before). This turns out to derive the correct acquisition order (Figure (26)).

(26) *Probability of correct parse over simulated time—affix-discovery module fed with output of AG Phonotactic word-discovery module*



It is somewhat odd that the model should work at all, since we earlier (Section 8.1) saw that in real English, Parse Reliability alone does not suffice for an adequate model.

The points that we draw from this result are as follows. First, unlike for the flat strategy, it *is* possible to construct a hierarchical model that captures the observed acquisition order. However, the anomalies of our model point to the need for a better word-finding component, one that would refrain from parsing [-ɪŋ] so often as a word. In Section 12.3, we explore what might be needed to construct such a model. With a better word-finding component, the hierarchical approach is likely to resemble the earlier modeling with real words (Section 8.3), which worked straightforwardly.

8.6. Further affixes learned by the hierarchical model

In Section 7.3, we critiqued flat models on the basis of their poor performance in positing affixes where none exist, such as [pi-] in *peekaboo*. Here, we address the same question for our proposed hierarchical model, trained on the real words of the Pearl-Brent corpus. Table (27) lists the 10 highest-scoring affix candidates (terminal sequences, length 1 or 2) output by the hierarchical model in descending order; in this case, we separate out the allomorphs [-z]

and [-s] as well as [-d] and [-t] (see Section 9.2). The scores represent the probability that the model would assign to the affixed parse in a novel form that lacks any other plausible affix.

(27) *Top 10 affixes learned by the hierarchical model (as trained on idealized data)*

Rank	*Affix*	*Type*	*Model score*	*Status*
1.	[-z]	suffix	0.9997	real, studied above
2.	[-ɪŋ]	suffix	0.999	real, studied above
3.	[-i]	suffix	0.966	real: diminutives, denominal adjectives
4.	[-t]	suffix	0.927	real, studied above
5.	[-s]	suffix	0.906	real, studied above
6.	[s-]	prefix	0.757	pseudo-affix
7.	[-d]	suffix	0.726	real, studied above
8.	[b-]	prefix	0.492	pseudo-affix
9.	[-ɻ]	suffix	0.476	real: agentive nouns, comparatives
10.	[k-]	prefix	0.297	pseudo-affix

In sum, the list comprises all five of the focus suffix allomorphs, two additional real suffixes, and three pseudo-affixes.

The real suffix [-i], which forms diminutives (*doggy, horsie*) and denominal adjectives (*messy, stinky*), appears to be a good candidate for early acquisition. Indeed, according to the model, it should be acquired ahead of [-s], [-t], and [-d]. In Section 9.3, we will assess our model against recent experimental findings about the acquisition of this suffix.

The suffix [-ɻ] forms agentive nouns (*drummer, dancer*) and comparative adjectives (*faster, darker*). [-ɻ] has a relatively low score in (27), leading us to expect later acquisition. Nevertheless, it is in the top 10, suggesting it might be informative to study it in future experimental work.

Are there any other real affixes which we should be considering? To be sure, English has a great number of affixes, documented in scholarly work such as Marchand (1969). Almost all of these, however, are rare or have learned status; they are almost certainly acquired post-infancy. Thus, we doubt there are any well-attested affixes that our model is failing to discover.

The three pseudo-affixes to which the model assigns the highest scores are [s-] (0.757, rank 6th), [b-] (0.492, 8th), and [k-] (0.297, 10th). These are posited because of an accident of the English lexicon, which includes for each of these pseudo-prefixes a large number of false friends; defined above as morphologically unrelated pairs that coincidentally display the putative affixation pattern. For example, the Pearl-Brent corpus contains 395 words beginning with [s], of which about 140, or 35%, are the "prefixed" member of a false-friend pair; we give a few examples in (28).

(28) *Some "false friends" for the pseudo-prefix [s-]*

False friend	*"Base"*	*Wrong analysis*
stuff	*tough*	[s[tʌf]]
send	*end*	[s[ɛnd]]
snap	*nap*	[s[næp]]
sleeve	*leave*	[s[liv]]
scare	*care*	[s[ker]]

Thus, the occurrence of at least a few errors of this kind is perhaps an inevitable consequence of distributional learning; after all, absent any semantic information, there is no compelling reason to reject the hypothesis that *snap* is prefixed. If [s-] is indeed learned as a prefix at some stage of distributional learning, we conjecture that this garden path is later retraced once the meanings of, for example, *snap* and *nap* are known. What is encouraging is that pseudo-affixes are not posited in great numbers; so the cost of trimming them back later is presumably modest compared with the benefit of grasping [-z/s], [-d/t], and [-ɪŋ] early on. Another encouraging fact is that for the nonaffix studied by Kim and Sundara (2021), namely, [-ʃ] in [bæbʃ], our hierarchical model assigns a very low probability score, 0.035.

For consistency, we also submitted our hierarchical model to the same peripheral-position test that we used earlier (Section 7.3) to evaluate flat models. As before, we used a 50% probability threshold to convert the gradient results of the hierarchical model into a binary decision. This renders the hierarchical model similar to the flat models, which likewise make only up-or-down decisions. The results we obtained are given in (29).

(29) *Affixes most frequently discovered by the hierarchical model—peripheral-position test*

Rank	*Affix*	*Count*	*Status*
1	[-t]	3292	real
2	[-i]	2970	real
3	[-s]	1373	real
4	[-z]	1273	real
5	[s-]	1150	pseudo-affix
6	[-d]	979	real
7	[-ɪŋ]	776	real

In fact, only seven affixes passed the test. These are precisely the affixes of (27) that exceeded the 50% threshold for inclusion; [b-], [-ɻ], and [k-], which are missing, are the affixes that fell short of 50%. No instances were found of the pseudo-affix [-ʃ]

It is also worth reexamining the particular cases in which the flat models (Section 7.3) assigned a high probability to a suspect parse on the basis of a single high-frequency word. Our hierarchical model consistently assigns very low probability to such parses; for instance, 3.1×10^{-7} for [[hɛn]ri] in *Henry*, 1.6×10^{-6} for [pi[kəbu]] in *peekaboo*, and 1.8×10^{-6} for [[mæn]di] in *Mandy*.

In sum, we find that comparing the lists of affixes discovered by flat versus hierarchical models is informative. Both classes of model are generally able to discover our focus affixes (though not in the right order, in the case of flat models). But they differ greatly in the pseudo-affixes they learn. Our hierarchical model generally posits a pseudo-affix only when forced to do so by a substantial set of false friends. In contrast, flat models tend to posit pseudo-affixes in a different way: they are greatly influenced by individual words that have high token frequency.

The "token problem" faced by flat models can be placed in a broader perspective, obtained from earlier research that likewise seeks to model experimental results, in this case obtained from adults. Numerous studies suggest that for experimentally gathered judgments that are

based on the lexicon, the use of type frequencies in modeling usually yields more accurate results; for phonotactics, see Albright (2009), Bailey and Hahn (2001), Coleman and Pierrehumbert (1997), Hayes and Wilson (2008), and Richtsmeier (2011); and for morphology and morphophonemics, see, for example, Albright (2002), Albright and Hayes (2003), Bybee (1995, 2001), Hayes and Londe (2006), Pierrehumbert (2001), and Goldwater (2007). In light of this evidence, it would make sense that infants should likewise use type frequency when learning affixes.

Our own hierarchical model fits the pattern established in these earlier studies, in that it works better when trained with types, not tokens. To show this, we retrained model (21), letting each training word be represented by its token frequency in the Pearl-Brent corpus, rather than just once, as in our main model. Under the token-frequency training regime, the model's ability to discover affixes was substantially diminished: the average probability of the correct parse was 0.344 versus 0.636 when trained with types.

9. Testing the hierarchical model on further affixes and allomorphs

An important purpose of an implemented model is to suggest novel experiments. This indeed has occurred in our research project: our hierarchical model was set up (and its weights fitted) solely with the goal of matching the three-way acquisition order [-z/-s] > [-ɪŋ] > [-d/-t]. But with the model up and running, we found that it generated new predictions that could be tested in the laboratory.

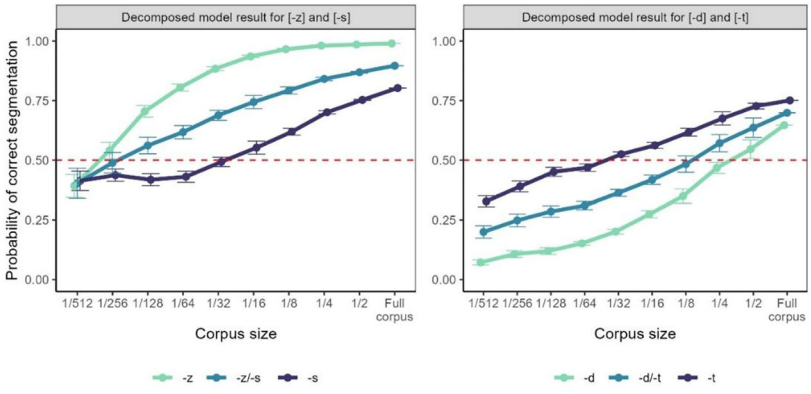
9.1. Early learning of [-i]

One prediction was that the suffix [-i] would be acquired relatively early. Sundara and Johnson (2024) undertook to investigate the acquisition of this suffix experimentally, using the same methods as before. They found that infants are able to recognize [-i] as a suffix by the age of 8 months. While this one experiment does not suffice to pin down the acquisition time precisely, the tested age of 8 months falls comfortably within the predictions of our model; see table (33) below.

9.2. Different acquisition trajectories for allomorphs

When we originally fitted them to data, our models *averaged together* the outcomes for distinct allomorphs: [-z] and [-s] are phonetically distinct but were treated as a single category, and likewise for [-d] and [-t]. We made this choice because these allomorphs were collapsed together in the experiments in Kim and Sundara (2021). However, it is possible that the different allomorphs are detected by infants at different times. (Since the infants probably do not know the meaning of the suffixes at this stage, the learning of one allomorph arguably cannot assist the learning of the others; from the infant's point of view, they are separate entities.) In fact, if we do *not* average the results, our models turn out to make quite different predictions about the two allomorphs, as seen in the graphs in (30).

(30) *Model results for [-z/-s] and [-d/-t], separating the allomorphs*



It can be seen that for most of the acquisition trajectory, [-z] is well ahead of its partner [-s], and the same holds true for [-t], leading [-d].

Thus, our model again generates predictions going beyond the data on which it was originally trained, an impetus for further empirical study. In unpublished work, Liang (2024), using methods identical to those used by Kim and Sundara (2021), shows that at 6 months, infants are able to detect [-z], but not [-s]. This result, too, falls within the range predicted by our model, as we will now show.

9.3. Putting the results together: All suffixes and allomorphs tested to date

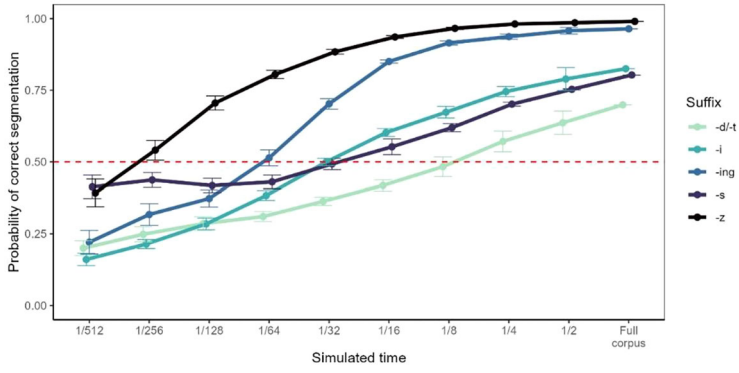
Our model (21) was trained under quite minimal conditions: it had to capture the coarse acquisition order [-z/-s] > [-ɪŋ] > [-d/-t], and its parameters were set solely with this goal in mind. The model generated predictions about the relative acquisition order of [-i] and of the suffix allomorphs [-z], [-s], [-t], and [-d], predictions which led to novel experimental work on these suffixes. The number of data points increased from 3 to 10. The full data presently available are summarized in (31).

(31) *Full data obtained to date*

Age	*Affix*	*Detected?*	*Source*
a. 6 months	[-z/-s]	yes	Kim and Sundara (2021)
b. 6 months	[-z]	yes	Liang (2024)
c. 6 months	[-s]	no	Liang (2024)
d. 6 months	[-ɪŋ]	no	Kim and Sundara (2021)
e. 6 months	[-d/-t]	no	Kim and Sundara (2021)
f. 6 months	[-ʃ]	no	Kim and Sundara (2021)
g. 8 months	[-ɪŋ]	yes	Willits et al. (2014); Kim and Sundara (2021)
h. 8 months	[-d/-t]	no	Sundara and Johnson (2024)
j. 8 months	[-i]	yes	Sundara and Johnson (2024)
i. 10 months	[-d/-t]	yes	Sundara and Johnson (2024)

We can assess how model (21) covers the new data by examining the trajectories it creates; these are shown in (32).

(32) *Predicted probability over time for correct parse under model (21)—full affix set*



As will be recalled, the key values extracted from these curves represent the point in time at which the probability assigned by the model comes to exceed 50%, so that the affixed parse becomes the model's majority preference. These, then, are the key predictions. In (33), we offer a synoptic view showing the compatibility of the model predictions with the full data. The words "yes" and "no" correspond to the data points given earlier in table (31). Each horizontal bold line represents the point at which model (21) crosses the 50% line for a particular affix, as seen in (32); vertical bars represent 95% confidence intervals. The vertical axis labels on the right side of the chart give a conjectured affiliation of corpus size with chronological age of experimental participants; for example, 1/128 is suggested as being affiliated with 6 months.

(33) *Synoptic view: experimental results compared with the predictions of model (21)*

Corpus size	[-z]	[-z/-s]	[-ɪŋ]	[-i]	[-s]	[-d/-t]	Age of infants tested
1/512							
1/256							
1/128	yes	yes	no		no	no	6 months
1/64							
1/32							
1/16			yes	yes		no	8 months
1/8							
1/4							
1/2						yes	10 months
Full corpus							

It can be seen the model is consistent with all of the currently available experimental data. The boldface lines fall in an orderly pattern, separating the cases of "not yet discovered" from "discovered" for all the affixes in question.

While we have emphasized the chronology of discovery, in the end of course our model must be evaluated against its ability to model all of the experimental findings together. Our point is methodological: the existence of an explicit model, based on limited data, led to appropriate experiments for further testing. So far, the model is compatible with all available experimental findings.

10. Discussion

10.1. Model architectures (hierarchical vs. flat, type vs. token)

With regard to the hierarchical modeling strategy, our studies have demonstrated *feasibility*: at least one model (Section 8.5) is able to capture the experimental observations summarized in (33). Our studies do not demonstrate *inevitability* of correct learning, since they depend on a particular setting of the model weights, but the qualitative patterning of our key metrics Terminus Frequency and Parse Reliability, shown in (24), demonstrates that our approach is feasible in principle.

In contrast, the flat strategy did not work well at all, irrespective of which particular model was employed. We have diagnosed this failure in two ways.

First, the flat models all proved (Section 7) to be overeager in detecting [-ɪŋ], wrongly learning it ahead of [-z/-s]. We conjectured that this error resulted from the greater resemblance of [-ɪŋ] to real words; it is longer in segment count than [-z/-s] and [-d/-t], and for the AG Phonotactic model, it is also relevant that [-ɪŋ] contains a vowel. In general terms, we think that looking for affixes using the same criteria employed to look for words is a likely source of error.

Second, the flat models fail because they learn from tokens, not types. This led to numerous predicted parses that we think are almost certainly erroneous, namely, the extraction of short substrings as affixes from single, highly frequent words, as in *[pi[kəbu]] *peekaboo* (Section 7.3); this type of error is avoided by the type-based hierarchical model, which treats all the tokens of *peekaboo* as a single data point. Beyond this, the hierarchical model itself does far better at finding the real suffixes when it is trained on types rather than tokens (Section 8.6). Under the hierarchical model, type-based learning has a natural origin, since the information used for affix discovery is not the incoming speech stream (as it must be for flat models), but the entries in the proto-lexicon. The plausibility of the hierarchical model is increased by independent evidence for the existence of a proto-lexicon in infants (Section 5.1), as well as evidence that adult speakers also apprehend lexical generalizations on a type, not token basis (Section 8.6).

10.2. "Reverse engineering" in acquisition research

Scholars such as Dupoux (2018) and Guest and Martin (2021) have made the case for the role of computational modeling in cognitive science. In every area, researchers confront the

problem that knowledge and computation within the human brain are concealed. The "reverse engineering" approach advocated by these authors is one way to confront the difficulty of research under such conditions.

In the present context, we judge that modeling work can, at the very least, serve as a basis for reducing uncertainty: it helps us to rule out approaches that are unlikely to be correct. For instance, we judge that what real children do in learning morphology is less likely to correspond to a flat model than to a hierarchical model, because even the best available flat models yield predicted patterns that are quite distinct from those observed in children. For the future, it is reasonable to hope that the development of a sequence of ever more-refined models, carried out in tandem with model-guided experimentation, will lead to a more precise understanding of what is happening in humans.

11. Predictions for future research

Only a few of the predictions made by our model have been tested; here we cover some others.

11.1. Suffixes learned as prefixes?

We have seen repeatedly that flat models tend to learn the suffix [-ɪŋ] as if it were a word. If, contrary to what we have claimed, infants likewise treat [ɪŋ] as a word, we might expect them to accept [ɪŋ] in a context where suffixes are not permitted—namely, at the beginning of an utterance. Suppose, then, that we devised an experiment in which infants of appropriate age are familiarized with passages that contain (for instance) an utterance-initial nonword of the form [ɪŋsum] *ingsoom*. We suggest that if, in the test phase, the infants recognize isolated *soom* [sum] as a word, this would constitute evidence that they have detected [ɪŋ] using something akin to a flat model—they cannot be using a hierarchical model of the type we advocate, for such models distinguish between prefixes and suffixes,[15] and the distributional evidence to support [ɪŋ] as a prefix is very weak.[16]

On the other hand, if infants do *not* detect *soom* in *ingsoom*, the result is ambiguous, for there are multiple explanations. One possibility is that, just as we have suggested, affixes are learned with a hierarchical model, which induces not just the segmental content of [-ɪŋ] but also the fact that it occurs directly after the stem. However, this result would not rule out *all* flat models, because infants might be learning a positional restriction, essentially syntactic in character, that prevents the "word" [ɪŋ] from occurring sentence-initially (a real-life model of this kind is the bigram model in Goldwater et al., 2009).

11.2. Further affixes to be investigated

The discussion above (see (27)) indicates that agentive/comparative [-ɻ] should be learned somewhat after the [-d] allomorph of [-d/-t]; this prediction has yet to be tested. No other affixes are predicted to be learned early. This itself is a model prediction: if (for example) an experiment showed that adverbial *-ly* [-li] or negative *un-* [ʌn-] were learned as early as [-d/-t], this would constitute counterevidence to the model.[17]

It might also be profitable to look at whether any of the pseudo-affixes learned by the models (notably [s-]; (27)) are learned as errors by infants—while we doubt that this would be the case, if such mislearning is ever detected, it would constitute very persuasive evidence for the general hypothesis of distributional learning.

A special case of pseudo-affixes are the ones extracted by flat models from frequent single words, such as [pi-] from *peekaboo*. We anticipate that no experimental evidence will ever be found to support the view that infants extract such affixes from the signal. If such evidence were found, it would constitute a powerful argument in favor of models that (like our flat models) depend on token frequency.

Beyond this, our hierarchical model makes general predictions about order of acquisition that go beyond the specific cases studied here. These predictions are not always what we might imagine in advance of modeling. Although it seems intuitive that infants should first learn affixes that correspond to frequent terminal strings, according to our model, this is not the whole story. An affix like [-ɪŋ], which has relatively low Terminus Frequency but excellent Parse Reliability, can be learned early. Moreover, we predict that strings like final [-ŋ], with excellent Terminus Frequency but negligible Parse Reliability, should not be acquired as suffixes.

11.3. Adult intuitions

People learn new words throughout their lives (Schwartzman et al., 1987; Ramscar, 2022), and many of these words have morphological structure—or sound like they might. It seems possible that the metrics of Terminus Frequency and Parse Reliability that we use might be detected in the behavior of adult speakers queried on the morphological status of novel words. For instance, [ˈtrɛpɪŋ] might be taken as the present participle of *trep*, or alternatively, as a (monomorphemic) surname (*Trepping*). We can imagine similar probes based on [-z/-s], like [spɔɪz]; and [-d/t], like [flaʊd].

Our hierarchical model, trained directly on the Pearl-Brent corpus, predicts that [-ɪŋ] forms should be perceived as bimorphemic more often than [-z/-s] forms, which in turn should be perceived as bimorphemic more often than [-d/-t] forms (see right edge of curves in (23)). We are unaware of any existing findings that test this specific prediction, but a substantial body of work has addressed similar questions in the domain of real words (Baroni, 2003; Creemers et al., 2023; Hay & Baayen, 2002, 2003). Also, tests of Terminus Frequency and Parse Reliability might be conducted by examining their performance against data from studies on artificial-language material, as in Lelonkiewicz, Ktori, and Crepaldi (2020).

12. Issues for further work

12.1. Generalizing Parse Reliability to languages with bound stems

The key element of Parse Reliability as implemented in our hierarchical model is its method of stem detection: to assess whether *x* is the morphological stem of [[*x*] *y*], our model searches the proto-lexicon for *x* as an independent word. This procedure is well suited to English, since

in English the morphological stem of a word virtually always occurs as a free form. However, in many languages, stems are not so easily found. For example, in Spanish, most stems are *bound*; for example, the word for "admire" has many different inflected forms (*admir-ar*, *admir-o*, *admir-as*, *admir-a*, etc.), every one of which is affixed; **admir* is not a legal form. To address such cases, the hierarchical model needs to be provided with a more general version of Parse Reliability: the parse [[*x*] *y*] must somehow be supported by evidence from other words (e.g., [[*x*] *z*]), and not just by the isolation form *x*. It is possible that an appropriate way to generalize Parse Reliability already exists, but we leave this issue to future research.

12.2. Beyond phonemic transcription

Our training data are given in phonemic transcription, a classical form of phonological representation widely employed in linguistics; it represents words in an idealized form from which the overt phonetic forms are in principle predictable by rule up to the point of free variation. However, in future research, it would be more realistic to employ a less-idealized phonetic transcription, incorporating token-specific detail. As Beech and Swingley (2023) demonstrate, this is likely to be a far greater challenge; the incorporation of a transcription that reflects low-level phonological processes make word detection harder. In addition, for our own purposes, it would also affect Parse Reliability values. For example, if the word *petting*, phonemically /pɛtɪŋ/, is pronounced with the tapped [ɾ] variant of /t/, as in North American English [ˈpɛɾɪŋ], then the corresponding hypothesized base will be [ˈpɛɾ], mismatching the isolation form [ˈpɛt]. Thus, a challenge for future work is to do the same sort of learning we have attempted, but with a corpus that is phonetically transcribed throughout (e.g., Khlystova, Chong, & Sundara, 2023; Pitt et al., 2007); or still more ambitiously, a corpus of speech waveforms.

12.3. Moving to joint learning

In our hierarchical model, learning proceeds sequentially: first, the utterances are divided into words, then the words are divided into morphemes, including affixes. This arrangement is not a logical necessity; we might alternatively have pursued **joint learning**, in which both learning tasks are addressed at once with information flow between the two. The joint learning strategy has proven effective in other domains (Dillon, Dunbar, & Idsardi, 2013; Feldman et al., 2013; Goldwater, 2018; Kemp, Perfors, & Tenenbaum, 2007; Martin et al., 2013).

We give an example showing why joint learning might be helpful here. In Section 8.5, we noted a serious problem for our hierarchical strategy involving [-ɪŋ]: all the word-discovery models we examined tended to parse [ɪŋ] as a separate word. This deprives the affix-discovery component of the [ɪŋ]-containing words it needs to learn the suffix [-ɪŋ] on schedule. In principle, we would like it to be the case that evidence arising during affix discovery (e.g., the high Parse Reliability score for [-ɪŋ]) could be useful at the word level: the very fact that [-ɪŋ] works well as a suffix should discourage the model from treating it as a word; this can be done, in principle, under joint learning, but not in our serial arrangement. Hence, it seems worth exploring the issue of joint learning further.

We note that Adaptor Grammars, which we have here used solely as the basis of word-discovery (Section 5.3), do in fact implement the joint-learning strategy, for they learn all of the user-defined levels in parallel. Unfortunately, we have not found a way to deploy this approach in a way that yields effective results. Our Adaptor Grammars can be set up to do fairly well at finding the content of one particular level (word or morpheme); but when they do this, the material found at the other levels emerges as linguistically nonsensical, not the meaningful level that we had been seeking.

To achieve a better joint model in the future, we suggest the following four practices may be helpful. First, as argued above (Sections 7.3 and 8.6), forming a proto-lexicon, and using it—not the raw input data—as the basis for learning morphology is likely to be more effective. Second, using distinct principles to carry out word detection versus affix detection (Section 7.2) is advisable. Third, we think it would be beneficial for the joint learning system to incorporate phonotactic learning as well, as in Adriaans and Kager (2010). Languages impose phonotactic restrictions at both the word level and the stem level, which can serve as cues for the learning of words and morphemes. For instance, the fact that [dl] is impossible stem-finally can be used to rule out morphological parses that have an impossible stem, for instance, *[[sidl]ɪŋ] for *seedling*. For word learning, it is well established that infants can make use of phonotactic information (e.g., Friederici & Wessels, 1993; Jusczyk, Houston, & Newsome, 1999; Mattys, Jusczyk, Luce, & Morgan, 1999; Johnson & Jusczyk, 2001; Mattys & Jusczyk, 2001; Gonzalez-Gomez & Nazzi, 2013). Fourth, after a certain point, the infant is abstracting syntactic knowledge from the input stream. With joint learning, this could also assist in the learning of affixes. For example, learning the syntactic category *verb* improves the basis for detecting [-ɪŋ] as a suffix, since [-ɪŋ] generally attaches to verbal stems.

Putting all these elements together, we suggest a future architecture for joint learning. We imagine an expanding proto-lexicon whose entries are annotated (and reannotated) for their morphological structure. The proto-lexicon is the domain from which the infant extracts crucial generalizations involving morphology and phonotactics. In parsing the input stream, the infant makes use of *all* her knowledge (lexical, morphological, phonotactic, syntactic) in finding the most likely parse, and uses the result to update her proto-lexicon. Further, as lexical entries are installed and updated, the infant's internal models of morphology, phonotactics, and syntax are likewise updated, and so the virtuous cycle continues.

Acknowledgments

We would like to thank Ocke-Schwen Bohn, Katherine Demuth, Emmanuel Dupoux, Naomi Feldman, Jon Gauthier, Laurel Perkins, our anonymous reviewers, and talk audiences in the Department of Linguistics at UCLA and the Department of Brain and Cognitive Sciences at MIT for helpful input in the preparation of this work. The research was supported by funding from NSF grant 2024498 to Megha Sundara and Bruce Hayes and by support from the UCLA Academic Senate.

Ethics statement

The work presented here does not involve human subjects research and, therefore, did not require ethics approval at our universities. No known conflicts of interest were involved. All figures were created by us and we grant permission for their use.

Notes

1 Some possible examples: Pearl and Sprouse (2013) use syntactic trees to distributionally learn "island constraints," and Wilson and Gallagher (2018) use distinctive features to distributionally learn phonotactics from sparse data.
2 We describe linguistic forms using the symbols of the International Phonetic Alphabet (https://www.internationalphoneticassociation.org/IPAcharts/inter_chart_2018/IPA_2018.html).
3 Mintz (2013) did *not* find evidence for learning of [-ɪŋ] in 8-month-old infants. This may have been because in the Willits et al., and Kim and Sundara studies, the targets were short and phonologically very simple.
4 Our use of a phonetically transcribed corpus follows mainstream practice, although the degree to which infants have segmental representations is an open research question (e.g., Jusczyk & Derrah, 1987). In principle, our research might be compatible with infants having presegmental representations, since we could make use of models of word segmentation that are based on the raw speech signal (e.g., Park & Glass, 2008; Kamper, Elsner, Jansen, & Goldwater, 2015; Algayres et al., 2022). However, the relative computational simplicity of segmentally based segmentation leads us to prefer it here.
5 For further counts, including of individual suffixes, see Section 8.1.1.
6 Such behavior become even more likely when the language being parsed has a richer morphological structure than English; see Loukatou, Stoll, Blasi, and Cristia (2022).
7 To be sure, the words of the proto-lexicon sometimes do bear meanings, as demonstrated by Bergelson and Swingley (2013). However, according to Bergelson and Swingley, the number of meaningful words is probably rather few (around 20 or so by 14 months of age); such numbers do not support morphological learning in any of the models we have considered.
8 For whether these processes operate sequentially, as the text above implies, or rather in parallel, see Section 12.3.
9 Precision is standardly defined as *correct guesses*/*guesses*, recall as *correct guesses*/*correct answers*, and F-score as the harmonic mean of precision and recall. For F-scores, in cases where recall was zero (no instances found), we entered a value of zero for precision, rather than Undefined (0/0). We likewise assigned an F-score of zero by fiat when both precision and recall came out as zero.
10 The full set of 27 models that we examined is given in the Supplementary Materials, Section A.

11 The F-scores in (11) and (12) can only be taken as a rough evaluation of the models, since the models differ in the amount of prior knowledge they assume—AG Phonotactic, for instance, relies on a particular preinstalled theory of syllable structure. There seems to be no straightforward way to include such factors in our evaluation, since the preinstalled content used by the various models is so different in character.
12 The research goal of Johnson, Pater, Staubs, and Dupoux (2015) was to discover both words and phonological processes simultaneously; since the latter are beyond the scope of this article, we turned off phonology for purposes of our learning simulations.
13 These summed values are not used in at this stage of modeling; rather, we analyze [-z] and [-s], and [-d] and [-t], separately and average the results. The Terminus Frequency values for the individual suffixes are: [-z] 12.7%, [-s] 7.5%, [-d] 7.1%, and [-t] 10.6%. The corpus type count is 3213.
14 Fewer than 700 data points are visible, since a great number cluster at or near the origin.
15 As do adults: Crepaldi, Rastle, and Davis (2010) show that stem effects on lexical decision (participants slow to reject [[*gas*]*ful*] as a word) evaporate when the affix is in the wrong position, as in [*ful*[*gas*]].
16 On the scattergram of (23), its coordinates are (0.0006, 0).
17 Of course, *-ly* and *un-* are learned eventually. However, since the distributional evidence for them is so weak, it seems likely that they are learned later on, when the child is able to make use of meaning.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311–333.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Albright, A. (2002). Islands of reliability for regular morphology: Evidence from Italian. *Language*, *78*, 684–709.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41.

Algayres, R., Ricoul, T., Karadayi, J., Laurençon, H., Zaiem, S., Mohamed, A., Sagot, B., & Dupoux, E. (2022). DP-Parse: Finding word boundaries from raw speech with an instance lexicon. *TACL*, *10*, 1051–1065.

Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in Psychology*, *3*, 232.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, *64*, 86–105.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*, 568–591.

Baroni, M. (2000). Distributional cues in morpheme discovery: A computational model and empirical evidence. Ph.D. dissertation, UCLA, Los Angeles, CA.

Baroni, M. (2003). Distribution-driven morpheme discovery: A computational/experimental study. In G. Booij & J. Marle (Eds.), *Yearbook of Morphology 2003* (pp. 213–248). Dordrecht: Springer.

Beech, C., & Swingley, D. (2023). Consequences of phonological variation for algorithmic word segmentation. *Cognition*, *235*, 105401.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., & Fibla, L. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*, 264–278.

Bernstein-Ratner, N. (1987). The phonology of parent–child speech. In K. E. Nelson & A. van Kleeck (Eds.), *Children's language* (Vol. 6) (pp. 159–174). Hillsdale, NJ: Erlbaum.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 31–44.

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

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Cohen-Goldberg, A. M. (2013). Towards a theory of multimorphemic word production: The heterogeneity of processing hypothesis. *Language and Cognitive Processes*, *28*, 1036–1064.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 49–56). Somerset, NJ: Association for Computational Linguistics.

Creemers, A., Chanchaochai, N., Tamminga, M., & Embick, D. (2023). The activation of embedded (pseudo-) stems in auditory lexical processing: Implications for models of spoken word recognition. *Language, Cognition and Neuroscience*, *38*(7), 966–982.

Crepaldi, D., Rastle, K., & Davis, C. J. (2010). Morphemes in their place: Evidence for position-specific identification of suffixes. *Memory & Cognition*, *38*, 312–321.

Davies, B., Rattanasone, N. X., & Demuth, K. (2017). Two-year-olds' sensitivity to inflectional plural morphology: Allomorphic effects. *Language Learning and Development*, *13*(1), 38–53.

Davies, B., Rattanasone, N. X., & Demuth, K. (2020). Acquiring the last plural: Morphophonological effects on the comprehension of /–əz/. *Language Learning and Development*, *16*, 161–179.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, *37*, 344–377.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

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*(4), 751.

Finley, S. (2018). Cognitive and linguistic biases in morphology learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*, e1467.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, *54*(3), 287–295.

Goldwater, S. (2018). Perceptrons and syntactic structures in models of language acquisition. Talk given at the Society for Computation in Linguistics. Salt Lake City.

Goldwater, S. (2007). Nonparametric Bayesian models of lexical acquisition. Doctoral dissertation, Brown University.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory* (Vol. 111, p. 120).

Golinkoff, R. M., Hirsh-Pasek, K., & Schweisguth, M. (2001). The reappraisal of young children's knowledge of grammatical morphemes. In J. Weissenborn & B. Hoehle (Eds.), *Approaches to bootstrapping: Phonological, syntactic and neurophysiological aspects of early language acquisition*. (pp. 167–188). John Benjamins.

Gonzalez-Gomez, N., & Nazzi, T. (2013). Effects of prior phonotactic knowledge on infant word segmentation: The case of non-adjacent dependencies. *Journal of Speech, Language, and Hearing Research*, *56*, 840–849.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*, 789–802.

Hammarström, H., & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, *37*(2), 309–350.

Hay, J., & Baayen, H. (2002). Parsing and productivity. In G. Booij & J. Marle (Eds.), *Yearbook of Morphology 2001* (pp. 203–235).

Hay, J., & Baayen, H. (2003). Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, *15*, 99–130.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, *23*(1), 59–104.

Howes, D. (1979). The naming act and its disruption in aphasia. In D. Aaronson & R. Rieber (Eds.), *Psycholinguistic research (PLE: Psycholinguistics)* (pp. 449–484). Psychology Press.

Johnson, K. E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL*-08 (pp. 398–406). Association for Computational Linguistics.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems 19* (pp. 641–648).

Johnson, M., Pater, J., Staubs, R., & Dupoux, E. (2015). Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 303–313).

Jones, G., Cabiddu, F., Andrews, M., & Rowland, C. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, 104232.

Jurafsky, D., & Martin, J. H. (in preparation). *Speech and language processing*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Jusczyk, P., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*, 648.

Jusczyk, P., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5818–5822).

Kemler-Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*, 111–116.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

Kim, Y., & Sundara, M. (2021). 6–month–olds are sensitive to English morphology. *Developmental Science*, *24*, e13089.

Khlystova, E., Chong, A. J., & Sundara, M. (2023). Phonetic variation in English infant-directed speech: A large-scale corpus analysis. *Journal of Phonetics*, *100*, 101267.

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language*, *115*, 104152.

Liang, K. (2024). The influence of phonotactics on suffix discovery in infancy. UCLA M.A. thesis. Retrieved from https://linguistics.ucla.edu/research/masters-theses/

Loukatou, G., Stoll, S., Blasi, D., & Cristia, A. (2022). Does morphological complexity affect word segmentation? Evidence from computational modeling. *Cognition*, *220*, 104960.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, M. Joanisse, & K. McRae (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 76–103).

Marchand, H. (1969). *The categories and types of present-day English word-formation*. München: Beck.

Marquis, A., & Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition*, *122*(1), 61–66.
Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*, 103–124.
Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*(2), 91–121.
Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465–494.
Mintz, T. (2013). The segmentation of sub-lexical morphemes in English-learning 15-month-olds. *Frontiers in Psychology*, *4*, 1–12.
Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564.
Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*, 24–34.
Oh, Y. M., Todd, S., Beckner, C., Hay, J., & King, J. (2023). Assessing the size of non-Māori-speakers' active Māori lexicon. *PLoS ONE*, *18*, e0289669.
Park, A. S., & Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 186–197.
Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*(2–3), 107–132.
Pearl, L., & Phillips, L. (2018). Evaluating language acquisition models: A utility-based look at Bayesian segmentation. In T. Poibeau, A. Villavicencio, L. Pearl, & L. Phillips (Eds.), *Language, cognition and computational models* (pp. 185–224). Cambridge: Cambridge University Press.
Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*, 23–68.
Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1–2), 73–193.
Pierrehumbert, J. (2001). Stochastic phonology. *GLOT*, *5*, 1–13.
Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech. Columbus, OH: Department of Psychology, Ohio State University.
Ramscar, M. (2022). Psycholinguistics and aging. *Oxford research encyclopedia of linguistics*. Retrieved from https://oxfordre.com/linguistics/
Richtsmeier, P. T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, *2*(1), 157–183.
Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
Schwartzman, A. E., Gold, D., Andres, D., Arbuckle, T. Y., & Chaikelson, J. (1987). Stability of intelligence: A 40-year follow-up. *Canadian Journal of Psychology*, *41*, 244–256.
Soderstrom, M., Wexler, K., & Jusczyk, P. W. (2002). English-learning toddlers' sensitivity to agreement morphology in receptive grammar. In B. Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 6th Annual Boston University Conference on Language Development* (Vol. 2, pp. 643–652). Cascadilla.
Sundara, M., & Johnson, M. (2024). Discovering inflectional and derivational suffixes in infancy. In Presentation at Boston University Conference on Language Development.
Todd, S., Youssef, C. B., & Vásquez-Aguilar, A. (2023). Language structure, attitudes, and learning from ambient exposure: Lexical and phonotactic knowledge of Spanish among non-Spanish-speaking Californians and Texans. *PLoS ONE*, *18*, e0284919.
Van Heugten, M., & Johnson, E. K. (2011). Gender-marked determiners help Dutch learners' word recognition when gender information itself does not. *Journal of Child Language*, *38*, 87–100.
Vitevich, M. S., & Storkel, H. L. (2012). Examining the acquisition of phonological word forms with computational experiments. *Language and Speech*, *56*(4), 493–527.

White, S. J., Drieghe, D., Liversedge, S. P., & Staub, A. (2018). The word frequency effect during sentence reading: A linear or nonlinear effect of log frequency? *Quarterly Journal of Experimental Psychology*, *71*, 46–55.
Willits, J. A., Seidenberg, M., & Saffran, J. R. (2014). Distributional structure in language: Contributions to noun-verb difficulty differences in infant word recognition. *Cognition*, *132*, 429–436.
Wilson, C., & Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry*, *49*, 610–623.
Zsiga, E. C. (2013). *The sounds of language: An introduction to phonetics and phonology*. Wiley.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary information