Investigating the Relationship Between Linguistic Representation and Computation through an Unsupervised Model of Human Morphology Learning*

Erwin Chan
University of Arizona
echan3@u.arizona.edu

Constantine Lignos
University of Pennsylvania
lignos@cis.upenn.edu

Abstract

We develop an unsupervised algorithm for morphological acquisition to investigate the relationship between linguistic representation, data statistics, and learning algorithms. We model the phenomenon that children acquire the morphological inflections of a language monotonically by introducing an algorithm that uses a bootstrapped, frequency-driven learning procedure to acquire rules monotonically. The algorithm learns a morphological grammar in terms of a Base and Transforms representation, a simple rule-based model of morphology. When tested on corpora of child-directed speech in English from CHILDES (MacWhinney, 2000), the algorithm learns the most salient rules of English morphology and the order of acquisition is similar to that of children as observed by Brown (1973). Investigations of statistical distributions in corpora reveal that the algorithm is able to acquire morphological grammars due to its exploitation of Zipfian distributions in morphology through type-frequency statistics. These investigations suggest that the computation and frequency-driven selection of discrete morphological rules may be important factors in children's acquisition of basic inflectional morphological systems.

Keywords: Language Acquisition; Morphology; Unsupervised Learning; Cognitive Modeling

1 Introduction

The relationship between linguistic representation and computation can be explored through models of unsupervised learning from a corpus. A system for unsupervised language learning embodies a formalism for

*This is the accepted manuscript of a paper published in *Research on Language and Computation*, Volume 8, Issue 2, pp. 209–238. The final publication is available at Springer via http://dx.doi.org/10.1007/s11168-011-9077-2.

linguistic representation, whose end shape is determined by the learning algorithm and the content of the input data. Unsupervised algorithms have been developed to learn a variety of levels of linguistic structure, such as word segmentation (Brent and Cartwright, 1996), morphology (Goldsmith, 2001, 2006), distributional part of speech classes (Schütze, 1993; Redington et al., 1998), syntactic constituency and dependency (Klein and Manning, 2004), and semantic word classes (Deerwester et al., 1998). These algorithms have shown that the statistical content of linguistic usage and principles of machine learning may, to some extent, explain how humans may be learning language.

An interesting application for unsupervised learning is in modeling child language acquisition. An algorithm that acquires linguistic structure in conditions similar to that of children may be informative about possible mental mechanisms for representing and processing language, and the contribution of the input. However, not any unsupervised algorithm is suitable as a model for child language acquisition: a cognitively-oriented algorithm should also model children's behaviors in language acquisition under appropriate input conditions. The benefit of this is that such requirements on the learning process impose additional constraints upon the learning architecture to be developed, therebhy leading to a deeper understanding of the principles involved in learning languages.

The particular problem that we choose to model is the observation that children acquire the set of morphological inflections for their language monotonically, in a sequence over time (Brown, 1973; Slobin, 1973; Hooper, 1979; Bybee, 1985; Slobin, 1985-1997; Dressler, 2005); see Table 4 for an example for English. Cross-linguistically, it appears that children's first inflections tend to be of nominative singular forms for nouns and adjectives, and third-person present tense singular forms or infinitive for verbs. Other inflections are acquired at a rate dependent upon the phonological and syntactic complexity of the morphological forms. Inflections that are easily identified and unambiguous tend to be acquired sooner than those that are ambiguous. Why do we see these behaviors in child language acquisition? We believe that a computational model may provide a principled explanation.

In this paper, building upon Chan (2008) and Lignos et al. (2009, 2010) we present a batch unsupervised algorithm for morphology learning in which the grammar is constructed monotonically. The algorithm acquires a *Base and Transforms model*, a simple rule-based model of regular morphological systems in which a word consists of a lexical stem and a grammatical suffix. Rule-based models of morphology in general (Chomsky and Halle, 1968; Halle, 1973) apply rewrite rules to a morphological base forms to produce the set of inflected forms of a lexeme. The benefit of employing a rule-based representation is that the monotonic growth of a grammar can be easily modeled. Beginning with an empty grammar, the algorithm acquires rules one at a time, without subsequently changing them.[1] The algorithm is applied to data of child-directed speech in English, and the order of acquisition of suffixes is compared to previous analysis in the language

acquisition literature (Brown, 1973).

To gain a better understanding of the learning model, we also analyze statistical distributions of morphological data in corpora and consider their implications for linguistic representations and learning. A key feature of the algorithm is how it exploits Zipfian distributions of morphology. The algorithm primarily involves greedy bootstrapping of a rule-based model of morphology through type-based computations. The ability of the algorithm to effectively acquire a morphological system is due to the fact that the input data in language usage is Zipfian-distributed.

With an understanding of the algorithm in terms of its structural and statistical learning biases, we have a computational account of children's behaviors in morphology acquisition: the monotonic acquisition of inflections by children may be attributed to the learning of rules through type-based computations over data from linguistic experience. This explanation relies on the assumptions that morphological data is Zipfian-distributed, and that children have the capacity to compute rule-based representations or their equivalent. The cognitive focus of our investigations therefore leads to insights about the relationship between linguistic representation, statistical distributions, and computation.

The rest of this paper is as follows. Section 2 reviews previous work in morphology induction, from the point of view of developing a cognitive model of morphology acquisition. Sections 3 and 4 present the algorithm and experiments on corpora. In Section 5 we examine the statistical characteristics of morphology in corpora and their implications for linguistic representation and learning. Section 6 discusses the implications of this work for language acquisition, theories of linguistic representation, and unsupervised learning techniques, and also model limitations and future work. Section 7 concludes.

2 Previous Work in Morphology Learning

Many different algorithms have been designed for morphology learning, involving a wide range of problem definitions, formalisms for linguistic representation, and learning techniques. As we desire an unsupervised algorithm for learning morphology that can also serve as a cognitive model of child language acquisition, several characteristics are desirable. First, the quantity of input data should be comparable with what children encounter in the early period of language acquisition.[2] Second, the linguistic representations that are assumed to be available in the input must be justified by the existence of other unsupervised algorithms for inducing those structures. For example, the assumption of discrete words as input to morphology induction may be justified by the body of research on the problems of inducing and segmenting waveforms into phonemes (e.g. Lin, 2005), and segmenting phoneme sequences into words (e.g. Gambell and Yang, 2004). Third, the inflections of a language should be acquired monotonically, and the particular order should be

consistent with children's acquisition. Other characteristics may be desirable as well for an algorithm to be a cognitive model, such as online processing of the input, but we will concentrate upon those mentioned above.

2.1 Supervised inflection learning models

A large number of models have been developed for single-inflection learning. Given a set of word pairs *(base$_i$, derived$_i$)*, where *base$_i$* is a base form, *derived$_i$* is a form derived from the base, and all *derived$_i$* are of the same inflection, the learning task is to acquire the ability to map new base forms to their corresponding derived forms, whether regular or exceptional. The prototypical single-inflection problem has been the acquisition of English past tense verbs. For example, the input to a past tense learner could be word pairs such as (in orthography) *(walk, walked)*, *(beat, beat)*, *(eat, ate)*, etc. "Past tense" models of learning and processing have been a topic of strong interest in cognitive science (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Pinker, 1999; Pinker and Ullman, 2002; McClelland and Patterson, 2002), and many learning systems have been developed for this problem (Golding and Thompson, 1985; Rumelhart and McClelland, 1986; Wothke, 1986; Ling, 1994; Daelemans et al., 1996; Mooney and Califf, 1996; Molnar, 2001; Clark, 2001, 2002; Albright and Hayes, 2002).

Despite the cognitive focus of "past tense" learning, it is problematic as a model for language acquisition. The learner begins with pairs of morphologically related words, each of one of two different inflections, and it knows which of the word forms is the base. What justifies making these assumptions? Would it be logically possible for an unsupervised algorithm be able to supply this information, before the learner begins the process of learning the rule or mapping? Current unsupervised algorithms would not be able to provide such information as a pre-processing step.

An issue not addressed by single-inflection learners is how to acquire the full set of inflections in a language. Most systems that address this problem are have a technological orientation (Johnson, 1984; Theron and Cloete, 1997; Yip and Sussman, 1997; Manandhar et al., 1998; Kazakov and Manandhar, 2001; Oflazer et al., 2001; Zajac, 2001; Plisson et al., 2004; Wicentowski, 2004; Stroppa and Yvon, 2005; Carlson, 2005; Shalonova and Flach, 2007; Dreyer et al., 2008). Like single-inflection learners, these systems assume knowledge of the base and the pairing of words with the base.

2.2 Unsupervised models of morphology learning with unrestricted data sets

Many unsupervised algorithms have been developed for inducing morphological structure from a raw corpus of words. While in theory unsupervised learning of morphology encompasses the above inflection learning

problems, in practice systems have concentrated on more basic tasks, such as discovering the morphemes within words, organizing word components into a grammar, or discovering morphologically related words. Below, we survey several classes of unsupervised systems; see Chan (2008) for a comprehensive review.

Segmenting words into morphemes. Harris (1955, 1970) proposes that phoneme and morpheme boundaries can be determined through statistical information, by counting letter successors at each position in a sequence. Harris' technique has been adopted and refined by many researchers, as early as Hafer and Weiss (1974), and as recently as Bordag (2007). Research has also been conducted on segmenting words for agglutinative languages (Creutz, 2003; Creutz and Lagus, 2004; Argamon et al., 2004; Hu et al., 2005a).

Discrete grammars of segmented words. Goldsmith (2001, 2006) develops *Linguistica,* a system to acquire a set of *signatures*, discrete paradigm-like data structures describing the morphological segmentation of the words of a corpus. Linguistica begins with a signature for the vocabulary of the input corpus. In an iterative process, improved grammars are proposed by various heuristics that propose alternative segmentations, causing signatures to split or merge. Grammars are assigned a numerical score according to a formula based on the Minimum Description Length principle; the least costly grammar is chosen.

Probabilistic grammars of segmented words. Bacchin et al. (2005) develops a probabilistic model and parameter estimation algorithm for segmenting words into stems and affixes. Snover and Brent (2001, 2002) formulate a Bayesian learning procedure for a probabilistic model consisting of paradigmatic classes of stems and suffixes. Goldwater et al. (2006) employ Gibbs sampling to estimate the parameters for a probabilistic model of morpheme segmentation in which the data distributions are generated by a Pittman-Yor process. Poon et al. (2009) learn log-linear models for segmentation through contrastive estimation and Gibbs sampling.

Probabilistic segmentation and rewrite rules. Naradowsky and Goldwater (2009) expand the algorithm of Goldwater et al. (2006) to include context-dependent spelling change rules. Both of these algorithms were tested on English verbs only, rather than all the words of a corpus.

Discrete rule-like representations. Several systems (Schone and Jurafsky, 2000, 2001; Freitag, 2005; Demberg, 2007; Plisson et al., 2007) acquire representations that account for wordform generation through string rewriting rules, a more powerful process than concatenation. However, these representations only state that pairs of words are morphologically related, and do not interpret them as consisting of a base and

a derived form, as is standard in supervised rule learning. Without the concept of a base, it is difficult to interpret the acquired representations as models for wordform generation.

2.3 Unsupervised induction of morphology and part of speech classes

Most work in unsupervised induction of morphology utilizes the spelling and frequency statistics of words to identify morpheme strings. Part of speech information can also be useful, and algorithms have been developed for induction of distributional word classes through sentential contexts (Schütze, 1993; Redington et al., 1998). There have been efforts to combine morphology induction and distributional induction of word classes (Parkes et al., 1998; Wicentowski, 2002; Clark, 2003; Higgins, 2003; Freitag, 2004, 2005; Hu et al., 2005b; Biemann, 2006; Dasgupta & Ng, 2007; Can and Manandhar, 2009).

There are several potential contributions of word class induction to morphology induction. One is to disambiguate syntactically ambiguous morpheme strings. For example, the suffix *-s* occurs on both nouns and verbs in English. Another is to identify allomorphic variants of strings; for example, we would like to know that the suffixes *-s* and *-es* in English orthography are both realizations of the same underlying morpheme (either a plural noun or present tense verb suffix). However, it has yet to be demonstrated that fine-grained morphological categories such as "second person plural future tense verb" can be induced.

2.4 Towards unsupervised learning of rules as a cognitive model

Previous work in unsupervised learning of morphology has been predominantly focused upon developing computational techniques for inducing different aspects of morphological structure from corpora, rather than cognitive modeling. Specific types of algorithms are incompatible with our goal of modeling children's monotonic acquisition of morphology. For example, consider any algorithm that iteratively refines a grammar for a corpus, whether by modifying the representations in discrete grammars (such as in Linguistica), or re-estimating parameters in a probabilistic model. If taken as a cognitive model, such an algorithm would incorrectly predict that children have morphological grammars that fully cover the input data at any point in time, even in the initial stages of acquisition.

It would be desirable to develop an unsupervised algorithm for learning a rule-based model of morphology. One reason is that cognitive modeling of the monotonic acquisition of inflections may be possible through such a representation. Another reason is more practical: in finite-state systems for morphological generation and processing (Sproat, 1992; Kaplan and Kay, 1994; Beesley and Kartunnen, 2003), rules are constructed by hand. While learning rules (or any other mechanism for mapping between forms) is straightforward in a supervised setting, the lack of annotated resources makes partially or fully unsupervised learning an

attractive alternative.

As an initial step towards an unsupervised rule learner, we simplify the definition of the learning problem. Instead of learning rules that specify both morphosyntactic and phonological features, we discover the morphological relationships among the words of a corpus, in terms of which are base forms and which are derived forms. For example, given the set of words *walk, walked, walking, talk, talked, talking*, we would like to learn that *walk* and *talk* are base forms for {*walked, walking*} and {*talked, talks*}. The relationships among these words can be represented by string rewrite rules indicating "add -*ed* to the base" and "add -*ing* to the base". In the following section, we develop a formalism and algorithm for discovering these relationships in an unsupervised manner. The identification of these string-level relationships in a rule-based model could potentially assist in the induction of more abstract featural representations (person, case, tense, etc.), which would ultimately be needed for a deeper linguistic analysis.

3 An Algorithm for the Unsupervised Learning of Rules

3.1 The Base and Transforms model

We now present a formalism for morphological derivations, called the Base and Transforms model, and a matched algorithm for the unsupervised acquisition of transforms and their associated word pairs. Both the model and algorithm were introduced by Chan (2008). The Base and Transforms model allows the representation of morphologically related words in a generative fashion by defining a base set of words and a set of transforms that can change the base into its derived forms. Transforms are acquired in a frequency-driven learning process, but the model itself is a discrete, non-probabilistic representation.

3.1.1 Base and derived forms

A set of morphologically related words can be represented as a single base form and one or more derived forms that can be created from the base through transforms. For example, the word *bake* will be the base form for the derived forms *baked*, *baking*, and *bakes*. We refer to any word produced by an inflectional or derivational morphological process as "derived." The algorithm can learn transforms that correspond to inflectional and derivational morphology, but it does not explicitly identify them as such.

3.1.2 Transforms

A transform is a rewrite rule applied a base to create a derived form that can operate at either the phonemic or the orthographic level. The current formulation of transforms is restricted to affixal morphology. A

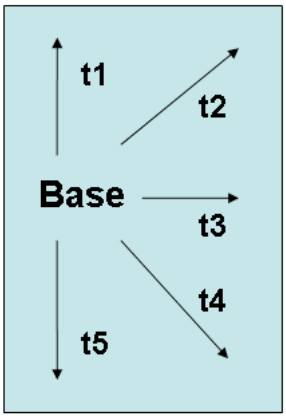

Figure 1: Generating a paradigm of six forms through a base and five transforms

transform is defined as two affixes *(s1, s2)*, where *s1* is removed from the base before concatenating *s2*. Thus to derive *baking* from *bake* we apply the suffix transform *(e, ing)*, removing *-e* from the base and then concatenating *-ing*. We represent a null affix as $, so that a purely additive affix is of the form *($, x)* and a purely subtractive one is of the form *(x, $)*, where x is not null.

A transform also has a corresponding word set, which is the set of base-derived pairs that the transform accounts for. The bases of a transform are the words that the transform can be applied to, and its derived words are the words created by applying the transform to its set of bases. Figure 1 illustrates how a morphological paradigm of six forms can be generated from a base and the application of five transforms to the base.

3.2 An algorithm for discovering transforms

We now present an unsupervised algorithm to discover the affixal morphology underlying the words of a corpus. It takes an unstructured set of words as input and returns a representation of the morphological grammar of the input corpus as represented by the Base and Transforms model.

The algorithm is an iterative, greedy procedure. In each iteration, it selects the transform that models as many word types as possible while meeting constraints for an acceptable transform. All word types in the corpus are placed in the Unmodeled set at the start of the algorithm's execution. As the algorithm acquires transforms, it places words in the Base or Derived word sets based on the function they serve in the learned transforms. The algorithm primarily uses the number of types that suffixes and transforms represent; token frequencies of words are only used to break ties. The operation of the algorithm requires several numerical parameters, including the maximum length of a suffix, the minimum size of a word after a suffix is removed, and thresholds for the minimum number of word types a transform can represent. Values

1. Place all words in the corpus in the Unmodeled set
2. Until a stopping condition is met:
   (a) Count suffixes in words of the Base and Unmodeled sets.
   (b) Hypothesize transforms from words in the Base and Unmodeled sets to words in Unmodeled.
   (c) Select the best transform.
   (d) Move the bases used in the transform to the Base set and the derived forms used by the transform to the Derived set.

Figure 2: The learning algorithm

for these parameters were set when developing the system on the Brown corpus.

An overview of the algorithm is given in Figure 2. Each word in the corpus belongs to one of three word sets at any point in execution: Base, Derived, or Unmodeled. All words begin in the Unmodeled set and are moved into Base or Derived as transforms are learned. The Base set contains the words that are used as bases of learned transforms. Similarly, the Derived set contains words that are derived forms of learned transforms. When proposing transforms, the algorithm creates word pairs whose bases are in the Base or Unmodeled set and whose derived forms are in Unmodeled. This results in a bootstrapping mechanism that encourages the reuse of existing bases for new transforms. The grammar created by the learner consists of the transforms learned and the base and derived words that they apply to.

We now present the learning loop of the algorithm in detail:

3.2.1 Count suffixes

Iterate over the words in the Base and the Unmodeled sets and count all of the suffixes of length 0-5 contained in each word, maintaining a separate count for the suffixes in the Unmodeled set and for the union of the Base and Unmodeled sets. For example, the word "hopeless" contains the suffixes (*-$*, *-s*, *-ss*, *-ess*, *-less*, *-eless*), and if it is only in the Base set those affixes would be counted toward the Base ∪ Unmodeled set's affix counts, but not the Unmodeled set's. A suffix is only counted if removing it leaves a sufficiently long stem, in this case three characters long. This length limitation exists to prevent the modeling of extremely short words that are likely closed-class or morphologically irregular words. Affixes are ranked by the number of types they appear in.

3.2.2 Hypothesize transforms

Hypothesize transforms of all combinations of the top 50 affixes as *s1* and *s2* of a transform. For example, from the common English suffixes *-$*, *-s*, and *-ing* the transforms *($, s)*, *(s, $)*, *($, ing)*, *(ing, $)*, *(ing, s)*, *(s, ing)* are hypothesized. For each hypothesized transform, check every word in the Base and Unmodeled sets that have the affix's *s1* and check whether the word that is the result of applying the transform to the word is in the Unmodeled set. If it is, add the base-derived pair to the word set of this transform. Transforms are ranked by the number of word pairs they account for, without regard to the frequency of the words in those pairs.

3.2.3 Select a transform

The highest ranked transform is selected, provided it meets the criteria for an acceptable transform. A transform should be rejected if it appears to be modeling a relationship between two forms that should both be derived forms, rather than a relationship between a base and a derived form. To reject these transforms, a transform must have a sufficiently low *overlap ratio*. A transform's overlap ratio is calculated as the ratio of the *stem overlap* to *base overlap*, defined as follows:

Base overlap. The base overlap is the number of base forms in the proposed transform that are base forms in the current grammar.

Stem overlap. The stem overlap is the number of base forms' stems (computed as the first four characters) in the proposed transform that are also stems of words in the Base set. The stem overlap is an approximation of the lexical similarity between two sets of words.

A high overlap ratio implies that the bases in the transform's word set are very similar to words in the Base set, but not members of it. The likely cause is that the bases used in the transform are derived forms of words in the Base set, and thus accepting the transform would cause the Base set to include multiple inflections of the same lexical category. This is undesirable as it results in inconsistent base forms. For example, in English the overlap ratio is used to reject the transforms *(ing, ed)* and *(ed, ing)* in the second iteration when running on the Brown corpus. Later, the more desirable rules *($, ed)* and *($, ing)* are learned, which reflect the desired base-derived relationship.

If the first and second ranked transforms account for the same number of types and are symmetric pairs, for example in English the transforms *($, s)* and *(s, $)*, a tie-breaking procedure is invoked:

Tie-breaking. For each of the transforms, count the number of base-derived pairs in which the base form is more frequent than the derived form. Choose the transform with the greater number of higher-frequency bases.

This tie-breaking procedure is typically only needed in the first iteration where the Base set is empty, as when the Base set is not empty, two symmetric transforms will represent a different number of types based on the number of words in the Base set that they are reusing.

3.2.4 Stopping condition

If there are no possible transforms remaining that account for five or more base/derived pairs, learning stops, as selection of any remaining transforms would only result in over-fitting the corpus.

4 Results

4.1 Evaluating the learner

In evaluating the output of the algorithm, we want to measure performance in two dimensions: the correctness of the morphological relationships identified and the proportion of morphological relationships in the language correctly identified. These measures translate to the traditional metrics of precision and recall. We compute both of these metrics using the CELEX Lexical Database (Baayen et al., 1995) as a gold standard.

Precision is computed as the proportion of base-derived pairs in the algorithm's output that are morphologically related in the gold standard. If a base-derived pair contains a word not present in the gold standard, the pair is ignored for the purpose of calculating precision. This affects some low frequency words that were not contained in CELEX and thus cannot be evaluated. Precision can be computed for an individual transform by considering its base-derived pairs or for the entire algorithm's output by considering the base-derived pairs of all transforms.

Recall is computed as the proportion of morphologically related word pairs identified in the gold standard that were identified as related in the algorithm's output. Two words (w_1, w_2) are related in the algorithm's output if the algorithm outputs the base-derived pair (w_1, w_2) or (w_2, w_1) or if the bases of w_1 and w_2 are the same. For example, if the algorithm learns the transforms *($, s)* and *(e, ing)* apply to *bake*, the words in the pair *(bakes, bake)* are related because there is a base-derived pair *(bake, bakes)*, and the words in the pair *(bakes, baking)* are related because they share the common base *bake*. The word pairs to be tested are created by generating all unordered pairs of morphologically related words in the gold standard also present in the corpus. For example, CELEX defines the following conflation set for the lemma *run*: *run, runs, ran,*

Corpus	**Types**	**Tokens**	**Trans. Learned**
Combined	7,174	730,328	23
Sarah	4,407	182,030	14
Adam	3,437	117,022	14
Nina	3,123	184,042	13
Peter	2,829	136,714	13
Naomi	2,511	52,760	9
Eve	1,935	57,760	9

Table 1: Token and type counts for CHILDES corpora used, ordered by type count

running. Thus we would test whether the following pairs of words are connected in the algorithm's output: *(run, runs), (run, ran), (run, running), (runs, ran), (runs, running), (ran, running)*. Each pair counts as a single hit regardless of the number of word pairs in the conflation set; there is no normalization for the number of words in a lemma.

4.2 Results on child-directed speech

To evaluate the algorithm's effectiveness as a model of language acquisition, we tested the algorithm on CHILDES corpora of English child-directed speech transcriptions (MacWhinney, 2000). Six children were chosen: Adam, Eve, Naomi, Nina, Peter, and Sarah. The corpora were pre-processed, removing any annotations and child utterances. Pronunciation data for all words in the corpus were obtained from CMUdict 0.6 (Weide, 1998), the Carnegie Mellon University Pronouncing Dictionary. If multiple pronunciations were found for a word, the first pronunciation was selected. Words for which no pronunciation could be found were removed from the corpus. The token and type counts of the CHILDES corpora used are given in Table 1. As shown in Table 1, the number of transforms learned in a corpus is directly proportional to the number of types present in the corpus. We present the algorithm's output when run on a corpus that combines of all children's data and on the individual corpus for each child.

4.2.1 Results on the combined corpus

The transforms learned when run on the combined corpus are given in Table 2 with example word pairs (in orthography), annotations for the most common morphological function, and type, token, and precision statistics for each transform. Each transform is given using ARPABET transcriptions, as formed from the transcriptions of each word returned by CMUdict. The algorithm ran for 23 iterations, achieving a cumulative precision of 93.07% and a recall of 76.03%.

The majority of the transforms correspond to common morphological rules in English, as shown by the

Iter.	Transform	Tokens	Types	Example	Morpheme	Precision
1	($, Z)	116591	518	trouble/troubles	Noun plural, Possessive, 3P Sg.	99.52%
2	($, IH.NG)	75830	284	land/landing	Present progressive	100.00%
3	($, S)	105930	195	ant/ants	Noun plural, Possessive, 3P Sg.	97.07%
4	($, IY)	21588	100	noise/noisy	Adjective derivation, Diminutive	69.23%
5	($, D)	24151	95	open/opened	Past tense	96.45%
6	($, T)	25720	89	step/stepped	Past tense	87.20%
7	($, ER)	45854	76	sing/singer	Agentive, Comparative	90.91%
8	($, AH.Z)	11501	58	fix/fixes	Noun plural, Possessive, 3P Sg.	100.00%
9	($, AH.D)	34326	29	lift/lifted	Past tense	98.21%
10	($, L.IY)	2836	20	bad/badly	Adverb derivation	100.00%
11	($, AH.N)	6091	19	hid/hidden	Past participle	72.73%
12	($, N)	1161	14	tore/torn	Past participle	81.25%
13	($, AH.L)	44171	13	what/what'll	Contraction with "will"	47.37%
14	($, AH)	3485	12	floor/flora	Spurious	0.00%
15	(AH.N, $)	4325	8	garden/guard	Spurious	33.33%
16	($, AH.S)	1116	7	fame/famous	Adjective derivation	42.86%
17	($, AH.N.T)	18750	7	could/couldn't	Contraction with "not"	37.50%
18	($, AH.T)	202	6	wall/wallet	Spurious	0.00%
19	(AH.L, L.IY)	250	6	passable/passably	Adverb derivation	100.00%
20	($, K)	8618	5	stay/steak	Spurious	0.00%
21	(IY, $)	1474	5	daddy/dad	Adjective derivation, Diminutive	60.00%
22	(AH.L, $)	2741	5	wiggle/wig	Spurious	0.00%
23	(T.IY, TH)	95	5	ninety/ninth	Ordinal derivation	100.00%

Table 2: Rules learned on English CHILDES data combined from six children

annotations in the "Morpheme" column. Because the algorithm operates at a phonemic level, allomorphs for each morpheme such as */Z/S/AH.Z/* for the noun plural are learned in multiple transforms. Also, in cases where multiple morphemes have the same phonemic representation, such as the plural and third person singular, a single transform may represent multiple morphemes. The most common regular verb inflections (plural, present progressive, past tense) are represented by seven of the first nine transforms learned.

In general, initially acquired transforms are more likely to represent linguistically reasonable morphological rules and attain a higher precision. Low type-frequency transforms are more likely to be spurious. For example, the transform *($, K)* is marked "spurious" because none of its base-derived pairs *(stay/steak, core/cork, stung/stunk, ming/mink, poor/pork)* contain morphologically related words. Spurious transforms begin to be learned at iteration 14. The transform (AH.N, $) demonstrates the limits of utilizing phonemic information without semantic information, containing word pairs such as *(garden/guard, kitten/kit, button/butt).*

Some transforms connect morphologically-related words, but do so by forming relationships between two forms that would ideally each be modeled as derived words. The transform *(T.IY, TH)*'s base-derived pairs *(forty/fourth, fifty/fifth, sixty/sixth, seventy/seventh, ninety/ninth)* connect morphologically related words but connect two forms that should each be derived from a common base. It would be more desirable to have transforms *($, T.IY)* to represent *nine/ninety* and *($, TH)* for *nine/ninth.* A similar phenomenon

Transform	**Adam**	**Eve**	**Naomi**	**Nina**	**Peter**	**Sarah**	**Mean**	**Std. Dev.**
($, Z)	1	1	1	1	1	1	1	0.00
($, IH.NG)	2	2	2	2	2	2	2	0.00
($, S)	3	3	3	3	3	3	3	0.00
($, T)	4	5	4	4	4	5	4.33	0.47
($, IY)	6	4	5	7	6	6	5.67	0.94
($, D)	7	8	6	5	5	4	5.83	1.34
($, ER)	5	6	7	8	7	7	6.67	0.94
($, AH.Z)	8	7	8	NL	8	8	7.8	0.40
($, AH.D)	9	NL	9	10	9	10	9.4	0.49
(AH.N, $)	NL	NL	NL	9	10	NL	9.5	N/A
($, AH.L)	12	9	NL	NL	11	9	10.25	1.30
($, N)	10	NL	NL	11	NL	NL	10.5	N/A
($, AH.N)	11	NL	NL	NL	NL	11	11	N/A
($, L.IY)	14	NL	NL	12	NL	12	12.67	0.94
(AH.L, $)	NL	NL	NL	13	NL	NL	13	N/A
($, AH.N.T)	13	NL	NL	NL	NL	13	13	N/A
($, K)	NL	NL	NL	NL	NL	14	14	N/A

Table 3: Order of rules on individual children's corpora within CHILDES

occurs for the derivational rule *(AH.L, L.IY)*, where we would prefer two derivational rules: *($, AH.B.AH.L) (pass/passable)*, and *($, AH.B.L.IY) (pass/passably)*. We may attribute the lack of these preferred rules to composition of the vocabulary of the small corpus.

The algorithm can also learn transforms that represent a base-derived relationship in a direction opposite than expected. The transform *(IY,$)* is learned after the more desirable *($, IY)* because its base words *(lady, monkey, daddy, lucky, puppy, Jenny)* have been placed in the Base set by other transforms (for example *lady/ladies, Jenny/Jenny's*) and thus cannot be derived by *(IY,$)*. This problem can be avoided by allowing words to move between the Base and Derived sets as the algorithm learns more rules, a technique discussed in detail by Lignos et al. (2009).

4.2.2 Results on individual children's corpora

Table 3 shows the transforms learned when the algorithm was tested on corpora for individual children. A transform is listed if it was learned from any of the six children's corpora. The order in which the transform was acquired is given for each corpus, along with a mean and standard deviation for the transform across all corpora. If a rule was not learned from a particular corpus, it is marked "NL," and if a rule was not learned from at least half of the corpora, its standard deviation is not given.

Whether a particular transform is learned from a particular corpus depends primarily on the number of word pairs that the transform can be applied to, which is largely determined by the number of word types in the corpus but is also affected by the bootstrapping effects of previous rules. Because of varying sizes of

Morpheme	Brown Average Rank	Corresponding Transforms	Mean Transform Rank
Present progressive	2.33	($, IH.NG)	2.00
Plural	3.00	($, Z/S/AH.Z)	3.93
Possessive	6.33	($, Z/S/AH.Z)	3.93
Past regular	9.00	($, D/T/AH.D	6.52
Third person regular	9.66	($, Z/S/AH.Z)	3.93
Contractible copula	12.66	($, Z/S)	2.0
Contractible auxiliary	14.00	($, D/AH.L)	8.04

Table 4: Brown (1973) English morpheme acquisition order and corresponding transforms

the children's corpora, the number of transforms learned before stopping varied from 9 to 14 rules, varying with the number of types in the corpus as shown in Table 1. Seven transforms were learned across all of the children's corpora: *($,Z), ($,IH.NG), ($,S), ($,T), ($,IY), ($,D), ($,ER).*

Because of varying corpus sizes, not all transforms are learned across all corpora. In the smaller corpora rarer morphological patterns may be indistinguishable from noise in the data. Thus in the early stages of acquisition, when the learner has only been exposed to a relatively small amount of data, only a few rules can be learned. As more words are observed, rarer morphological patterns can rise above the noise and be learned, as shown by the larger number of rules learned in the combined corpus.

The relatively consistent order of rule learning between the children's corpora suggests that even on small, disjoint data sets the algorithm reliably produces similar learning orders. As shown by the standard deviations in Table 3, the order is more consistent for the higher frequency rules; the less frequent rules are affected more by characteristics of individual corpora and data sparsity. In order to evaluate this consistency, Spearman's rank correlation was computed pair-wise between the learning orders on individual children's corpora. Because each rule must have a valid rank in every corpus to compute the correlation, only the first seven rules, as ordered in Table 3, could be used in the ranking for each corpus. Of the 15 correlations computed, 13 were significant ($p \leq .05$), and two (Adam/Sarah, Eve/Sarah) were marginally significant ($p \leq .07$). The correlation coefficients for the 15 correlations ranged from .75 to 1.

In addition to learning morphological rules in an consistent manner, an accurate model of morphological acquisition would learn the rules of regular morphology in an order similar to that of children. We compare our results to those of Brown (1973), who manually analyzed child-directed speech transcripts for Adam, Eve, and Sarah. In Table 4, we present the English acquisition sequence for regular, suffixal morphemes, as analyzed by Brown (1973). For each morpheme, we list the corresponding transforms that the algorithm learns. Because of phonological variation in many English morphemes, multiple transforms are needed to represent a single morpheme, such as the past tense, and in some cases each transform also can represent the surface form of multiple morphemes, such as the transform *($, Z)*. For each transform, we give a ranking

corresponding to the order in which the transform was learned across the six children's corpora. When a morpheme corresponds to a single transform, the mean rank of the transform as given in Table 3 is used. When multiple transforms map to a particular morpheme, the mean of their mean ranks is given.

Although it is difficult to perform a direct comparison between the phonemic rules the learner learns and the morphemes noted by Brown, it can be seen by using the mean ranks of transforms the sequential order of acquisition of present progressive, plural, possessive, and past regular is correctly predicted by the algorithm. The main inconsistency between the algorithm and children's order of acquisition is in the present progressive, third person regular, and contractible copula. These are acquired in three separate morphological rules by children, each of which has three surface phonological forms. The algorithm, however, acquires them in three separate phonological transforms, one for each phonological variant of the morpheme.

The transforms for */Z/*, */S/*, and */AH.Z/* are learned relatively early by the algorithm (mean rank 3.93), whereas children acquire their corresponding inflections (plural, possessive, past regular, and contractible copula) somewhat later (mean rank 10.55). There is, however, a principled reason for this discrepancy. In addition to identifying surface forms of morphemes, children are also acquiring the syntactic uses of morphemes and determining allophones. It is known in the acquisition literature (Slobin 1973, Slobin 1985-1997) that acquisition of homophonous morphemes is delayed since a child must sort out the different syntactic functions of morphemes, whereas acquisition of unambiguous morphemes is faster. The English allophone */Z/S/AH.Z/* is three ways ambiguous with respect to underlying function, which is why it is not acquired earlier as predicted by the algorithm. As the algorithm only looks at the frequencies of surface phonological forms, we should not expect the order of morpheme acquisition to be the same as children for ambiguous morphemes.

The parallels in order of acquisition between children and the output of the algorithm have several implications for our understanding of the processes behind children's acquisition. First, the existence of a monotonic order of acquisition of morphemes may result from frequency-driven learning over rules, as modeled in the architecture of the learning algorithm. Second, even though there is individual variation in lexical content of what different children hear, there is enough statistical regularity in the morphological distributions of the language for there to be consistencies in predicted order of acquisition across children. Since there are also consistencies in order of acquisition observed across children, it is possible that the discovery of rule-based linguistic patterns across words constitutes a major component of children's acquisition mechanism.

Iter.	Transform	Tokens	Types	Example	Morpheme	Precision
1	($, s)	3387	249217	size/sizes	Noun plural, 3P Sg.	99.52%
2	($, ed)	1000	73238	pitch/pitched	Past tense	96.46%
3	($, ing)	796	57616	tutor/tutoring	Present progressive	97.04%
4	($, 's)	715	87070	sister/sister's	Possessive	100.00%
5	($, ly)	636	58025	dead/deadly	Adverb derivation	98.59%
6	($, d)	581	33074	value/valued	Past tense	99.40%
7	(e, ing)	408	25385	smoke/smoking	Present progressive	99.36%
8	(y, ies)	218	13020	humanity/humanities	Noun plural, 3P Sg.	100.00%
9	($, y)	159	86326	snow/snowy	Adjective derivation	89.16%
10	($, es)	137	8876	match/matches	Noun plural, 3P Sg.	94.33%
11	($, er)	111	18852	strong/stronger	Comparative, Agentive	91.30%
12	($, e)	80	21483	cut/cute	Spurious	14.29%
13	(e, y)	76	3572	stone/stony	Adjective derivation	99.17%
14	(e, ion)	71	1316	estimate/estimation	Noun derivation	97.27%
15	(t, ce)	69	2313	deviant/deviance	Noun derivation	100.00%
16	($, al)	61	2786	orbit/orbital	Adjective derivation	85.90%
17	(on, ve)	57	2134	meditation/meditative	Adjective derivation	97.18%
18	($, n)	56	81718	grow/grown	Past participle	75.00%
19	($, ic)	50	3464	realist/realistic	Adjective derivation	83.02%
20	(ion, ed)	41	835	elevation/elevated	Adjective derivation	95.18%
21	(r, d)	39	915	muffler/muffled	Adjective derivation	87.30%
22	(t, $)	31	10541	budget/budge	Spurious	12.50%
23	($, r)	30	72560	true/truer	Comparative, Agentive	97.87%
24	(er, ing)	27	1584	drummer/drumming	Present progressive	91.84%
25	($, ion)	25	789	extract/extraction	Noun derivation	80.00%

Table 5: Rules learned on the Brown corpus

4.3 Results on written text

Table 5 summarizes the algorithm's output when run on the Brown corpus (Francis and Kucera, 1967), a diverse collection of American English written text. We present these results to show that the algorithm performs well on larger orthographic corpora in addition to smaller phonemic datasets of child-directed speech.[3] Because the number of word types in the Brown corpus (48,056) is much larger than even the combined child-directed corpus (7,174), the minimum number of word pairs per transform was raised to 25 to obtain a similar number of transforms as for the combined child-directed corpus. The algorithm ran for 25 iterations, achieving a cumulative precision of 97.39% and a recall of 80.78%.

As with the child-directed speech corpus, the algorithm succeeds in learning many of English's most salient inflectional and derivational rules. Because of the larger number of derivational morphemes used in written text, more derivational transforms are acquired from the Brown Corpus than from child-directed speech. As in the child-directed corpus results, a number of derived-derived transforms are learned: *(on, ve), (ion, ed), (er,ing).* Often these derived-derived transforms are learned because of a missing or nonexistent common base, such as *divise* for *division/divisive.* They can also result when the application of a transform results in orthographic changes that the algorithm cannot model. For example, when *-er* or *-ing* are added

to *drum*, there is orthographic gemination of the *m*. As a result, the transform *($, ing)* cannot model *drum/drumming*, and similarly *($, er)* cannot model *drum/drummer*. With many forms like *drummer* and *drumming* still umodeled, the algorithm selects a rule to model them, and since the orthographic gemination appears in both derived forms, *(er, ing)* can be used to model the relationship. Building some "slack" into transform application can allow these orthographic geminates to be handled along with non-geminate cases, and this is explored by Lignos et al. (2009).

5 Consequences of Statistical Distributions in Morphology

Having described the learning algorithm and demonstrated that its order of acquisition of transforms is similar to children's acquisition of morphology, we would like to understand *why* it works. In this section we explore the statistical characteristics of morphology in corpora, and consider their implications for learning linguistic representations. Through a simplified version of the learning algorithm, we explain how type-based computations and greedy acquisition of rules are especially well-suited for Zipfian distributions in morphology.

A particularly important question is why we would choose a rule-based model of morphology in the first place. Given the data statistics of a corpus, does the choice of linguistic representation matter? We show that Zipfian distributions and sparse data in morphology are highly compatible with a rule-based representation, from the point of view of computational efficiency in learning.

5.1 Statistical characteristics of morphology

Corpora of several languages were examined to look for cross-linguistic commonalities in the statistical distribution of morphology in language usage. Every word in a corpus was first converted to a common format of a lemma and inflection, as indicated by a fine-grained morphosyntactic part-of-speech tag; for example, the word "slept" would be represented as *(sleep, verb-past-tense)*. This allowed us to abstract away from processes in language (such as allomorphy, syncretism, and lexical ambiguity) that obscure the underlying forms of words, and made it easier to compare different languages. Lemmas and tags (which indicated inflectional morphology) were obtained from corpus annotations or taggers.

Figure 3 illustrates the distribution of word frequencies according to lemma and inflection for the verbs in a Spanish corpus. Lemma and inflection axes have been sorted according to token frequency, such that the word in the most-frequent lemma and most-frequent inflection appears in the top corner of the figure. By plotting word frequencies in a log scale, it can be seen that the distributions of lemmas and inflections are approximately Zipfian[4] (Zipf, 1935, 1949; Newman, 2005), characterized by a small number of highly

frequent units, a larger number of moderately frequently units, and a very large number of infrequent units. Zipfian distributions are familiar in language from the frequency distributions of words and many other types of constructions.

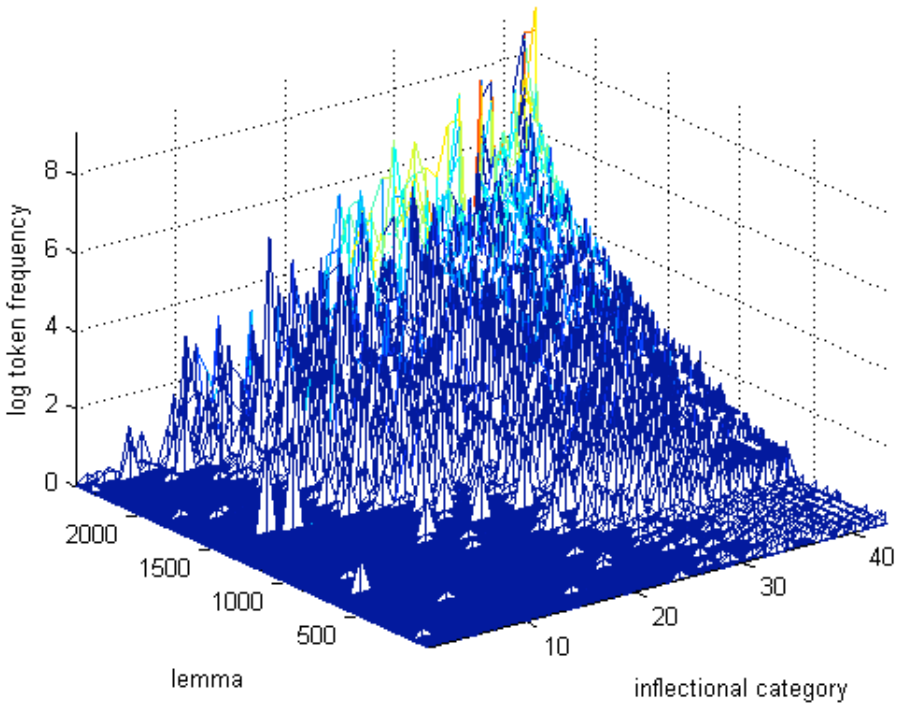

Figure 3: Log token frequency of Spanish verbs, categorized by lemma and inflection

Corpora of both text documents and child-directed speech transcriptions in different languages exhibit similar Zipf-like distributions of lemmas and inflections. The combination of these distributions creates a sparse data problem in morphology: in languages with more than a few inflections, for the set of lemmas in the corpus, the vast majority of logically possible word forms will not be found in a corpus. This leads to a challenging learning problem: how would it be possible to gain the ability to have knowledge of all possible word forms? What kind of morphological grammar could be efficiently learned given sparse, Zipfian-distributed data?

5.2 Morphology learning given sparse, Zipfian-distributed data

Next, we consider the implications of statistical distributions of data for particular linguistic formalisms and associated learning algorithms. We compare rule-based and paradigm learning algorithms, and to simplify the problem somewhat, we restrict the learning scenario to the inflections of one lexical category. It will be

Language	Millions of tokens	# Total verb forms in corpus	Max # verb forms for any lemma	% Saturation
Czech	2.0	72	41	56.9
Greek	2.8	83	45	54.2
Hebrew	2.5	33	23	69.7
Slovene	2.4	32	24	75.0
Spanish	2.6	51	34	66.7
Swedish	1.0	21	14	66.7
Catalan	1.7	45	33	73.3
Italian	1.4	55	47	85.5
CHILDES Spanish	1.4	55	46	83.6
CHILDES Catalan	0.3	39	27	69.2
CHILDES Italian	0.3	49	31	63.3

Table 6: Sparseness of verb paradigms in corpora. *Saturation* is the percentage of forms accounted for by the verb lemma with the *most* forms in a corpus.

seen that learning in the presence of Zipfian-distributed lemmas and inflections seems to favor rule-based learning over paradigm-based learning algorithms, from the point of view of computational efficiency. Then, we consider alternative hypothetical distributions in which paradigms are favored over rules. This illustrates the relationship between data statistics and theories of linguistic representation.

5.2.1 Learning through paradigms

We first discuss the relevance of data distributions for paradigm learning algorithms. As shown in Table 6, when there are sufficiently many inflections in a language, there is never *any* lemma that appears in a full set of forms.[5] This may be attributed to the Zipfian distribution of inflections, and the improbability of the joint occurrence of all morphological forms for a particular lemma.

To illustrate the relevance of this for learning, consider a model of morphological learning in which the possible forms of a lemma are determined by assigning the lemma to its paradigmatic class. By identifying the most-frequent lemma, its set of forms could serve as an "exemplar" that less-frequent lemmas would be associated with. We should not expect this hypothetical algorithm to work in practice due to the lack of full paradigms in real-world morphological data. The statistical distribution of morphological data thus makes it a nontrivial problem to learn a paradigmatic representation. For a paradigm-based learning algorithm to work, more sophisticated procedures would be needed; one example is Linguistica (Goldsmith, 2001, 2006).

5.2.2 Learning through rules

An alternative linguistic representation is a rule-based model of morphology, consisting of a lexicon of base forms and a set of rules that can be applied to base forms to generate full paradigms. In learning a rule

to generate a single inflection, the acquisition of each inflection may proceed independently of others given a base inflection, as demonstrated by research in single-inflection learning (Section 2.1). The concept of paradigms of forms does not play a role in rule learning, and therefore the lack of full paradigms of forms in data is not an issue.

Given the Zipfian distribution of inflections, an efficient method for acquiring the set of rules would be to take the most type-frequent inflection as the base, and learn rules for the rest of the inflections in order of decreasing type frequency. This is a simplification of the algorithm in Section 3, and is illustrated in Figure 4. With this greedy learning strategy, the set of rules acquired at any point would be the statistically best approximation of the morphological system, by maximizing the amount of data accounted for, given a specific number of rule structures.

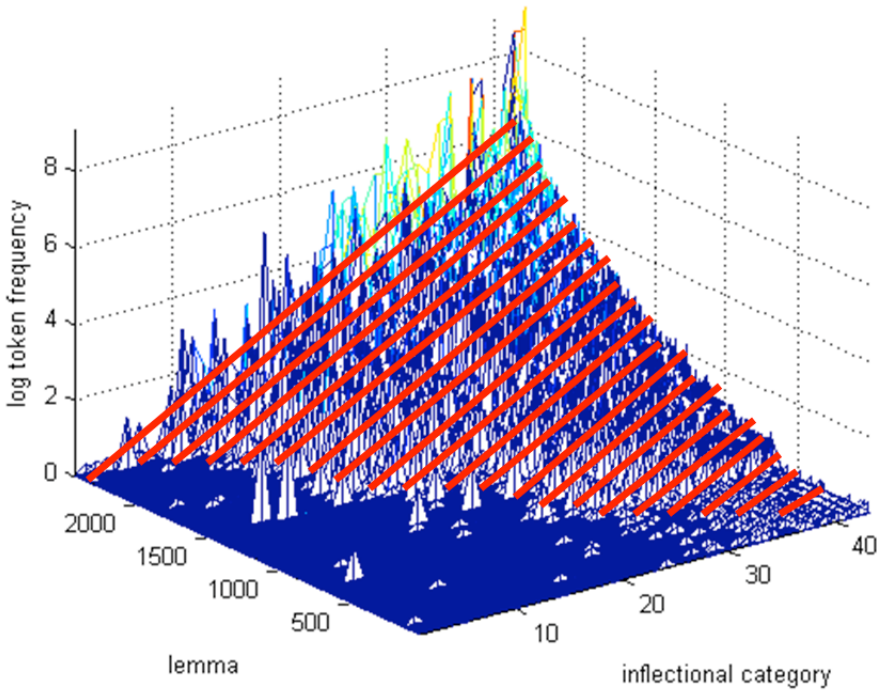

Figure 4: Acquiring a rule-based model from Zipfian-distributed lemmas and inflections, one inflection at a time

5.3 Morphology learning under alternative data distributions

Above, we reasoned about how a rule-based model is more suitable than paradigm-based models for morphological learning given Zipfian-distributed lemmas and inflections. It is also the case that some alternative statistical distributions of data would favor entirely different representations and learning algorithms. For example, consider a hypothetical situation with Zipfian-distributed lemmas and uniformly-distributed inflec-

tions (Figure 5). The rule-learning algorithm would fail, as it would not be able to decide on a base, since all inflections are equally type-frequent. Paradigm learning algorithms, however, could be quite appropriate, as the distribution of inflections would cause there to be full paradigms for the high-frequency lemmas of the language. For example, the previously described hypothetical algorithm for discovering an exemplar paradigm would succeed in these circumstances.

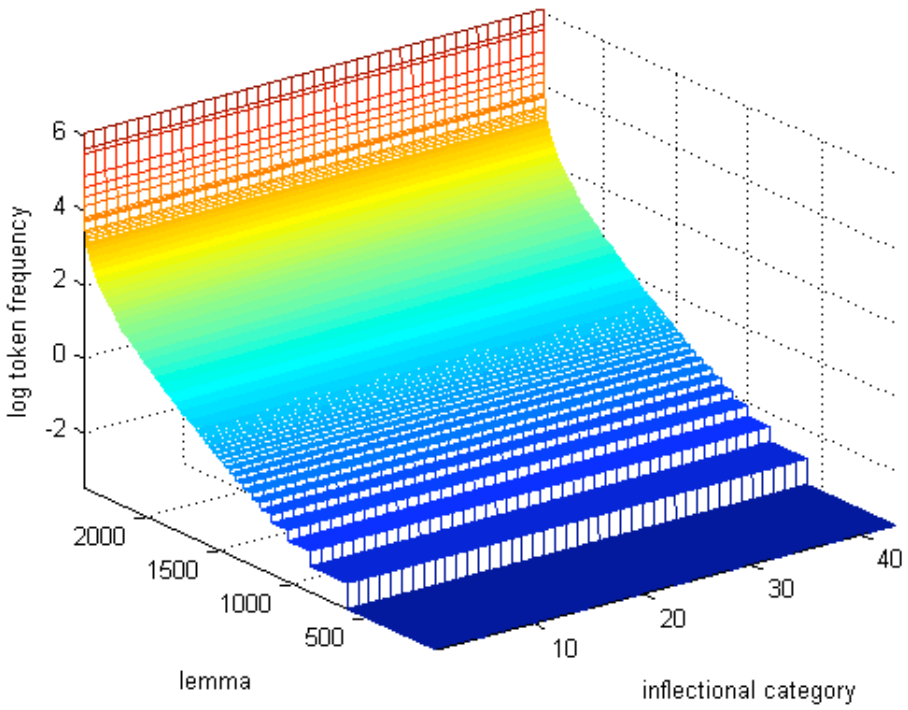

Figure 5: Zipfian distributed lemmas and uniformly distributed inflections. Frequencies are from the Spanish corpus, and are averaged over lemmas.

As another example of the dependence between statistical distribution of data and learning linguistic representations, consider a situation where lemmas and inflections are both uniformly distributed. Assuming a fixed quantity of lemmas and a sufficiently large corpus, every possible word form, each word having the same frequency. In this case, the simplest learning algorithm for determining the full set of word forms would be one that merely memorized each word in the input. It would not be necessary to utilize more sophisticated representational mechanisms such as rules or paradigms.

5.4 Data distributions and the development of learning algorithms

To conclude this section, the apparent dependence between statistical distributions and linguistic representations suggests that it is important to investigate the properties of the input data in the task of designing

a learning system. Knowledge of data distributions can help one to select linguistic representations and formulate learning algorithms over those representations. In terms of the learning algorithm of Section 3, Zipfian distributions in language provide a principled explanation of why it should be able to acquire a reasonable morphological grammar from a corpus of words: the construction and monotonic selection of rule-like structures through type-based computations is a computationally efficient algorithm for approximating the morphology of a language, given Zipfian-distributed lemmas and inflections. That is, the algorithm has an implicit assumption that the input data is Zipfian-distributed in the first place.

6 Discussion

We began with the goal of developing an unsupervised algorithm for morphology learning as a model of children's acquisition of inflectional morphology. The resulting system is significant not only for how it specifically models children's monotonic acquisition of morphological inflections, but also with respect to other larger issues. In this section we discuss implications of the learning model for cognitive processing of language, design methodology for unsupervised learning, and the relationship between language, data statistics, and computation. We then discuss limitations of the model, and consider ways in which it could be improved.

6.1 Implications for cognitive processing in language acquisition

The monotonic acquisition of inflections by children is a phenomenon that has been mostly unexplained in the language acquisition literature. Slobin (1973) proposes that general patterns of behavior in the course of language acquisition are due to cognitive learning biases or "operating principles", and describes a number of heuristics that are relevant to morphology (such as "pay attention to the ends of words"). While we agree with this view, a proper explanation of language acquisition in terms of learning principles requires an algorithmically formulated model.

The ability of our algorithm to acquire transforms one at a time is consistent with children's behavior, and experiments on child-directed speech show that the algorithm is predictive of the particular order of acquisition, to a certain degree. These results support the hypothesis that children's behavior in language acquisition is likely to be due to a computational process. The specific details of the algorithm, as a solution to the abstract computational problem, help us understand the types of cognitive processes that may be involved in morphology acquisition. Children may be employing greedy learning and type frequency-based computations in morphology acquisition, which in our experiments leads to a linguistically reasonable morphological grammar for languages with simple morphological systems, due to Zipfian distributions in

language.

Laboratory experiments have shown evidence that children are capable of performing type-based computations (Gerken, 2006; Gerken and Bollt, 2008). This kind of conclusion, however, does not show *how* such information can be used in the acquisition of an entire language. For such a task, a corpus-based computational model can be a beneficial method of investigation. Corpus-based computation modeling can test whether a learning procedure works in the presence of actual linguistic data, which is Zipfian-distributed and sparse; direct experimentation, on the other hand, typically involves small sets of artificial linguistic stimuli. In this way, computational modeling can make unique contributions to our understanding of linguistic psychology, despite being indirect evidence of humans' abilities.

An alternative to type-based learning is found in usage- or item-based theories of language acquisition (e.g. Goldberg, 1995; Tomasello, 2003; Ninio, 2006). In such approaches, children are theorized to learn languages through the acquisition of individual highly-frequent words or constructions. Processes of analogy relate less-frequent items to more frequent ones. Type-based learning, in comparison, is much more computationally efficient, by forming generalizations from the input. The implementation of a type-based learning model for morphology, along with the aforementioned experiments demonstrating the ability of children to make type-based generalizations, together provide converging evidence that children acquire language by computing abstract linguistic properties of the input.

6.2 Implications for theories of linguistic representation

One of the fundamental questions in linguistics and cognitive science is the nature of the mental representation of language. For morphology, a multitude of formalisms have been employed in systems for computational processing: stem-suffix representations in unsupervised learning, rule-based finite-state models (Sproat, 1992; Kaplan and Kay, 1994; Beesley and Karttunen, 2003), paradigm-based models (Corbett and Fraser, 1993; Forsberg and Ranta, 2004), and connectionist models (Rumelhart and McClelland, 1986), for example. Many of these formalisms have parallels in linguistics, such as Item-and-Arrangement (Hockett, 1954), rule-based (Chomsky and Halle, 1968; Halle, 1973), and paradigm-based (Stump, 2001) theories of morphology.

Given the wide range of linguistic and computational theories of morphology, which one of these is to be preferred as a theory of mental processes? More precisely, which representation (with associated learning and processing algorithms) is an adequate high-level description of the neural processes underlying language? In some respects one might think that this question does not matter, since it is often the case that one formalism can be represented in terms of another. For example, Karttunen (1998) shows how Optimality Theory (Prince and Smolensky, 1993) with a finite number of constraint violations can be translated to a finite-state model,

and Karttunen (2003) makes a similar statement for Stump's Realizational Morphology (Stump, 2001).

While different formalisms may be equivalent in generative capacity, it is not necessarily the case that they would perform equally well in a learning scenario. The embedding of a linguistic formalism within the larger context of a learning algorithm and real-world data places restrictions on which formalisms could be employed for computationally efficient learning and modeling of human behavior. A formalism that is overly complex may suffer from sparse data, as was shown with full paradigmatic tables of words. In contrast, in our algorithm for morphology learning, the transform (a simplified version of a morphophonological rewrite rule) is a suitable representation for learning under sparse data conditions, since it allows for the exploitation of Zipfian distributions in morphology through type-based computations. For morphology, then, there is a close relationship between the structural and statistical aspects of language acquisition that seems to favor a rule-based theory.

6.3 Implications for unsupervised learning of natural language

In addition to modeling human morphology acquisition, the cognitive orientation of this work has led to innovations in computational techniques for unsupervised learning. The most notable feature of our approach is the role of the statistical characteristics of morphology in the design of the algorithm. The main procedure of the learning algorithm is greedy acquisition of discrete structural representations. The ability of the algorithm to acquire transforms that are linguistically reasonable is due to Zipfian distributions in morphological data. Stated differently, the algorithm has an implicit bias for Zipfian-distributed data.

Investigation of statistical distributions of data is important in order to determine the optimality of an algorithm, given the input. A thorough understanding of the input data could also lead to the development of simpler algorithms. Zipfian distributions in language, while well-known in computational linguistics, do not often play a large role in the design of algorithms. Certainly, previous work in morphology learning has included heuristics that would be expected to succeed due to Zipfian distributions; for example, Schone and Jurafsky (2001) and Demberg (2007) select rule-like structures according to high type frequency, and Yarowsky and Wicentowski (2000) and Wicentowski (2002) utilize the difference in relative frequency between inflections. Our work, however, exploits statistical distributions much further; Zipfian distributions are a principal factor in the selection of the linguistic representation and design of the learning algorithm.

A different approach to the use of data statistics in unsupervised learning is exemplified by Goldwater et al. (2006) and Narodowsky and Goldwater (2009). In these works, there is an explicit model for the generation of Zipfian-like statistical distributions through a Pitman-Yor process. The parameters of the statistical model are estimated as part of a procedure that learns a probabilistic model of morphological

structure. In comparison, in our work, we have designed the learning model around the fact that linguistic data is Zipfian-distributed; it was not necessary to explicitly include a mathematical model of the data distribution. An implicit statistical bias of this sort therefore can allow for a system that is more parsimonious in its architecture.

One of the common techniques in unsupervised learning is iterative optimization of a grammar. For example, Linguistica (Goldsmith, 2001, 2006) constructs a series of discrete grammars over a number of iterations. The grammar that is selected for the next iteration is the one that decreases description length as much as possible. Iterative methods are also employed in probabilistic models for numerical optimization of parameters (Snover and Brent, 2001; Snover et al., 2002; Bacchin et al., 2005; Goldwater et al., 2006; Narodowsky and Goldwater, 2009; Poon et al., 2009). Greedy selection of the structures of a grammar (as in our algorithm) has a computational advantage over iterative optimization techniques. The search space of grammars is vastly reduced: a greedy learning procedure can acquire the structures of a grammar one at a time. This may be compared with searching over the *sets* of structures and parameters that comprise a full grammar of a language, a process that is sensitive to local maxima in the search space.

6.4 Model limitations and future work

The morphology learning algorithm presented here could be improved in a number of ways. First, the learning model could be enhanced to identify strings corresponding a wider range of morphological phenomena besides suffixation. Several extensions to the basic learning model have already been developed in Lignos et al. (2009) for identifying prefixes and multi-step derivational morphology (such as *bankers* = *bank* + *er* + *s*). Detection of rules for English irregular verbs would require the ability to detect vowel changes in addition to suffixation. While it may seem undesirable to develop specific procedures for specific types of morphology, it is not possible to implement a truly "knowledge-free" system, due to the exponential number of possible string relationships; some linguistically-motivated learning bias is necessary (Gildea and Jurafsky, 1996).

Second, a more linguistically accurate rule-based model of morphology would include phonologically conditioned rules and abstract morphosyntactic categories. For example, we would like to know that a transform (*$, s*) refers to the concept of "plural noun" (in one case) and occurs on base forms ending in a voiceless consonant. The induction of abstract features for morpheme strings is a very challenging problem, as discussed in Section 2.3. If words could be assigned to morphosyntactic categories, it would be possible to induce phonological contexts of rule application (as in previous work in supervised rule learning), so that rules for phonologically-conditioned change could be separated from rules for morphological affixation.

Third, additional techniques could be employed for learning morphology. Children have access to syn-

tax, semantics, the visual scene, etc., and computational procedures approximating such information could potentially be incorporated into the learning model. In the current experiments, though, we have sought to restrict the computational procedures employed, in order to analyze the contribution of limited information sources in greater detail. We would not expect that the main techniques employed for learning inflectional morphology (i.e., type-based computation of rules and greedy selection) would be applicable to all additional types of morphological phenomena. Statistical analysis may help to reveal what other computational procedures would be needed.

7 Conclusion

In this paper we have presented an unsupervised algorithm for morphology induction as a cognitive model of language acquisition. The specific phenomenon that we sought to model is the observation that children acquire the morphological inflections of their language monotonically. The algorithm accomplished this through greedy, bootstrapped learning of transforms in a base-and-transforms formalism for morphology, a rule-based form of representation. When tested on child-directed corpora of English, the algorithm approximately predicted the order of acquisition of inflections in children.

Investigations of frequency distributions of morphology in corpora led to an understanding of the relationship between linguistic representation and input data statistics. A rule-based representation supplemented with type-based computations and greedy search make it possible to exploit Zipfian distributions of lemmas and inflections for computationally efficient learning. This is more difficult with representations that make incorrect statistical assumptions, such as full paradigms of forms. Children's monotonic acquisition of inflections may thus be explained as being a result of statistically optimal approximation of the input in learning, given a predisposition for a rule-based model of linguistic representation.

In conclusion, the goal of modeling human language acquisition through a computational model has led not only to a precise explanation for a behavioral phenomenon in children's acquisition, but also to a deeper understanding of the relationship between linguistic representation, input data statistics, and computational principles of learning.

Notes

1 The correction of children's overgeneralizations of English past tense verbs (e.g. *eated*/*ate*) is not necessarily a counterexample to monotonicity. It can be modeled by the addition of a rule for a small number of irregular words, and the switch of rule membership of a lexeme to this new rule. The previously

applying default past tense rule (add *-ed*) would still exist for other words in the vocabulary. The learner we present here does not attempt to model the acquisition of irregular verbs or the accompanying "U-shaped" learning patterns (see Marcus et al., 1992) observed in children.

2 According to Dan Swingley (p.c.), in an analysis of the Brent corpus of CHILDES (MacWhinney, 2000), mothers of 9 to 15-month old children spoke to them at an average rate of 1737.5 words per hour. Assuming 4 hours of interaction per day, this amounts to approximately 2.5 million words a year. The corpora used in this work are within this size.

3 For a comparison of the algorithm's performance on orthographic corpora against other unsupervised techniques, see Kurimo et al. (2009).

4 It is atypical to describe the distribution of inflections as Zipfian, since they constitute a relatively small, finite set. However, it is technically proper to do so by viewing Zipf's law as a probability mass function $f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$, where N is the (finite) number of elements in the distribution, k is the rank of a particular element, and s is the constant term in Zipf's law.

5 The following corpora were used for preparation of Table 6: Catalan: Màrquez et al. (2004); CHILDES languages: MacWhinney (2000); Czech: Hajic et al. (2006); Greek: Linguistic Data Consortium (1994); Hebrew: Itai and Wintner (2008), made available by the Knowledge Center for Processing Hebrew; Italian: Baroni and Ueyama (2006); Slovene: Erjavec (2006); Spanish: Graff and Gallegos (1999); Swedish: Gustafson-Capková and Hartmann (2006). The following taggers were used: Greek: Papageorgiou et al. (2000); Hebrew: Segal, (1999); adult Spanish, child-directed Spanish, Catalan, and Italian: Carreras et al. (2004).

References

Albright, A. & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. (In *Proceedings of the Special Interest Group on Computational Phonology*)

Argamon, S., Akiva, N., Amir, A. & Kapah, O. (2004). Efficient unsupervised recursive word segmentation using minimum description length. (In *Proceedings of the International Conference on Computational Linguistics*)

Baayen, R. H., Piepenbrock, R. & van Rijn, H. (1996). The CELEX2 lexical database (CD-ROM). (Philadelphia, PA: Linguistic Data Consortium)

Bacchin, M., Ferro, N. & Melucci, M. (2005). A probabilistic model for stemmer generation. *Information Processing and Management, 41*, 121-137

Baroni, M. & Ueyama, M. (2006). Building general- and special-purpose corpora by web crawling.

(In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*

Beesley, K. & Karttunen, L. (2003). Finite state morphology. (Stanford, CA: CSLI Publications)

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. (In *Proceedings of the Association for Computational Linguistics*)

Bordag, S. (2007). Elements of knowledge-free and unsupervised lexical acquisition. Dissertation, University of Leipzig

Brent, M. & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93-125

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

Bybee, J. L. (1985). Morphology: A study of the relation between meaning and form. (Amsterdam: John Bejamins)

Can, B. & Manandhar, S. (2009). Unsupervised learning of morphology by using syntactic categories. In Working Notes for the Cross Language Evaluation Forum (CLEF), MorphoChallenge

Carreras, X., Chao, I., Padró, L. & Padró, M. (2004). FreeLing: an open-source suite of language analyzers. (In *Proceedings of the Language and Resources Evaluation Conference*)

Carlson, L. (2005). Inducing a morphological transducer from inflectional paradigms. (In Arppe, A., Carlson, L., Lindn, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H. & Yli-Jyr, A. (Eds.), *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday. Stanford, CA: CSLI Publications)*

Chan, E. (2008). Structures and distributions in morphology learning. Dissertation, University of Pennsylvania.

Chomsky, N & Halle, M. (1968). The sound pattern of English. (New York: Harper & Row)

Clark, A. (2001). Learning morphology with pair hidden markov models. (In *Proceedings of the Student Workshop at the 39th Annual Meeting of the Association for Computational Linguistics*)

Clark, A. (2002). Memory-based learning of morphology with stochastic transducers. (In *Proceedings of the Association for Computational Linguistics*)

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. (In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*)

Corbett, G. G. & Fraser, N. M. 1993. Network morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics, 29*, 113-42

Creutz, M. (2003). Unsupervised segmentation of words using prior distributions of morph length and frequency. (In *Proceedings of the Association of Computational Linguistics*)

Creutz, M. & Lagus, K. (2004). Induction of a simple morphology for highly-inflecting languages. (In *Proceedings of the Special Interest Group in Computational Phonology*)

Cucerzan, S., & Yarowsky, D. (2003). Minimally supervised induction of grammatical gender. (In *Proceedings of the Joint Conference on Human Language Technology and the Third Meeting of the North American Chapter of the Association for Computational Linguistics.*

Daelemans, W., Berck, P. & Gillis, S. (1996). Unsupervised discovery of phonological categories through supervised learning of morphological rules. (In *Proceedings of the 16th International Conference on Computational Linguistics*)

Dasgupta, S. & Ng, V. (2007). Unsupervised part-of-speech acquisition for resource-scarce languages. (In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*)

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407

Demberg, V. (2007). A language-independent unsupervised model for morphological segmentation. (In *Proceedings of the Association for Computational Linguistics*)

Dressler, W. U. (2005). Morphological typology and first language acquisition: some mutual challenges. (In Booij, G., Guevara, E. Ralli, A. Sgroi, S, & Scalise, S. (Eds.), (In *Morphology and Linguistic Typology, Online Proceedings of the Fourth Mediterranean Morphology Meeting*, Catania, 21-23 September 2003, University of Bologna. http://morbo.lingue.unibo.it/mmm)

Dreyer, M., Smith, J. & Eisner, J. (2008). Latent-variable modeling of string transductions with finite-state methods. (In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*)

Erjavec, T. (2006). The English-Slovene ACQUIS corpus. (In *Proceedings of the Language and Resources Evaluation Conference*)

Forsberg, M. & Ranta, A. 2004. Functional morphology. (In *Proceedings of the International Conference on Functional Programming*)

Francis, W. N. & Kucera, H. (1967). Computing analysis of present-day American English. (Providence, RI: Brown University Press)

Freitag, D. (2004). Toward unsupervised whole-corpus tagging. (In *Proceedings of the International Conference on Computational Linguistics*)

Freitag, D. (2005). Morphology induction from term clusters. (In *Proceedings of the Conference on Computational Natural Language Learning*)

Gambell, T. & Yang, C. (2004). Statistical learning and universal grammar: modeling word segmentation. In *Proceedings of the International Conference on Computational Linguistics*)

Gerken, L. A. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition, 98*, B67-B74

Gerken, L. A. & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development, 4*, 228-248

Gildea, D. & Jurafsky, D. (1996). Learning bias and phonological rule induction. *Computational Linguistics, 22*, 497-530

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. (Chicago: University of Chicago Press)

Golding, A. R. & Thompson, H. S. (1985). A morphology component for language programs. *Linguistics, 23*, 263-284

Goldsmith, J. A. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27*, 153-198

Goldsmith, J. A. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering, 12*, 1-19

Goldwater, S., Griffiths, T. L. & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. (In Weiss, Y., Schllkopf, B. & Platt, J. (Eds.), *Advances in Neural Information Processing Systems, 18*. Cambridge, MA: The MIT Press)

Graff, D. & Gallegos, G. 1999. Spanish newswire text, volume 2. (Philadelphia, PA: Linguistic Data Consortium)

Gustafson-Capková, S. & Hartmann, B. 2006. Manual of the Stockholm Umeå Corpus version 2.0. (Stockholm: Department of Linguistics, Stockholm University)

Hajic, J., et al. (2006). Prague Dependency Treebank 2.0, CDROM, LDC2006T01. (Philadelphia, PA: Linguistic Data Consortium)

Halle, M. (1973). Prolegomena to a theory of word-formation. *Linguistic Inquiry, 4*, 3-16

Hafer, M., & Weiss, S. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10, 371-385

Harris, Z. (1955). From phoneme to morpheme. *Language, 31*, 190-222

Harris, Z. (1970). Papers in structural and transformational linguistics. (Dordrecht: D. Reidel)

Higgins, D. (2003). Unsupervised learning of Bulgarian POS tags. (In *Workshop on Morphological Processing of Slavic Languages*)

Hockett, C. F. (1954). Two models of grammatical description. *Word, 10*, 210-231

Hooper, J. B. (1979). Child morphology and morphophonemic change. *Linguistics, 17*, 21-50. (Also in J. Fisiak (Ed.), *Historical morphology* (pp. 157-187). The Hague: Mouton)

Hu, Y., Matveeva, I., Goldsmith, J. A. & Sprague, C. (2005a). The SED heuristic for morpheme discovery: a look at Swahili. (In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*)

Hu, Y., Matveeva, I., Goldsmith, J. A. & Sprague, C. (2005b). Using morphology and syntax together in unsupervised learning. (In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*)

Itai, A. & Wintner, S. 2008. Language resources for Hebrew. *Language Resources and Evaluation, 42* 77-98

Johnson, M. (1984). A discovery procedure for certain phonological rules. (In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics*)

Kaplan, R. & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics, 20*, 331-378

Karttunen, L. (1998). The proper treatment of optimality in computational phonology. (In *Proceedings of FSMNLP'98. International Workshop on Finite-State Methods in Natural Language Processing*)

Karttunen, L. (2003). Computing with realizational morphology. (In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 2588 of *Lecture Notes in Computer Science* (pp. 205216). Heidelberg: Springer-Verlag)

Kazakov, D. & Manandhar, S. (2001). Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning, 43*, 121-162

Klein, D. & Manning, C. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. (In *Proceedings of the Association for Computational Linguistics*)

Kurimo, M., Virpioja, S., Turunen, V. T., Blackwood, G. W. & Byrne, W. (2009). Overview and results of Morpho Challenge 2009. (In *Working Notes for the CLEF 2009 Workshop*)

Linguistic Data Consortium. 1994. ECI Multilingual Text. CDROM, LDC94T5. (Philadelphia, PA: Linguistic Data Consortium)

Lignos, C., Chan, E., Marcus, M. P. & Yang, C. (2009). A rule-based unsupervised morphology learning framework. (In *Working Notes for Cross-Linguistic Evaluation Forum, MorphoChallenge*)

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

Lin, Y. (2005). Learning features and segments from waveforms: a statistical model of early phonological

acquisition. Dissertation, UCLA

Ling, C. X. (1994). Learning the past tense of English verbs: the symbolic pattern associator vs. connectionist models. *Journal of Artificial Intelligence Research, 1*, 202-229

MacWhinney, B. (2000). The CHILDES-Project: Tools for analyzing talk. Second edition. (Hillsdale, NJ: Erlbaum)

Manandhar, S., Džeroski, S. & Erjavec, T. (1998). Learning multilingual morphology with CLOG. (In *Proceedings of Inductive Logic Programming (ILP), 8th International Conference. Lecture Notes in Artificial Intelligence, 1446* (pp. 135-144). Heidelberg: Springer Verlag)

Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 54(4), 1-182

Màrquez, L., Taulé, M., Marti, A., Garcia, M., Real, F. & Ferrés, D. (2004). Senseval-3: The Catalan lexical sample task. (In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*)

McClelland, J. L. & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Science, 6*, 465-74

Molnar, R. A. (2001). Generalize and sift as a model of inflection acquisition. Masters thesis, Massachusetts Institute of Technology.

Mooney, R. J. & Califf, M. E. (1996). Learning the past tense of English verbs using inductive logic programming. (In Wermter, S., Riloff, E. & Scheler, G. (Eds.), *Symbolic, connectionist, and statistical approaches to learning for natural language processing.* Heidelberg: Spring Verlag)

Naradowsky, J. & Goldwater, S. (2009). Improving morphology induction by learning spelling rules. (In *Proceedings of the International Joint Conference on Artificial Intelligence*)

Newman, M. E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*, 323-351.

Ninio, A. (2006). Language and the learning curve: A new theory of syntactic development. (Oxford: Oxford University Press)

Oflazer, K., Nirenburg, S. & McShane, M. (2001). Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics, 27*, 59-85

Papageorgiou, H., Prokopidis, P., Giouli, V., & Piperidis, S. (2000). A unified POS tagging architecture and its application to Greek. (In *Proceedings of the Language and Resources Evaluation Conference*)

Parkes, C., Malek, A. M. & Marcus, M. P. (1998). Towards unsupervised extraction of verb paradigms from large corpora. (In *Proceedings of the Sixth Workshop on Very Large Corpora*)

Pinker, S. (1999). Words and rules: The ingredients of language. (New York: HarperCollins)

Pinker, S. & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73-193

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

Plisson, J., Lavrac, N. & Mladenic, D. (2004). A rule based approach to word lemmatization. (In *SiKDD 2004 at multiconference IS-2004, Ljubljana, Slovenia*)

Poon, H., Cherry, C., & Toutanova, K. (2009). Unsupervised morphological segmentation with log-linear models. (In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference)

Prince, A. & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report, Rutgers University Center for Cognitive Science and Computer Science Department, University of Colorado at Boulder. Also published by Blackwell Publishers, 2004

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. (In McClelland, J. L., Rumelhart, D. E. & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume II.* Cambridge, MA: The MIT Press)

Schone, P. & Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. (In *Proceedings of the Conference on Computational Natural Language Learning*)

Schone, P. & Jurafsky, D. (2001). Knowledge-free induction of inflectional morphologies. (In *Proceedings of the North American Chapter of the Association for Computational Linguistics*)

Schütze, H. (1993). Part-of-speech induction from scratch. (In *Proceedings of the Association for Computational Linguistics*)

Segal, E. 1999. Hebrew morphological analyzer for Hebrew undotted texts. Master's thesis, Technion, Israel Institute of Technology, Haifa.

Shalonova, K. & Flach, P. (2007). Morphology learning using tree of aligned suffix rules. (In *Proceedings of the Workshop on Challenges and Applications of Grammar Induction*)

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. (In Ferguson, C. A. & Slobin, D. I. (Eds.), *Studies of child language development.* New York: Holt, Rinehart & Winston)

Slobin, D. I. (Ed.) (1985) (2 vols.), (1992), (1997) (2 vols.). The crosslinguistic study of language acquisition. (Hillsdale, NJ: Erlbaum)

Snover, M. & Brent, M. (2001). A Bayesian model for morpheme and paradigm identification. (In

Proceedings of the Association for Computational Linguistics)

Snover, M., Jarosz, G. & Brent, M. (2002). Unsupervised learning of morphology using a novel directed search algorithm: taking the first step. (In *Proceedings of the Special Interest Group in Computational Phonology*)

Sproat, R. (1992). Morphology and computation. (Cambridge, MA: The MIT Press)

Stroppa, N. & Yvon, F. (2005). An analogical learner for morphological analysis. (In *Proceedings of the Conference on Computational Natural Language Learning*)

Stump, G. T. (2001). Inflectional morphology: A theory of paradigm structure. (Cambridge: Cambridge University Press)

Theron, P. & Cloete, I. (1997). Automatic acquisition of two-level morphological rules. (In *Proceedings of the Conference on Applied Natural Language Processing*)

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. (Cambridge, MA: Harvard University Press)

Weide, R. (1998). The Carnegie Mellon Pronouncing Dictionary [cmudict. 0.6]. (Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi- bin/cmudict)

Wicentowski, R. (2002). Modeling and learning multilingual inflectional morphology in a minimally supervised framework. Dissertation, Johns Hopkins University

Wicentowski, R. (2004). Multilingual noise-robust supervised morphological analysis using the WordFrame model. (In *Proceedings of Special Interest Group on Computational Phonology (SIGPHON)*)

Wothke, K. (1986). Machine learning of morphological rules by generalization and analogy. (In *Proceedings of the International Conference on Computational Linguistics (COLING)*)

Yarowsky, D. & Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. (In *Proceedings of the Association for Computational Linguistics*)

Yip, K. & Sussman, G. J. (1997). Sparse representations for fast, one-shot learning. A.I. Memo No. 1633, Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Zajac, R. (2001). Morpholog: constrained and supervised learning of morphology. (In *Proceedings of the Special Interest Group in Computational Phonology*)

Zipf, G. K. (1935). The psycho-biology of language, an introduction to dynamic philology. (Cambridge, MA: The Riverside Press)

Zipf, G. K. (1949). Human behavior and the principle of least effort. (Cambridge, MA: Addison-Wesley)