Computational Models of Morphological Learning

Jordan Kodner

Department of Linguistics, and
Institute for Advanced Computational Science
Stony Brook University
Stony Brook, NY 11790
jordan.kodner@stonybrook.edu
https://orcid.org/0000-0001-8632-9859

October 14, 2021

1 Summary and Keywords

A computational learner needs three things: Data to learn from, a class of representations to acquire, and a way to get from one to the other. Language acquisition is a very particular learning setting that can be defined in terms of the input (the child's early linguistic experience) and the output (a grammar capable of generating a language very similar to the input). The input is infamously impoverished: as it relates to morphology, the vast majority of potential forms are never attested in the input, and those that are attested follow an extremely skewed frequency distribution. Learners nevertheless manage to acquire most details of their native morphologies after only a few years of input. That said, acquisition is not instantaneous nor is it error-free. Children do make mistakes, and they do so in predictable ways which provide insights into their grammars and learning processes.

The most elucidating computational model of morphology learning from the perspective of a linguist is one that learns morphology like a child does, that is, on child-like input and along a child-like developmental path. This article focuses on clarifying those aspects of morphology acquisition that should go into such an elucidating computational model. Section 2 describes the input with a focus on child-directed speech corpora and input sparsity. Section 3 discusses representations with focuses on productivity, developmental paths, and formal learnability. Section 4 surveys the range of learning tasks that guide research in computational linguistics and NLP with special focus on how they relate to the acquisition setting. And to conclude, Section 5 presents a summary of morphology acquisition as a learning problem with Table 4 highlighting the key takeaways of this article.[1]

Keywords: morphology, computational linguistics, child language acquisition, corpus linguistics, child development, natural language processing, learning theory

2 The Input

Defining the input is crucial for characterizing a learning problem, as kinds of input may admit dramatically different learning trajectories and conditions of learnability. The input to the morphological acquisition learning problem has three general characteristics: first, it is acoustic, second, it is very sparse and extremely skewed, and third, it only contains unlabeled positive evidence. Phonological learning and word segmentation are challenging problems which have spawned large bodies of computational research in their own right (see Räsänen (2012) Jarosz (2019) for reviews), but the acoustic nature of the input is not a focus in the morphology literature. Rather than navigating phonological learning, computational approaches to morphological acquisition instead assume that some amount of phonology has already been learned and word segmentation is complete and accurate. I will adopt that convention here. I will also focus on inflectional morphology.

The acoustic input is of course paired with the child's observations of the world around them. These observations are important for learning, but the role they play in the acquisition process is complicated. The learner connects linguistic input to scenes in the environment in order to work out meanings that can be related to concrete experiences, but these scenes are usually highly ambiguous (Medina et al., 2011), and curtailed exposure to these scenes, such as in the case of congenitally blind children, hardly hampers learning (Landau et al., 2009). The takeaway here for morphology acquisition is that the linguistic input remains paramount over other sorts of input. One need not provide a computational learner with a fully featured situational representation when "good enough" semantic representations should suffice to learn morphology. In practice, this "good enough" representation is extracted distributionally from text under the guise of *word embeddings* currently popular in natural language processing (Levy et al., 2015; Wang et al., 2020) or by explicit annotation of semantic features like person and tense as in the Universal Dependencies treebank (UD; Zeman et al., 2021) or UniMorph projects (McCarthy et al., 2020) as substitutes for what the child would have to learn situationally.

2.1 Collecting Input

The preferred way to gather input for computational acquisiton systems is to extract it from corpora of child-directed speech (CDS) such as those available in the CHILDES database (MacWhinney, 1991). There are CDS corpora available for several languages, though the CHILDES collection of English corpora is by far the largest: it contains about 13 million tokens, which can stand in for a year or more of input. The CDS in the Adam, Eve, and Sarah sections of the Brown corpus (Brown, 1973)[2] in particular has been widely studied. CDS (in text form) has a higher token/type ratio than other genres and a shorter mean utterance length. Since most of it is dialogue, second and first person are much more frequent than they are in typical NLP corpora. Many of the corpora in CHILDES have been automatically lemmatized and tagged for morphological features.

It is common when extracting word lists or lexica from CDS to filter out types with low token frequencies as a kind of normalization, often with once in a million as the threshold (following Nagy and Anderson (1984) and Yang (2016)). This removes rare or data set-specific items that most children probably do not experience. Applied to a few million tokens of CDS, it also tends to yield a lexicon roughly the size of a three-year-old child's, thus approximating the lexical knowledge from which morphology is acquired. Non-child-directed corpora can often be substituted for CDS in morphological acquisition studies since they are tend to be similar in key ways: if a non-CDS corpus such as Universal Dependencies is filtered to achieve a lexicon of comparable size to a CDS-derived lexicon, it will tend to share similar lexical contents and distributional properties as CDS-derived lexica. This means that non-CDS can be reasonably substituted for CDS when CDS is not available, as is the case for the vast majority of the world's languages (Kodner, 2019, 2020).

Children are well on their way to acquiring their native morphology by age three, which implies that a few million words of CDS is sufficient for a child to acquire morphology. This turns out to be more or less invariant in respect to morphological complexity. English learners acquire the language's relatively limited verbal morphology between two and three (Brown, 1973; Kuczaj, 1977), but so do learners of languages with richer inflectional morphology including Swahili (Deen, 2005), Italian (Guasti, 1993; Caprin and Guasti, 2009), and Turkish (Aksu-Koç, 1985) among others. Children in this age group still have small vocabularies, consisting of only a few hundred up to around thousand types by age three. There is quite a bit of individual variation within these bounds, but the trend holds up well across the languages which have been studied, including English, German, and Mandarin, among many others (Anglin et al., 1993; Fenson et al., 1994; Hart and Risley, 1995; Bornstein et al., 2004; Szagun et al., 2006).

In short, if a corpus contains, say, a thousand types with some kind of semantic annotation, a good human-like computational morphology system should be able to acquire morphology from it. If it contains, say, a thousand noun types, a thousand verb types, and so on, that should be more than enough to learn all the productive patterns of a language's morphology and it worth considering whether it is appropriate to trim it.

2.2 Input Sparsity

The challenge of morphological acquisition is further heightened by the extreme sparsity and skew that define the input. Much has been made about the presence of Zipfian and other long-tailed distributions in language, and in particular the lexicon and morphological exponence (Zipf, 1949; Miller, 1957; Howes, 1968; Baayen, 1993; Chan, 2008; Yang, 2013; Lignos and Yang, 2018). Following a Zipfian distribution, lexical frequencies are proportional to the inverse of their frequency rank. That is, the second most frequent item should be about half as common as the most frequent, the third most frequent item should be about a third as frequent, and so on. Such a distribution is dramatically skewed, with most items lying on a long thin tail.

From the perspective of a morphological learner, this means that most roots will only appear fleetingly in the input, and if a root only appears, say, once or twice, it follows that it can only appear in one or two forms. Chan (2008) showed that the proportion of items' morphological paradigms that are attested in a given corpus, their *paradigm saturations*, also follow long-tailed distributions. That is, a few roots appear in many of their possible forms (they have high paradigm saturation), while the majority of them appear in only a tiny fraction of their possible forms. This is correlated with paradigm size, but the curve scales with corpus size, so using more data will not truly alleviate the sparsity. "Paradigm" here is used descriptively to refer to the set of inflections that a given root can potentially take, so this notion of paradigm saturation is applicable regardless of one's particular theoretical framework. An important implication of severe paradigm sparsity is that learning models which require all or most forms to be attested cannot be workable in practice for most languages (see Chan (2008, ch. 3) for a critique of such models).

Figure 1 shows paradigm saturation plots for verbs from English CDS (Brown, 1973; Brent and Siskind, 2001; MacWhinney, 1991), Spanish CDS (Fernald and Marchman, 2012), and German CDS (Behrens, 2006), along with Finnish and Turkish from Universal Dependencies (Zeman et al., 2021). These are arranged by paradigm size. Note that the distribution becomes sparser as paradigm size increases. There are no fully saturated paradigms in the data set except for in English.

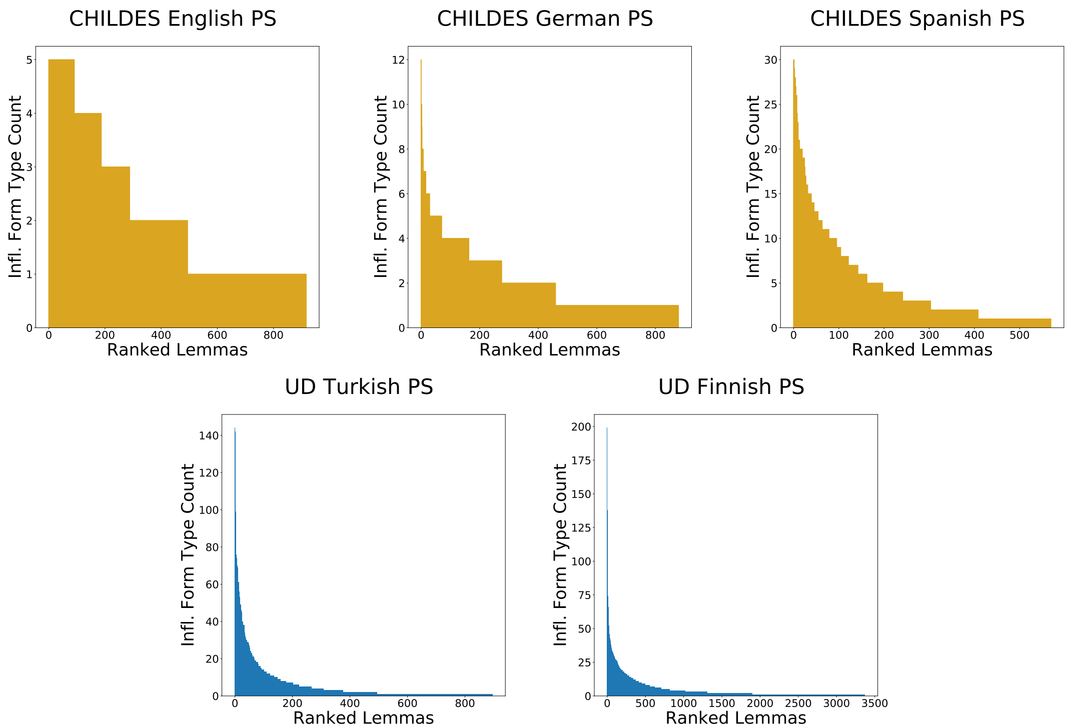


Figure 1: Verb paradigm saturation plots for English, German, and Spanish CDS and Finnish and Turkish from Universal Dependencies. Maximum height indicates the largest attested number of forms according to the corpus's annotation scheme.

2.3 Positive Evidence

Even young children are relatively accurate in their morphological productions despite their small working vocabularies and the extreme sparseness and skew of the input. It is still a puzzle exactly why children are so good at this. One initially plausible solution to the problem would be that children also leverage negative feedback knowingly or unknowingly provided by their caregivers during acquisition and not just the positive evidence described in the previous section. If true, this would greatly simplify the learning task. However, it is well understood that children receive virtually no actionable negative evidence during the acquisition process (Brown and Hanlon, 1970; Braine, 1971; Bowerman, 1988; Marcus, 1993). Explicit negative evidence from caregivers, be it corrections or other behavioral cues, is relatively rare, is often not focused on the grammar but discourse, and is above all unreliable and noisy. Caregivers do not necessarily notice when a child produces an error and may not bother to correct it, and even if they do perceive an error, it may actually be a misunderstanding on their part. Most importantly, children ignore it anyway even when it is clearly provided. Consider this example from Cazden (1972) cited in Marcus et al. (1992) of a child who is steadfastly oblivious to the adult's corrections:

Child: My teacher holded the baby rabbits and we patted them.
Adult: Did you say your teacher held the baby rabbits?
Child: Yes.
Adult: What did you say she did?
Child: She holded the baby rabbits and we patted them.
Adult: Did you say she held them tightly?
Child: No, she holded them loosely.

This *no negative evidence problem* is greatly exacerbated by the sparsity of the input. Most words, most categories, and most forms will not appear often enough for the child to learn a robust enough distribution to overcome noisy negative feedback. Concretely, Marcus (1993) calculated how many times a sentence would have to be misproduced and corrected to overcome noisy feedback– 85 times under the most charitable assumptions– and concluded that few sentences would ever be produced enough times for that to happen. The same argument can be made in morphology: the vast majority of forms will simply not be uttered often enough, if at all, for the child to extract a statistically reliable negative signal. To make matters worse, this assumes that the relative probabilities of negative feedback given an error or no error is known to the child in the first place, which it is not. This presents the child with an impossible task.

This is of serious concern for any theory of grammar and greatly constrains what approaches the child learner may take. Utterances are not labeled as "grammatical" or "ungrammatical," so an approach that simply learns to classify forms as grammatical or not is not feasible. Supervised classification approaches, those which learn to assign inputs into categories based on labels provided during training, are one of the mainstays of machine learning, but they cannot possibly be what a child employs. There is one further implication: the lack of labels leaves generalization as one of the core pieces of the acquired language competence. With no explicit negative examples to stake out the limits of what can be said, the learner must develop a clear notion of productivity based on positive evidence alone instead in order to limit the scope of morphological processes.

Given the unworkability of explicit *direct* negative evidence, Braine (1971) suggested *indirect* negative evidence, or cues that would permit the child to infer that absence of evidence is indeed evidence of absence, as a fall back. The implications of indirect negative evidence in grammar learning were later explicated in Chomsky (1981, ch. 1). This intuition has been expressed under many guises including usage-based preemption (MacWhinney, 2004; Stefanowitsch, 2008) and mathematically formalized Bayesian methods (Perfors et al., 2010). Unfortunately, it is of questionable utility for morphology acquisition for much the same reason that direct negative evidence is. In the long tail of the Zipfian distribution, most words only appear rarely and in few of their possible forms. Thus a child typically cannot distinguish a suspicious absence of a form (the indirect negative evidence for its ungrammaticality) from a chance omission from their input sample (Yang, 2017). Absence of evidence is not evidence of absence, especially not when most things are unevidenced.

3 Productivity and Representation

Children clearly acquire some kind of cognitive representations from their sparse and skewed input, and exactly what these representations are has been subject to decades of research. Many distinct formalisms have been proposed over the decades (see Stewart (2015) and Audring and Masini (2018) for summaries), however many of these may actually be formally equivalent, in which case the distinctions between them are matters of practical convenience – how neatly they conceptually interface with other language-related questions.

Acquisition and computational research probably cannot single out a particular theory as the correct one, but they do provide unique insights into the overarching characteristics of plausible representations. Child productions, especially their errors or novel productions, serve as a window into the grammar's organization, and formal mathematical theories of representation and learnability provide hard results as to what kinds of systems are or are not learnable.

3.1 Productivity

Productivity is traditionally described as the ability of some pattern or rule to be extended to new circumstances. Since generative capacity is so central to language, and since the input is so sparse, working out when to productively apply generalizations is one of the most important tasks for the child during acquisition.

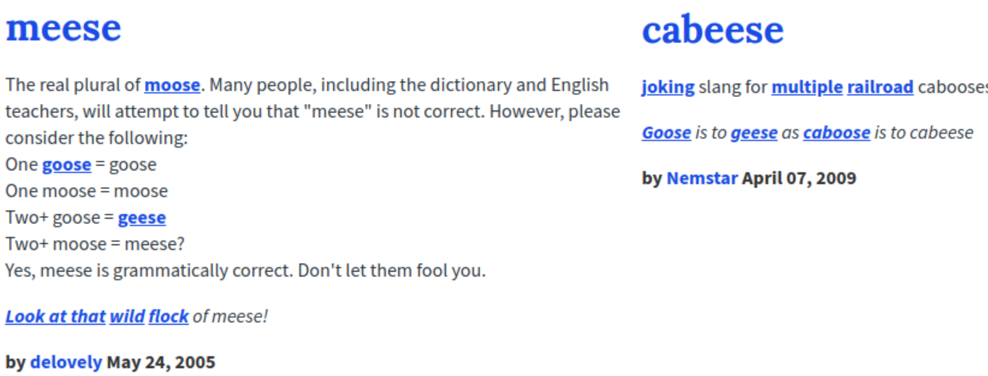
This is not an all or nothing proposition. Patterns may be productive unconditionally (a global default), or may be restricted according to certain phonological or semantic conditions. For example, the English *-s* plural in its various phonologically conditioned allomorphs is clearly a global default, while German plurals in *-(e)n* are productive but only for feminine nouns (Zaretsky and Lange, 2015). That said, not all patterns with apparently generalizable conditions are actually productive. For example, the common *sing-sang*, *ring-rang*, *swim-swam*, *drink-drank* pattern for English past tense is probably not productive for adults even though it is amenable to phonological description (Yang, 2016, ch. 4). Additionally, productive patterns often have exceptions, such as *fling-flung*, *go-went*, and *tell-told* for English past tense and *goose-geese*, *child-children*, and *sheep-sheep* for English noun plurals. The child needs to determine the productive patterns in their language despite these exceptions while also remembering to account for the exceptions.

Productivity is correlated with a pattern's frequency but is by no means a mere matter of frequency or probability matching. It may well be the case that a frequent pattern is unproductive or that a less frequent pattern is productive instead, as is the case for German plurals where the relatively infrequent *-s* plural is the global default (Clahsen, 1990; Marcus et al., 1995). It may also be the case that there is no default as in the case of paradigmatic gaps (Gorman and Yang, 2019). These appear here and there including for Polish genitives (Dąbrowska, 2001) and Russian perfects (Halle, 1973). Learners of course do not know in advance which of these potential complexities will manifest in their particular languages.

Perhaps the most well known way that productivity in children has been studied has been through the Wug test (Berko, 1958) and its many successors (e.g., Prasada and Pinker, 1993; Marcus et al., 1995; Albright and Hayes, 2003; Klafehn, 2003; Oseki et al., 2019), which test whether a speaker, or particularly a young learner, has internalized a productive pattern by whether or not they can apply it to novel items. If the pattern is represented as a productive in the learner's grammar, it should be possible for the learner to extend it to new items. However, if it is unproductive, listed or stored in most theories, the learner is not expected to apply it to those items. The original study presented children with images of novel objects (e.g., a strange chick-like *wug*) and actions (e.g., a man *loodging*, an action no child would have witnessed in real life). After learning the word, children were prompted to produce plurals, past forms, diminutives, third person singulars, comparative and superlative adjectives, and other forms of English inflectional and derivational morphology.

Berko found that even preschoolers correctly inflected most items a majority of the time. However, performance was far from perfect particularly in two situations. First, preschoolers struggled with the phonologically conditioned syllabic allomorphs of the third person singular (/-əz/) and past (/-əd/) with fewer than half producing the correct forms. Second, and most relevantly to productivity, children struggled with minority patterns. For example, only two of 86 children produced *glang* or *glung* as the past of *gling*, while *glinged* was produced at a rate comparable to *ricked* or *melted*. This word was chosen to identify whether children acquired a productive pattern along the lines of common *sing-sang* or *sting-stung* verbs. Their failure to produce these forms indicated that the pattern was not productive for them, unlike *-ed*, *-s*, or *-ing*.

It is worth commenting on a difference between children and adults here. While children showed near-categorical results in the Berko (1958) study, older subjects did not– over half of them readily analogized the *sing* or *sting* pattern to *gling-glang* or *gling-glung*. It is unclear exactly why this discrepancy exists– there may or may not be a difference in the grammars of young learners and adults– but the methodology of the wug test certainly has an effect.. Adults and children appear to approach the wug test differently (Schütze, 2005), with many adults treating it as a game that requires clever analogies (Derwing and Baker, 1977). For a concrete example, consider *meese* and *cabeese*, joke plurals for *moose* and *caboose* by analogy with *goose-geese*. Figure 2 presents excerpts from urbandictionary.com[3] that conveniently lay out the thought process behind the forced analogy. To spoil the joke, *meese* and *cabeese* are funny because they are not real plurals. *Geese* is an exception, not something you are supposed to generalize. At any rate, novel coinings have consistently adopted regular *-ed* pasts. For example, the past forms of *Bing* and *bling* are *Binged* and *blinged*, not **Bang* or **blung*. See Yang (2020) for further discussion.


meese

The real plural of moose. Many people, including the dictionary and English teachers, will attempt to tell you that "meese" is not correct. However, please consider the following:
One goose = goose
One moose = moose
Two+ goose = geese
Two+ moose = meese?
Yes, meese is grammatically correct. Don't let them fool you.

Look at that wild flock of meese!

by delovely May 24, 2005

cabeese

joking slang for multiple railroad cabooses

Goose is to geese as caboose is to cabeese

by Nemstar April 07, 2009

Figure 2: English irregular plural analogy as a joke.

3.2 Child Errors

Though children are certainly excellent at acquiring their native languages, they are far from perfect at it. They do make some errors[4] during development, and these errors are not random. When developing a computational model of morphological acquisition, it may be more informative to mirror learner errors than to achieve maximum accuracy. These errors, both individual instances and general trends, provide a window into children's morphological representations as they develop and learning proceeds.

Two major classes of errors are misapplications of productivity and omissions. Regarding the first class, there is a well-known asymmetry between over-regularization errors and over-irregularization errors in child productions (Pinker and Prince, 1994). The former can be explained as an over-application of productive patterns (*go*-*goed*) and is relatively common in child productions. The latter, such as an over-application of non-productive vowel mutation (e.g., *fry*-**frew* cf. *fly*-*flew*) is quite rare. Various studies have estimated the rate of over-regularizations in child English past tense productions to be between 8% and 10% (Maratsos, 2000; Yang, 2002; Maslen et al., 2004) while over-irregularization is under 0.2% (Xu and Pinker, 1995). Similar patterns are found in other languages as well, for example, for German past participles about 10% of productions are over-regularization with the productive *-t* suffix, while less than one percent are over-irregularizations with the unproductive *-(e)n* (Clahsen and Rothweiler, 1993). Less than 5% of error productions are over-regularization in children's Spanish verb productions, but much fewer, under 0.01%, can be seen as over-irregularization (Clahsen et al., 1992; Mayol, 2007). See Marcus et al. (1992, ch. 4) and Lignos and Yang (2018) for more discussion.

Over-regularization errors often follow a pattern of *U-shaped learning* in which children begin by making few such errors, rapidly enter a phase in which they produce a substantial number of errors, then gradually taper off to adult-like performance (Ervin and Miller, 1963; Bowerman, 1982; Pinker and Prince, 1988; Plunkett and Marchman, 1991). For English past tense, very young children may accurately produce irregulars such as *went* or *felt* for a time before suddenly producing over-regularized tokens such as **goed* and **feeled* (Marcus et al., 1992; Prasada and Pinker, 1993). This is evidence for a change in representation– early on, children lacks a productive past tense form and so memorizes regulars like they do irregulars, later, they realize that *-ed* is productive and begin applying it widely, and finally, they work out which words are exceptions to the productive pattern as they mature.

While this U-shaped pattern is not the only possible developmental trajectory– that depends on the input and specific linguistic pattern being acquired– it is both common and revealing. Computational researchers may demonstrate that their models admit a U-shaped learning trajectory as evidence in favor of their approaches (e.g., Rumelhart and McClelland, 1986; Plunkett and Marchman, 1991; Belth et al., 2021). Whether or not it actually manifests as U-shaped, children employ a non-monotonic learning strategy. At least in some circumstances, such non-monotonic strategies are provably necessary for learnability (Carlucci and Case, 2013), though it remains to be seen exactly how those proofs would be adapted for the acquisition setting (Section 3.3).

Another asymmetry can be found in the prevalence of *omission* errors over *commission* errors. The former refers to missing morphological information such as the substitution of a bare stem for an inflected form, while the latter refers to the substitution of one morphological form for another, such as the first person in lieu of the second. The phenomenon of root infinitives may be connected to errors of omission: in many languages including German (Poeppel and Wexler, 1993) and French (Ferdinand, 1996), children may produce infinitives instead of finite forms where finite forms are expected, then the surfacing of the infinitive is modeled in the syntax. In some languages these infinitive productions are much rarer (Italian; Guasti, 1993) or nearly absent altogether (Swahili; Deen, 2005). Languages with agglutinative inflectional morphology show that omission is not all or nothing. For example, Deen (2005) describes four different omission patterns in Swahili-learning children aged two years and two months to three years and one month. These patterns along with productions with no omissions and their rates summarized across the four children in the study are summarized in Table 1.[5]

Omission Type	**#**	**%**
SA-T-V-IND (no omission)	557	41.5
∅-T-V-IND	484	36.1
SA-∅-V-IND	114	8.5
∅-∅-V-IND (bare stem)	171	12.7
INF-V-IND (root inf.)	16	1.2

Table 1: Omission patterns for Swahili finite verbs without object marking. Summarized from Deen (2005). SA: subject agreement, T: tense, V: verb root, IND: indicative mood, INF: infinitive.

3.3 Representations and Formal Learnability

Learnability is a question investigated in computational learning theory, a branch of computer science and mathematics that investigates how systems ‘learn’ representations from data (Jain et al., 1999; Mohri et al., 2012). It provides the formal underpinning of all kinds of learning, from machine learning to human language acquisition (Niyogi, 2006; Clark and Lappin, 2012; Heinz, 2016). This article does not provide a detailed discussion on formal learnability, but does touch on a few relevant topics in this section.

In formal mathematical applications and engineering, it is common to represent morphological processes as compositions over finite state transducers (FSTs; Roark et al., 2007; Gorman and Sproat, 2021), which are objects that consist of states and transitions between those states. They extend the classic finite state automata of formal language theory with the addition of output strings on the transitions, allowing them to map between forms. As an example, Figure 3 presents an FST that performs a simplified mapping between Spanish singular and plural nouns.

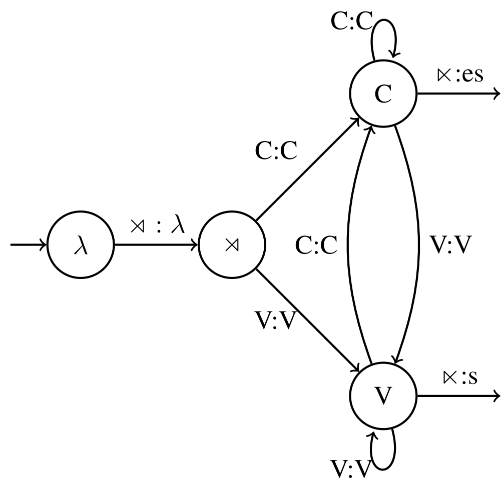


Figure 3: Spanish plurals end in *-s* if the singular ends in a vowel and *-es* otherwise. ⋉ and ⋊ are string start and end symbols. λ represents an empty string. The transducer returns each character (or phone) of the input string plus *-s* or *-es* as appropriate.

From an engineering perspective, the strengths of this approach are flexibility, interpretability, and verifiability. Composition of FSTs or related formalisms is flexible enough to capture (nearly) all morphological processes, both affixation and complex non-concatenative processes. FSTs are interpretable, and therefore debuggable, because one can audit a transducer by tracing through a path of states for any input. They are verifiable because they are well defined mathematical objects which are subject to formal proofs of validity. Several programming libraries are available for implementing FSTs including Xerox Finite-State Tools (Beesley and Karttunen, 2003), OpenFst (Allauzen et al., 2007), Foma (Hulden, 2009), and Pynini (Gorman, 2016). Other classic technologies, including Two-Level Morphology (Koskenniemi, 1983) are formally equivalent to FSTs or nearly so (Roark et al., 2007, ch. 4).

FSTs and the like can be equivalently conceived of as functions which themselves can be composed to yield more complex morphological operations. An agglutinative form like the Swahili SA-T-V-IND pattern in Table 1 might be accomplished by composing three functions, one that applies subject agreement, one that applies tense, and one that applies mood. From a formal perspective, one- and two-way FSTs capture exactly the rational and regular relations respectively, well-defined mathematical classes. Properly characterizing morphology in terms of regular relations (or a proper subset thereof) opens it up to mathematical proofs of formal learnability. See the classic Oncina et al. (1993) algorithm and de la Higuera (2010) and Heinz et al. (2015a) for reviews. Most, or maybe all morphological processes belong to a proper subset of the regular languages (Chandlee, 2017). Some subsets of these languages are provably learnable *in the limit* from positive evidence alone (Oncina et al., 1993; Chandlee, 2014; Jardine et al., 2014), that is, they are *Gold learnable* (Gold, 1967).

Given the lack of actionable negative feedback during language acquisition, the positive evidence-only assumption is a realistic one. Learning in the limit is less so, because the learner may require an arbitrarily large number of inputs before settling on the intended grammar. This can be overly permissive given that children receive input which is curtailed in many ways. On the other hand, the success criterion is overly strict, since children are not actually expected to converge on exactly the same grammar(s) that generated their finite input. They should achieve something very close to it, but it need not be identical. One reason for this is variation in the input. Variation, both interpersonal and intra-personal, are reoccurring themes in linguistics, and few if any learners receive input generated by just a single grammar.[6]

Gold learnability, its relatives, and work built on them have been criticized for their lack of realism and inapplicability to the acquisition setting (e.g., Nowak et al., 2002; Niyogi, 2006). Though warranted to an extent, such criticisms may be too strong. First, these classic learning settings remain better defined and more manageable from a mathematical perspective. See (Heinz et al., 2015b; Heinz, 2016) for additional discussion of learning paradigms. Second, language acquisition is a rather esoteric learning setting (summarized in Table 4), and it has yet to be sufficiently formalized in purely mathematical terms. Well-defined frameworks present a way to understand the behavior of learning algorithms independently of the frameworks themselves.

It should be noted that FSTs and FST operations are of a different kind from those typically proposed by theoretical morphologists. They are a formalism for expressing a class of relations or mappings rather than a system that is reified cognitively. They can be said to capture the *computational* (in the sense of Marr (1982)) properties of any equivalently expressive theoretical formalism, possibly including a given theoretical account. FSTs were chosen as an example computational formalism here and are studied in the field because they lend themselves to precise statements about expressivity and learnability and because they are convenient to implement and audit.

It is worth mentioning another family of computational formalisms that have proven quite successful in the NLP community but seem to have less to contribute to discussions of learnability. *Distributed representations*, gained a following in morphological learning through the work of connectionist psychologists and linguists in the 1980s (e.g., Dell, 1986; Rumelhart and McClelland, 1986; Seidenberg and McClelland, 1989). Rather than operating on discrete symbols, the representations in connectionist models were distributed across many "neurons" in architectures inspired by the organization of the brain. More recently, these neural networks have been greatly enhanced and expanded by the machine learning and NLP fields, forming the basis for modern *deep learning* (LeCun et al., 2015; Goldberg, 2017). Deep learning models have slowly propagated from engineering back towards linguistics and cognitive science.

While increasingly popular with the growth of deep learning, distributed representations remain contentious as a cognitive model of language. See Pater (2019) and the several responses to it for an up to date discussion. In particular, they do not offer the rigorous mathematical advantages that FSTs and other lower-expressivity formalisms do from the perspective of learning theory (Rawski and Heinz, 2019), as they are both extremely powerful (more powerful than the minimum required to represent a human grammar) and also severely lacking in interpretability due to the distributed nature of the representations and massive size of the networks. Together, these curtail their usefulness as a mode of explanation. But theoretical discussion aside, they still struggle to achieve human-like performance in the acquisition setting despite their significant power (see Section 4.2.2 for further discussion).

4 Implementations and Data Sets

The field's ultimate bounty, a complete end-to-end algorithmic implementation for morphological acquisition from naturalistic acoustic input through to human-like representations and productions is still out of reach. Not only do models fall short in their performance, but there is still surprisingly little agreement as to what kinds of models should even be employed. That said, significant progress continues to be made on the various components of morphological acquisition. This section surveys morphological acquisition as it has been divided up by the computational linguistics and NLP communities with an eye towards the characterization of acquisition provided in the previous sections.

Several simplifying assumptions are typically made in morphological learning systems. First, input usually comes in the form of segmented text, either running text or word lists. Text may be presented in native orthography or transcribed phonologically, and word lists may or may not come with token frequency information. Second, semantic information may be omitted altogether, it may be annotated as feature tags, or it may be induced distributionally from co-ocurrences in running text. Third, the goal is often not to learn a grammar *per se* as the child does, but rather to produce a human-interpretable analysis or to map features to a correct form. This might be described as an E-language approach rather than I-language approach to learning in the sense of Chomsky (1986), differentiating it from much modern non-computational work in acquisition and theory.[7] The problems that interest the computational morphology learning community (which lies largely within NLP) could be divided into two categories. *Morphological analysis* is concerned with systems that learn how to recognize or analyze morphological forms or patterns. *Morphological generation* is focused on the production of morphological forms or on extracting productive generalizations over which forms are produced.

4.1 Morphological Analysis

Morphological analysis is a broad term that can capture all tasks involved in morphological segmentation or the pairing of form and meaning. Segmentation in the strictest sense is just the division of words into morphemes, or equivalently, the identification of morpheme boundaries within words. As in the case of FSTs, morphological operations may be conceived of as the successive application of functions or processes that build morphological forms. There are many such processes including affixation at the edges or middle of a word, reduplication, and stem transformations. Of these, only edge-affixation is available to the simplest concatenation-based models. Others explicitly model derivations as a series of morphological operations, which allows for wider cross-linguistic coverage (Schone and Jurafsky, 2001; Narasimhan et al., 2015; Soricut and Och, 2015; Luo et al., 2017; Xu et al., 2018).

Segmentation learning systems have a very long history (See Hammarström and Borin (2011) for a long survey). Harris (1954) proposed a model based on transitional probabilities which continues to be heavily cited in computational morphology papers. Though less often cited, it was actually implemented on a CDC mainframe by Philip Rabinowitz several years later (Harris, 1970), and while performance was poor by modern standards, it is easy to sympathize with Harris's optimism at the time.

More recently, there was a blossoming of segmentation models during the 2000s supported by the Morpho-Challenge tasks (Kurimo et al., 2010) yielding a wide range of models (e.g., Creutz and Lagus, 2005; Monson et al., 2007; Virpioja et al., 2009; Lignos, 2010). Morpho-Challenge largely standardized the task, facilitating comparison between models. Its data sets, which are still available online[8] and continue to be used (Narasimhan et al., 2015; Eskander et al., 2016; Xu et al., 2018) are word lists drawn from web data and contain thousands of items generally distributed according to natural highly skewed data sets. In that way, they stand in reasonably well for acquisition, at least if they can be filtered. One downside is that they contain a very large amount of noise drawn from languages other than the target, mis-encoded Unicode, and other non-language text.

The original Morpho-Challenge task was truly unsupervised and provided only word forms as input for learners. This is an excellent stress test for determining which aspects of morphology can be learned from form alone but is unrealistically restrictive: some amount of semantics can be approximated by leveraging distributional information in running text which still maintaining the unsupervised setting. Segmentation on CDS-derived wordlists performs reasonably well, suggesting that the word forms alone are a sufficient signal for most morphology learning (Lignos et al., 2010). Distributional information can also be leveraged to construct partial paradigms as an intermediate step towards segmentation or as a goal unto itself(Goldsmith, 2001; Xu et al., 2018, 2020). Such approaches, including (Narasimhan et al., 2015) which used running text to supplement the Morpho-Challenge data sets, add a degree of realism, since children also experience language as utterances rather than wordlists. That said, they train on much more data than is available to the young learner and so do not serve as acquisition models.

Several other segmentation data sets exist besides Morpho-Challenge, which are often created for specific use cases such as for low-resource languages (Mott et al., 2020) or for languages that require more complex annotation schemes. There are data sets of various kinds available for several Arabic varieties, for example, all of which must contend with the family's pervasive non-concatenative morphology (Maamouri and Bies, 2004; Maamouri et al., 2012; Khalifa et al., 2018).

The main conceptual alternative to segmentation is feature tagging. A feature tagging system learns to assign semantic features such as those encoding person/number, tense/aspect/mood, or inflectional class, to inflected forms. For example, English *walked* and *ran* may both be tagged with a PAST feature rather than attempting to attribute any part of the word form to the past tense meaning. These features need to be annotated and provided to the learning system, so this approach requires more supervision at a minimum than segmentation does. The two most important feature annotated data sets at the time of writing are UniMorph (McCarthy et al., 2020) and Universal Dependencies (UD; Zeman et al., 2021). Both projects are available for a large and ever growing set of languages, and they provide lemmas and feature sets for inflected forms, albeit with incompatible schemes (Table 2). UD also provides dependency parses, but not all language corpora provide lemmatization. However, one advantage of UD over UniMorph is that its running text allows one to extract distributional information and measure data sparsity.

Language	**Lemma**	**Inflected**	**UD Features**	**UniMorph Features**
Finnish	työpaikka	työpaikkoja	Case=Par|Number=Plur	N;PRT;PL
Spanish	cruzar	cruzó	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	V;IND;PST;3;SG;PFV
Turkish	gir(mek)	gireceğim	Aspect=Perf|Mood=Ind|Number=Sing|Person=1|Tense=Fut	V;IND;FUT;1;SG;POS;DECL

Table 2: Comparison of UniMorph and Universal Dependencies semantic feature annotations in Finnish, Spanish, and Turkish. They provide *gir* and *girmek* respectively as the lemma for this Turkish example.

In addition to UniMorph and UD, corpora within CHILDES often contain lemmatization and morphological feature annotations on their `%mor` tiers. These can often be compared to the transcription lines to extract inflected forms as well. CHILDES annotations are useful because they gives much the same information as the larger feature annotated corpora but also provide distributional information associated with actual samples of the learner's linguistic input. The paradigm saturation Figure 1 was generated from both CHILDES and UD data.

Feature annotation, as opposed to segment annotation, has gained popularity in recent years because it is amenable to direct string-to-string mapping favored by modern neural generation systems. For several years now, UniMorph has been the de facto standard for the annual (CoNLL-)SIGMORPHON shared tasks, which include several generation subtasks (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Kann et al., 2020). See each year's summary paper for descriptions of several models.

The features provided in these annotation schemes together with lemmatizations implicitly define morphological paradigms. These can be leveraged for *paradigm discovery* or at least used as gold standards for unsupervised paradigm discovery. This task is concerned with grouping word forms into morphologically related sets, or alternatively, grouping morphological processes into sets that apply to homogeneous sets of words. In the first sense, the paradigm relates to all the potential forms of a particular lemma, for example, English verbs take up to five forms (*ride*, *rides*, *riding*, *rode*, *ridden* or *jump*, *jumps*, *jumping*, *jumped*, *jumped*). In the second sense, it refers to the abstract paradigm itself, perhaps represented as a collection of functions corresponding to morphological processes. For example, many Afro-Asiatic languages inflect verbs for three persons (1, 2, 3), two or three numbers (SG, PL(, DU)), and two genders (M, F). One function for each and their compositions could describe the paradigm.

Entirely unsupervised paradigm discovery based on form alone poses a significant challenge. Keeping Zipfian paradigm saturation in mind, the vast majority of stems will only be attested with a tiny fraction of their paradigm for any morphologically rich language. Furthermore, the elements of a paradigm may be ambiguous as to part-of-speech. Borrowing an example from Xu et al. (2018) and Xu et al. (2020), *-er* is part of a set of patterns that apply to adjectives (comparative) as well as one that applies to nouns (agentive), while *-s* applies both to verbs and nouns. If a low paradigm saturation item is only attested with *-s*, a learner presented with a wordlist or a short and syntactically ambiguous CDS utterance does not know whether that item is a verb that should also accept *-ing* or if it is a noun that should not.

Leveraging distributional information renders this problem significantly more manageable (Parkes et al., 1998; Goldsmith, 2001; Chan, 2006; Narasimhan et al., 2015; Xu et al., 2018), since contextual information can disambiguate part-of-speech. Even extremely simple distributional information can be beneficial for morphological learning. For example, Parkes et al. (1998) shows that tabulating right and left co-occurences of an English word type is often sufficient to classify it by both syntactic category and inflectional category even if word form is totally ignored. Experimental evidence shows that children are indeed sensitive to such local co-occurences (Mintz, 2003). Nevertheless, state-of-the-art performance in unsupervised paradigm discovery remains quite low, even when part-of-speech is provided up front (Jin et al., 2020). More work has been done on semi-supervised paradigm discovery with seed sets or some amount of annotated data (Dreyer and Eisner, 2011). The unsupervised version is an extremely challenging task, and developing a model that achieves it in a naturalistic acquisition setting would immensely improve our understanding of child language development.

For the child learner, paradigms, in the sense of groupings of related forms, are important for overcoming input sparsity: if a child encounters a word in one corner of its paradigm, it is likely to be licit in other parts of the paradigm too, so the child can infer unencountered forms. However, this assumes that the child has already discovered the paradigm. Once features and a paradigm are inferred, they can be combined with segmentation to perform a more complete analysis that assigns features to individual morphemes or transformations. An intensional representation of this mapping constitutes a grammar not dissimilar from those studied in generative linguistics. Such systems built on input distributions extracted from CHILDES and features and lemmatizations taken from UniMorph have been developed recently (Payne et al., 2021; Belth et al., 2021), but the full pipeline from paradigm discovery to morpheme-feature mapping is still elusive.

4.2 Morphological Generation

Morphological generation is concerned with producing word forms under certain conditions rather than just recognizing and analyzing them. This is of course something that humans do when we speak, and it is a critical engineering task for natural language generation of languages with any amount of productive morphology. Most morphological generation can be classified as some kind of inflection task, mapping a lemma or stem to a particular form. Models of generalization or productivity are also categorized as generation here because they describe and audit the patterns that are used to generate new forms.

4.2.1 Inflection

As a computational morphology task, *inflection* refers to the production of a form given a lemma and a set of semantic features, for example, the pair (*goose*, PL) should yield *geese*. Many minimally supervised inflection models have been proposed, including classic connectionist models (e.g., Rumelhart and McClelland, 1986) and those drawing from advancements in NLP (e.g., Mooney and Califf, 1995; Yarowsky and Wicentowski, 2000). More recently, a significant body of research on this task has been developed supported by the (CoNLL-)SIGMORPHON shared tasks and UniMorph. Most recent approaches have been neural systems that perform string-to-string mapping. That is, they eschew any sort of underlying representations including segmentation and instead learn a more holistic mapping from the surface input string (the lemma plus the feature tags) to the output string (the inflected form). Each year's shared task has presented some variant of the problem with more or less data. It is best to refer to each year's summary paper for details (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Kann et al., 2020). A more challenging version of this task posed in 2018 does not present the feature tags but rather sentence context. This forces the learner to infer the relevant semantics distributionally, which is more similar to the actual use case in machine translation or natural language understanding. An example from the task paper is reproduced in (1). The system is presented with the lemma *dog* but not the tag PL. Rather, it has to infer that it must produce the plural form given the sentence context.

(1)
The	_____	are	barking
the/DT	dog	be/AUX+PRES+3PL	bark/V+V.PTCP

Lemmatization might be seen as the reverse of inflection, since it requires recalling a lemma given an inflected form. It can be aided with feature annotation as well, for example, the pair (*geese*, PL) should yield *goose*. A related task, *stemming* can be thought of as an ersatz lemmatization that simply chops off endings to create "good enough" uniform forms to be fed into downstream NLP tasks. To explain the difference, consider the forms, *carry*, *carrying*, *carries*, and *carried*. A lemmatizer should reduce all of these to *carry*, while a stemmer, such as the classic Porter Stemmer (Porter, 1980)[9] might reduce them to *carri*. The latter is never attested in English, but that may not matter if it is only used internally for some engineering task down the road.

Reinflection maps from inflected form to inflected form laterally rather than through a lemma. This takes a triple as input, for example, a reinflection from the English past to progressive might look like (*sang*, PAST, PRES.PROG) and would yield *singing*. *Paradigm completion* can be thought of as a kind of reinflection task as well: the system is first trained on paradigms of a known size and shape (greatly simplifying the problem of paradigm discovery). Then at test time, the system fills in missing cells from a paradigm that is provided to it via reinflection from the filled cells. This was posed as part of the 2017 and 2019 shared tasks.

On the whole, modern systems perform quite well on inflection and reinflection tasks in terms of output accuracy. In their analysis of the 2017 CoNLL-SIGMORPHON shared task, Gorman et al. (2019) find that many errors can actually be attributed to the data sets themselves rather than the models: sometimes the model produced a valid variant form absent from the gold standard, the gold standard was compiled incorrectly, or the source data for the gold standard was itself incorrect. Relatively few errors were the very unusual "silly" errors that characterized earlier connectionist work. Some were orthographic errors and the majority not attributable to the data sets were allomorphy errors. Some of these are reminiscent of what a child might do, for example incorrectly guessing irregular German plurals. On the other hand, some allomorphy errors were not child-like. One system over-applied Spanish diphthongization (*o* to *ue* and *e* to *ie*), a frequent but largely unpredictable pattern which children *under*-apply if anything (Mayol, 2007).

The findings of Gorman et al. (2019) highlight the progress that neural approaches have made since the early connectionist days where "silly" errors were very common. They do not, however, tell us whether these models are constructing these forms in a human-like way. The classic connectionist models were unduly subject to frequency effects, and it seems that modern neural models still are, as seen in the spurious over-application of Spanish diphthongization. The next section will return to this point in its discussion of productivity and German nouns.

4.2.2 Generalization

The generation tasks discussed so far have all been focused on achieving correct surface forms rather than a human-like grammar as one might acquire from the data during acquisition. However, the grammar is more than just a mapping of semantics to forms. Studies of developmental trajectories (e.g., U-shaped learning) and experiments (e.g., the classic Wug test) use productive generalization to provide a window into the grammar. Furthermore, the sparsity of early linguistic input necessitates productivity. The learner will not encounter most possible forms and therefore must be able to generate rather than just recall them.

While there is a fair amount of convergence on what constitutes productivity at a high level (See Bauer (2001) for a review), there is quite a large degree of divergence when it comes to the details. Even the definition of productivity can be slippery and hard to pin down. Many researchers use it to describe what might be called "cognitive productivity" or the speakers' internal drive to make and employ generalizations (Rumelhart and McClelland, 1986; Albright, 2003; O'Donnell, 2015; Yang, 2016). Computational models have been central to the development of our understanding of cognitive productivity. Four models will be briefly discussed here. Though their particulars differ, sometimes dramatically, they crucially agree that there must be some critical mass or threshold of evidence for the learner to acquire productivity.[10]

First, connectionist networks and their younger and bigger cousins, deep neural networks, rely on distributed representations to encode linguistic knowledge and do not make an explicit distinction between productive and unproductive or regular and irregular patterns. Rather, productivity is seen as an emergent tendency for some patterns to be generalized. The feed-forward connectionist model of Rumelhart and McClelland (1986) kicked off the Past Tense Debates (McClelland and Patterson, 2002; Pinker and Ullman, 2002, for a review) when it was purported to learn to correctly inflect English past tense and follow a realistic U-shaped developmental trajectory. This was criticized by Pinker and Prince (1988) who observed that the network was actually frequency matching and that U-shaped learning was only achieved by training on irregulars first then flooding the network with regulars, an input distribution never observed in a child's natural environment.

Connectionist and neural models in the following decades have taken advantage of engineering advances along the way, allowing for more naturalistic string-to-string mappings and variable length inputs as well as more accurate frequency matching (e.g., Plunkett and Marchman, 1993; Hare and Elman, 1995; Bullinaria, 1997; Plunkett and Juola, 1999; Seidenberg and Plaut, 2014; Kirov and Cotterell, 2018). More recently, Kirov and Cotterell (2018) applied a modern encoder-decoder (ED; Bahdanau et al., 2014) to the task and found that it solved most of the practical issues of earlier models altogether. In line with neural generation models in general, it achieved quite high accuracy on a computational wug test following Albright and Hayes (2003). Unfortunately, this type of model still primarily probability matches (Corkery et al., 2019; McCurdy et al., 2020; Beser, 2021). The technical improvements have not overcome the basic linguistic failing of the early connectionist models.

McCurdy et al. (2020) demonstrate this by evaluating the ED model's behavior on German where the productive pattern is not the most frequent (Table 3). This has long been seen as a critical test case for connectionist models (Köpcke, 1988; Marcus et al., 1995; Clahsen, 1999) because it decouples the concepts of "most productive" and "most frequent." ED overwhelmingly prefers *-e*, which has the widest distribution among the possible suffixes. While *-(e)n* is more frequent overall, its numbers are highly concentrated among feminine nouns and those ending in schwa (Sonnenstuhl and Huth, 2002). *-s*, which is much less frequent, is default plural that the model should have generally predicted for unknown words, at least non-feminine ones (Clahsen, 1990; Marcus et al., 1995), but the ED model does not pick up on that. It extracted patterns from the training data but did not behave like a human.

Plural suffix	**% of all**	**% of neuter**
-(e)n	37.3	3.2
-e	34.4	51.9
-∅	19.2	21.5
-er	2.0	10.6
-s	4.0	7.7
other	2.1	5.1

Table 3: German nominative plural suffix frequency in UniMorph. Adapted from Corkery et al. (2019)

A more general takeaway from this back and forth is a word of caution against applying the standard NLP train-test evaluation scheme to cognitive modeling. It is conventional in NLP to divide evaluation data into a training set and test set drawn from the same distribution, to train a model on the former, and then to test on the latter. Good performance on the test set suggests that the model has learned the distribution of the training data without over-fitting. However, the goal in modeling acquisition is not just to learn and accurately extend the distribution of forms in the training data, it is also to generalize in a human-like way (or better, for human-like reasons), regardless of how that relates to probabilities. Compared to feed-forward connectionist models, deep learning models achieve better performance in an engineering sense, but they reveal little about how humans acquire morphology.

In contrast to the connectionists, many models have been proposed which build morphological forms by explicitly encodable productive rules, patterns, or parts (Pinker et al., 1987; Pinker and Prince, 1988; Clahsen et al., 1992; Mooney and Califf, 1995; Albright, 2003; O'Donnell et al., 2011; Yang, 2016). The Minimum Generalization Learner (MGL) (Albright, 2002; Albright and Hayes, 2003) uncovers "islands of reliability" in which productive rules can be learned. An MGL learner may, for example, uncover a *sing-sang* or *sting-stung* rule on the basis of stem changing English past tense. The model works from the bottom up, creating many narrowly defined rules and and joining them into more general ones if the data allows for it. In order to evaluate the model, MGL may be compared against the results of adult Wug tests on English (Albright, 2002) and other languages as well (Japanese; Klafehn, 2003; Oseki et al., 2019).

The MGL employs a notion of reliability defined in (2) as the number of forms that a rule derives (hits) divided by the number forms that it could potentially derive (scope). The reliability of rules with smalls scopes are penalized by transforming them into a confidence score. Perhaps the greatest downfall of the MGL is that it requires complete or nearly complete paradigms during training, which is not realistic for languages with even moderately sized paradigms (Chan, 2008, ch. 3).

(2) **MGL Reliability**:
Reliability $\hat{p}$ is the number of forms that a rule derives (hits) divided by the number of forms it could potentially derive (scope):

$$\text{reliability } \hat{p} = \frac{\text{hits}}{\text{scope}} \tag{1}$$

Fragment Grammars (FG; O'Donnell et al., 2011; O'Donnell, 2015) present an alternative theory of productivity in which forms may be parsed into some number of reusable fragments (equivalent to productive rules) or stored whole (as for exceptions). For example, the word *agreeability* may be stored as a single fragment, it may be divided into *agree* and a reusable *-ability*, or further into *-abil-* and *-ity*. Learning is conceived of as a Bayesian inference problem in which the learner must decide whether forms are better represented as fragments or stored.

In contrast to the models just described, the Tolerance Principle (TP Yang, 2016) positions itself specifically as an evaluation metric by which a learner decides whether a hypothesized pattern is productive over some domain or is not. As in other rule-based models, the TP casts rules as productive and stored items as exceptions. The core of the TP is the *tolerance threshold*, the number of exceptions below which it becomes more efficient to hypothesize a generalization than to list items. (3) provides a formulation of the Tolerance Principle. The tolerance threshold θ_N is defined as the number of known types that a generalization should apply to divided by its natural logarithm.

(3) **Tolerance Principle**:
If R is a productive rule applicable to N candidates, then the following relation holds between N and e, the number of exceptions that could but do not follow R:

$$e \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N}$$

The tolerance threshold was derived according to observations of the child learner's input and learner behavior. It assumes a generally Zipfian input distribution and performs *better* on small child-sized data than on larger data sets. It also finds support in psycholinguistics experiments run on children (Schuler, 2017; Emond and Shi, 2020). Slotted into other learning models as an evaluation metric (Belth et al., 2021; Payne et al., 2021), TP-based models perform well on data derived from CHILDES, achieving high accuracy and a U-shaped learning trajectory on English past tense, for example, while also outperforming neural models such as that of Kirov and Cotterell (2018).

5 What a Computational Learner Needs

A computational learner needs data to learn from, a class of representations to acquire, and a way to get from one to another. Morphology acquisition is a very specific learning problem that is defined by these components, and not all morphological learning is morphological acquisition. Table 4 summarizes the key characteristics of the morphological acquisition problem in semi-formal terms. The more of these that a computational learning model achieves, the better it directly models child learners acquiring their native morphologies.

Component	Characteristics
Input size	**Finite**; tens of millions of tokens
	Most tokens are irrelevant (e.g., function words or unsegmented early input stream)
	Learner "knows" a few hundred types → Most learning is on the basis of these types
Input distribution	**Predictable**; highly sparse and generally Zipfian.
	Most forms will not be attested during learning
	High frequency inputs maybe be disproportionately irregular
	Sparsity correlated with paradigm size
Learning path	**Non-monotonic**; Often follows U-shaped learning path
	Errors are overwhelmingly over-regularization or omission
Successful outcome	Learner ultimately produces outputs consistent with the community
	→ Learners converge on extensionally "similar" but not necessarily identical grammars

Table 4: Informal summary of morphology acquisition as a formal learning problem

From studies of acquisition corpora, we know that the input is both very sparse and very skewed. Lexical items and inflected forms are predictably distributed according to long-tailed Zipfian distributions across languages, and children apparently learn most of their native morphologies on the basis of only a few hundred types with no help from negative evidence. Nevertheless, they learn more accurately than any artificial system so far. Their occasional errors of over-regularization or omission and their behavior in the laboratory show us glimpses of how they acquire language in such averse conditions, but exactly how they do it is still a puzzle, a puzzle which computational approaches are in a prime position to solve.

Further Reading

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

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The Greedy and Recursive Search for Morphological Productivity. In *Proceedings of the 43th Annual Conference of the Cognitive Science Society*, pages 2869–2875.

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

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

Chan, E. (2008). *Structures and distributions in morphological learning*. PhD thesis, University of Pennsylvania, Philadelphia, PA.

Deen, K. U. (2005). *The acquisition of Swahili*, volume 40. John Benjamins Publishing.

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

Heinz, J. (2016). Computational theories of learning and developmental psycholinguistics. In Lidz, J., Synder, W., and Pater, J., editors, *The Oxford Handbook of Developmental Linguistics*, chapter 27, pages 633–663. Oxford University Press, Oxford, UK.

Jain, S., Osherson, D., Royer, J. S., and Sharma, A. (1999). *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change).* The MIT Press, 2nd edition.

Lignos, C. and Yang, C. (2018). Morphology and language acquisition. *Cambridge handbook of morphology*, pages 765–791.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3922–3931.

O'Donnell, T. J. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. MIT Press, Cambridge, MA.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Rumelhart, D. E. and McClelland, J. L. (1986). *On learning the past tenses of English verbs*. MIT Press, Cambridge, MA.

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

Yang, C. (2016). *The Price of Linguistic Productivity*. MIT Press, Cambridge, MA.

References

Aksu-Koç, A. A. (1985). The acquisition of Turkish. *The Cross-linguistic Studies of Language Acquisition. Vol. 1: The Data*, pages 839–876.

Albright, A. (2003). A quantitative study of Spanish paradigm gaps. *WCCFL 22 Proceedings. Cascadilla Press, Somerville, MA*, pages 1–14.

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

Albright, A. C. (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFST: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Anglin, J. M., Miller, G. A., and Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186.

Audring, J. and Masini, F. (2018). *The Oxford handbook of morphological theory*. Oxford University Press.

Baayen, H. (1993). On frequency, transparency and productivity. In *Yearbook of morphology 1992*, pages 181–208. Springer, Dordrecht.

Baayen, R. H. and Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, pages 69–96.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bauer, L. (2001). *Morphological productivity*, volume 95. Cambridge University Press.

Beesley, K. R. and Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

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

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The Greedy and Recursive Search for Morphological Productivity. In *Proceedings of the 43th Annual Conference of the Cognitive Science Society*, pages 2869–2875.

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

Beser, D. (2021). Falling Through the Gaps: Neural Architectures as Models of Morphological Rule Learning. In *Proceedings of the 43th Annual Conference of the Cognitive Science Society*, pages 1042–1048.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, 75(4):1115–1139.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In *Language acquisition: The state of the art*, pages 319–346. Academic Press.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In Hawkins, J. A., editor, *Explaining language universals*, pages 73–101. Basil Blackwell, Oxford.

Braine, M. D. (1971). On two types of models of the internalization of grammars. In Slobin, D., editor, *The ontogenesis of grammar: A theoretical symposium*, volume 1971, pages 153–186. Academic Press, New York.

Brent, M. R. and Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):33–44.

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

Brown, R. and Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. *Cognition and the development of language*.

Bullinaria, J. A. (1997). Modeling reading, spelling, and past tense learning with artificial neural networks. *Brain and Language*, 59(2):236–266.

Caprin, C. and Guasti, M. T. (2009). The acquisition of morphosyntax in Italian: A cross-sectional study. *Applied Psycholinguistics*, 30(1):23–52.

Carlucci, L. and Case, J. (2013). On the Necessity of U-Shaped Learning. *Topics in cognitive Science*, 5(1):56–88.

Cazden, C. B. (1972). *Child language and education.* Hold, Rinehart & Winston, New York.

Chan, E. (2006). Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology*, pages 69–78. Association for Computational Linguistics.

Chan, E. (2008). *Structures and distributions in morphological learning*. PhD thesis, University of Pennsylvania, Philadelphia, PA.

Chandlee, J. (2014). *Strictly local phonological processes*. University of Delaware.

Chandlee, J. (2017). Computational locality in morphological maps. *Morphology*, 27(4):599–641.

Chomsky, N. (1981). *Lectures in government and binding*. MIT Press Cambridge, MA.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Clahsen, H. (1990). Constraints on parameter setting: A grammatical analysis of some acquisition in stages in German child language. *Language Acquisition*, 1(4):361–391.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and brain sciences*, 22(6):991–1013.

Clahsen, H. and Rothweiler, M. (1993). Inflectional rules in children's grammars: Evidence from German participles. In *Yearbook of morphology 1992*, pages 1–34. Springer.

Clahsen, H., Rothweiler, M., Woest, A., and Marcus, G. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45:225–255.

Clark, A. and Lappin, S. (2012). Computational learning theory and language acquisition. *Philosophy of linguistics*, 14:445–475.

Corkery, M., Matusevych, Y., and Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. *arXiv preprint arXiv:1906.01280*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., Kann, K., Mielke, S., Nicolai, G., Silfverberg, M., et al. (2018). The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. *arXiv preprint arXiv:1810.07125*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., et al. (2017). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *arXiv preprint arXiv:1706.09031*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Dąbrowska, E. (2001). Learning a morphological system without a default: The Polish genitive. *Journal of child language*, 28(3):545–574.

de la Higuera, C. (2010). *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.

Deen, K. U. (2005). *The acquisition of Swahili*, volume 40. John Benjamins Publishing.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283.

Derwing, B. L. and Baker, W. J. (1977). The psychological basis for morphological rules. In Macnamara, J., editor, *Language learning and thought*, pages 85–110. Academic Press, New York.

Dreyer, M. and Eisner, J. (2011). Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 616–627, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Emond, E. and Shi, R. (2020). Infants' rule generalization is governed by the Tolerance Principle. In *45th Annual Boston University Conference on Language Development*.

Ervin, S. M. and Miller, W. R. (1963). Language development. *Child Psychology*, pages 108–143.

Eskander, R., Rambow, O., and Yang, T. (2016). Extending the use of Adaptor Grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Ferdinand, A. (1996). *The development of functional categories: The acquisition of the subject in French*. ICG Printing, Dortrecht.

Fernald, A. and Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child development*, 83(1):203–222.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

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

Gorman, K. (2016). Pynini: A Python library for weighted finite-state grammar compilation. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 75–80.

Gorman, K., McCarthy, A. D., Cotterell, R., Vylomova, E., Silfverberg, M., and Markowska, M. (2019). Weird inflects but ok: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151.

Gorman, K. and Sproat, R. (2021). *Finite-State Text Processing*, volume 14. Morgan & Claypool Publishers.

Gorman, K. and Yang, C. (2019). When nobody wins. In *Competition in Inflection and Word-Formation*, pages 169–193. Springer.

Guasti, M. T. (1993). Verb syntax in Italian child grammar: Finite and nonfinite verbs. *Language Acquisition*, 3(1):1–40.

Guy, G. R. and Boyd, S. (1990). The development of a morphological class. *Language variation and change*, 2(1):1–18.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic inquiry*, 4(1):3–16.

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

Hare, M. and Elman, J. L. (1995). Learning and morphological change. *Cognition*, 56(1):61–98.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Harris, Z. S. (1970). Morpheme boundaries within words: Report on a computer test. In *Papers in Structural and Transformational Linguistics*, pages 68–77. Springer.

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

Heinz, J. (2016). Computational theories of learning and developmental psycholinguistics. In Lidz, J., Synder, W., and Pater, J., editors, *The Oxford Handbook of Developmental Linguistics*, chapter 27, pages 633–663. Oxford University Press, Oxford, UK.

Heinz, J., de la Higuera, C., and van Zaanen, M. (2015a). *Grammatical Inference for Computational Linguistics*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Heinz, J., De la Higuera, C., and Van Zaanen, M. (2015b). Grammatical inference for computational linguistics. *Synthesis Lectures on Human Language Technologies*, 8(4):1–139.

Howes, D. (1968). Zipf's Law and Miller's random-monkey model. *The American Journal of Psychology*, 81(2):269–272.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Jain, S., Osherson, D., Royer, J. S., and Sharma, A. (1999). *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change)*. The MIT Press, 2nd edition.

Jardine, A., Chandlee, J., Eyraud, R., and Heinz, J. (2014). Very efficient learning of structured classes of subsequential functions from positive data. In *International Conference on Grammatical Inference*, pages 94–108. PMLR.

Jarosz, G. (2019). Computational modeling of phonological learning. *Annual Review of Linguistics*, 5:67–90.

Jin, H., Cai, L., Peng, Y., Xia, C., McCarthy, A., and Kann, K. (2020). Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Kann, K., McCarthy, A., Nicolai, G., and Hulden, M. (2020). The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. *arXiv preprint arXiv:2005.13756*.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Al Kaabi, M. (2018). A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Klafehn, T. (2003). *Emergent properties of Japanese verbal inflection*. PhD thesis, University of Hawaii at Manoa.

Kodner, J. (2019). Estimating child linguistic experience from historical corpora. *Glossa: A Journal of General Linguistics*, 4(1):122.

Kodner, J. (2020). *Language Acquisition in the Past*. PhD thesis, University of Pennsylvania.

Köpcke, K.-M. (1988). Schemas in German plural formation. *Lingua*, 74(4):303–335.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*, volume 11. University of Helsinki, Department of General Linguistics Helsinki, Finland.

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

Kurimo, M., Virpioja, S., Turunen, V., and Lagus, K. (2010). Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American-English*. Brown University Press, Providence, RI.

Landau, B., Gleitman, L. R., and Landau, B. (2009). *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lignos, C. (2010). Learning from unseen data. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 35–38.

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

Lignos, C. and Yang, C. (2018). Morphology and language acquisition. *Cambridge handbook of morphology*, pages 765–791.

Luo, J., Narasimhan, K., and Barzilay, R. (2017). Unsupervised learning of morphological forests. *Transactions of the Association for Computational Linguistics*, 5:353–364.

Maamouri, M. and Bies, A. (2004). Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 2–9, Geneva, Switzerland. COLING.

Maamouri, M., Bies, A., Kulick, S., Tabessi, D., and Krouna, S. (2012). Egyptian Arabic Treebank DF Parts 1-8 V2.0 - LDC catalog numbers LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21.

MacWhinney, B. (1991). *The CHILDES language project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of child language*, 31(4):883.

Maratsos, M. (2000). More overregularizations after all: New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language*, 27(1):183–212.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., and Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, New York.

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

Mayol, L. (2007). Acquisition of irregular patterns in Spanish verbal morphology. In Nurmi, V. and Sustretov, D., editors, *Proceedings of the twelfth ESSLLI Student Session*, pages 1–11, Dublin.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3922–3931.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., et al. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244.

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

McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. *arXiv preprint arXiv:2005.08826*.

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019.

Miller, G. A. (1957). Some effects of intermittent silence. *The American journal of psychology*, 70(2):311–314.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Monson, C., Carbonell, J., Lavie, A., and Levin, L. (2007). Paramor: Finding paradigms across morphology. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 900–907. Springer.

Mooney, R. J. and Califf, M. E. (1995). Induction of first-order decision lists: Results on learning the past tense of English verbs. *Journal of Artificial Intelligence Research*, 3:1–24.

Mott, J., Bies, A., Strassel, S., Kodner, J., Richter, C., Xu, H., and Marcus, M. (2020). Morphological segmentation for low resource languages. In *International Conference on Language Resource and Evaluation (LREC)*.

Nagy, W. E. and Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, pages 304–330.

Narasimhan, K., Barzilay, R., and Jaakkola, T. (2015). An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. MIT press Cambridge, MA.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617.

O'Donnell, T., Snedeker, J., Tenenbaum, J., and Goodman, N. (2011). Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

O'Donnell, T. J. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. MIT Press, Cambridge, MA.

Oncina, J., García, P., and Vidal, E. (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):448–458.

Oseki, Y., Sudo, Y., Sakai, H., and Marantz, A. (2019). Inverting and modeling morphological inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 170–177.

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

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Payne, S. R., Kodner, J., and Yang, C. (2021). Learning morphological productivity as meaning-form mappings. *Proceedings of the Society for Computation in Linguistics*, 4(1):177–187.

Perfors, A., Tenenbaum, J., and Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37:607–642.

Pinker, S., Lebeaux, D. S., and Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, 26(3):195–267.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Pinker, S. and Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. *The reality of linguistic rules*, 321:51.

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

Plunkett, K. and Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23(4):463–490.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

Plunkett, K. and Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69.

Poeppel, D. and Wexler, K. (1993). The full competence hypothesis of clause structure in early German. *language*, pages 1–33.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*.

Prasada, S. and Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and cognitive processes*, 8(1):1–56.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, 54(9):975–997.

Rawski, J. and Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to Pater. *Language*, 95(1):e125–e135.

Roark, B., Sproat, R., and Sproat, R. W. (2007). *Computational approaches to morphology and syntax*, volume 4. Oxford University Press.

Rumelhart, D. E. and McClelland, J. L. (1986). *On learning the past tenses of English verbs*. MIT Press, Cambridge, MA.

Schone, P. and Jurafsky, D. (2001). Knowledge-free induction of inflectional morphologies. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–9.

Schuler, K. D. (2017). *The acquisition of productive rules in child and adult language learners*. PhD thesis, Georgetown University.

Schütze, C. T. (2005). Thinking about what we are asking speakers to do. In Kepser, S. and Reis, M., editors, *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 457–485. Mouton de Gruyter, Berlin.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Seidenberg, M. S. and Plaut, D. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive science*, 38 6:1190–228.

Sonnenstuhl, I. and Huth, A. (2002). Processing and representation of German-n plurals: A dual mechanism approach. *Brain and Language*, 81(1-3):276–290.

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

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19(3):513–531.

Stewart, T. W. (2015). *Contemporary morphological theories: A user's guide*. Edinburgh University Press.

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

Virpioja, S., Kohonen, O., and Lagus, K. (2009). Unsupervised morpheme discovery with allomorfessor. In *CLEF (Working Notes)*.

Wang, S., Zhou, W., and Jiang, C. (2020). A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

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

Xu, H., Kodner, J., Marcus, M., and Yang, C. (2020). Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681.

Xu, H., Marcus, M., Yang, C., and Ungar, L. (2018). Unsupervised morphology learning with statistical paradigms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 44–54.

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford University Press, Oxford.

Yang, C. (2013). Who's afraid of George Kingsley Zipf? or: Do children and chimps have language? *Significance*, 10(6):29–34.

Yang, C. (2016). *The Price of Linguistic Productivity*. MIT Press, Cambridge, MA.

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24(2):100–125.

Yang, C. (2020). Saussurean rhapsody: Systematicity and arbitrariness in language. In *The Oxford Handbook of the Lexicon*. Oxford University Press, USA.

Yarowsky, D. and Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 207–216.

Zaretsky, E. and Lange, B. P. (2015). No matter how hard we try: Still no default plural marker in nonce nouns in modern high German. In *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, pages 153–178.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranzabe, M. J., Arıcan, B. N., Arnardóttir, H̑., Arutie, G., Arwidarasti, J. N., Asahara, M., Aslan, D. B., Ateyah, L., Atmaca, F., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Balasubramani, K., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Barkarson, S., Basmov, V., Batchelor, C., Bauer, J., Bedir, S. T., Bengoetxea, K., Berk, G., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Bjarnadóttir, K., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Braggaar, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cassidy, L., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cesur, N., Cetin, S., Çetinoğlu, Ö., Chalub, F., Chauhan, S., Chi, E., Chika, T., Cho, Y., Choi, J., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Cristescu, M., Daniel, P., Davidson, E., de Marneffe, M.-C., de Paiva, V., Derin, M. O., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dinakaramani, A., Di Nuovo, E., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eiche, S., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Facundes, S., Farkas, R., Fernanda, M., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerardi, F. F., Gerdes, K., Ginter, F., Godoy, G., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grobol, L., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Güngör, T., Habash, N., Hafsteinsson, H., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Hanifmuti, M. Y., Hardwick, S., Harris, K., Haug, D., Heinecke, J., Hellwig, O., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Huber, E., Hwang, J., Ikeda, T., Ingason, A. K., Ion, R., Irimia, E., Ishola, Ọ., Ito, K., Jelínek, T., Jha, A., Johannsen, A., Jónsdóttir, H., Jørgensen, F., Juutinen, M., K, S., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Kara, N., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Köksal, A., Kopacewicz, K., Korkiakangas, T., Kotsyba, N., Kovalevskaitė, J., Krek, S., Krishnamurthy, P., Kuyrukçu, O., Kuzgun, A., Kwak, S., Laippala, V., Lam, L., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Levina, M., Li, C. Y., Li, J., Li, K., Li, Y., Lim, K., Lima Padovani, B., Lindén, K., Ljubešić, N., Loginova, O., Luthfi, A., Luukko, M., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Marşan, B., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsuda, H., Matsumoto, Y., Mazzei, A., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Mischenkova, K., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Mojiri Foroushani, A., Molnár, J., Moloodi, A., Montemagni, S., More, A., Moreno Romero, L., Moretti, G., Mori, K. S., Mori, S., Morioka, T., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Nakhlé, M., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nevaci, M., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nourian, A., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Onwuegbuzia, E., Osenova, P., Östling, R., Øvrelid, L., Özateş, Ş. B., Özçelik, M., Özgür, A., Öztürk Başaran, B., Park, H. H., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perkova, N., Perrier, G., Petrov, S., Petrova, D., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Rama, T., Ramasamy, L., Ramisch, C., Rashel, F., Rasooli, M. S., Ravishankar, V., Real, L., Rebeja, P., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Rögnvaldsson, E., Romanenko, M., Rosa, R., Roșca, V., Rovati, D., Rudina, O., Rueter, J., Rúnarsson, K., Sadde, S., Safari, P., Sagot, B., Sahala, A., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Sanıyar, E., Särg, D., Saulīte, B., Sawanakunanon, Y., Saxena, S., Scannell, K., Scarlata, S., Schneider, N., Schuster, S., Schwartz, L., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shishkina, Y., Shohibussirri, M., Sichinava, D., Siewert, J., Sigurðsson, E. F., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Skachedubova, M., Smith, A., Soares-Bastos, I., Spadine, C., Sprugnoli, R., Steingrímsson, S., Stella, A., Straka, M., Strickland, E., Strnadová, J., Suhr, A., Sulestio, Y. L., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tan, M. A. C., Tanaka, T., Tella, S., Tellier, I., Testori, M., Thomas, G., Torga, L., Toska, M., Trosterud, T., Trukhina, A., Tsarfaty, R., Türk, U., Tyers, F., Uematsu, S., Untilov, R., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van der Goot, R., Vanhove, M., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Vlasova, N., Wakasa, A., Wallenberg, J. C., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Widmer, P., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamashita, K., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yenice, A. B., Yıldız, O. T., Yu, Z., Žabokrtský, Z., Zahra, S., Zeldes, A., Zhu, H., Zhuravleva, A., and Ziane, R. (2021). Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.

Notes

[1]I am grateful to Caleb Belth, Spencer Caplan, Hossep Dolatian, Jeffrey Heinz, Mitch Marcus, Charles Yang, and OUP's anonymous reviewers for their feedback on this article.

[2]Not to be confused with the Brown University Standard Corpus of Present-Day American English (Kučera and Francis, 1967), a classic NLP data set also called "the Brown Corpus."

[3] `https://www.urbandictionary.com/define.php?term=meese`, and `https://www.urbandictionary.com/define.php?term=cabeese` (Accessed April 1, 2021). Urban Dictionary is a website that crowdsources definitions for slang.

[4]Novel or innovative child productions are often called "errors." There are some problems with this characterization. First, it is important to distinguish between performance errors and competence errors, but it is often difficult or impossible to tell them apart in children. A performance error is certainly a mistake, but competence error in this case would merely be difference between an adult's grammar and the child's. Second, variation is common in language, so there may be multiple learning targets presented to the child, and a production consistent with one adult's may nevertheless be an error relative to another. Which adult should they be compared against? Third, sparsity rears its head again. If the child has never heard some adult form, then from the child's perspective there is nothing to compare to in the first place. A child who has never encountered *forsook* as the past of *forsake*, for example, could not possibly guess the correct form, while an attempt at **forsaked*, which an over-regularization, indicates that the child does have a productive *-ed*.

[5]Note though that Swahili-speaking adults do rarely produce finite forms without subject agreement or tense marking, but at a much lower rate than the children. Such reduced forms might appear in a child's input, albeit rarely. It is unclear whether children internalize these or produce reduced forms *de novo*, but in any case, they are clearly not probability matching their input.

[6]Note that learning in the limit does not necessarily guarantee that learners will all eventually acquire identical grammars anyway since the space of human grammars includes many with identical extensions. Examples of this would have to be generally "asymptomatic," but there are some examples that can be detected through weak signals. One potential morpholollgical example comes from Guy and Boyd (1990), a sociolinguistic study which investigates the rate of T/D-deletion (final coronal obstruent lenition) of "semi-weak verbs." These are verbs whose pasts are irregular but do end in t/d (e.g., *tell-told*, *sleep-slept*. Since the rate of T/D-deletion differs between mono-morphemes and forms with regular past *-ed* it can be used as a diagnostic of underlying representations. In their investigation of adults, some deleted for semi-weak verbs at a rate similar to regular verbs and some at a rate more similar to mono-morphemes. Though it was not discussed in the original sociolinguistics paper, this suggests to a computational acquisition researcher that the speakers' representations of semi-weak verbs differ–some speakers segment them and some do not– and this is only uncovered by the indirect analysis. Speakers do not converge even after a lifetime of input.

[7]This conception of computational morphology, a mechanical characterization of surface patterns without reference to a cognitively realized grammar that generates them is reminiscent of classic Structuralist views on linguistic analysis. A more detailed discussion of this parallel is unfortunately too ambitious for the present article.

[8]`http://morpho.aalto.fi/events/morphochallenge/`

[9]The Porter Stemmer has been reimplemented countless times over the last four decades, and it and its variants are still in common use. Implementations in several programming languages, both popular and esoteric, can be found at `https://tartarus.org/martin/PorterStemmer/`

[10]On the other hand, the term in the sense of Baayen (Baayen, 1993; Baayen and Renouf, 1996, *et seq.*) might be described as "corpus" or "descriptive" productivity because it measures the tendency for a form or pattern to be generalized in the output and not in the speaker. Descriptive productivity is relevant to learning only inasmuch as it further clarifies distributions in the input.