Integrating semantics into developmental models of morphology learning

Abigail L. Tenenbaum, Mika Braginsky, Roger P. Levy
{abit, mikabr, rplevy}@mit.edu
Department of Brain and Cognitive Sciences, MIT

Abstract

A key challenge in language acquisition is learning morphological transforms relating word roots to derived forms. Traditional unsupervised algorithms find morphological patterns in sequences of phonemes, but struggle to distinguish valid segmentations from spurious ones because they ignore meaning. For example, a system that correctly discovers "add /z/" as a valid morphological transform (*song-songs*, *year-years*) might incorrectly infer that "add /ah.t/" is also valid (*mark-market*, *spear-spirit*). We propose that learners could avoid these errors with a simple semantic assumption: morphological transforms approximately preserve meaning. We extend an algorithm from Chan and Yang (2008) by integrating proximity in vector-space word embeddings as a criterion for valid transforms. On a corpus of child-directed speech, we achieve both higher accuracy and broader coverage than the purely phonemic approach, even in more developmentally plausible learning paradigms. Finally, we consider a deeper semantic assumption that could guide the acquisition of more abstract, human-like morphological understanding.

Keywords: language acquisition, morphology, development, semantics.

Introduction

Relating word roots to derived forms poses a key challenge in learning language and occurs in early years of development (Berko, 1958). Our goal in this work is to build computational models of this process. We focus on algorithms that learn morphological transforms from an unannotated vocabulary. Most previous unsupervised models extract morphemes only from patterns and statistics in word strings or phoneme sequences. Approaches of this sort that are based on minimum description length (e.g. Goldsmith, 2001) or Bayesian principles (e.g. Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Goldwater, Griffiths, & Johnson, 2009) produce strong results, but fail to emulate children's learning in several important ways.

First, these algorithms can struggle to distinguish valid segmentations from spurious ones. For example, an algorithm that learns the suffix /z/ from observing *song*, *songs*, *year*, and *years* in the vocabulary, might also extract the coincidental suffix /ah.t/ from seeing *mark*, *market*, *spear*, and *spirit*. This issue fundamentally comes from only considering superficial patterns, rather than underlying meanings.

Another limitation is the gap between ideal statistical models (high resource learners) and cognitive plausibility for children (low resource learners). Prior approaches have typically been built around complicated algorithms that learn over many iterations, have high computational demands, and track a large number of independent parameters. Some of these previous models (e.g. Goldwater et al., 2009; Goldsmith, 2001) are explicitly intended as accounts of language development at the computational level (in the sense of Marr (1982)), but our goal here is to move towards algorithmic models of morphology induction that are cognitively plausible. That is, we want not only to model learning in terms of the problem being solved or the function being optimized, but also to specify more of the cognitive mechanisms or processes by which this learning might occur.

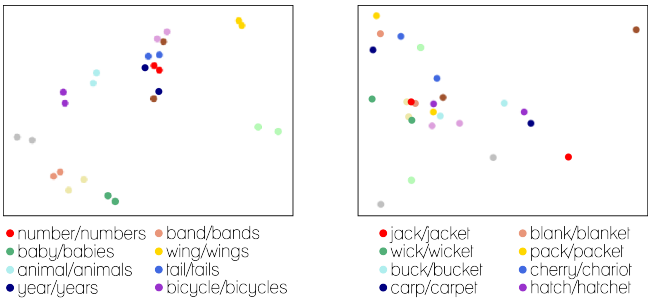

Figure 1: Pairs of morphologically related words appear nearby in t-SNE reduced semantic space (left), while spurious pairs have less consistently similar meanings (right).

Our proposal builds on previous work by Chan and Yang (2008) and Lignos, Chan, Yang, and Marcus (2010). Motivated by evidence (Brown, 1973) that children may instead learn morphological rules one at a time through on-line hypothesis formation, Chan and Yang (2008) developed an algorithm that iteratively extracts morphological transforms by seeking pairs of suffixes with many stems in common. Lignos et al. (2010) investigated the results of running this algorithm on child-directed speech, aiming to bridge the gap between high-powered and developmental models of morphology learning.

Their approach constitutes a significant step towards cognitive plausibility in that the algorithm is simple, intuitive, and yields a trajectory that better matches developmental data. However, their algorithm still accepts some spurious morphological transforms, and requires sufficiently high computational resources to track over 1000 possible transforms at a time. Thus it neither fully bridges the gap nor escapes the tendency to make mistaken generalizations.

Our contribution is to extend the Lignos et al. algorithm in a way that responds to both of these issues. The key insight is that children have access to linguistic and experiential data beyond a simple vocabulary, since they hear words used in context and have some sense of their meanings. To help the learning model avoid spurious transforms, we extend the purely phonemic algorithm with a simple semantic assumption: **valid morphological transforms should roughly preserve meaning**. More formally, we approximate mean-

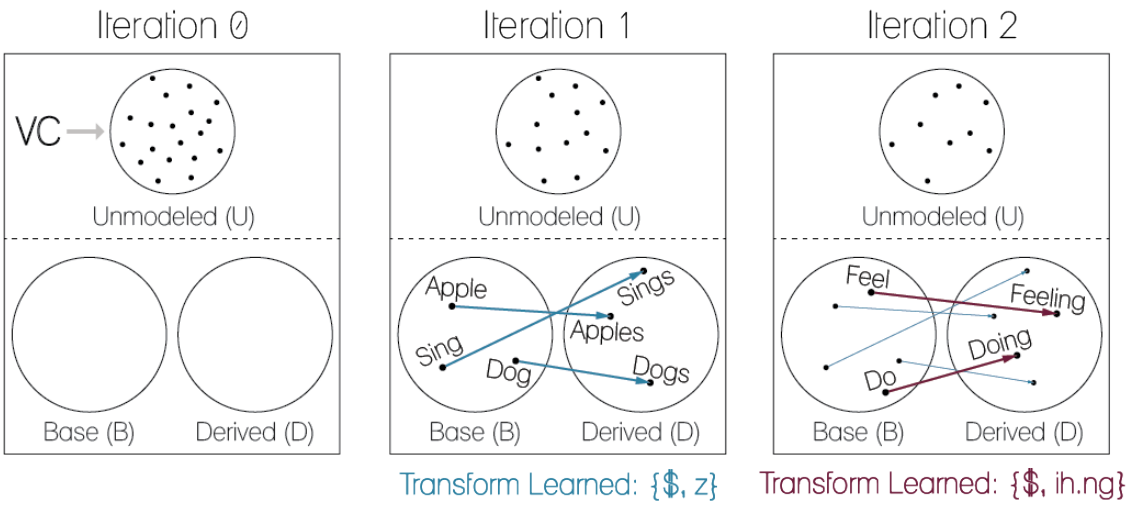

Figure 2: Overview of the algorithm. All wordpairs begin in the Unmodeled (U) set. A new transform is learned with each iteration, and the word forms it connects are moved into the Base (B) and Derived (D) sets.

ing with distributional word embeddings, and we assume that valid transforms should relate each base form to a derived form that is sufficiently nearby in the embedding space. We use vector representations generated by the GloVe algorithm (Pennington, Socher, & Manning, 2014), trained on co-occurrence statistics.

To get some intuition for why this assumption makes sense, consult Figure 1. In both panels, we use t-SNE dimensionality reduction (Maaten & Hinton, 2008) to plot the 300-dimensional GloVe vectors of 34 words. On the left, we see 17 pairs of words that are related by the valid suffix /z/, while on the right we see 17 pairs related by the spurious suffix /ah.t/. As expected, the valid transform relates semantically similar wordpairs – e.g. the singular and plural forms of various nouns – while the spurious one seems uncorrelated with semantic similarity.

In the rest of this paper, we first describe our algorithm in detail and then present results addressing two central questions. Does incorporating semantic information help an unsupervised algorithm learn morphology using ideal (high resource) settings? Do semantics support more developmentally plausible (low resource) settings for the algorithm? We close by discussing model limitations and proposing a deeper semantic assumption that could help in the acquisition of more abstract, human-like morphological understanding.

Methods

In this section we describe our model (which is largely similar in structure to that of Lignos et al.), highlighting where and how we incorporate semantic information. Our algorithm takes as input a subset of the corpus of child-directed speech from the CHILDES database (MacWhinney, 2000). Initially, it extracts all the word types uttered by adult speakers, encodes them phonemically according to the CMU Pronouncing Dictionary (Weide, 1998), and places them in the Unmodeled set. On each iteration, the algorithm learns a single transform of the form $\{s_1, s_2\}$, meaning "remove the suffix sequence s_1 and add the suffix sequence s_2." For example, the transform "remove the empty string $ and add /z/," which is almost always learned first, would be written {$, z}. As the algorithm iteratively discovers new morphological transforms, the Unmodeled (U), Base (B) and Derived (D) sets are dynamically updated to reflect the relationships learned between words. The composition of these sets determine which transforms can be learned in future iterations. A high-level overview of this process is shown in Figure 2.

We now outline the steps (1-3) involved in one iteration of the algorithm. Table 1 summarizes the parameters.

1. Hypothesize Transforms First, we extract all suffixal phoneme sequences of length 0-4 from among the word types in the Unmodeled set and rank them in frequency. Next, we form candidate transforms $\{s_1, s_2\}$ from all pairs of suffixes s_1 and s_2 chosen from the **N** most frequent suffixes. For each transform hypothesized thus, we list all the permitted base/derived wordpairs that it explains. The parameter **P** determines which sets (out of U, B, and D) the proposed base and derived forms of a wordpair may come from. To illustrate, consider a scenario where **N**=3 and the most frequent suffixal sequences are $, /z/, and /ih.ng/. In this case, the list of (nontrivial) candidate transforms would be {$, z}, {z, $}, {$, ih.ng}, {ih.ng, $}, {z, ih.ng}, and {ih.ng, z}.

2. Select Transform The next step involves our central modification of the Lignos et al. algorithm. Specifically, for each base/derived wordpair, we evaluate the cosine distance between the GloVe representation of the base form and the GloVe representation of the derived form. We filter out hypotheses that are not sufficiently close in semantic space, either at the level of entire transforms (**L**=Coarse) or at the level of individual wordpairs within transforms (**L**=Fine). If **L**=Coarse, we let Δ for each hypothesized transform be the

average of cosine distances δ across all of the permitted wordpairs that the transform explains. We then discard transforms with $\Delta \geq \mathbf{T}$. If **L**=Fine, we keep all transforms but discard wordpairs with cosine distance $\delta \geq \mathbf{T}$.

In either case, we then discard transforms that don't meet a threshold of overlap ratio (as described in Lignos et al., 2010, p. 5) and rank the remaining transforms according to the number of wordpairs that they explain, with ties broken by token counts (Chan & Yang, 2008, p. 107). Finally, if the top candidate transform explains $\geq \mathbf{W}$ wordpairs, we add it to the list of learned transforms. Otherwise, the algorithm terminates.

Continuing with the example introduced in step 1, suppose our **VC** is the above paragraph. Then {$, z} and {z, $} explain the wordpairs *transform* → *transforms*, *explain* → *explains*, and *transforms* → *transform*, *explains* → *explain*, respectively. Since every word starts in U on the first iteration, these wordpairs all follow the form U→U, and so they are all permitted. The other candidate transforms explain no wordpairs. Supposing that $\mathbf{W} \leq 2$, our algorithm would learn {z, $}, since the base forms of {z, $} (*transforms* and *explains*) appear more often than those of {$, z} (*transform* and *explain*).

3. Update Sets For each base/derived wordpair explained by the newly learned transform, we move the derived form to the Derived set and the base form to the Base set (unless doing so would move it out of the Derived set).

Developmental Parameter Settings

The full parameter space shown in Table 1 spans a wide range of learning paradigms. To study the algorithm at its best (high resource), we set **N**=50 and **EC**=6B while exploring all combinations of the other parameters (**VC**, **T**, **L**, **W**, and **P**). Separately, we vary our developmental parameters (**VC**, **EC**, and **N**) to better approximate a child-plausible (low resource) learner in a variety of ways (again, covering the space of combinations of the other parameters).

- To model different levels of linguistic exposure, we vary the corpus of word types, **VC**, between Brown+ (∼700K tokens) and Full (∼7M tokens).

- Because children have limited understanding of word meanings, we vary the size and source of the corpus on which our GloVe vectors are trained, **EC**, as a proxy for different levels of semantic experience. In addition to using word vectors pretrained on 6B tokens from Wikipedia, we train four of our own embeddings on smaller subsets of the Wikipedia corpus to serve as more realistic co-occurrence measures. We also train one embedding on child-directed adult speech extracted from the CHILDES corpus for potentially the most child-plausible semantic representations of all these.

- Since children have limited memory and processing capacity, and the algorithm hypothesizes $\mathbf{N}(\mathbf{N}-1)/2$ transforms on each iteration, we explore performance with a range of smaller values for **N**. Storing 3, or even 105 possibilities (for **N**=3 and **N**=15, respectively) seems more developmentally plausible than storing 1225 possibilities (for **N**=50).

Table 1: Summary of parameters (and abbreviations).

Parameter	Range of values investigated	Lignos' Setting
Vocabulary Corpus (**VC**)	Brown+ (CHILDES corpora Adam, Eve, Sarah, Naomi, Peter, Nina), Full (all NA-English corpora)	Brown+
Embedding Corpus (**EC**)	3M, 10M, 20M, 50M, 6B (Wikipedia tokens), CHILDES (all NA-English corpora)	–
Number of Top Suffixes (**N**)	3, 5, 10, 15, 20, 50	50
Semantic Threshold (**T**)	Values vary across different combinations of **EC** and **L**: 10-20 values for each pair	–
Thresholding Level (**L**)	Coarse (screened at the level of transforms), Fine (screened at the level of individual wordpairs)	–
Wordpair Threshold (**W**)	3, 4	5
Permitted Wordpairs (**P**)	Static (U→U, B→U, U→B), Nonstatic (above and B→B, D→U, U→D)	U→U, B→U

Results

To evaluate the performance of our algorithm under one parameter setting, we determine how many valid transforms it learns and count the rest as spurious. We hand-coded transforms as valid if they connected at least three correct base/derived wordpairs in a semantically consistent way, regardless of what other, possibly spurious wordpairs they also explained. For example, we count {t, s} as valid because it explains the wordpairs *important* → *importance*, *intelligent* → *intelligence*, and *patient* → *patience*, even though it also includes spurious wordpairs like *print* → *prince*.

In Table 2, we compare the transforms found by Lignos et al. with the results of our most successful or revealing parameter settings. We further devote one subsection each to discussing our initial questions: First, we consider whether integrating semantic information improves the performance of an ideal (high resource) learner, and second, we investigate the extent to which even very limited representations of meaning could offset significant decreases in learners' computational power.

High Resource Learner

Figure 3 shows the results of running our modified algorithm under conditions that are otherwise comparable to those of

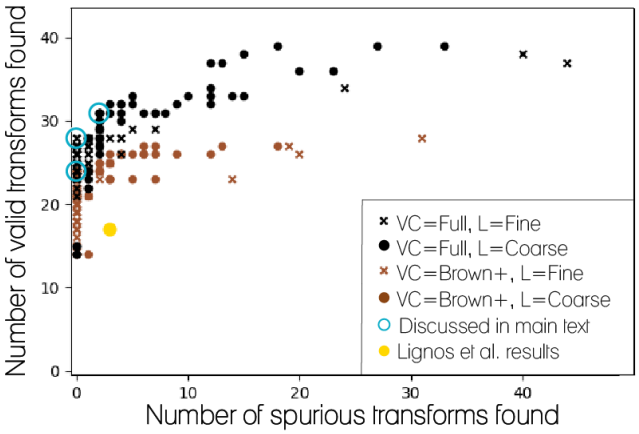

Figure 3: Summary of results for the ideal (high resource) learner in terms of hits (valid transforms, y-axis) and false alarms (spurious transforms, x-axis). Each point represents one setting of the algorithm's parameters, and key parameter choices are highlighted in point shape and color.

Lignos et al. We set two of our developmental parameters to their highest resource values (**EC**=6B and **N**=50), then let **T**, **L**, **W**, and **P** vary within each color to produce a distribution of points. The points in the top left of the plot represent the most successful runs because they maximize the number of valid transforms found while minimizing the number of spurious ones found. We see two overall trends.

First, our semantically informed algorithm markedly outperforms Lignos et al. when acting as a high resource learner. Where Lignos et al. find 17 valid and 3 spurious morphological transforms, we find 24 valid and 0 spurious transforms (see Table 2, Column H1) using the same Vocabulary Corpus as they do (**VC**=Brown+). With a strictly larger Vocabulary Corpus (**VC**=Full), our algorithm extracts 28 valid transforms without finding a single spurious one (see H2). We also find a set of parameter values that maximizes the number of valid transforms learned (31) while still accepting no more than 2 spurious ones (see H3).

Second, we notice a fundamental performance tradeoff between coarse and fine thresholding. Applying a coarse threshold allows us to find larger numbers of transforms, but usually at the cost of admitting many spurious transforms. For some learning goals this may be appropriate, but applying a fine threshold often appears to be more reliable in that it enables the algorithm to maximize valid transforms found while also making no mistakes. We take this to suggest that fine thresholding better models children's learning, and return to that point in the discussion.

Low Resource (Developmental) Learner

Each plot in Figure 4 shows the effects of varying one developmental parameter – Vocabulary Corpus, Embedding Corpus, or Number of Top Suffixes – while keeping the other two at their highest resource settings (**VC**=Full, **EC**=6B, **N**=50). Compared to the high resource learner (pictured in black throughout), overall performance declines as we lower

Table 2: Comparison of the transforms found by our algorithm to those found in Lignos et al. Green boxes represent valid transforms learned, while red boxes represent spurious transforms learned.

	Lignos	*High Resource*			*Low Resource*		
N =	50	50	50	50	15	15	3
VC =	Brown+	Brown+	Full	Full	Full	Full	Brown+
L =	–	Coarse	Fine	Coarse	Fine	Fine	Fine
		(H1)	(H2)	(H3)	(L1)	(L2)	(L3)
{$, z}	■	■	■	■	■	■	■
{$, ih.ng}	■	■	■	■	■	■	■
{$, s}	■	■	■	■	■	■	■
{$, iy}	■	■	■	■	■	■	■
{$, d}	■	■	■	■	■	■	■
{$, t}	■	■	■	■	■	■	■
{$, er}	■	■	■	■	■	■	■
{$, n}	■	■	■	■	■	■	■
{$, ah.n}	■	■	■	■	■	■	
{$, ah.d}	■	■	■	■	■	■	
{$, ah.l}	■		■	■	■	■	
{$, l.iy}	■	■	■	■	■		
{$, ah.z}	■	■	■	■			
{$, ah.s}	■		■	■			
{$, ah.n.t}	■						
{ah.l, l.iy}	■						
{t.iy, th}	■						
{$, ih.z}		■	■	■			
{$, th}		■	■	■			
{$, ah.s.t}		■	■	■			
{$, ah.b.ah.l}		■	■	■			
{t, s}		■	■	■			
{t, sh.ah.n}		■	■	■			
{$, m.ah.n.t}		■	■	■			
{d, t}		■	■	■	■	■	
{d, ih.ng}		■		■		■	
{d, z}						■	
{$, f.ah.l}		■					
{ah.n.d, $}		■					
{$, t.iy}		■	■				
{t, ih.ng}			■				
{$, ih.d}			■	■			
{s.ah.n, t.ah.d}			■	■			
{$, ah.n.s}			■	■			
{$, ey.sh.ah.n}			■	■			
{$, ih.v}				■			
{z, iy}				■			
{z, er}				■			
{er, ih.ng}				■	■		
{er, s}					■		
{er, z}					■		
{er, d}					■		
VALID	17	24	28	31	17	14	8
{t, iy}					■		
{$, ah}	■				■		
{$, ah.t}	■						
{$, k}	■						
{k, f}				■			
{ow, ah.n}				■			
SPURIOUS	3	0	0	2	2	0	0

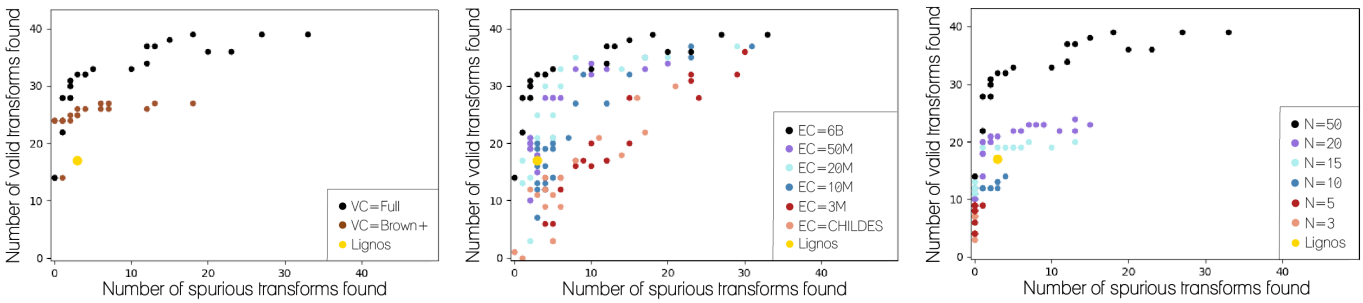

Figure 4: Trends in results for a low resource learner as a function of Vocabulary Corpus (left), Embedding Corpus (middle), or **N** (right), the number of suffixes considered in forming candidate transforms. We let **T**, **L**, **W**, and **P** vary within each color.

resources in any of the three dimensions, though the curves retain the same basic shape even as the values become more child-plausible.

There are low resource parameter settings that match the level of success achieved with no access to semantic information. Learning only from child-directed speech (**VC**=Full and **EC**=CHILDES) represents the setup closest to the linguistic exposure and semantic experience of a young child. Using these input corpora, our algorithm with **N**=15 can find the same number of valid transforms (and one fewer spurious transform) as Lignos et al. do with **N**=50 (see Table 2, Column L1). Our setting for **N** represents a much lower demand on processing capacity and memory. With the same developmental parameters, we also find a set of values for **T**, **L**, **W**, and **P** that maximizes the number of valid transforms learned (14) without admitting any spurious ones (see L2). Even using the lowest resource setting for each developmental parameter (**VC**=Brown+, **EC**=CHILDES, **N**=3), our algorithm learns 8 of the most common valid transforms, including the top 7 found by Lignos et al., without making any mistakes (see L3).

Discussion

We see that incorporating semantic information into models of morphology learning enables high resource learners to achieve greater coverage and accuracy over the space of morphological transforms. Moreover, these models support developmental (low resource) learning that is both equally successful to and more child-plausible than the purely phonemic approach.

Our work is related to several recent proposals in the NLP community which improve morphology learning by incorporating semantic information of some kind (e.g. Goldwater et al., 2009; Soricut & Och, 2015). What distinguishes our work is that we take a lower resource approach to the problem, seeking to build a more developmentally plausible model, both in terms of the training data and the computational efficiency of the algorithm. In contrast, the NLP approaches might produce fuller or deeper analyses, but they also rely on complex statistical calculations and large amounts of data which may or may not be available to children.

Following Chan and Yang (2008) and Lignos et al. (2010), we focus on learning suffixes, as they are the simplest and earliest emerging forms of English morphology. But high-powered NLP approaches are able to learn additional structures, including prefixes and other word components that children eventually learn. It will be important in future work to build developmentally plausible models of how children acquire these aspects of language as well.

We chose to encode semantic information using vector embeddings trained on adult speech or text data, but one could reasonably object that the linguistic knowledge encoded in a Wikipedia corpus is inaccessible to children of the age that our model is trying to capture. Even training vector embeddings on the CHILDES corpus as a whole has the potential to confuse the linguistic experience of older children with that of younger children. However, we used these corpora only to approximate the large amount of linguistic input that children get from their parents on a daily basis, whether through direct interaction or indirect observation. A valuable step in future work would be to explore ways of getting a closer and more precise proxy for the linguistic exposure of differently aged children.

One could also reasonably object to our use of vector embeddings in the first place, on the grounds that co-occurrence statistics only scratch the surface of the rich representations of meaning that even young children can access. However, we are not committed to this choice as an account of how children actually acquire and represent semantic knowledge. Rather, as above, we use vector embeddings to supplement the vocabulary list only as a proxy for children's limited understanding of word meanings and any other general semantic information conveyed through gesture or context. In the future, it would be interesting to explore using richer semantic representations that could also be learnable from the information available to children.

Assessing Our Model As A Developmental Account

Since the motivation of our work was to propose an algorithm for morphology acquisition that better matches the resources available to children, it is especially instructive to compare our results with empirical findings in language development.

Brown (1973) describes several basic inflectional morphemes that emerge earliest in child language (between

months 27 and 40): present progressive, present tense, noun plural, and past tense. Even given minimal resources, our algorithm finds phonological subtypes of each of these main morphemes. For example, of the four forms that the noun plural can take, we find the two most common, {$, z} and {$, s}, at **N**=3. The other two, {$, ah.z} and {$, ih.z}, are found only with larger values of **N**.

More generally, Brown's empirical observations highlight two important ways in which our model fails to capture the richness of children's morphological understanding. First, rather than learn individual transforms that apply to specific, known words, we ultimately want an algorithm that groups base stems into broad types according to the suffixes they support.

An extension of our model should be able to learn, for instance, that the stem *lie* belongs to the group defined by the set of suffixes [$, z, d, ih.ng] because *lie*, *lies*, *lied*, and *lying* are observed in the given vocabulary. The algorithm should also recognize that *plough* belongs to this same group of base stems because it shares certain key semantic properties with them – in this case, that they're all verbs. This would allow the learner to hypothesize that *ploughs*, *ploughed*, and *ploughing* (in their phonetic forms) are simply words that it hasn't seen yet.

Further, such an algorithm would ideally group all non-exceptional verbs together. Under our current phonemic approach, however, words like *plough*, *state*, and *splice* would all belong to different groups (defined by [$, z, d, ih.ng], [$, s, ih.d, ih.ng], and [$, ih.z, t, ih.ng], respectively). A more realistic (and successful) model for developmental morphology acquisition should recognize that these differences are superficial and seek to learn the underlying semantic transforms themselves, rather than their varied phonemic manifestations.

Model Extension: Semantic Transforms

Our current algorithm acquires morphology on the transform level, but a more nuanced treatment would instead aim for a token level understanding of segmentation.

We have already explored one step in this direction: Applying the semantic threshold at the level of wordpairs (**L**=Fine) yields significantly purer transforms than applying it at the level of transforms. For example, doing so allows our model to learn valid patterns like *important* → *importance*, *intelligent* → *intelligence*, and *patient* → *patience* while also avoiding some spurious wordpairs like *print* → *prince*.

Going forward, we intend to consider an extension of our model along these lines that could help the algorithm refine its results on the token level, and also has the potential to address some of the developmental limitations discussed in the previous section. We propose a stronger version of our original semantic assumption: **valid morphological transforms should connect pairs of base and derived forms that are offset in a consistent direction in semantic space.**

To implement this idea, we first take all the wordpairs within a single valid transform and find the GloVe difference vector between the base and derived form of each wordpair. We then perform agglomerative clustering on those vectors. Since the spurious wordpairs aren't offset by the transform in a consistent semantic direction, this should help us filter them out. For instance, the difference vectors of *important* → *importance*, *intelligent* → *intelligence*, and *patient* → *patience* might vaguely align, while that of *print* → *prince* might point in a completely different direction.

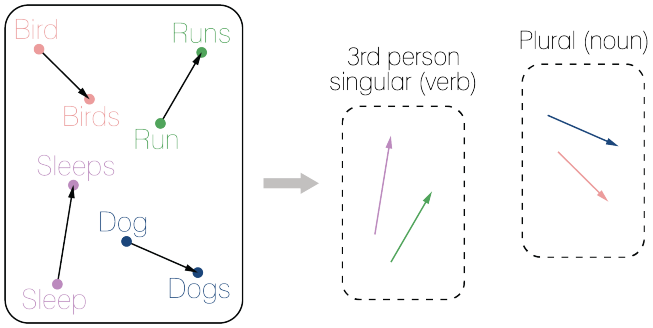

Figure 5: Schematic of how agglomerative clustering on GloVe difference vectors in semantic space could allow the learner to identify semantically based morphological transforms.

This same approach could also help us separate semantically distinct transforms that appear identical at the level of phonemes. Agglomerative clustering of difference vectors models how learners isolate semantically distinct sub-transforms within a single phonemic transform, and then identify groups of phonemically distinct sub-transforms that together represent a single underlying semantic transform.

Figure 5 demonstrates how this idea would work in principle with several wordpairs explained by the transforms {$, z} and {$, s}. We hope that *sleep* → *sleeps* and *run* → *runs*, as verbs, end up in the same cluster because their semantic difference vectors point in similar directions, even though on the superficial phonemic level, *sleep* → *sleeps* seems more similar to *bird* → *birds*, since they are both explained by the transform {$, s}.

Indeed, when we run an agglomerative clustering algorithm on the GloVe difference vectors of all the wordpairs explained by {$, z} and {$, s} (as outputted by H3) and choose a distance threshold that sorts them into three groups, the resulting clusters correspond approximately to: nouns becoming plural, verbs becoming 3rd person singular, and spurious wordpairs.[1] These results are visualized (again using t-SNE dimensionality reduction) in Figure 6, where each point represents the difference vector of a base/derived wordpair. Notice that the verbs and nouns each form systematic clusters, while the spurious wordpairs are spread seemingly at random.

In this case, we chose a distance threshold that yielded three clusters in order to most clearly isolate the morphological structure that we hypothesized would be present in

[1] We use Ward's minimum variance method for cluster analysis and set the distance threshold at T=30.

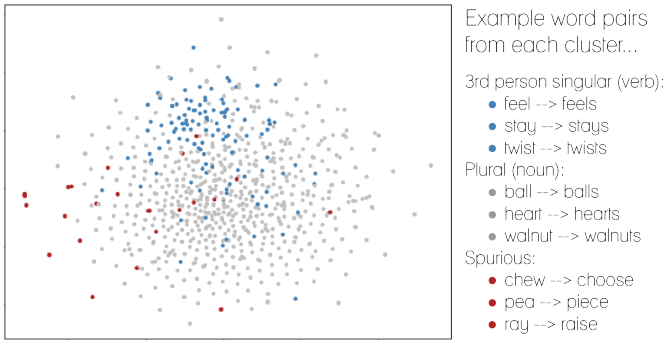

Figure 6: Visualization of how semantically coherent transforms cluster in t-SNE reduced word embedding space. Each point corresponds to the difference vector between a base and derived form connected by {$, z} or {$, s}.

and potentially extractable from this semantic space. It remains an open question whether this threshold can be learned or emerge automatically.

A future version of our morphology learning model could incorporate such a clustering mechanism either as a post-processing step to filter out spurious wordpairs, or integrated into the algorithm itself to find the underlying semantic structure within and between morphological transforms.

Conclusion

Morphemes are the smallest chunks of language that have meaning, so they are a natural place for structure and meaning to come together. Here we investigated the interaction between meaning and structure in a computational model of language acquisition and showed that even relatively simple forms of semantic representation substantially increase the accuracy, coverage, and efficiency of our model. The advantage that incorporating semantic information provides in learning morphological structure remains even in more developmentally plausible learning conditions.

Acknowledgements

We would like to thank the anonymous reviewers for valuable comments that helped improve the paper. We are also grateful to M. Belledonne, P. Qian, and J. Gauthier for technical help along the way.

References

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2-3), 150–177.

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

Chan, E., & Yang, C. D. (2008). Structures and distributions in morphology learning. *A dissertation in Computer and Information Science, University of Pennsylvania*.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics*, *27*(2), 153–198.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Lignos, C., Chan, E., Yang, C., & Marcus, M. P. (2010). Evidence for a morphological acquisition model from development data. In *Proceedings of the 34th Annual Boston University Conference on Language Development* (Vol. 2, pp. 269–280).

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579–2605.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: W.H. Freeman and Company.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Soricut, R., & Och, F. J. (2015). Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1627–1637).

Weide, R. L. (1998). *The CMU Pronouncing Dictionary.* Retrieved from `http://www.speech.cs.cmu.edu/cgibin/cmudict`