The Greedy and Recursive Search for Morphological Productivity

Caleb Belth[1] (cbelth@umich.edu), Sarah Payne[2] (paynesa@sas.upenn.edu), Deniz Beser[3] (beser@isi.edu), Jordan Kodner[4,5] (jordan.kodner@stonybrook.edu), Charles Yang[2,6] (charles.yang@ling.upenn.edu)

[1]Department of Computer Science and Engineering, University of Michigan
[2]Department of Linguistics and Department of Computer and Information Science, University of Pennsylvania
[3]Information Sciences Institute, University of Southern California
[4]Department of Linguistics, Stony Brook University
[5]Institute for Advanced Computational Science, Stony Brook University
[6]Department of Psychology, University of Pennsylvania

Abstract

As children acquire the knowledge of their language's morphology, they invariably discover the productive processes that can generalize to new words. Morphological learning is made challenging by the fact that even fully productive rules have exceptions, as in the well-known case of English past tense verbs, which features the *-ed* rule against the irregular verbs. The Tolerance Principle is a recent proposal that provides a precise threshold of exceptions that a productive rule can withstand. Its empirical application so far, however, requires the researcher to fully specify rules defined over a set of words. We propose a greedy search model that automatically hypothesizes rules and evaluates their productivity over a vocabulary. When the search for broader productivity fails, the model recursively subdivides the vocabulary and continues the search for productivity over narrower rules. Trained on psychologically realistic data from child-directed input, our model displays developmental patterns observed in child morphology acquisition, including the notoriously complex case of German noun pluralization. It also produces responses to nonce words that, despite receiving only a fraction of the training data, are more similar to those of human subjects than current neural network models' responses are.

Keywords: linguistics; language acquisition; morphology; computational modeling

Introduction

The acquisition of English past tense is one of the most extensively studied problems in cognitive science (McClelland & Patterson, 2002; Pinker & Ullman, 2002). Yet its simplicity—a single, and numerically dominant, rule (*-ed*) along with a list of irregular verb exceptions—is hardly representative of the complexity of the world's morphological systems (Comrie, 1989). In comparison, German noun pluralization is a more challenging test for morphological learning theories. The plural is formed by five suffixes, but even the *least* frequent, *-s*, is productive and applies to novel nouns (e.g., *iPhones*), while the other four suffixes are also productive for sub-categories of nouns characterized by their gender and phonological properties (Wiese, 1996). Nevertheless, children exhibit remarkable proficiency of their native language's morphology at a very early age; see (Lignos & Yang, 2016) for a cross-linguistic review.

Children's discovery of productive rules is also attested experimentally when presenting young children with nonce words in a Wug test (Berko, 1958) and when new words (e.g., *google-googled*) enter a language. As a result, traditional linguistic approaches to language acquisition often make use of productive rules (Chomsky & Halle, 1968). Rumelhart and McClelland (1986) proposed an impactful neural-network (NN) model, suggesting that NNs might be capable of exhibiting rule-like behavior despite having no explicit representation of rules. Ensuing critiques of the model by Pinker and Prince (1988) and of connectionist models of cognition more generally by Fodor and Pylyshyn (1988) sparked what came to be known as the "Past Tense Debate." Due to the extensive advancements in NN architecture in natural language processing and machine learning, the debate has recently been revived by Kirov and Cotterell (2018) with some initially promising improvements over early models (in particular, more realistic accuracy). However, these accomplishments have been challenged, especially when compared against human behavioral results on Wug tests in English past tense (Corkery, Matusevych, & Goldwater, 2019) and German plurals (McCurdy, Goldwater, & Lopez, 2020).

The Tolerance Principle (TP) may provide another solution to some of these challenges (Yang, 2016). It provides a tipping point for rule productivity based on the level of rote-memorized exceptions that the rule needs to withstand. A mathematical consequence of the TP is that rules defined over smaller vocabularies can tolerate a larger fraction of exceptions. This property is attractive. First, it makes the TP a promising candidate for modeling early language development, when children's vocabulary size is quite limited. Second, if the learner fails to acquire a broad rule defined over a vocabulary, the vocabulary can be partitioned into subsets, within which narrower rules may be identified by recursive application of the TP. In this work, we propose an abductive learning model of morphology, which we call *Abduction of Tolerable Productivity* (ATP). ATP is a search procedure that recursively hypothesizes rules and evaluates productivity, until all words in the vocabulary are either accounted for by productive rules or listed as exceptions.

Background

Psychological and Developmental Considerations

Computational models of cognition, which inevitably make simplifying assumptions, must still operate within the boundaries established by empirical research. We review several important lines of results from child language acquisition, which serve as the design specifications for computational models of morphological learning.

Data size One of the most remarkable characteristics of child language acquisition is how small the vocabulary on which they learn is. All English-learning children acquire the productive use of past tense by age three (Kuczaj, 1977), some even before age two (Brown, 1973). Likewise, German-learners show over-regularization of suffixes such as *-(e)n* and *-e* before or around age two (Mills, 1986; Elsen, 2002). At such an early stage of development, children have a very modest vocabulary: a two-year-old English learner's total vocabulary is 500 at most, and a three-year-old's vocabulary has an upper limit of just over 1000; most children's vocabulary size is considerably smaller (Fenson et al., 1994; Hart & Risley, 1995). See (Bornstein et al., 2004; Szagun, Steinbrink, Franik, & Stumper, 2006) for similar findings in other languages. Thus, a developmentally plausible cognitive model should be able to learn the core morphology of inflection from datasets containing only a few hundred words.

Productivity and Wug Tests The classic study by Berko (1958) was the first systematic demonstration that young children apply productive rules to form noun plurals (*wug-wugs*) and past tense verbs (*rick-ricked*). Berko also found that children nearly categorically resist the analogical use of irregular forms, even for words very similar to existing irregulars. For example, only 1 of the 86 children in the study produced *glang* for the novel verb *gling*. The categorical status of productivity is also strongly confirmed in naturalistic production. One of the first quantitative longitudinal studies (MacWhinney, 1978) noted that regular forms are rarely analogized into irregular forms, while irregular forms are often regularized. More recent quantitative studies estimate the rate of past tense over-regularization at 8-10% of all past tense forms (Maratsos, 2000; Maslen, Theakston, Lieven, & Tomasello, 2004). By contrast, children almost never over-extend an irregular form (e.g., *bite-bote* from *write-wrote*, *fry-frew* from *fly-flew*): the most comprehensive study places the error rate at 0.2% (Xu & Pinker, 1995).

The Tolerance Principle

The Tolerance Principle (TP) is a cognitively-motivated, theoretical tipping point that makes quantifiable predictions about when a linguistic process is used productively. It is inspired by studies of lexical processing and the hypothesis that children use a morphological process productively when it is computationally more efficient to do so. The TP only depends on two quantities: N—the number of words in the rule's scope—and e, the number of exceptions. For instance, the -s process for English pluralization applies to singular nouns and has exceptions like *children, sheep, fish*, etc. In our model, rules are of the form $r : A \Longrightarrow C$ where A is the antecedent and C the consequent; N measures how many times A applies and e measures the number of times C fails to follow from A. Given a rule r with a scope of size N and e exceptions, the TP states that

$$r \text{ is productive iff } e \leq \theta_N \triangleq \frac{N}{\ln N}. \tag{1}$$

The TP has consistently made accurate predictions on when children accept rules as productive and when they do not, as confirmed in artificial language learning experiments (Schuler, Yang, & Newport, 2016; Koulaguina & Shi, 2019) with precisely controlled conditions. Productivity under the TP is categorical, which mirrors children's morphological use reviewed above. It is also parameter-free: the two values N and e are word counts directly from the training data that require neither parameter tuning nor statistical fitting.

Critically for complex morphological systems, the TP can be applied recursively: if a rule r is unproductive over its current scope, its scope can be recursively narrowed. For example, none of the five German noun plural suffixes covers a sufficiently large number of nouns to tolerate the rest as exceptions: the learner will attempt to organize nouns into subcategories—defined by morphological gender and the phonological form of the noun—to recursively search for productivity within. As we will see, this divide-and-conquer strategy is fully automated in our search procedure.

Model

Abductive Search for Productivity

The recursive application of the TP lends itself to a Peircean abductive learning procedure. Given data in pairs, such a procedure hypothesizes rules that map one set to the other (e.g., lemmas to their inflected forms). If no such rule is productive via the TP, then it subdivides the words according to some feature—that is, it refines the hypothesis over more narrow scopes—and recursively tries again.

We propose just such a procedure: *Abduction of Tolerable Productivity* (ATP). Its input is a set $\mathcal{X}$ of **instances** in the form $(\ell, \mathcal{F}, \text{I})$, where ℓ is a lemma, $\mathcal{F} \subseteq \Omega$ is a set of features from the **feature space** Ω, and I is the inflected form corresponding to lemma ℓ and features $\mathcal{F}$; for instance (`walk`, {`3`, `SG`, `PST`}, `walked`), where the example features `3`, `SG`, `PST` carry the information that the inflection is 3rd person, singular, and past tense.

ATP recursively grows a decision tree, where each instance's "label" is the morphological change that produces the inflection; the resulting tree thus encodes a map from lemma and features to inflection. In principle ATP could model any type of inflectional morphology, but the inflections modeled in this paper involve only suffixation, in which case an instance's "label" is the suffix that is concatenated to the lemma. For clarity, we describe ATP in terms of suffixation.

At each recursive level, the decision of which feature to split on is selected to maximize consistency: the relative frequency of the most frequent suffix that the instances with that feature take. That is, it splits on $\hat{s} = \arg\max_{s \in \Omega} = \frac{f_{\sigma_{\max}(\mathcal{X}_s)}}{|\mathcal{X}_s|}$, where $\mathcal{X}_s = \{(\ell, \mathcal{F}, \text{I}) \in \mathcal{X} : s \in \mathcal{F}\}$ is the set of instances with feature s and $f_{\sigma_{\max}(\mathcal{X}_s)}$ is the frequency of the most frequent suffix in $\mathcal{X}_s$. The split is formed by recursing separately on those instances with feature s (i.e., $\mathcal{X}_s$) and those without it (i.e., $\mathcal{X} \setminus \mathcal{X}_s$).

If *productivity* is defined as the frequency of the most frequent suffix in $\mathcal{X}_s$, then this is one possible formulation of Yang (2016)'s proposed *Principle of Maximize Productivity*, which he described as "Pursue rules that maximize productivity." Furthermore, it is motivated by findings that consistent patterns promote generalization and category formation in learning (Gerken, 2006; Reeder, Newport, & Aslin, 2013).

Features The set of features provided as input is expanded to include regularities in the ending of lemmas. For instance, the null suffix for German nouns is predominantly used for non-feminine lemmas ending with a schwa followed by *l*, *r*, or *n* (Wiese, 1996). Regularities of this sort are extracted at each recursive level as the shortest lemma endings that productively predict a suffix in $\mathcal{X}$ (the training instances at that level).

To do so, ATP considers suffixes one-by-one and, for each suffix, it considers the endings of lemmas that take that suffix, starting with the shortest ending. It caches any ending where the number of lemmas that do have the ending but do *not* take the suffix is tolerably low; effectively, *ending* $\Longrightarrow$ *suffix* passes the TP. For example, enough English verb lemmas ending in [k] take the [-t] suffix to pass the TP; similarly for [p], [ʃ], and other voiceless segments. We denote the lemmas that take any such endings as $\mathcal{E}$ (e.g., $\mathcal{E}$ = {lemma : lemma ends in [k] or [p] or ... or [ʃ]} in the English verb example). If $\mathcal{S}$ consists of the lemmas whose inflections take the suffix being considered, then ATP tests the TP for both $N_1 = |\mathcal{E}|, e_1 = |\mathcal{E} \setminus \mathcal{S}|$ and $N_2 = |\mathcal{S}|, e_2 = |\mathcal{S} \setminus \mathcal{E}|$. If both tests pass the TP, the endings contributing to $\mathcal{E}$ are added to the feature space Ω. The purpose of these tests is to establish that there is a productive relationship between the endings and the suffix. In the running example, all words that end in a voiceless segment other than [t] take the [-t] suffix, and all words that take the [-t] suffix end in a voiceless segment, so both tests pass. Further examples of the result can be seen in Figs. 3-4, where sets of lemma endings appear in brackets, separated by "|" for *logical or*. This has the interpretation of picking out the lemmas that end in any such ending (i.e., $\mathcal{E}$).

Thus, at the end of the lemma ending extraction, the feature space Ω consists of whatever features were provided as input, together with the newly added sets of endings. ATP is agnostic to the nature of a feature, such as whether it has to do with content or form. It considers them all equally, seeking to maximize consistency, as described above.

Moreover, alternative approaches to incorporating or discovering features could be used without changing the fundamental recursive, abductive search procedure. Such adaptions could be useful in morphological processes beyond suffixation, where ATP could make use of features relevant to processes such as infixation and stem change within the same abductive search procedure.

Recursive and Base Cases Once a split is performed, the node's children are formed recursively on the partitioned set. Features that are equivalent across all instances in $\mathcal{X}$ are ignored since they are completely uninformative. The base case of this recursion is reached when the most frequent suffix at the node passes the TP or when there are no more features to split on. The path to a node with a productive suffix is the rule (e.g., Figs. 3-4), and all exceptions to the rule are memorized by storing them at the corresponding node. If a node is reached where no suffix is productive and there are no more features to split on, the node's instances are memorized.

Inflection Production

Once trained, ATP can use its acquired knowledge to produce the inflected form corresponding to a lemma and its features. It does so by traversing the decision tree to a leaf. If the leaf has a productive suffix, ATP produces the inflected form via the rule. If there is no productive suffix at the leaf, ATP makes an analogical guess by retrieving the memorized lemma at that node with the smallest character-level Hamming distance, padding 0's if necessary to make the two strings the same length. ATP uses the learned inflected form for the nearest-neighbor lemma to produce the inflection for the target lemma. In some cases—such as when nonce words are presented—some features may be unknown. In this case, an inflection is produced by traversing all logically compatible paths and using the most specific (i.e., deepest in the tree) compatible rule. A path is *logically compatible* if none of the features it specifies are contradicted by the word's known features. For instance, if the gender of a nonce word is unknown, and ATP encounters a branch point in the tree that is determined by gender, it takes both branches.

It is worth pointing out that analogical guessing forces an answer in order to quantitatively evaluate our model on test data. However, in many cases of language use, speakers may refuse to produce any form when they do not have an applicable productive rule, as in the well-known case of morphological gaps (Yang, 2016).

Code We make the code for ATP, along with instructions for using it on new data, available online.[1]

Evaluation

We evaluate the developmental plausibility of ATP with respect to the order in which rules are acquired and its quantitative performance on realistic data. We further evaluate how the regularities ATP discovers and its productions on Wug tests correspond to children's discoveries and productions.

Data and Setup

We briefly discuss the data that we use in our experiments, and name each dataset for reference.

CHILDES-DE contains 442 German nominative singular/plural pairs of child-directed speech from the Leo (Behrens, 2006) CHILDES corpus. Features encode gender, one of Ω = {`feminine`, `masculine`, `neuter`}, and

[1] https://github.com/cbelth/ATP-morphology.

frequency of words is known. We removed umlauts, as these follow a separate morphological process from suffixation. For developmental experiments, we sampled subsets of size 400, weighted by frequency, each sample modeling a child's 400 word vocabulary. The findings (reviewed earlier) that age-two children have productive plural morphology suggests that suffixes are learnable on vocabularies of this size or even smaller. While the CHILDES-derived words are used for training, we test on CELEX words. Morphological knowledge is generally acquired during childhood but must be able to generalize to other words in the lexicon.

CHILDES-EN was constructed from 6,539 word-inflection pairs extracted from child-directed English (MacWhinney, 2000). There are 3,321 plural nouns (23 irregular), 1,494 past tense verbs (120 irregular), and 1,724 progressive verbs (the exceptionless *-ing*). As reviewed earlier, children's vocabulary size during morphological learning is very modest. We thus log-binned words into 20 bins based on frequency and sampled 50 from each, simulating a child's vocabulary growth from 50 to 1K words. We repeated 100 times with different random seeds to simulate 100 children. Features are one of $\Omega = \{\texttt{progressive}, \texttt{past}, \texttt{plural}\}$.

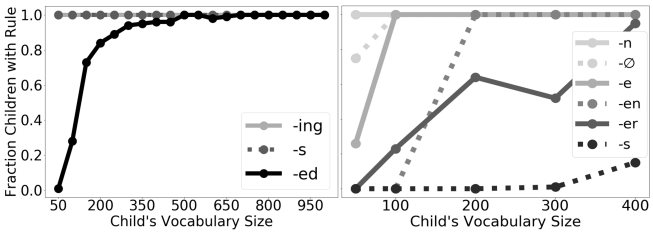
CELEX-EN contains just the stem/past tense pairs from CHILDES-EN, intersected with CELEX (Baayen, Piepenbrock, & Gulikers, 1996) for word frequency. For cross-validation, we formed 10 random 1000/100/200 train/dev/test splits, and for developmental experiments, we test on the top n items, $n \in \{100, 200, 400, 600, 800, 1000\}$. To simulate children, we added random jitters between 0 and 5 to the frequencies. This has the Zipfian effect of scrambling which low-frequency items appear in each learner's training data.

Setup Throughout the experiments, we use orthography for German and IPA transcriptions for English, following McCurdy et al. (2020)'s use of orthography and Kirov and Cotterell (2018)'s use of phonological transcriptions. Token frequencies are used to construct realistic datasets, but only type frequencies are used in learning and evaluation. We use paired t-tests at a 0.95 confidence level for statistical analysis. When testing Kirov and Cotterell (2018)'s ED model, we follow their setup, using an identical RNN implementation, trained for 100 epochs, with batch size 20. Both encoder and decoder RNNs are bidirectional LSTMs (Schuster & Paliwal, 1997) with two layers, 100 hidden units, and a vector size of 300.

Developmental Results

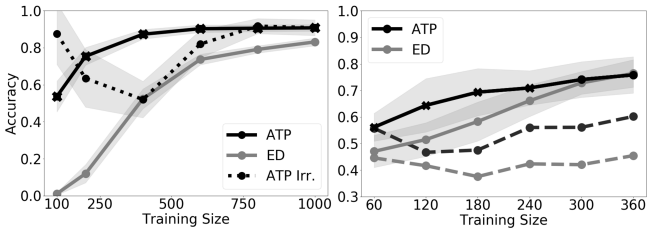
Order of Acquisition We study the order in which ATP discovers productive processes in English and German to assess its consistency with children's developmental patterns on both English pluralization, verb past tense, and verb present participle (CHILDES-EN) and German pluralization (CHILDES-DE).

Figure 1 shows that ATP acquires the *-s* for pluralization (e.g., `book, books`) and *-ing* for the verb present participle



(a) English **(b)** German

Figure 1: Fraction of "children" (ATP runs on different data) who have learned each major suffixation rule. The order of acquisition closely follows child development. Legend order and shade of gray match acquisition order. All suffixes are consistently acquired, except German *-s* ($\approx 20\%$).



(a) English **(b)** German

Figure 2: Accuracy on English and German vs. ED. ATP outperforms ED by a stat. sig. amount on English and the first four training sizes for German. English accuracy on irregulars in the training data (dotted line) demonstrates the classic U-shape. The dashed line on German is the performance when the models are presented lemmas without gender.

(e.g., `walk, walking`) on the smallest dataset (50 pairs). In contrast, the exception-laden *-ed* of the past tense takes longer to overcome the irregulars, but is consistently acquired by around 500 words. These patterns align well with the order of morpheme acquisition by English-learning children (Brown, 1973).

On German pluralization, which has five primary suffixes (*-(e)n*, *-e*, *-∅*, *-er*, *-s*; we separate *-n* and *-en*), ATP acquires -n immediately (on 50 words), which closely follows "A" in Elsen (2002)'s diary study, who learned *-e(n)* words most quickly. By 100 words, *-∅* and *-e* have been acquired; again matching "A," where the rate of learning of *-∅* and *-e* words is virtually identical. Other longitudinal studies have also shown the very early acquisition of these suffixes (Gawlitzek-Maiwald, 1994). The *-er* suffix is learned by 95% of simulated children by 400 words. In Elsen (2002), *-s* and *-er* words were learned at similar rates, while other studies (Köpcke, 1998; Bittner, 2000; Szagun, 2001) also find that *-s*, while productive, generally emerges later. In our data, *-er* is attested 15 times, while *-s* only 8 times; Elsen (2002) reports that A's *s*-over-regularizations jump at 2;1, when her vocabulary contains around 50 *-s* words, which may explain the difference.

Performance We evaluate quantitative performance on CELEX-EN and CHILDES-DE in terms of how accurately the model generates inflections for held-out (`lemma`, `features`) pairs as the training data is progressively grown, simulating vocabulary growth. We compare to Kirov and Cotterell (2018)'s ED model. Results are in Fig. 2.

On English past tense, ATP outperforms ED by a statistically significant amount on all training sizes. Furthermore, when evaluated on irregulars that it has seen during training, ATP over-applies learned rules before correcting, exhibiting the developmental regression (Kuczaj, 1977; G. Marcus, Pinker, Ullman, Hollander, & Xu, 1992) that is a hallmark of the English past tense acquisition (dotted line). Before the model learns the productivity of *-ed*, all training verbs are memorized (irregulars and regulars alike). Once the productive rule is learned, the model erases the memory of rule-following verbs, which no longer require storage. The two drops in irregular training accuracy correspond with the two largest jumps in test accuracy, demonstrating the acquisition of a rule (likely /-t/ and /-d/; see Fig. 3). Throughout learning, all test errors were over-regularizations of rules (unless guessing occurred due to no rule having yet been acquired), matching the regularization vs. irregularization contrast in child acquisition (Berko, 1958; MacWhinney, 1978; Xu & Pinker, 1995).

On German pluralization, ATP outperforms ED by a statistically significant amount on the first four datasets (60-240 words) and is statistically indistinguishable after that point. Though the training data contains gender, the models can also be tested with novel nouns without gender information. Plural suffixation is not only conditioned on gender but also the phonological properties of nouns (Wiese, 1996; Zaretsky & Lange, 2015). ATP always outperforms ED without gender, indicating that it is more capable of extracting the phonological regularities in the German system.

In addition to its accuracy, ATP runs in seconds on all dataset sizes, compared to minutes for ED.

We note that the setup of this evaluation differs in certain respects from that used in the neural modeling literature. In particular, each number in Fig. 2 is the average *test* accuracy of *fully-trained* models on each of the training sets of that size (with the exception of the dotted U-shape line, as discussed above). In contrast, learning curves in works like Kirov and Cotterell (2018) report *training* accuracy of *partially-trained* models by epoch. Moreover, in such works, the model's final test-accuracy is an average only over different random initializations of the model, holding the training data constant. We chose our setup because our focus is on modeling development, and evaluating periodically on held-out test data is a more faithful measure of a model's developmental trajectory.

Acquired Knowledge

Discovered Rules ATP learns rules explicitly, represented in a decision tree. The output trees in Figs. 3-4 have transparent linguistic interpretations.

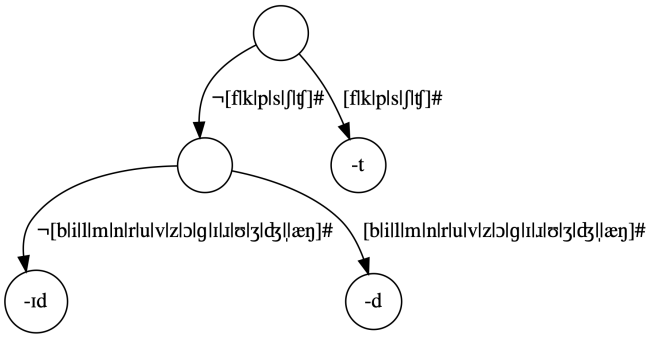


Figure 3: ATP decision tree for CHILDES-EN past tense. IPA symbols in brackets are separated by "|" to indicate that a lemma that ends in any of the listed endings follows (or does not follow, if preceded by "¬") the branch.

Table 1: Correlations with human production results (bold = stat. sig.). Training ATP on just 400 words of child-directed speech yields higher correlation with human productions than training a NN on 22x times the data (McCurdy et al., 2020).

	Neuter			Unknown		
	%R	%NR	ρ	%R	%NR	ρ
-(e)n	0.17	0.04	-0.26	0.19	0.23	**0.43**
-e	0.27	0.35	-0.14	0.45	0.62	0.01
-∅	0.11	0.0	**0.55**	0.07	0.00	**0.55**
-er	0.44	0.17	**0.53**	0.29	0.0	**0.46**
-s	0.01	0.44	0.3	0.01	0.15	**0.64**
other	0.00	0.00		0.00	0.00	

The English tree (Fig. 3) correctly characterizes the *-ed* rule for English past tense: the first branch to the [-t] node splits on voiceless stem endings, the next branch to [-d] splits on voiced stem endings, and [-ɪd] captures the remaining instances.

The German tree (Fig. 4) captures all five primary suffixes (*-(e)n*, *-e*, *-∅*, *-er*, *-s*). The *-s* suffix—well-known to be the most idiosyncratic—is picked up at the deepest point in the tree, conforming to its role as the rule of last resort (G. F. Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995).

This particular English tree was learned from 1000 words (one simulated child) of CELEX-EN, and the German tree was learned from the 442 German nouns in CHILDES-DE.

Wug Test Production The Wug test can be used to assess the morphological knowledge acquired by both humans and computational models. G. F. Marcus et al. (1995) carried out a Wug test study with 24 nonce words, divided into Rhyme (rhymes with familiar German words) and Non-Rhyme (unfamiliar) nonce words. The same stimuli were used by Zaretsky and Lange (2015) and McCurdy et al. (2020). McCurdy et al. (2020) compared the predictions of Kirov and Cotterell (2018)'s ED neural network model to human productions. We passed the same 24 nonce words to ATP trained on CHILDES-DE. This experiment follows McCurdy et al. (2020): it computes (a) the fraction of productions for each

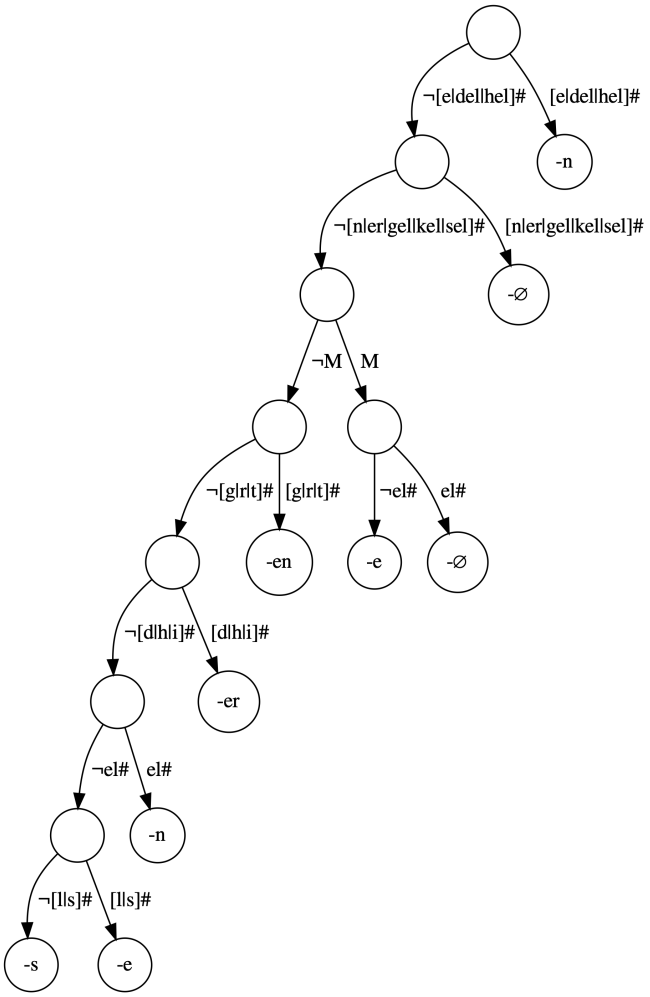


Figure 4: ATP decision tree for CHILDES-DE. "M" indicates nouns with masculine gender.

suffix, divided into Rhyme (R) and Non-Rhyme (NR) and (b) Spearman's rank correlation (ρ) between the production probabilities of human and ATP productions. To simulate multiple people, we ran ATP on 500 samples of size 400 (by frequency) from CHILDES-DE. As in McCurdy et al. (2020), we treated each run as a model of a human, and computed the production probabilities for a particular word-suffix pair as the fraction of models that produced that suffix for that word.

When presenting ATP nonce words with unknown gender (right columns of Tab. 1), ATP correlates with human productions statistically significantly for all suffixes except *-e*. This is in contrast to the ED model, which shows no correlation for any suffix (McCurdy et al., 2020). Moreover, ATP was trained on a realistic 400 words of child-directed speech—over which the core inflectional morphology is learned—while ED was trained on 8.7K words, or 22x the data. Furthermore, both Zaretsky and Lange (2015) and McCurdy et al. (2020) found that *-(e)n* and *-s* are used more frequently for non-rhyme words than rhyme words. ATP again matches this behavior, and ED did not (McCurdy et al., 2020).

When presenting nonce words with unknown gender to ATP, the correlation with human performance is higher than when presenting with neuter gender (left columns). The participants in McCurdy et al. (2020) were presented nonce words with the neuter determiner *das*. However, the impact of this is unknown. As noted earlier (Wiese, 1996), the suffix choice is conditioned on both gender and phonology, and a conflict generally arises only in non-feminine gender (nearly all feminine nouns add *-(e)n*). When they conflict, as in the test stimuli, human subjects may persist with a gender-conditioned rule, or they may eschew gender altogether and rely on phonological similarity to existing words (Zaretsky & Lange, 2015). Our model makes a testable prediction regarding this open question that can be pursued in future research.

Discussion

ATP is not limited to morphology acquisition. Further research could investigate its use in learning phonology, syntax, or anything where linguistic generalizations are learned. Children's adeptness at language acquisition is a constant reminder of how much knowledge can be learned from tiny amounts of evidence. The decision trees that ATP learns leave a step-by-step trace of what was learned and how. They thus provide explicit places to look for the steps children take in acquiring language.

Acknowledgments

This work was supported by an NSF GRFP to CB. We thank the participants of the 2020 distributional learning seminar at the University of Pennsylvania for helpful discussion.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The celex lexical database (cd-rom)*.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and cognitive processes*, *21*(1-3), 2–24.

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

Bittner, D. (2000). Gender classification and the inflectional system of German nouns. *Trends in Linguistics Studies and Monographs*, *124*, 1–24.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., … Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child development*, *75*(4), 1115–1139.

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

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Cambridge, MA: MIT Press.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL* (pp. 3868–3877).

Elsen, H. (2002). The acquisition of German plurals. In *Morphology 2000: Selected papers from the 9th morphology meeting, vienna, 25-27 february 2000* (Vol. 218, p. 117).

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Gawlitzek-Maiwald, I. (1994). How do children cope with variation in the input? the case of German plurals and compounding. In R. Tracy & E. Lattey (Eds.), *How tolerant is Universal Grammar? essays on language learnability and language variation* (p. 225-266). Tübingen: Niemeyer.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Baltimore, MD: Paul H Brookes Publishing.

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651-665.

Köpcke, K.-M. (1998). The acquisition of plural marking in English and German revisited: Schemata versus rules. *Journal of Child Language*, *25*(2), 293-319.

Koulaguina, E., & Shi, R. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, *26*(4), 416–435.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589-600.

Lignos, C., & Yang, C. (2016). Morphology and language acquisition. In G. Hippisley Andrew R. abd Stump (Ed.), *The Cambridge handbook of Morphology* (p. 765-791). Cambridge: Cambridge University Press.

MacWhinney, B. (1978). *The acquisition of morphophonology*. University of Chicago Press.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Maratsos, M. (2000). More overregularizations after all: New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen and Xu. *Journal of Child Language*, *27*, 183-212.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57 4*, 1-182.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*, 189-256.

Maslen, R., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research*, *47*(6), 1319-1333.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*(11), 465-472.

McCurdy, K., Goldwater, S., & Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. In *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL* (pp. 1745–1756).

Mills, A. (1986). *The acquisition of gender: A study of English and German*. Berlin: Springer.

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.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Science*, *6*(11), 456-463.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30–54.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs.

Schuler, K. D., Yang, C., & Newport, E. L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *Cogsci* (Vol. 38, pp. 2321–2326).

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language*, *21*, 109-141.

Szagun, G., Steinbrink, C., Franik, M., & Stumper, B. (2006). Development of vocabulary and grammar in young german-speaking children assessed with a german language development inventory. *First Language*, *26*(3), 259-280.

Wiese, R. (1996). *The phonology of German*. Oxford: Clarendon.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*(3), 531-556.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

Zaretsky, E., & Lange, B. P. (2015). No matter how hard we try: Still no default plural marker in nonce nouns in modern high german. In *A blend of malt: Selected contributions from the methods and linguistic theories symposium* (pp. 153–178).