

## Further Documentation and Discussion

Supplementary Materials to Breiss et al., “Modeling how suffixes are learned in infancy”

### A. Full list of flat models

This list augments the list of five flat models discussed in §5.3 of the main article.

Model and source	Word F-score	Affix F-score
<i>JPSD Maxent</i> (Johnson et. al 2015), $d = 1.55$	0.860	0.376
<i>Adaptor Grammar</i> , Phonotactic (see main text)	0.780	0.191
<i>JPSD Maxent</i> (Johnson et. al 2015), $d = 1.64$	0.760	0.577
<i>Adaptor Grammar</i> , U-T-Seg (see main text)	0.751	0.400
<i>PUDDLE</i> (Monaghan et al. 2012)	0.724	0.028
<i>BatchOpt</i> learner, bigram (Philips and Pearl 2015)	0.699	0.347
<i>DiBS</i> (Daland and Pierrehumbert 2011)	0.688	0.080
<i>JPSD Maxent</i> (Johnson et. al 2015), $d = 1.44$	0.668	0.048
<i>Adaptor Grammar</i> , U-X-T-X-X-Seg (see main text)	0.661	0.405
Goldwater et al. (2009) — unigram	0.633	0.227
Goldwater et al. (2009) — bigram	0.632	0.310
<i>BatchOpt</i> learner, unigram (Philips and Pearl 2015)	0.622	0.291
<i>Adaptor Grammar</i> , U-X-T-X-Seg (see main text)	0.616	0.473
<i>Adaptor Grammar</i> , U-T-X-seg (see main text)	0.614	0.109
<i>Transitional Probability</i> (Saksida 2017) <sup>1</sup>	0.529	0.313
<i>OnlineOpt</i> , unigram (Philips and Pearl 2015)	0.503	0.210
<i>OnlineSubopt</i> , unigram (Philips and Pearl 2015)	0.477	0.240
<i>OnlineSubopt</i> , bigram (Philips and Pearl 2015)	0.473	0.322
<i>OnlineOpt</i> , bigram (Philips and Pearl 2015)	0.463	0.306
<i>Adaptor Grammar</i> , U-T-X-X-Seg (see main text)	0.449	0.054
<i>OnlineMem</i> , unigram (Philips and Pearl 2015)	0.443	0.331
<i>OnlineMem</i> , bigram (Philips and Pearl 2015)	0.100	0.226
Random baseline	0.100	0.199

Sources (where not include in main bibliography)

Daland, Robert, and Janet B. Pierrehumbert (2011). Learning diphone-based segmentation. *Cognitive Science* 35, 119-155.

Phillips, Lawrence and Lisa Pearl (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science* 39, 1824-1854.

Saksida, Amanda, Alan Langus, and Marina Nespors (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science* 20, e12390.

---

<sup>1</sup> We report the results for this model using the Relative algorithm and Mutual Information as the dependency measure; other settings generally performed worse.

## B. The peripheral position test

In section §7.3 we briefly describe an ad hoc test used to check what sort of affixes a flat model may be regarded as discovering, and use this test to indict these models with the charge of frequently discovering erroneous pseudo-affixes, such as the “prefix” [pi-] in *peekaboo* (i.e. [pi[kəbu]]). In this section we describe the test and our results in more detail.

The premise of the test is that a very short morpheme posited by a flat model at the beginning of an utterance can plausibly be interpreted as a prefix, and likewise a very short morpheme posited by a flat model at the end of an utterance can plausibly be interpreted as a suffix. For example, in (1) the utterance-final position of [-z] strongly suggests that it should be considered a suffix.

### (1) Illustration of the peripheral-position test

[ʌʔoðætsmamiz]	from Pearl-Brent training set ( <i>uh-oh, that's Mommy's</i> )
[ʌʔo ðæts mami z]	as parsed by JPSD MaxEnt <sub>1.55</sub> model
[mami-z]	assumed as a suffixed parse

We add two refinements, intended to avoid penalizing a model for learning a wrong suffix in cases where a different diagnosis is more appropriate. First, we do not penalize a model for detecting a real word that happens to be edge-adjacent and short; compare the two cases below, from JPSD MaxEnt<sub>1.55</sub>.

### (2) True words not classified as affixes

- a. *It measures how tall you are.* [ɪt mɛʒɹz haʊtəl ju ʌ.ɪ] short real word — counted as correct
- b. *Are you getting that guitar?* [ɑr ju ɡet ɪŋ ðæt ɡɪt ʌ.ɪ] pseudo-suffix — counted as wrong

Second, we do not penalize a model for cases in which the parsed peripheral material is better construed as a pseudo-stem to which a real affix has been attached. An example parse is [jɛ s ɪt ɪz] *yes it is*, from AG U-X-T-X-Phon: it would seem unfair to treat [jɛ-] as a mislearned prefix, since the element that follows it is in fact a real suffix; hence the most reasonable complete parse is [[jɛ]s] (i.e., -s [-s] suffixed to *ye* [jɛ]), not \*[jɛ[s]] (*ye-* [jɛ-] prefixed to *s* [s]).

We first offer two “sanity checks,” intended to show that the peripheral-position test does in general diagnose entities that could reasonably be called affixes. In (3), we give the results of the peripheral-position test (ten best candidate affixes) for two of our flat models. Note that the length of the Pearl-Brent corpus is 28,391 utterances.

(3) *Real affixes and pseudo-affixes most frequently discovered by two flat models — peripheral position test*

	<i>a. JPSD MaxEnt<sub>1.64</sub></i>				<i>b. AG U-X-T-X-Phon</i>		
<i>Rank</i>	<i>Affix</i>	<i>Count</i>	<i>Status</i>		<i>Affix</i>	<i>Count</i>	<i>Status</i>
1.	[-z]	674	real		[-z]	627	real
2.	[-ɪŋ]	642	real		[-s]	572	real
3.	[-s]	345	real		[-i]	543	real
4.	[-əl]	318	pseudo-affix		[-ɪŋ]	471	real
5.	[s-]	129	pseudo-affix		[-t]	411	real
6.	[-t]	124	real		[s-]	211	pseudo-affix
7.	[-ən]	105	pseudo-affix		[w-]	201	pseudo-affix
8.	[-i]	103	real		[-ə]	182	pseudo-affix
9.	[pi-]	90	pseudo-affix		[-e]	168	pseudo-affix
10.	[-di]	87	pseudo-affix		[-ri]	167	pseudo-affix

The test seems fair in the sense that most of the affixes that infants know are indeed in the top ten: [-ɪŋ], both allomorphs of [-z/-s], and the [-t] allomorph of [-d/-t]. An exception is that in neither model is the [-d] allomorph of [-d/-t] in the top ten; this appears to be characteristic of all the flat models we examined.<sup>2</sup> It is also worth noting that according to the peripheral-position test, the flat models generally do *not* recognize the pseudo-suffix [-j] (as in *babsh* or *dopsh*) studied by Kim and Sundara (2021); for example, zero tokens were found by JPSD MaxEnt<sub>1.64</sub> and just 13 by AG U-X-T-X-Phon (the models that emerged as the two best performers in §5.3).

In the main text, our focus was on the tendency of the flat models to treat erroneous strings in common words as if they were affixes, as in the prefix [pi-] extracted from the common word *peekaboo* ['pikəbu]. The results reported there are repeated here in more detail, listing the model responsible for the error and the relevant frequency counts in the corpus. A frequent pattern is the detection of erroneous affixes that occur in a name, probably of a family member or pet, that is a locally-frequent word. We find that all five of our flat models locate incorrect affixes from such sources.

(4) *Implausible affixes learned by the flat models*

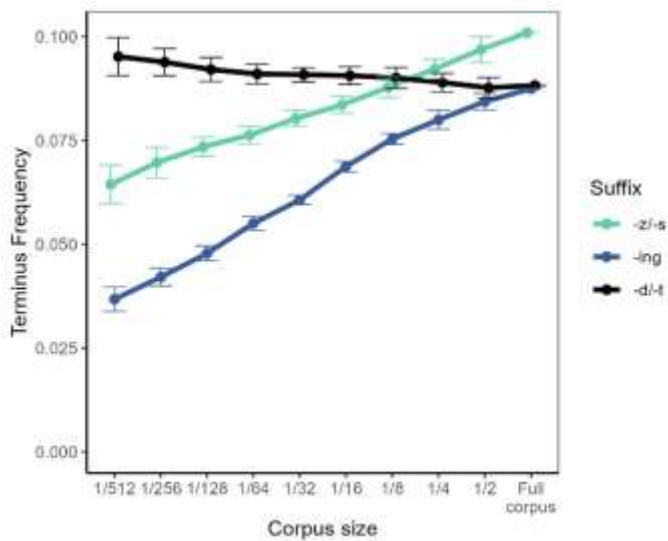
- JPSD MaxEnt<sub>1.64</sub>: [pi-] is the ninth-ranked affix in table (3) above; it is largely based (77/87 tokens) on *peekaboo*. [-di] is obtained almost entirely from *Mandy* (57/65 tokens, 14th).
- AG U-X-T-X-Phon: [-ri] is obtained from *Henry* [[hen]ri] (152/167, 10th).
- JPSD MaxEnt<sub>1.64</sub>: [-di] is obtained from *Mandy* (57/65, 14th).
- AG U-X-T-X-X-Phon: [-ri] is obtained from *Henry* (167/178, 7th).
- AG U-T-Phon: [-ən] is obtained from *Dillon* (51/90, 8th), and [æl-] from *Alhexander* (46/58, 11th).
- JPSD MaxEnt<sub>1.55</sub>: [-dɹ] is obtained from *Alexander* (53/68, 8th).

<sup>2</sup> Rank of [-d] in the peripheral-position test for the five flat models: AG U-X-T-X-X-Phon 15<sup>th</sup>, AG U-X-T-X-Phon 17<sup>th</sup>, AG U-T-Phon 17<sup>th</sup>, JPSD MaxEnt<sub>1.64</sub> 34<sup>th</sup>, JPSD MaxEnt<sub>1.55</sub> not discovered.

### C. Testing a model with just Terminus Frequency

In the main text (§8.1) we put forth a tentative model in which candidate affixes are chosen solely for having high Terminus Frequency. We noted there a simple reason for rejecting such a model, namely, that the non-affix [-ŋ] has a substantially higher Terminus Frequency than the true affix [-ɪŋ]. Here, we give another reason to reject such a model, namely that it fails to predict the correct acquisition order for the three affixes [-z/-s], [-d/-t], and [-ɪŋ]. This is shown below in (5); the time slices shown are the same as those used in §6 of the main text.

(5) *How Terminus Frequency changes over simulated time for three suffixes*

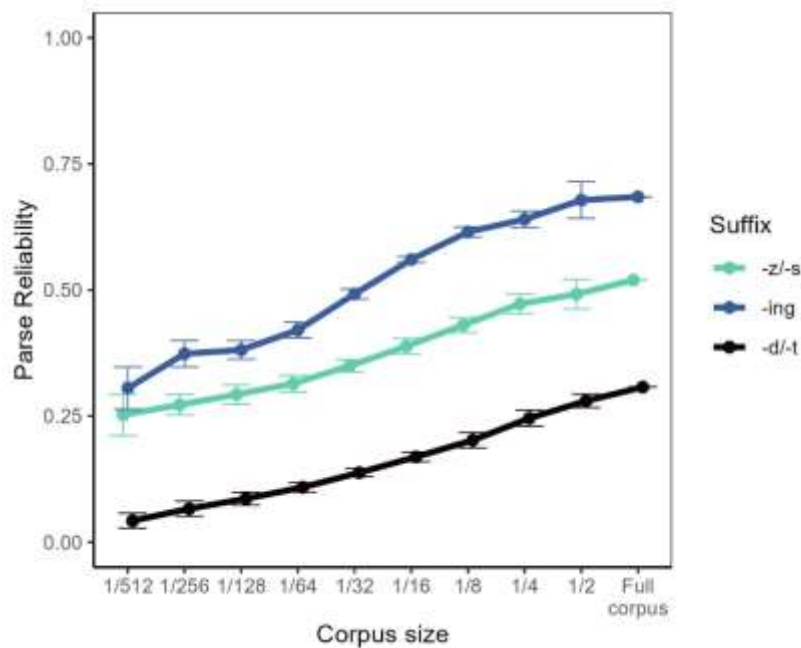


As the figure shows, if Terminus Frequency were the only relevant factor, we would expect [-d/-t] to be learned very early, with [-z/-s], and then [-ɪŋ] gradually catching up with it; the correct order is [-z/-s], [-ɪŋ], [-d/-t].

### D. Testing a model with just Parse Reliability

In the main text (§8.1) we put forth a tentative model in which candidate affixes are chosen solely for having high Parse Reliability. The values of our three target affixes over simulated time for Parse Reliability are given in (6).

(6) How Parse Reliability changes over simulated time for three suffixes

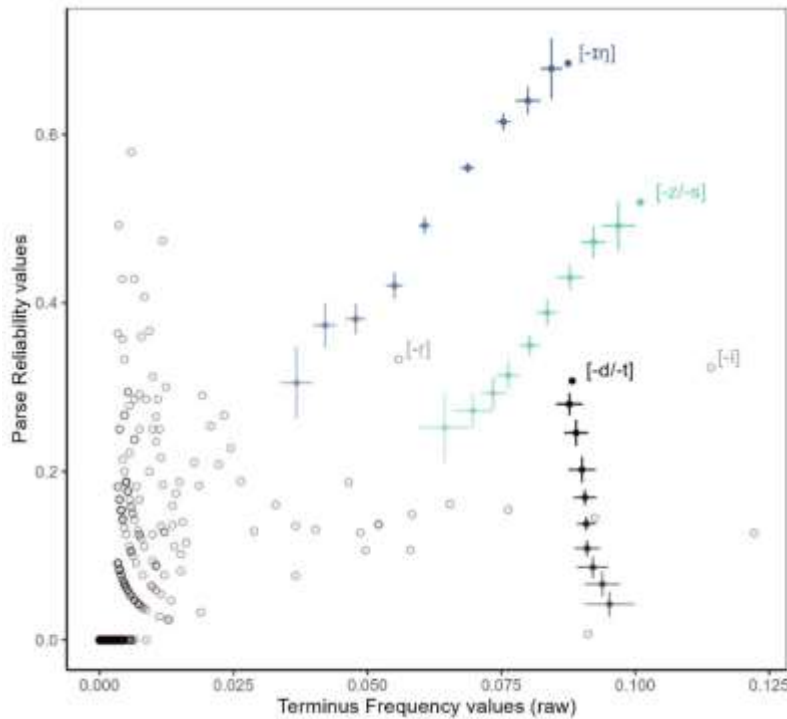


The patterning is different from (5), but no better at explaining acquisition order: although [-d/-t] is placed low where it should be, relying solely on Parse Reliability makes the wrong prediction that [-ɪŋ] should be learned before [-z/-s].

### E. Trajectory of real affixes over simulated time through a space defined by Terminus Frequency and Parse Reliability

Figure (7) plots the Terminus Frequency and Parse Reliability of our three main target affixes [-z/-s], [-ɪŋ], [-d/-t] over simulated time across the ten steps of simulated time. It can be seen that the three affixes gradually emerge from the L-shaped region occupied by non-affixes into the privileged upper right quadrant occupied solely by real affixes.

(7) Trajectory of Terminus Frequency and Parse Reliability for the three target affixes

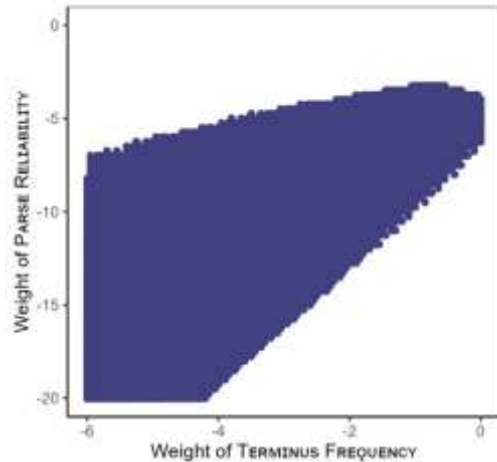


Of course, the remaining affix candidates also pursue trajectories through the space of Figure (7); our checking suggests that throughout the ten time slices, they generally do not depart from the L-shaped region, so we give only their final positions.

## F. The set of feasible weights

We noted above that the weights of our three features are hand-chosen to make the model work. Here we show there is not one single set of weights for these features that uniquely derives the experimental results; rather, a reasonably broad range does so. This is shown in (8), which gives the projection of the region onto the PARSE RELIABILITY/TERMINUS FREQUENCY plane.

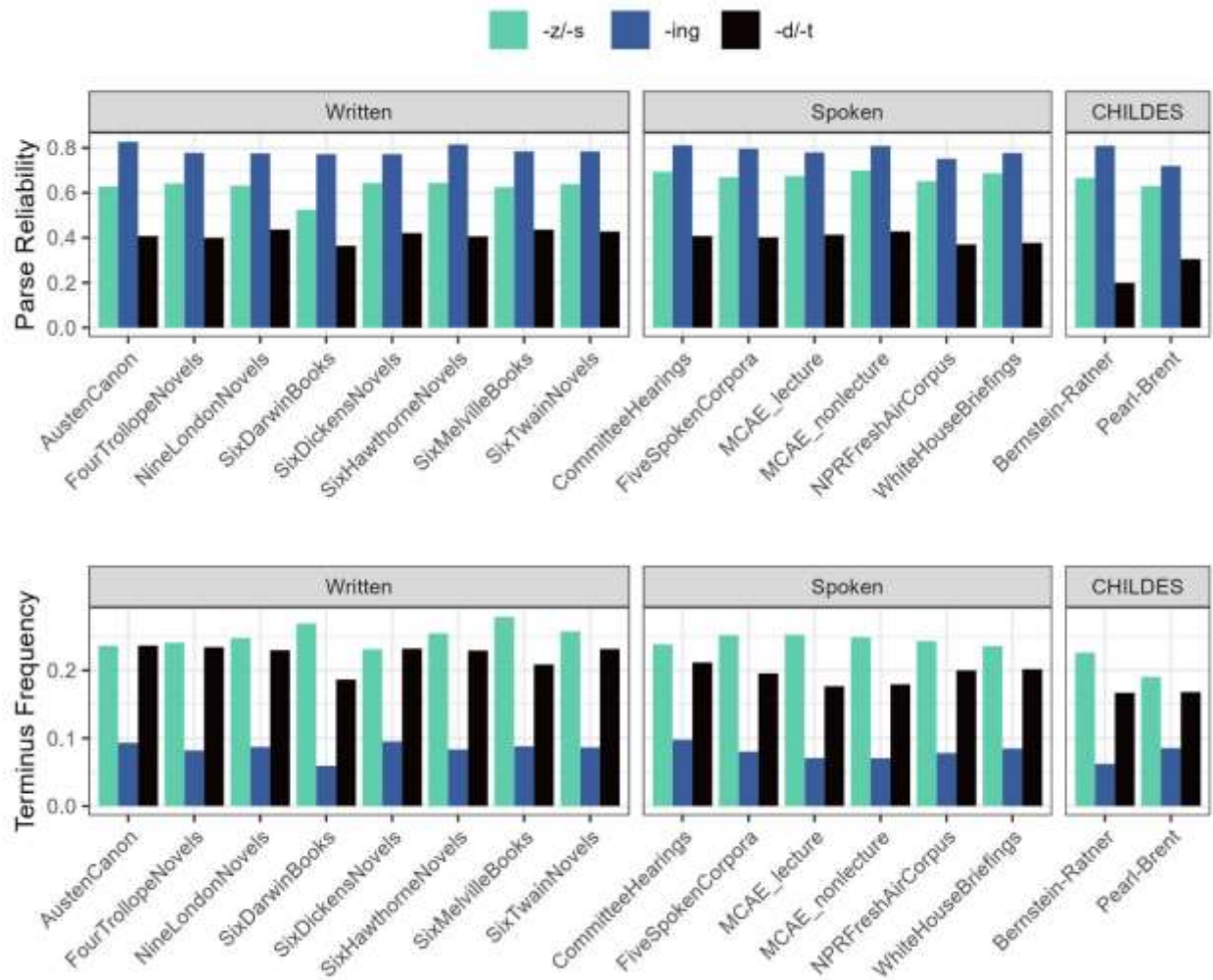
(8) Region of weight space in which the model discovers the suffixes in correct order



**G. Are Terminus Frequency and Parse Reliability values stable across corpora?**

For our hierarchical model to be plausible, we need to show that the observed values for Terminus Frequency and Parse Reliability in our target suffixes are not an idiosyncrasy of the Pearl-Brent corpus, but are characteristic of English as a whole, observable in multiple corpora. To this end we calculated Terminus Frequency and Parse Reliability for our target affixes in the 14 corpora (about 700K words each) employed in Breiss and Hayes (2020). While the results are not entirely comparable to those from Pearl-Brent (since we had to work with orthography rather than phonemic transcription), they are remarkably self-consistent: for all 14 corpora the ranking of suffixes by Parse Reliability was the same, namely  $[-\text{ɪŋ}] > [-\text{z}/-\text{s}] > [-\text{d}/-\text{t}]$ ; for 12/14 corpora the ranking for Terminus Frequency was  $[-\text{z}/-\text{s}] > [-\text{d}/-\text{t}] > [-\text{ɪŋ}]$ , with only tiny deviations in the other two cases. We also checked a widely-studied corpus of child-directed speech, used in Goldwater et al. (2009) and deriving originally from Bernstein-Ratner (1987) and Brent and Cartwright (1996).<sup>3</sup> The values for all 16 corpora are given in Figure (9).

<sup>3</sup> The corpus is currently obtainable at <https://homepages.inf.ed.ac.uk/sgwater/resources.html>.

(9) *Terminus Frequency and Parse Reliability across corpora for three suffixes*

The two corpora of child-directed speech give values that are reasonably comparable to the 14 adult corpora, especially in light of the fact they are considerably smaller (Pearl-Brent 97K; Bernstein-Ratner 36K). We conclude that the Terminus Frequency and Parse Reliability values for the Pearl-Brent corpus that we model are representative.

Breiss, Canaan and Bruce Hayes (2020). Phonological markedness effects in sentence formation. *Language* 96, 338-370.

## H. What drives the changes in Terminus Frequency and Parse Reliability over simulated time?

Our modeling results depend on the shape of the curves in Figures (5) and (6) above, which depict the evolution of Terminus Frequency and Parse Reliability values over simulated time. Since these curves are essential to our modeling, we offer some diagnosis for why they take on the forms observed.



The sharply different patterns of change in Terminus Frequency for the three affixes in Figure (5) of this document — notably the steep slope for [-ɪŋ] — can be explained as follows. For reasons given in §7.3 and §8.6, we employ type frequency throughout our modeling, hence also in preparing Figures (5) and (6). Of course, types can only be learned by encountering individual tokens. The suffix [-ɪŋ] has a rather low token frequency (token/type ratio 8.6; compare 15.1 for [-z/-s] and 29.3 for [-d/-t]), so it takes longer for the infant to encounter the tokens needed to learn the types of words suffixed with [-ɪŋ]. Or, to put it another way, the early-flattening curves for [-z/-s] and (especially) [-d/-t] in Figure (5) reflect saturation, as late-added tokens often fail to add to the type count.

For Parse Reliability, seen in Figure (6), the values ascend roughly in parallel. In adult English, [-ɪŋ] has the highest Parse Reliability, followed by [-z/-s], followed by [-d/-t]. The values ascend because as corpus size increases we have a greater chance of having encountered  $x$  when we see  $[[x] y]$ . We see no principled reason for the ranking among affixes to differ with corpus size.

## I. Is our hierarchical model a good affix discoverer?

It would not be sensible for us to offer our hierarchical model (§8) as an account of experimental results if, outside the context of these data, it were a generally poor performer at the task of affix discovery. To assess this issue, we constructed a gold-standard annotated version (with affixes identified) of the Pearl-Brent corpus and assessed our model for how well it could identify the affixes therein. The system performed fairly well: the average probability assigned to the correct parse (which, depending on word, will be  $[[x] y]$ ,  $[y [x]]$ , or  $[x]$ ), was 0.636. This can be broken down as follows: when the correct parse is affixed, the average probability is 0.823; and when the correct parse is monomorphemic, the average probability is 0.522.<sup>4</sup> Each of the three features contributes significantly to this outcome, according to a likelihood ratio test. Unsurprisingly, better performance can be obtained by specifically tuning the weights to match the gold standard.<sup>5</sup>

We note, however, that the best performance on adult gold-standard data is not necessarily the best performance when the goal is to model infant learning. This was shown by Larsen et al. (2017), who examine word-discovery models of the kind we use in our flat approach, and find that the models that do best on their gold-standard data are not necessarily the ones that best learn the particular words believed to be internalized first by children. Our findings parallel this, in that the weights that best match the acquisition order of suffixes are not the same weights that perform best by our gold-standard criterion.

Larsen, Elin, Alejandrina Cristia, and Emmanuel Dupoux. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. *Proceedings of Interspeech* 2198-2202.

---

<sup>4</sup> The disparity between 0.823 and 0.522 suggests that the model “overparses”; it is eager to find any affixes present. This reflects the large negative magnitude we chose for the feature PREFER MONOMORPHEMIC, in the interest of fitting the experimental data. The implicit claim is that infants are liberal affix-discoverers, who prioritize the discovery of relevant affix candidates over avoiding erroneous ones.

<sup>5</sup> Specifically, if we optimize the weights for learning the gold-standard parses, the three probability values mentioned in the text rise to 0.731, 0.803, and 0.600 respectively.

## J. Should other features be included as well?

In the course of developing the hierarchical model in §8, we explored features other than TERMINUS FREQUENCY and PARSE RELIABILITY that seemed potentially helpful, in the sense that including them improved the model's performance at finding affixes in our gold-standard corpus. These features include (a) a reward for  $[[x]y]$  parses in which  $x$  is long; (b) a penalty for individual  $[[x]y]$  parses where  $x$  does not occur in the corpus (compare PARSE RELIABILITY, which generalizes across *all* cases with  $y$ ); and (c) a measure of "attachment capacity," whereby  $y$  is likely to be a suffix to the extent that there is a high probability that a word of the form  $x$  will be accompanied in the corpus by a word of the form  $xy$ . Evaluated against our gold-standard data, this model assigns an average probability of 0.792 to correct parses overall, of 0.843 to correct suffixed parses, and 0.696 to correct nonsuffixed parses (compare 0.636, 0.823, 0.522 for our proposed model, discussed in section H). It is also capable (with suitable weights) of capturing the experimental results reviewed here. For purposes of presentation, however, we prefer our simple three-feature model, since the basis on which it works is directly observable in Figures (24) of the main text and (7) above.