Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute?

Micha Elsner[1], *Andrea D. Sims*[1], *Alexander Erdmann*[1], *Antonio Hernandez*[1], *Evan Jaffe*[1], *Lifeng Jin*[1], *Martha Booker Johnson*[1], *Shuan Karim*[1], *David L. King*[1], *Luana Lamberti Nunes*[2], *Byung-Doh Oh*[1], *Nathan Rasmussen*[1], *Cory Shain*[1], *Stephanie Antetomaso*[1], *Kendra V. Dickinson*[2], *Noah Diewald*[1], *Michelle McKenzie*[1], *and Symon Stevens-Guille*[1]

[1] Department of Linguistics, The Ohio State University

[2] Department of Spanish and Portuguese, The Ohio State University

ABSTRACT

We survey research using neural sequence-to-sequence models as computational models of morphological learning and learnability. We discuss their use in determining the predictability of inflectional exponents, in making predictions about language acquisition and in modeling language change. Finally, we make some proposals for future work in these areas.

Keywords: morphology, computational modeling, typology

1 INTRODUCTION

Theoretical morphologists have long appealed to notions of learning, or learnability, to explain language change and the varied typological patterns of the world's languages. The high-level argument is simple: all natural languages must be learned, and "unlearnable" linguistic systems cannot survive. Therefore, the learning mechanism provides constraints on what sorts of languages can exist in the world. In the realm of morphology, however, it has not proven simple to define

learnability (or, as it is often described, *morphological complexity*). Different theories offer different ideas of what must be learned in order to acquire a morphological system, and how to measure the difficulty of the learning problem for a particular language.

In doing so, they have sometimes used computational models of the learner to buttress their claims. In many cases, their tools for model-building draw on the rich tradition of morphological processing within computational linguistics. Computational linguists construct models of morphology not only as direct contributions to linguistic research, but as engineering solutions to the low token/type ratios of languages with large inflectional paradigms; such models have been applied to language generation, machine translation, and other tasks. In recent years, a particular model from the machine translation community, the *neural sequence-to-sequence* model, has grown in popularity for morphological tasks. Sequence-to-sequence models are now being applied, not only as engineering solutions, but also as theoretically interesting models of morphological complexity.

This paper provides an overview of both theoretical and computational work in this framework. Beginning with an overview of morphological complexity, and the different proposals for how it can be measured, we show that sequence-to-sequence models are a natural fit for the Word and Paradigm model and its notion of Integrative Complexity. We present some criticisms of previous implementations of Integrative Complexity, and explain how sequence-to-sequence modeling has already begun to address them. However, we spotlight several areas where the framework, as currently conceived, falls short. We go on to describe some important open questions to which it might be applied in the future.

2 THEORETICAL FOUNDATIONS

Work in computational morphology has often been concerned with engineering questions, rather than with modeling speakers' morphological knowledge. Whether recent computational models can be profitably applied to theoretical questions thus depends on the extent to which the structure of a model reflects principles of morphological theory. To see the issues at hand, we begin with an overview of two

theoretical positions, the Item-and-Arrangement (IA) family of theories, and the Word-and-Paradigm (WP) family (names proposed by Hockett (1954); see Blevins (2016) for a historical overview). These contrasting positions set forth different concepts of what morphological complexity is, and how it might shape morphological typology.

2.1 *Learnability and typology in morpheme-based models*

IA models take the *morpheme* as the fundamental unit of analysis,[1] and describe inflectional systems in terms of their syntagmatic structure – that is, the associations between stems, affixes and meanings. For models built upon this assumption, it is natural to define the language learner's task as acquisition of the morpheme inventory and the syntagmatic rules for composing morphemes into words. This, in turn, tends to lead to a focus on the size of morphological systems, what Ackerman and Malouf (2013) call a language's 'enumerative complexity' (E-complexity). Quantitative measurement of this kind of morphological complexity has a long history, going back to Greenberg (1960).

One typological generalization that has been approached from an IA perspective is that languages tend to have far fewer inflection classes than they could, given their number of allomorphs. Table 1 shows a simplified example from Icelandic. Two allomorphs are shown

[1] Traditional IA models define morphemes as lexical bundles of minimal form and minimal meaning. This is consistent with the principle of 'incrementalism' (Stump 2001), according to which concatenating a morpheme to a stem adds the morpheme's form to the word and simultaneously adds its meaning, with meaning broadly construed to include morphosyntactic and morphosemantic values. The meaning of a word should thus be fully determined by the meanings of its parts plus their order of combination. However, as Blevins (2013: 436) points out, "…ideas tend to outlive the traditions that initially hosted them and mutate during their own lifespans". Incrementalism has proven too restrictive, and starting in the early 1990's (Anderson 1992; Halle and Marantz 1993) it was largely replaced by 'realizationalism', which postulates that operations on form, such as concatenating an affix to a stem, are licensed by the morphosyntactic properties of a word. Formal operations thus *realize* the meaning of a word rather than adding to a word's meaning. Some modern theories, such as Distributed Morphology (Halle and Marantz 1993; Harley and Noyer 2003), adopt realizationalism while retaining other IA/morpheme-based assumptions, such as the primary importance of concatenative operations and syntagmatic (stem-affix) relations in morphological structure. These theories are 'lexical-realizational' in the terminology of Stump (2001).

Table 1: Select inflected forms of Icelandic verbs: three paradigm cells with two allomorphs each. 1 and 2 are logically possible but unattested classes

	GRÍPA	KALLA	*1	*2
	‘grasp’	‘shout’		
1SG.PST	greip	kall-aði	X	X-aði
2SG.PST	greip-st	kall-aðir	X-aðir	X-st
3SG.PST	greip	kall-aði	X	X-aði

for each of three paradigm cells. (The Icelandic verb system has both more paradigm cells and more classes, but these few forms are sufficient for illustration.) Based on the forms shown, there could mathematically be as many as $2\times2\times2 = 8$ inflection classes, if the allomorphs of different paradigm cells were independent of each other. And as the number of allomorphs and paradigm cells grows, the number of possible classes – and thus the potential E-complexity of the inflectional system – increases rapidly. Yet allomorphs tend not to be independent of each other: this is why it is useful to talk about inflection *classes*. Indeed, in Icelandic verbs the 1SG.PST zero allomorph (as in GRÍPA) is never found in the same paradigm as the 2SG.PST allomorph *-aðir*. Likewise, 1SG.PST *-aði* (as in KALLA) is never found with 2SG.PST *-st*, and so on. While eight classes are potentiated by these allomorphs, the shown allomorphs in fact group into two classes – the minimum possible number. Moreover, this is representative of a strong tendency cross-linguistically for the actual number of classes observed in a language to be far fewer than the mathematically possible number of classes (Carstairs 1987).[2] This raises the question: Why?

A number of morphological theories attempt to explain this and other constraints by appealing to learnability. In an IA framework, there is often an assumption that more inflection classes and larger paradigms make languages more difficult to learn, and that inductive learning biases must therefore serve to constrain the learner’s hypothesis space. Perhaps most famously, the No Blur Principle (Carstairs-McCarthy 1994), later revised as Vocabular Clarity (Carstairs-McCarthy 2010), posits that for each paradigm cell, only one allomorph can “...fail to identify inflection class unambiguously” (1994:742). In other words, there can be only one ‘default’ (class-unspecified) form for each paradigm cell. This proposed constraint is

[2] Apparent exceptions include Burmeso and Nuer (Baerman 2012; Baerman *et al.* 2017).

rooted in the idea that learning biases must serve to constrain learners' hypotheses about allomorph distributions. Specifically, Carstairs-McCarthy views No Blur as a byproduct of the Principle of Contrast (Clark 1987), the idea that in lexical acquisition, children are biased towards hypothesizing that a difference in word form corresponds to a difference in word meaning. Extending the Principle of Contrast from words to inflectional allomorphs, Carstairs-McCarthy defines inflection class membership as meaning in the relevant sense. He proposes that the observed cross-linguistic restriction on the proliferation of inflection classes is indirectly caused by an inductive bias that pushes child language learners towards positing that each suppletive allomorph either belongs to a different inflection class, or does not bear inflection class meaning. From an IA perspective there is logic to this extension, since morphemes (including any suppletive allomorphs) are taken to be the units of storage in the lexicon – the level at which form and meaning are related.

Important here are ways in which Carstairs-McCarthy's assumptions about the nature of morphological knowledge shape his conceptualization of the relationship between learning and inflectional typology. The first thing to observe is that Carstairs-McCarthy posits a fundamental distinction between concatenative and non-concatenative morphological processes; No Blur applies only to concatenative allomorphs. While this distinction is motivated theory-internally in IA models, it has no clear independent motivation (Stump 2001). Minimally, this raises questions about why the Principle of Contrast should constrain learners' hypotheses about allomorph distributions *only* for concatenative morphology. We know of no empirical evidence supporting this assumption.

Second, Halle and Marantz (2008) point out that cross-linguistically, inflection classes often group hierarchically into macroclasses, yet such distributions virtually require individual allomorphs to belong to multiple classes. They focus on the empirical problems that such patterns create for No Blur, but even if we set these aside, we can observe that Carstairs-McCarthy's hypothesis about how learning might shape morphological typology reflects an IA emphasis on morphemes as isolable form-meaning units and the syntagmatic (stem-affix) dimension of structure. It seems to imply that paradigmatic relationships among words/classes (e.g. whether they fully separate,

hierarchically grouped, cross-classifying, etc.[3]) are irrelevant to questions of learnability, beyond what is dictated by No Blur. However, this is an open question.

Finally, although No Blur does not place an absolute limit on the number of inflection classes in a language,[4] it will generally predict that the actual number of classes in a language is substantially smaller than the potential number of its classes, capturing the cross-linguistic tendency observed above.[5] Since allomorphs realizing different paradigm cells must be class-specific for the most part, No Blur indirectly captures the grouping of allomorphs into classes as a byproduct of the learner's acquisition of the morpheme inventory.

Blevins (2004) and Ackerman and Malouf (2015) argue that No Blur is derivative of the paradigmatic structure of inflectional morphology, and paradigmatically-structured learning of morphology. We turn to this perspective in the following section.

2.2 *Learnability and typology in word-based models*

Word-and-Paradigm morphology offers an alternative link between learnability and inflectional typology. WP is in many respects the oldest framework for inflectional theory, reflected in the traditional pedagogical approach to describing classical languages' inflectional systems in terms of their principal parts. (A lexeme's principal parts are those inflected forms that together suffice to deduce all of the lexeme's inflected forms, i.e. its full paradigm of surface word-forms.) Such models take the word as the basic unit of morphological structure and analyze inflectional meaning as being instantiated via paradigmatic contrasts – that is, contrasts between the forms filling different inflectional cells.[6] The learner's task, therefore, is to understand the relationships between the forms of each lexeme. We focus here on

[3] For further discussion from the typological perspective, see Dressler *et al.* (2006) and Brown and Hippisley (2012).

[4] Unlike its predecessor, Paradigm Economy (Carstairs 1987), which directly defines constraints on the number of classes that a language can have.

[5] Except that there are languages that violate the various formulations of the constraint, whether stated as Paradigm Economy, No Blur, or Vocabular Clarity (Halle and Marantz 2008; Müller 2007; Þorgeirsson 2017; Stump and Finkel 2013).

[6] WP models differ in the extent to which the abstract concept of the paradigm is considered to be a metaphor, emergent structure, or a reified theoretical primi-

the *abstractive* WP framework, in which the key relationships are directly between surface forms.[7] Abstractive models take as their starting point the idea that the inflected forms of a lexeme are interpredictable to some (potentially substantial) degree. Or, as Wurzel (1989: 114) stated it, "...inflectional paradigms are, as it were, kept together by implications. There are no paradigms (except highly extreme cases of suppletion) that are not based on implications valid beyond the individual word, so that we are quite justified in saying that inflectional paradigms generally have an implicative structure."

Importantly, in WP models there is no requirement that paradigm cells be realized by some segmentable phonological form (a classical morpheme); cells can also be realized by non-concatenative morphological operations that alter the phonological form of the stem. These operations can encompass, for instance, root-and-pattern morphology in Semitic languages, tonal morphology in Bantu languages, and German ablaut. Moreover, WP models make no explicit or implicit assumptions that there should be a one-to-one correspondence between morphological form and meaning. They accommodate insertion of material (exponents) with no obvious meaning,[8] multiple exponence, in which a meaning is signaled by multiple morphological pieces, and zero exponence, in which there is no phonological change corresponding to a change in meaning. WP models are thus "inferential-

tive and direct object of study. We touch on this interesting question in Section 6, in the context of what we call the Paradigm Cell Discovery Problem.

[7] As with IA models, modern WP models differ in many respects from classical WP models, reflecting in part the adoption of goals and principles from modern generativism (Blevins 2016; Matthews 1972). Importantly here, modern WP models can be divided into *constructive* and *abstractive* types (Blevins 2006). Constructive models (Anderson 1992; Stump 2001) characterize the morphological structure of a word in terms of form operations applied to lexically-stored stems to produce surface inflected forms. In contrast, abstractive models (Blevins 2016; Bochner 1993; Albright 2002a) describe the morphological structure of a word in terms of form operations applied to one or more surface word-forms to produce another.

[8] For example, verbal inflection classes in many Indo-European languages are organized around so-called 'theme vowels' (e.g. [a] vs. [e] vs. [i] in Spanish *am-a-r* 'love-TV-INF', *ten-e-r*, 'have-TV-INF', and *part-i-r* 'depart-TV-INF'), which serve to mark the inflection class of the verb but do not bear any syntactically- or semantically-relevant meaning.

Table 2: Indicative present forms of two Icelandic verb classes. Some allomorphs are the same in both classes

	GRÍPA	KALLA
	'grasp'	'shout'
1SG.PRS	gríp	kall-a
2SG.PRS	gríp-ur	kall-ar
3SG.PRS	gríp-ur	kall-ar
1PL.PRS	gríp-um	köll-um
2PL.PRS	gríp-ið	kall-ið
3PL.PRS	gríp-a	kall-a

realizational" (Stump 2001), also sometimes called "a-morphous" (Anderson 1992).

Given its postulation that surface forms serve as the bases for other surface forms, abstractive WP models must contend with the question of how hard it is for speakers to predict an unobserved surface word-form for a lexeme, given some other word-form(s) in the paradigm; this is the Paradigm Cell Filling Problem (PCFP) (Ackerman *et al.* 2009). For illustration, we return to the simplified Icelandic example introduced earlier. Table 2 shows that while the two verbs represent different classes, some allomorphs are the same for both. If a speaker encounters a new verb in the 3PL.PRS with allomorph *-a*, the distribution of allomorphs engenders some amount of uncertainty regarding what some of the present and past tense forms of the verb are. (For the latter, see Table 1 in the previous section.) When a word is subject to the PCFP, there may thus be some amount of ambiguity regarding its inflection class membership. Morphological complexity, then, is taken to be the difficulty of this problem for a speaker or learner of some particular language.

As noted above, in an IA framework the complexity of an inflectional system tends to be conceptualized in terms of the size of its paradigms and the number of its classes (i.e. its E-complexity). However, we know of no clear evidence that languages with high E-complexity are more challenging to learn. There are many clear examples of natural languages with large paradigms – Kibrik (1998) famously observed that in principle, verbs in the Nakh-Daghestanian language Archi can have more than 1.5 million forms each. Moreover, historical change may increase, rather than reduce, the E-complexity of an inflectional system, even though we might predict that this renders languages less learnable. To give just one example, in the Iranian language Zazaki, phonological and syntactic competition among ezafe

forms has resulted in the development of a complex system of nominal inflection (with upwards of 144 paradigm cells) that includes rampant fusionality and syncretism. This stands in stark contrast with closely related languages in which a more agglutinative system can still be observed (Karim 2019). It is admittedly harder to provide example languages with large numbers of inflection classes, since the number of classes identified for an inflectional system depends heavily on analytic assumptions (Parker 2016), rather than being directly empirically observable. But such examples have certainly been proposed: for example, as many as 115 classes of Russian nouns (Parker 2016). Ultimately, these arguments raise doubts about whether metrics based on E-complexity are directly related to the learnability of morphological systems.

However, there is no particular reason that an E-complex language should have a difficult PCFP; it is not the *number* of forms that matters but their *predictability*. Thus, the WP model offers a different formulation of complexity, which Ackerman and Malouf (2013) term "Integrative complexity" (I-complexity). As the number of allomorphs in an inflectional system grows, the potential I-complexity of the system grows, but to the extent that the inflected forms of lexemes are interpredictable, it is possible for the actual I-complexity to remain low (Ackerman and Malouf 2013). Studies have attempted to measure the complexity in this sense for various languages' inflectional systems using set-theoretic (Stump and Finkel 2013) and information-theoretic (Ackerman and Malouf 2013; Stump and Finkel 2013; Bonami and Beniamine 2016; Sims and Parker 2016; Cotterell *et al.* 2018a) measurements.[9] These studies have generally focused on the role of abstractive paradigmatic relations – conceptually, proportional analogy – in solving the PCFP.

Like IA models, WP models have attempted to explain typological patterns by appealing to learnability, and in WP frameworks the PCFP has been intimately connected to typological questions. In particular, abstractive WP models propose that the cross-linguistic restriction on

[9] A separate line of investigation has found that information-theoretic measurements of inflectional paradigmatic relations predict speakers' response times in lexical decision tasks (Milin *et al.* 2009; Moscoso del Prado Martín *et al.* 2004), suggesting the relevance of abstractive-type relations to morphological processing.

the proliferation of inflection classes that we noted in the preceding section is caused by the need for learners to be able to solve the PCFP.

Ackerman and Malouf (2013) estimate the average difficulty of the PCFP in a language by its average conditional entropy: abstracting away from stems, they calculate the average unpredictability of the exponent (affix) realizing one paradigm cell, given the exponent for another paradigm cell of the same lexeme. This implementation of proportional analogy does not capture any predictiveness that derives from similarity among whole words, an issue that we return to below, but it does offer a quantification of how difficult it is to predict inflectional exponents from other exponents (for example, Icelandic 2SG : *-ir* :: 3SG : *-i*). Finding low average conditional entropy in each of ten languages (generally, less than 1 bit, equivalent to or better than a coin toss), they conclude that the PCFP is not exceptionally difficult in these languages and propose a typological universal, the Low Entropy Conjecture: "...enumerative morphological complexity is effectively unrestricted, as long as the average conditional entropy, a measure of integrative complexity, is low..." (436). Their conclusion is consistent with that of Stump and Finkel (2013), who find that an inflectional system's average cell predictor number – a set-theoretic measure of the number of dynamic principal parts required to determine the inflected form corresponding to a given paradigm cell – tends to be low.[10]

Ackerman and Malouf connect the PCFP to the Low Entropy Conjecture via learnability:

> If low entropy is the correct measure for explaining the implicational organization of paradigms, rendering complex systems learnable, then this makes a prediction about types of systems that we do not find… [T]here are no known fully suppletive systems in the languages of the world. This absence is easily explicable given the low entropy conjecture and its facilitating function for learnability: a completely suppletive system is one in which no form bears an implicational

[10] In dynamic principal parts analysis, the paradigm cells identified as principal parts need not be the same from one inflection class to another. This contrasts with static principal parts analysis, in which the cells that function as principal parts are identified for an inflection class system as a whole, rather than on a class-by-class basis.

> relation with any other form, and thus there is no useful patterned organization reflective of low entropy. (Ackerman and Malouf 2013: 454)

The logic here seems to be that inflected forms that are not learnable on the basis of implicative relations are likely to be subject to analogical changes that make them more predictable.

Moreover, word tokens have Zipfian distribution, so most inflected forms of most lexemes are sparsely attested in both adult speech (Baayen 2001) and child-directed speech (Lignos and Yang 2018). This suggests that speakers of morphologically rich languages encounter the PCFP on a regular basis and throughout their lifetimes (Bonami and Beniamine 2016). From the perspective of a WP theory, we might posit that token frequencies of word-forms play an important role in defining which lexemes and paradigm cells are subject to the PCFP. Words that are of sufficiently high frequency as to be directly stored and retrieved from memory are not subject to the PCFP (although, as Bybee (1995) observes, memory is not a dictionary; speakers may use relationships within the lexicon to aid retrieval even of word forms for which they have extensive experience). If we assume a memory-rich model of word storage, abstractive WP models thus seem to predict that learnability will have the potential to enforce low I-complexity only when words cannot be retrieved directly from memory and are therefore subject to the PCFP.

Taken together, these distributional facts raise further questions about the connection between the learnability of individual forms and the I-complexity of inflectional systems. The model of morphological knowledge that is assumed by abstractive WP models, combined with the Zipfian distribution of word tokens, predicts that the PCFP should be more challenging in some languages than others, with implications for morphological typology. Recent work by Cotterell *et al.* (2018a) moves in the direction of exploring these implications. Based on calculations for thirty-one languages, they argue that inflectional systems may be high in either E-complexity or I-complexity, but not both. The presumption is that in inflectional systems with small paradigms, each paradigm cell will be attested more often on average, as compared to an inflectional system with large paradigms, all else being equal. Essentially, the same amount of semantic 'space' (and thus, presumably, usage) is being divided among more inflected forms in the lat-

ter. This leads to cross-linguistic differences in the extent to which the PCFP presents challenges in learning. We might expect fewer constraints on I-complexity in languages where the PCFP presents less of a challenge (i.e. languages with small paradigms, and thus low E-complexity). Cotterell *et al.* argue that this prediction is borne out in the data.

Abstractive WP models also predict language-internal differences: some subparts of an inflectional system may be more challenging for the PCFP than others. This suggests the need for fine-grained measures of learnability, as well as more nuanced hypotheses about how learnability might shape morphological typology.

3 THE NEED FOR FINE-GRAINED MEASURES OF LEARNABILITY

As discussed above, I-complexity (the difficulty of the PCFP in a particular language) has been proposed as a measure of morphological learnability. Average conditional entropy is often taken as a formal model of I-complexity. In turn, average conditional entropy is often computed by segmenting word forms into two substrings: a 'stem' and an 'exponent' or 'affix',[11] and applying four-part morphological analogy based on pairs of exponents (Ackerman and Malouf 2013). For example, Icelandic 2SG *-ir* implies 3SG *-i*. (In Section 3.2 we note differences between this notion of proportional analogy and how the concept is often employed in historical linguistics.) However, both the choice of average entropy, and calculating entropy based only on pairs of exponents, have been criticized. In this section, we argue that continued progress towards understanding learnability-based constraints will require a more fine-grained understanding of how inflectional distributions are learned, because the difficulty of predicting a lexeme's entire paradigm (i.e. learning how to use it as a speaker) is not a direct function of the difficulty of predicting an individual form (the PCFP). Moreover, looking only at exponents can miss regularities in

[11] The segmentation process is often implemented using computational string alignment; see Beniamine *et al.* (2018) for a discussion. Stump and Finkel (2013) prefer to call the lexically-specific part of the form the 'theme' and the inflectionally-specific part the 'distinguisher', since these may not correspond to linguistically justified morphological analyses.

some inflectional systems which we believe human learners are able to exploit. Section 4 discusses how some of these troublesome cases can be addressed with more sophisticated computational models.

3.1 *Criticisms of averaging*

Averaging can conceal large differences in the predictability of individual cells; the average does not indicate whether every pairwise PCFP in a given language is somewhat unpredictable, or whether some PCFPs are very easy while others are very difficult. Thus, average conditional entropy can group together languages for which the underlying network of predictability relationships between cells are in fact quite different. There is some evidence that indeed, the typology of cell-to-cell predictability values is quite diverse: Stump and Finkel (2013) find that there is substantial cross-linguistic variation in the number of dynamic principal parts required to predict *all* inflected forms of lexemes, i.e. full paradigms, in contrast to the relatively uniform number of dynamic principal parts needed to predict a single inflected form. While this does not invalidate the average as an overall measure of morphological complexity, it does call for more sophisticated tools which can distinguish between these different kinds of systems.

Systems characterized by recurring partials are one example in which the individual PCFPs have predictability values far from the average. Recurring partials are groups of cells that divide the paradigm into implicatively coherent subsets. Within each subset, cells predict one another especially well, but they predict cells outside their subset poorly. The more deeply these subsets divide the paradigm, the more principal parts will be required to reproduce it.[12] Averaging the pairwise values suggests that the PCFP in these systems has an intermedi-

[12] In the extreme case, where interprediction is perfect within subsets and impossible between them, predicting the full paradigm requires exactly as many principal parts as there are subsets. Large numbers of principal parts may also be required to describe languages with cross-classifying inflection class subsystems, such as Chiquihuitlán Mazatec (Jamieson 1982), Russian (Brown *et al.* 1996) and Greek (Sims 2006). In such languages, different dimensions of inflectional exponence (e.g. suffixes, inflectional stress, stem extensions) vary semi-independently of one another, so that forms which predict one inflectional dimension may not be sufficient to predict another.

ate difficulty, but this is deceptive; each individual pairwise decision is either very easy, or very hard. Cotterell *et al.* (2018a) level the same criticism from a mathematical point of view, observing that averaging the pairwise entropy values does not compute the joint entropy of the distribution but generally yields an overestimate.

Even in cases where averaging across paradigm cells yields a good description of the system's difficulty, averaging across words, or classes of words, may not. Words with high token frequency are more likely to be irregular (Bybee 2003; Corbett *et al.* 2001) – that is, they belong to inflection classes with relatively few members (low type frequency). Stump and Finkel (2013) propose that these irregulars contribute disproportionately to the difficulty of the PCFP, the so-called Marginal Detraction Hypothesis. This property holds for a variety of languages, although seemingly not universally (Sims and Parker 2016). These classes expose a trade-off between predictability and predictiveness; exponents that are unpredictable by virtue of being irregular (and thus associated with a specific class) tend to be highly predictive of the other inflected forms of the same lexeme by virtue of this same fact (Finkel and Stump 2009). Again, the average is deceptive, failing to distinguish systems with a few highly irregular classes from systems in which every word is slightly unpredictable.

3.2 *Criticisms of simplistic implementations of morphological analogy*

We now turn to criticism of four-part morphological analogy based only on pairs of exponents as a measure of predictability. We start by observing that this type of "analogy" is not quite the same concept as analogy in historical linguistics (Hock and Joseph 1996: 10) or exemplar models (Skousen 1989). The issue has to do with abstracting away from stems and modeling only the relationship among exponents. Analogical inflectional change is not always sensitive to similarities between whole word forms,[13] but sometimes it clearly is. It

[13] Hock (1991: 172) suggests that the spread of English plural *-s*, e.g. *kine* to *cows*, should be considered a result of proportional analogy on the model of, e.g., *stone-stones*. This extension of *-s* to new words was not dependent on the overall phonological similarity of the words which gained plural *-s* to existing words with *-s*. Regularizations of this sort have sometimes been treated as simplification of the rule system (Kiparsky 1968), as something distinct from analogical change, but a distinction between rule-based change and analogical change is not even

is well known that in English, some classes of irregular verbs have attracted the occasional new member based on whole word similarity (e.g. the historically weak verb *string* changed to the *string-strung* pattern on the model of *swing-swung, sting-stung, sling-slung*, etc.). In historical linguistics (also in exemplar models), analogy is thus generally conceptualized as based on relationships among whole words.

The analogical computations of Ackerman and Malouf (2013) rely only on similarity between pairs of exponents (distinguishers), without taking stems (themes) into account. While tractable methodologically, this leads to a number of issues. We present cases in which inflectional forms are predictable, or partly predictable, on the basis of whole-word information. Importantly, the issue is not just that morphological analogy can overestimate the difficulty of the PCFP, but that the overestimation problem is likely to be larger for some languages than for others, so that the overly simplistic implementation based on pairs of exponents give an unrealistic description of the typological space. Baerman (2014) suggests that there is a typologically interesting class of languages in which information beyond what is captured by this narrow notion of morphological analogy contributes heavily to determining exponence; such a conjecture is difficult to test at a large scale without a model which is capable of exploiting these regularities as it learns to predict inflectional forms.

Several previous studies have emphasized the importance of stem information to predicting exponence in particular inflectional systems. Verb conjugation class membership in Italian (Albright 2002b) and English past tense verb forms (Albright and Hayes 2002; Bybee and Moder 1983; Rumelhart and McClelland 1986) are included among many other cases. It is thus necessary to attend to the syntagmatic dimension when modeling the predictability of inflectional exponence. In fact, Baerman (2014) argues that in Võro, a variety of Estonian, inflectional exponents are predictable *predominantly* from stem shape (specifically, how stem alternants are distributed in the paradigm), and that the exponents of other inflected forms of the same lexeme are uninformative. Morphological stem shape can also matter: the inflection class that a lexeme belongs to may be predictable from its

---

possible in some theoretical frameworks and we see no good motivation for it in this case.

derivational morphology. For example, in Croatian, abstract nouns derived from adjectives with *-ost* always belong to the feminine Class III, even though nouns whose stems end in a consonant normally fall into the masculine Class I. Capturing this syntagmatic dimension of predictability requires a model to be sensitive to stem shape.

Another property not captured by analogy is the re-use of affixes in different parts of the paradigm: in Kashmiri, the same set of suffixes express the remote past in one inflection class and the recent past in another (Stump and Finkel 2015). This is one example of what Baerman *et al.* (2017) call 'distributional' systems, in which inflection class distinctions are instantiated not by different exponents themselves, but by the distributions of exponents among paradigm cells. Analogy-based entropy calculations treat the different paradigm cells as separate random variables; the PCFP becomes harder when two classes share the same affix in the same paradigm cell (since this makes it harder to predict the realizations of other cells). Occurrences of the same affix in *different* cells are not modeled, either as a source of potential confusion or a regularity which the learning mechanism can exploit.

Finally, some systems have predictable relationships between cells, even where the content of those cells is unpredictable. For instance, a small number of Croatian nouns, for instance JAJE 'egg', have weak stem suppletion in oblique singular cases but not direct singular cases (*jaj-e* 'egg-NOM.SG' but *jajet-u* 'egg-DAT.SG'). This arguably makes the PCFP easier for learners of Croatian – faced with a new noun, they may be unable to guess whether it is (weakly or strongly) suppletive or what its suppletive stem may be. However, there are no Croatian nouns where a suppletive stem applies to a miscellaneous collection of singular and plural cells. So learners can predict to some extent the set of cells in which suppletive forms are allowed to appear.

To sum up, predictability based on simplistic implementations of morphological analogy captures only the interpredictability of exponents, not whether the interpredictable forms are coherent in any sense. To us, it seems unlikely that learnability constraints operate at the level of inflectional systems as a whole, even though this is the level at which the Low Entropy Conjecture and similar proposals have been formulated. It seems more likely that any upper bound on how

complex inflectional systems can be is an emergent property that derives (at least partly) from the learnability of individual inflected forms in the context of their local relationships to other inflected forms. The distributional patterns highlighted in this section identify deficiencies with previous single-measure estimates of I-complexity, and show the need for a more fine-grained approach. Sophisticated tools are therefore needed to model acquisition at this fine-grained level, and to explore its implications for morphological typology.

4 COMPUTATIONAL TASKS AND METHODS

Studies of the PCFP, whether using simple or sophisticated tools, are inherently statistical in nature; their conclusions depend on the data and on how well the data can be modeled. Thus, the advent of larger datasets and better systems for computational morphology are well-placed to make theoretical contributions to this field of study. In the next section, we discuss "morphological reinflection" as a computational formalization of morphological predictability. This approach is theoretically underpinned by abstractive WP models, making it well suited to investigation of the PCFP and the typological questions that stem from it. We summarize arguments that this formalization captures forms of predictability which are accessible to human learners, but were not measured by previous formal models such as that of Ackerman and Malouf (2013). By changing the way we estimate predictability to more closely conform to the human learner, we have the potential to change our current understanding of the morphological typology of the world's languages and its relationship to learnability.

In addition to refining our estimates of morphological complexity, we discuss ways in which computational models can be used to more precisely locate potential sources of learning difficulty within inflectional morphology. In other words, the models can tell us not only whether a particular *system* is easy to learn, but which *forms*, *classes* or other elements of the system contribute to its difficulty, and what errors we might expect an imperfect learner to make in acquiring them. We discuss these issues in the following sections.

4.1 *Reinflection with sequence-to-sequence models*

(Re)inflection tasks[14] involve converting one surface inflected word-form into a target form of the same lexeme, as illustrated in Table 3 for the German noun *Aak* (a kind of boat) and verb *aalen* 'to hunt eels, to relax' (Cotterell *et al.* 2016). The morphosyntactic values of the target are known, so reinflection amounts to predicting an inflected form for a known paradigm cell, given another known form. The task is thus equivalent to the PCFP and fits well into the theoretical framework of WP morphology. Of course, this equivalence comes with a few methodological caveats: most reinflection models are trained on orthographic, rather than phonemically transcribed data. And the datasets used are traditionally word lists, which do not reflect the token or type frequencies of the natural language, an issue we discuss in detail in Section 5. The ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON) has sponsored a series of such tasks for shared use (Cotterell *et al.* 2016, 2017, 2018c), motivating the recent development of highly effective reinflection systems.

Table 3: Example German reinflection problems from SIGMORPHON 2016 tasks 1 and 3. In task 1, above the line, the input form is a citation form; in task 3, below, the input form is arbitrary

Input	Target features	Target
Aak	pos = N, case = NOM, gen = NEUT, num = SG	Aak
aalen	pos = V, tense = PST	geaalt
Aakes	pos = N, case = NOM, gen = NEUT, num = SG	Aak
aaltet	pos = V, tense = PST	geaalt

Broadly speaking, these systems fall into the machine learning framework of sequence-to-sequence modeling. Such models map one sequence of tokens to another sequence, which is potentially longer or shorter. The source and target tokens need not line up one-to-one, a useful property (perhaps an indispensable one) for modeling morphology. Older reinflection models in the sequence-to-sequence framework

[14] Properly speaking, "inflection" suggests the prediction of variable target forms from a fixed base, as in SIGMORPHON task 1, while "reinflection" suggests the prediction of one arbitrary form from another, as in tasks 2 and 3. For convenience, we discuss both task settings as "reinflection".

operated by inducing string edit rules (Durrett and DeNero 2013; Albright 2002b) or transductions (Nicolai *et al.* 2015). Current models use a *neural network* sequence-to-sequence (encoder-decoder) framework, which was devised for machine translation (Sutskever *et al.* 2014; Bahdanau *et al.* 2014), but which later also proved capable of learning inflectional morphology (Faruqui *et al.* 2016; Kann and Schütze 2016; Aharoni and Goldberg 2017; Malouf 2017).

Sequence-to-sequence models can be thought of as relying on whole-word analogy, just as exemplar models (Skousen 1989) do. But unlike traditional exemplar models, they induce their own, implicit, similarity function between examples. The models project the input form and featural specification of the desired output into a latent space described by a set of numerical features. The space is "latent" in that its features have no pre-specified interpretations, but reflect whatever information about the input the system finds most useful for producing the correct output. This space defines the implicit similarity metric, by clustering related inputs near one another. From its latent representation, the system can extract the output character by character. This architecture does not enforce a clear separation between stems and affixes, nor between phonological and morphological conditioning; because the output is produced character-by-character, rather than assembled from concatenated pieces, sequence-to-sequence models can learn to capture dependencies between each output character, the preceding outputs, and the *entire* input string.

This representational flexibility makes neural sequence-to-sequence models appealing as formal models of the WP morphological framework. Because the latent representation does not make implicit assumptions about morphemes as one-to-one form-meaning mappings, they are capable of learning nonconcatenative morphological processes, as discussed in Section 2.2. Cotterell *et al.* (2018a) point this out as an advantage of the "a-morphous" sequence-to-sequence framework over approaches relying on morpheme segmentation. Sequence-to-sequence models are also capable of learning stem-affix relationships, both morphological and phonological, as discussed in Section 3. Faruqui *et al.* (2016) illustrates this for the case of Finnish vowel harmony (see also Corkery *et al.* 2019); many earlier models with explicit morpheme segmentation were forced to represent this process as suppletive allomorphy (for example, positing the two phonological

variants of the inessive suffix, *-ssa* and *-ssä*, as suppletive allomorphs), which could lead to overassessment of the system's complexity (Stump and Finkel 2015). But the sequence-to-sequence model learns a generalizable harmony rule. Similarly, sequence-to-sequence models can learn a fully general process of reduplication (Prickett *et al.* 2018) where earlier models would be forced to memorize a separate rule for each reduplicating substring.

Kirov and Cotterell (2018) argue that these models remedy many of the shortcomings of earlier neural networks for inflection (Rumelhart and McClelland 1986), which were criticized (Pinker and Prince 1988; Pirrelli *et al.* 2015; Lignos and Yang 2018) both for inaccurate predictions, and for using representations which obscured the sequential nature of the input and were thus incapable of learning common typological patterns such as position class systems.

Due to these advantages, sequence-to-sequence models have been applied to several theoretical questions. Cotterell *et al.* (2018a) use them to measure I-complexity without relying on simplistic variants of four-part morphological analogy. Cotterell *et al.* (2018b) use them to simulate acquisition-based change in predicting the regularization of English past tense verbs. In doing so, projects like these implicitly assume that the SIGMORPHON reinflection task can be treated as a model of morphological acquisition and the sequence-to-sequence neural net as a model of the learner.

Thinking of sequence-to-sequence models in this way opens up several questions which deserve further attention. First is the issue of qualitative evaluation: which parts of the morphological system can the model acquire, and which does it struggle with? Second is the interpretation of the reinflection task as a model of acquisition: what sorts of datasets are appropriate and what sorts of task settings are realistic as representations of the data that humans (adults or children) encounter and the tasks that they face? Third is the question of how the model can be used to predict language change and language typology. What kinds of languages are predicted to be learnable, how does the input affect morphological learning, and what are the consequences for typological distributions, especially of inflectional systems?

5 EVALUATING MODELS' MORPHOLOGICAL KNOWLEDGE

The performance of reinflection models is generally measured in terms of percentage accuracy of the predicted output strings.[15] This yields a single number per language – usually a fairly high one. For the purpose of comparing different model variants, this kind of evaluation is sufficient, but as a tool for understanding the morphological system, it falls short. In fact, this kind of evaluation is vulnerable to the criticisms of other measurements of morphological complexity which attempt to distill the question of "how learnable" a system is into a single number (Section 3). A few other studies (Gorman *et al.* 2019; King *et al.* 2020) make the same criticisms and propose techniques for error analysis similar to the one outlined below. Malouf (2017) and Corkery *et al.* (2019) perform a different, and complementary, analysis of what is learned, by plotting reduced-dimensional projections of the model's latent space.

We argue here for a fine-grained, linguistically sophisticated analysis of model errors. Counterintuitively, we argue that such an informative error analysis requires the very theoretical concepts which the model was designed to do without: "inflection class" and "exponent". As a simple case study, we present an experiment using Latin nouns, one of the best-studied examples of an inflection class system (Carstairs-McCarthy 1994; Stump and Finkel 2015; Beniamine *et al.* 2018). Latin nouns inflect for case (6 cases, not counting the rare locative) and number (singular and plural) for a total of 12 paradigm cells per noun. We use a Pytorch (Paszke *et al.* 2017) implementation of Kann and Schütze (2016)[16] on the nouns from Latin Unimorph (Kirov *et al.* 2018),[17] training for 5 epochs (passes over the training data) of stochastic gradient descent. The dataset contains 8342 nominal paradigms; we hold out 10% of the lexemes as a development set and another 10% for testing. This training/testing procedure is the

[15] The Levenshtein edit distance on character strings is also reported, but exact match is more useful; under an edit metric, a system can achieve fairly high scores by copying the input form, and the performance of this baseline varies across languages based on the relative lengths of stems versus affixes.

[16] `https://github.com/DavidLKing/MED-pytorch`

[17] Unimorph 1.0, accessed 5 December 2018.

one used in SIGMORPHON 2016 (Cotterell *et al.* 2016), and is the most standard from a machine learning point of view. It has been criticized on cognitive grounds, since it gives the learner access to complete paradigms for 90% of the words, a point we return to in our discussion of acquisition below.

As in SIGMORPHON task 1, we inflect a citation form to produce the other forms. The choice of input form is important for task performance; Stump and Finkel (2015) show that Latin nouns have 4 static principal parts; in other words, that if a single set of paradigm cells must predict the output cell, it would have to consist of 4 members per lexeme. We instead use the NOM.SG, which is the conventional dictionary head word, but not a particularly predictive paradigm cell – thus, the system will be forced to guess at the class memberships for some of the words.[18] Our measure of success is exact match accuracy.

The accuracy of the model on test is 86%; the accuracy on the validation data is 93%. This is comparable with previously reported results using this model and this type of task (though far from the current state of the art – see Cotterell *et al.* 2018c), and is high enough to render the model useful for some practical applications in language generation (King and White 2018). But what is the system getting wrong? A first attempt at an error analysis (on the development set) is to compute error counts by paradigm cell (Table 4). This shows that the VOC.SG has much lower error than the other cells, which is unsurprising since VOC.SG is identical to NOM.SG unless the noun ends in *-us* or *-ius*.[19] But the remaining errors are distributed relatively evenly across the cells.

[18] It is theoretically relevant whether morphological systems have defined bases, and if so, whether the base is one surface form (Albright 2002a), many (as assumed by some abstractive WP models; Kann *et al.* 2017), or an abstract stem (as assumed, more or less, by constructive WP models; Cotterell *et al.* 2015; Stump 2001). This question could be evaluated in this framework. If the stem is abstract, it is also theoretically relevant how to decide what is part of the stem and what is part of inflectional exponence (Spencer 2012; Beniamine *et al.* 2017). Our choice here is theoretically unmotivated and merely represents a common way of constructing a reinflection task.

[19] Why does the system ever make errors in the (easy) vocative singular? Some of these reflect inconsistency in the training data, as also noted by Gorman *et al.* (2019), but the system also applies two generalizations somewhat inconsistently: Nouns in *-us* take *-e*, but not all nouns ending in *-s* do so: *rinoceros* 'rhinoceros'

Case	SG errors	PL errors
NOM	–	59
GEN	56	71
DAT	58	57
ACC	55	62
ABL	61	57
VOC	28	57

Table 4:
Errors by paradigm cell

We can move beyond this by examining the inflectional microclasses. Taking for granted that Latin nominal inflections are entirely suffixing, and without any model of phonological alternations, we construct a function to assign each lexeme to a class: we compute a stem by taking the longest common initial string across all inflected forms of the lexeme, and a signature which is a list of 'suffixes' (that is, everything except the stem) ordered by paradigm cell. This procedure is a simplification of existing approaches designed for more complex problems (Gaussier 1999; Goldsmith 2001; Durrett and DeNero 2013) and for calculations of inflectional complexity (Stump and Finkel 2013).[20] The system finds 272 classes in the full dataset; the most common class in our data is exemplified by GUTTA 'drop'.[21] (Many of the smaller "classes" represent compound words like *dies Martis* 'Tuesday' for which the assumption of suffixation is erroneous.)

Of these classes, 79 occur in the development set, but only 11 of them are represented by 9 or more lexemes. Table 5 shows each one with its error rate. Interestingly, several of the most common classes have error rates of around 1%, while the 5 least frequent have rates between 10–20%. This is consistent with the observation that inflected forms within small classes are unpredictable as a result of irregularity. It can also be construed as consistent with the Marginal Detraction Hypothesis (Stump and Finkel 2013), inasmuch as small classes are disproportionately responsible for whatever total amount of unpredictability is found in the inflectional system. Again, it is possible to

rather than *rinocere*; nouns in *-ius* take *-ī* rather than *-e*: *sēcrētārī* 'secretary' rather than *sēcrētārie*.

[20] As noted earlier, Stump and Finkel (2013) call the initial string the 'theme' and the remainder the 'distinguisher' to make it clear that while these subparts of the word may correspond to a stem and affix in theoretical terms, they need not.

[21] With suffixes *-īs, -ā, -ās, -am, -īs, -ae, -ārum, -ae, -ae, -a, -ae, -a*.

look deeper by examining the actual erroneous outputs. In each case, we compute a suffix for the errorful form and ask which microclasses (if any) it *could* have belonged to.[22] Many of the erroneous forms cannot be assigned to any microclass and cannot therefore be easily interpreted. These include cases where the system produces sequences with no linguistic interpretation (*honōrificābilitūdinitās* to *honōrififiilisitūditās*) – an occupational hazard of running recurrent neural networks on long strings – and other cases where the stem is incorrectly copied.

In the remaining instances, the word receives an ending that is legitimate elsewhere in the language but not for the lexeme in question. These include mispredictions of stem elements that are neutralized in the NOM.SG: the NUTRIX 'nurse' class contains nouns whose NOM.SG ends in *-ix* and whose stem ends in *-ic*. Members of this class are frequently misinflected like HARUSPEX 'diviner', which has *-ex* in the nominative and *-ic* elsewhere,[23] and GREX 'herd' (not shown in the chart), whose stem ends in *-eg*.[24] They also include cases where the nominative form does not identify the inflection class: the SENATUS 'Senate' class (part of the traditional 4th declension) has an *-us* suffix in NOM.SG and an *-u* stem vowel. The *-us* suffix is also consistent with the much more common class of *-o*-stem nouns like ASELLUS 'donkey.DIM', and the system is not always capable of telling the difference.[25]

While the overall performance number tells us little about what is difficult in Latin nominal inflections, and the featural analysis not much more, it is possible to learn something about the system by examining the sequence-to-sequence model errors. A natural next step is to ask whether these errors tell us something about how the system is acquired (did Roman infants learn the SENATUS class later than the

[22] In Table 5, the microclasses are labeled with the citation form of an arbitrary member. When calculating confusions between classes, we restrict ourselves to alternative classes including at least 10 lexical items. A single error may be consistent with multiple overlapping classes, so the counts of 'confused classes' can exceed the total errors.

[23] Ex. *faicēs* for *faecēs* 'dregs.ACC.PL'.

[24] Ex. *quincungēs* for *quincuncēs* 'five-twelfths.ACC.PL'.

[25] Ex. *quassīs* for *quassibus* 'shaking.ABL.PL' and *Scōtibus* for *Scōtīs* 'Scot.ABL.PL'.

ASELLUS class) and how stable it might be (should we predict that the SENATUS class eventually merged into ASELLUS). We address these questions below.

Table 5: Errors for common microclasses

Class	Lexemes	Forms	Errs	Err. rate	Frequently confused classes
GUTTA	171	1881	27	0.014	none (27)
GRAVITATIO	170	1870	28	0.015	GREMIUM (9), LEXICON (9)
GREMIUM	145	1595	5	0.003	none (3), MINUTAL (1)
ASELLUS	87	957	43	0.045	none (16), SENATUS (10)
IMPERATOR	44	484	4	0.008	LITTUS (3), none (1)
GRAVITAS	41	451	32	0.07	none (11), ASELLUS (10)
NUTRIX	18	198	39	0.197	none (9), HARUSPEX (9)
GUTTUR	10	110	33	0.3	IMPERATOR (10), none (8)
SENATUS	9	99	11	0.11	ASELLUS (10), MYTHOS (10)
HOSTIS	9	99	23	0.232	none (2)
GYMNAS	9	99	25	0.253	none (11), GRAVITATIO (8)

6 ACQUISITION

Cotterell *et al.* (2018b) suggests that sequence-to-sequence models can function as cognitive models of infant language learners (though see Corkery *et al.* (2019) for some differences in behavior for nonce words). But to use a sequence-to-sequence model as a credible stand-in for the human infant, we must determine what the input for acquisition of morphology looks like – the right representation and learning algorithm cannot tell us anything if it is supplied with the wrong data. From the computational point of view, this question divides more or less neatly into two parts: first, what is the distribution of lexemes and paradigm cells in the input? And second, what information (phonological, syntactic or otherwise) is available to the learner when they hear a form?

The answer to the first question is conceptually well-known: both lexical items and paradigm cells have a Zipfian distribution (Blevins *et al.* 2017; Lignos and Yang 2018). In informal terms, natural language consists of many repetitions of the most common words, interspersed with a large population of rare words which appear a few

times each. Roughly the same is true for inflections: a natural corpus contains many repetitions of the most-used cells, but rarely-used cells are sparse in the data, and, in general, found only with common words. These distributions interact on the semantic level, so that (for instance) paired body parts like hands and eyes are commonly attested in the dual, while the dual forms for nouns like "nose" and "tooth" are relatively rare (Tiersma 1982; Bybee 1995).

Simulations in the sequence-to-sequence framework are just beginning to engage with this issue, perhaps because appropriate training data can be difficult to acquire. The datasets released by the SIGMORPHON shared tasks do not reflect the Zipfian frequency distributions of natural language. The 2016 dataset was chosen at random from Wiktionary, while the 2017 provided fewer examples, with complete paradigms for only a few words in each language. Systems trained on these datasets tend to learn from and be evaluated mostly on rare words. They have little incentive to learn about rare inflection classes, even where these contain extremely common words that make up a large percentage of child input. As already mentioned, frequency appears to be critical to the acquisition and diachronic stability of these small classes (Bybee 1995). Cotterell *et al.* (2018b), in their study of English irregulars, instead provide the system with data balanced by token frequency. This forces the system to learn irregulars like *go* ∼ *went*.

But the straightforward choice to balance the system by token frequency is also problematic, since many theories propose that learners are more sensitive to type frequency (Bybee 2003; Yang 2017; Goldwater *et al.* 2006). Typically, such theories suggest that generalization of a pattern to new items depends on the number of types to which it applies, while retention of a pattern for observed items depends on the number of tokens experienced. Bayesian models like adaptor grammars (Johnson *et al.* 2006) are capable of interpolating between types and tokens by using a separate "memory" component to store high-token-frequency training items, while some tokens of each type (logarithmically many) are treated as evidence for a general base distribution. The same process has been proposed as a model of morphological processing (Bertram *et al.* 2000; Baayen 2007). In modeling terms, two alternatives suggest themselves. One possibility is to use a conventional model, but provide it with a dataset in which a word

type whose frequency is t has $\log(t)$ tokens. The other is to add a memory component which can use the neural model as a Bayesian prior (Kawakami *et al.* 2018).

Regardless of the particular theoretical choices a researcher wishes to make, any attempt to study a real language via simulation requires access to high-quality data. Here, it is important to note that none of the SIGMORPHON datasets, nor the newer and larger Unimorph dataset (Kirov *et al.* 2018), provide an adequate set of lexical items for preparing Zipfian datasets in a large set of languages. The German Unimorph 2.0 dataset, for instance,[26] lacks paradigms for the copula *sein*, the auxiliary verbs *können, möchten, sollen, wollen*, and some commonly used content words in the child-directed inventory: *hören* 'hear', *essen* 'eat', *Hund* 'dog', etc. Many of these words exist in derived or compounded form (for instance, *Dachshund, Kampfhund, Schweinhund* are all represented), but this is unhelpful when attempting to construct a dataset which matches the token frequency of natural language, since none of these derivatives is particularly frequent. Unfortunately, the spotty coverage of high frequency words for German appears to be typical of the Unimorph datasets.

Thus, although Unimorph is an important resource for understanding morphological systems across a wide variety of languages, it is of limited use for simulations of language acquisition that seek to account for the role of frequency distributions. The easiest current option for creating frequency-matched datasets is to scavenge morphologically tagged forms from the Universal Dependencies syntactic datasets (Nivre *et al.* 2016). These do not have complete paradigms and do not represent child-directed speech; nonetheless, their coverage of commonly used forms such as auxiliary verbs is reasonably complete.

The second question, the issue of what information is available to the learner when they hear a form, is more complex. The SIGMORPHON reinflection problem, in its hardest form (task 3) is intended to model something like a "wug"-test (Berko 1958), in which an already somewhat proficient speaker of a language hears a novel word and

[26] Downloaded from `https://unimorph.github.io/` on 29 October 2019, with 179339 forms and 15060 lemmas.

then tries to produce some form of the word themselves. An English speaker, for instance, might participate in the following conversation (Berko 1958):

> A: This is a man who knows how to *spow*. He did the same thing yesterday. What did he do yesterday?
> B: Yesterday he _ (SPOW).

Here, B's role in the conversation requires them to produce a form of the abstract lexical item SPOW. In order to do so, they must guess that A's production *spow* is the V.NFIN form, and then infer the corresponding V.PST. But this description of B's mental processes assumes a relatively mature grammar of English, in which B already knows that *how to* _ is a good context for the nonfinite English verb, that English marks the past tense differently from the nonfinite, but that it is not necessary to mark person or number in the past tense, etc. All that is missing is the exponence, that is, the actual surface form which occupies the cell V.PST, thus the conventional description of this task as a paradigm cell *filling* problem.

For the developing language learner, however, this problem setup assumes too much. The learner does not start off knowing which features of the context are relevant to determining the form of SPOW, or which abstract features are active for the desired output form, or even which surface forms in dialogues like this belong to the same lexeme! This more complex problem can be viewed as one of paradigm cell *discovery*.[27] But how can a Paradigm Cell Discovery Problem (PCDP) be modeled computationally?

One possibility is the cloze task described in the SIGMORPHON 2018 shared task (Cotterell *et al.* 2018c). In this task, the output slot is described in terms of a sentence frame rather than a set of abstract features. However, the sentence frame representation has a serious problem in that it does not always specify the semantic features of the output. In the dialogue above, it is clear from the auxiliary verb *have* that the output has to express PST tense. But in many sentences, the morphological marker is the only expression of the property – in the sentence "The sun SHINE on the TREE", both *shines* and *shone*, and

[27] Boyé and Schalchli (2019) independently raise essentially the same issue, which they call the Paradigm Cell Finding Problem.

both *tree* and *trees*, are acceptable. SIGMORPHON 2018 deals with this by allowing both answers to be accepted. But from a learning standpoint, this is not reasonable; one response is presumably more faithful to the world context, the conversational common ground, and the speaker's own mental representation of the event than the other. It is not clear how such problems ought to be addressed. Visual language grounding (Kamper *et al.* 2017: among others) is a good source of information about objects and their properties, but probably cannot be used to learn features of verbs such as IRREALIS or REMOTE PAST, since abstract verb semantics are mostly inaccessible from visual context alone (Gillette *et al.* 1999; Papafragou *et al.* 2007). An effective solution to this problem probably cannot depend solely on learning semantic/surface correspondences, but requires attention to the structure of the surface morphological system itself (the *morphome*; cf. Maiden 2005; Aronoff 1994). In a morphomic analysis, the learner tries to determine how many different exponences each lexeme seems to have and their distributions, without necessarily assuming that each one corresponds to a coherent set of semantic meanings. Dreyer and Eisner (2011) models this process by clustering surface forms into lemmas and paradigm cells, at the same time inducing a set of morphological processes which relate the cells. But Dreyer's "cells" are purely formal groupings with no syntactic or semantic interpretation.

At the same time, the learner must do a realizational analysis using grounding and linguistic context to determine what external factors seem to license inflectional variations. The two analyses may not match; in some cases, as with complex tense/aspect distinctions, the surface differences between two forms may be much more salient than the distributions. In cases of syncretism, on the other hand, the surface forms are identical across two paradigm cells which nonetheless express different abstract features, and this can only be noticed on the basis of distributional evidence. The goal of a PCDP model must be to reconcile the two analyses by determining the interface between them.

A PCDP model, therefore, cannot be evaluated solely on the basis of a cloze task, since this will fail to test the model's ability to distinguish between too many feature pairs in most contexts. It should also function as a morphological part of speech tagger and can be evaluated

in that respect.[28] Given a form in context, it should be able to categorize it consistently by labeling it as an instance of an abstract paradigm cell, and perhaps even assigning it some latent semantic dimensions. These can be compared with the results of existing unsupervised POS taggers (Christodoulopoulos *et al.* 2010).

One way forward might be to augment existing grammar induction models for untagged word strings (Seginer 2007; Jin *et al.* 2018; He *et al.* 2018) to assemble words into morphological paradigms. These models are already constructed to predict correlations between words at the sentence level by positing syntactic relationships between them; equipping them with a model for morphological variation (Dreyer and Eisner 2011; Silfverberg *et al.* 2018) would allow them to model the morphosyntactic interface. Taking the nominative/accusative cases for example, if the case markings are relatively consistent and regular, and the statistical properties of the relationship between the subject/object and the verb are also relatively consistent, the grammar induction system should find distinguishing these cases beneficial to predicting sentence structure. It is unclear how well such an approach would work. Current grammar induction systems do not always induce linguistically plausible grammars. For systems that induce phrase structure grammars, morphological agreement features must be conveyed by aggressively subcategorizing syntactic categories (Petrov *et al.* 2006), which greatly increases the size of the model to be induced. Nevertheless, a combined model of this type might serve as a useful baseline for the PCDP.

7 CHANGE

Models of acquisition test how well a single learner can discover the rules of the system, given data produced by actual speakers of the language. But the language learners of today are the language users of tomorrow; a natural extension of the learning simulation is to make the output from one generation of computational learners serve as

[28] Taggers which use fine-grained, multidimensional tags to indicate all the morphosyntactic properties of a particular word token are generally trained in the supervised setting (Chrupała *et al.* 2008; Müller *et al.* 2013); for this task, it would be necessary to apply this fine-grained standard of evaluation to unsupervised models.

training data for another, observing how the system changes over time (Kirby and Hurford 1997, 2002). Such iterated learning experiments have been used to study the emergence and disappearance of irregular forms in both simulated (Ackerman and Malouf 2015; Parker *et al.* 2019) and real (Hare and Elman 1995; Cotterell *et al.* 2018b) datasets, and to study the spread and loss of different languages or linguistic features in a social network (Abrams and Strogatz 2003; Castelló *et al.* 2013).

Iterated language change simulations tend to take one of two perspectives on language change, modeling change as arising either from acquisition or from usage. Models of acquisition-based change treat most differences between generations as cases of imperfect acquisition: due either to data sparsity or to biased hypothesis selection, the "children" do not acquire the same language as the "parents". Sparsity and a preference for regularity lead the system to regularize, eliminating irregular forms and merging inflection classes (Kalish *et al.* 2007; Reali and Griffiths 2009); this is also a typical outcome in Ackerman and Malouf (2015), though some simulations do lead to large numbers of inflection classes. A preference for distinctiveness, on the other hand, can lead to the maintenance or even the creation of irregularity, since irregular forms are compact and easy to recognize (*go* ∼ *went* rather than *goed*; Dale and Lupyan 2012). In any case, these models see change as primarily arising from learning.

Many such models make the same incorrect prediction: morphological change should be rapid and common, and it should work to eliminate "non-functional" parts of the system, such as inflection class, which do not correspond to any abstract meaning. In fact, the typological pattern is the opposite; many real morphological systems have these non-functional elements, and while individual words may move from class to class, the classes can be remarkably stable across long periods of historical time. As Harris (2008: 66) says, "there is apparently no need of repair; the system works and can be acquired... there is nothing about our innate endowment that demands that a language simplify". Although some elements of a morphological system may take years to reach adult-like competence (Xanthos *et al.* 2011; Forshaw *et al.* 2017), given enough exposure, learners will eventually produce it with high fidelity. The preference for over-regularization observed in child learners may be a relatively temporary phase of

development (Maratsos 2000; Ambridge *et al.* 2013; Joseph 2011) which does not normally cause sweeping changes in the adult system (for an opposing viewpoint, see Huang and Pinker 2010). Additionally, the Natural Morphology framework (Dressler 2003; Wurzel 1989, 2000) emphasizes the role of languages' system-defining structural properties in making their morphological systems conservative when it comes to language change (Wurzel 1989: 104). From this perspective, the pressure in language change is towards greater system congruity, not necessarily towards elimination of inflection classes or other "non-functional" parts of the system.

Sociolinguistic models, on the other hand, see change as primarily a result of biased language usage, with significant language changes occurring throughout the lifetime (Labov 2007). In this kind of model, users make both conscious and unconscious decisions about what language features to use (Milroy 2007). For instance, in a model of language change in Spain (Castelló *et al.* 2013), agents in a social network speak either prestigious Castilian or stigmatized Galician. Speakers may switch languages in either direction, based on how many of their neighbors in the network they will be able to communicate with and how socially prestigious they will become. For most network topologies, Galician will eventually be lost entirely. This is not an effect of learning biases: in a model of this type, there is no difference in learnability between the systems; rather, it is taken for granted that agents could, in principle, acquire either system perfectly, if it proved to be worth the social investment.

In the case of morphological systems, change is likely to be a combination of both learnability and prestige and other social factors. While any linguistic variability is likely to gain some amount of social evaluation, some morphological variables within a population seem relatively unmarked (perhaps because they apply primarily to unfamiliar words; Dąbrowska 2008), while others attract widespread attention and stigmatization. This is the idea behind Labov's division of socially-relevant linguistic variables into indicators, markers, and stereotypes (Labov 1971, 2001). At the same time, however, the system may provide the learner with varying degrees of evidence for the different forms. These conflicting pressures are probably responsible for selecting among the possible outcomes. For example, Jutronic (2001) describes competition between two dialects of Croatian in the

city of Split; for instance, the local dialect form *profešuri* competes with standard *profesora* to realize 'professor.PL.GEN'. Dialect contact of this kind can eventually converge on one of the two original systems, but can also give rise to a more complex system in which both variants are analyzed as morphological markers (Trudgill 2011: 27).

In many cases, however, the real impact of social factors is to cause changes to the morphology indirectly, through their impact on other linguistic subsystems. For instance, socially conditioned phonological change can cause a reorganization of the inflectional system. On the one hand, sound change can destroy morphological distinctions, by merging or eliminating affixes. Where distinctions are not leveled outright, it can change which elements of the surface string act as markers for a morphological feature, raising a phonological alternation to the level of an exponent. On the other hand, processes of phonological reduction and grammaticalization can create new morphemes, as in the evolution of the French adverbial suffix *-ment* from the Latin noun *mente* 'mind.ABL' (Joseph 2003).

The real impact of learning biases in morphological change may be felt primarily in determining how a language reacts to this kind of disruption. Harris (2008) argues that the Georgian pattern in which the same case endings indicate different semantic roles for different classes of verbs (Series I vs II) results from the historical development of the Kartvelian languages from true ergative languages to split-ergative alignment. The Series II verbs began as a productive antipassive construction which was lost along with ergativity, but became "frozen" in the language as a morphologically complex relationship between classes of verbs and surface case markers. Harris argues that while some languages undergoing this kind of change converge on a single consistent set of case markers, the salience of the Georgian markers prevented this kind of mis-learning. In other words, as the language changed, the older meaning of the construction became too opaque for learners to acquire it, creating the potential for two eventual outcomes: one in which the new surface pattern persisted and one in which it was regularized. It was in this situation that the phonological distinctiveness of the markers themselves (and perhaps other acquisition biases) became important in determining which system would be learned.

In other examples, the external pressures on the system come from bilingualism or adult language learning. A wide variety of studies involving bilinguals can be interpreted as demonstrating these kinds of change, which can lead the system either to lose or to gain morphological features. Dorian (1978) shows the loss of features in a dying variety of Scots Gaelic in the process of being replaced by English. The Gaelic system retains a variety of exponents of the plural and gerund, even in the last generation of speakers, but morphological processes involving *features* which English does not use (for instance lengthening) were lost. On the other hand, Lefebvre (1996) shows the introduction of new features by bilingual speakers who expect a system to express certain abstract features, and recruit L2 features as surrogate markers. This is the case in Haitian Creole, where the tense/aspect/modality, pronominal and nominal systems have been interpreted as relexification of its substrates (mainly Ewe and Fongbe) using a French superstratum. For example, Ewe has the morpheme *wò* indicating the third person plural pronoun and the plural in noun phrases. Lefebvre argues that the Haitian Creole morpheme *yo* encodes both notions and that it reflects substratum influence. While the presence of L2 speakers has been suggested as a pressure towards less enumerative morphological complexity (Trudgill 2011; Dale and Lupyan 2012; Frank and Smith 2018), understanding the actual impact of a learner population might require more insight into their L1 system and how their learned representations can be adapted to fit the state of the L2, as well as into the social circumstances under which they learn.

Whether simulating L1 or L2 learning, sequence-to-sequence models provide an interesting platform for detailed and realistic learning simulations. But these simulations need to move beyond training the system on corpora reflecting synchronic steady states, then analyzing the errors to predict incipient large-scale restructuring towards some imagined typological ideal. Typological variety is the product of the historical paths down which languages travel (Harris 2008; Anderson 2004): typologically rare morphological systems occur when a particular change (phonological, social or otherwise) interacts with a particular morphological system. By incorporating data from outside the realm of morphology, we can hope to create better models of diachronic change and better understand the circumstances under which these rare systems arise.

CONCLUSIONS

The main goal of this paper has been to argue that sequence-to-sequence models hold out the possibility not only of improvements in practical tasks, but also of real advances in morphological theory and typology. Alongside this promise comes the necessity to think harder about our experimental setups. We have put forward possible improvements in how the models are evaluated, in what tasks they are trained to perform and in how we extrapolate from a single learner to a community of socially motivated language users. In particular, we have argued for error analyses in terms of paradigm cells and inflectional classes, rather than dataset-wide accuracy. We have proposed using Zipfian datasets and replacing, or at least supplementing, Paradigm Cell Filling with Paradigm Cell Discovery. And we have suggested that models of morphological change reach beyond the morphological system to incorporate factors such as prestige, bilingualism and sound change.

ACKNOWLEDGMENTS

This paper grew out of a joint seminar, co-taught by Micha Elsner and Andrea Sims, in the Ohio State University Department of Linguistics in fall 2018. We thank our chair, Shari Speer, for making it possible for us to work together on a course. We also thank three anonymous reviewers for their comments.

REFERENCES

Daniel M. ABRAMS and Steven H. STROGATZ (2003), Linguistics: Modelling the dynamics of language death, *Nature*, 424(6951):900.

Farrell ACKERMAN, James P. BLEVINS, and Robert MALOUF (2009), Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter, *Analogy in grammar: Form and acquisition*, pp. 54–82.

Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: The low conditional entropy conjecture, *Language*, 89(3):429–464.

Farrell ACKERMAN and Robert MALOUF (2015), The No Blur Principle effects as an emergent property of language systems, in Anna E. JURGENSEN, Hannah SANDE, Spencer LAMOUREUX, Kenny BACLAWSKI, and Alison ZERBE, editors, *Proceedings of the Forty-First Annual Meeting of the Berkeley Linguistics Society*, pp. 1–14, Berkeley Linguistics Society.

Roee AHARONI and Yoav GOLDBERG (2017), Morphological inflection generation with hard monotonic attention, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2004–2015.

Adam ALBRIGHT (2002a), *The identification of bases in morphological paradigms*, Ph.D. thesis, Department of Linguistics, University of California, Los Angeles.

Adam ALBRIGHT (2002b), Islands of reliability for regular morphology: Evidence from Italian, *Language*, 78(4):684–709.

Adam ALBRIGHT and Bruce HAYES (2002), Modeling English past tense intuitions with minimal generalization, in *Proceedings of the Sixth Meeting of the Association for Computational Linguistics Special Interest Group in Computational Phonology in Philadelphia, July 2002*, pp. 58–69.

Ben AMBRIDGE, Julian M. PINE, Caroline F. ROWLAND, Franklin CHANG, and Amy BIDGOOD (2013), The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure, *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):47–62.

Stephen R. ANDERSON (1992), *A-morphous morphology*, Cambridge University Press.

Stephen R. ANDERSON (2004), Morphological universals and diachrony, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of morphology 2004*, pp. 1–17, Springer.

Mark ARONOFF (1994), *Morphology by itself: Stems and inflectional classes*, MIT Press.

R. Harald BAAYEN (2001), *Word frequency distributions*, Kluwer.

R. Harald BAAYEN (2007), Storage and computation in the mental lexicon, in Gonia JAREMA and Gary LIBBEN, editors, *The mental lexicon: Core perspectives*, pp. 81–104, Elsevier.

Matthew BAERMAN (2012), Paradigmatic chaos in Nuer, *Language*, 88(3):467–494.

Matthew BAERMAN (2014), Covert systematicity in a distributionally complex system, *Journal of Linguistics*, 50(1):1–47.

Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2017), *Morphological complexity*, Cambridge University Press.

Dzmitry BAHDANAU, Kyunghyun CHO, and Yoshua BENGIO (2014), Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*.

Sacha BENIAMINE, Olivier BONAMI, and Joyce MCDONOUGH (2017), When segmentation helps: Implicative structure and morph boundaries in the Navajo verb, in *Proceedings of the First International Symposium on Morphology*, pp. 11–15.

Sacha BENIAMINE, Olivier BONAMI, and Benoît SAGOT (2018), Inferring inflection classes with description length, *Journal of Language Modelling*, 5(3):465–525.

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

Raymond BERTRAM, Robert SCHREUDER, and R. Harald BAAYEN (2000), The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26(2):489–511.

James P. BLEVINS (2004), Inflection classes and economy, in Gereon MÜLLER, Lutz GUNKEL, and Gisela ZIFONUN, editors, *Explorations in nominal inflection*, pp. 51–95, Mouton de Gruyter.

James P. BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):511–573.

James P. BLEVINS (2013), American descriptivism ('structuralism'), in Keith ALLAN, editor, *The Oxford handbook of the history of linguistics*, pp. 419–437, Oxford University Press.

James P. BLEVINS (2016), *Word and paradigm morphology*, Oxford University Press.

James P. BLEVINS, Petar MILIN, and Michael RAMSCAR (2017), The Zipfian paradigm cell filling problem, in Ferenc KIEFER, James P. BLEVINS, and Huba BARTOS, editors, *Perspectives on morphological organization: Data and analyses*, pp. 141–158, Brill.

Harry BOCHNER (1993), *Simplicity in generative morphology*, Mouton de Gruyter.

Olivier BONAMI and S. BENIAMINE (2016), Joint predictiveness in inflectional paradigms, *Word Structure*, 9(2):156–182.

Gilles BOYÉ and Gauvain SCHALCHLI (2019), Realistic data and paradigms: The paradigm cell finding problem, *Morphology*, 29(2):199–248.

Dunstan BROWN, Greville G. CORBETT, Norman FRASER, Andrew HIPPISLEY, and Alan TIMBERLAKE (1996), Russian noun stress and Network Morphology, *Journal of Linguistics*, 34:53–107.

Dunstan BROWN and Andrew HIPPISLEY (2012), *Network morphology: A defaults-based theory of word structure*, Cambridge University Press.

Joan BYBEE (1995), Diachronic and typological properties of morphology and their implications for representation, in *Morphological aspects of language processing*, pp. 225–246, Erlbaum Hillsdale.

Joan BYBEE (2003), Mechanisms of change in grammaticization: The role of frequency, in Brian D. JOSEPH and Richard D. JANDA, editors, *The handbook of historical linguistics*, pp. 602–623, Blackwell.

Joan Bybee and Carol Moder (1983), Morphological classes as natural categories, *Language*, 59(2):251–270.

Andrew Carstairs (1987), *Allomorphy in inflexion*, Croom Helm.

Andrew Carstairs-McCarthy (1994), Inflection classes, gender, and the principle of contrast, *Language*, 70(4):737–788.

Andrew Carstairs-McCarthy (2010), *The evolution of morphology*, Oxford University Press.

Xavier Castelló, Lucía Loureiro-Porto, and Maxi San Miguel (2013), Agent-based models of language competition, *International Journal of the Sociology of Language*, 221:21–51.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman (2010), Two decades of unsupervised POS induction: How far have we come?, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 575–584, Association for Computational Linguistics.

Grzegorz Chrupała, Georgiana Dinu, and Josef Van Genabith (2008), Learning morphology with Morfette, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Eve V. Clark (1987), The principle of contrast: A constraint on language acquisition, in Brian MacWhinney, editor, *Mechanisms of language acquisition*, pp. 1–33, Erlbaum.

Greville G. Corbett, Andrew Hippisley, Dunstan Brown, and Paul Marriott (2001), Frequency, regularity and the paradigm: A perspective from Russian on a complex relation, in Joan Bybee and Paul J. Hopper, editors, *Frequency and the emergence of linguistic structure*, pp. 201–226, John Benjamins.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater (2019), Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3868–3877, Association for Computational Linguistics, Florence, Italy, doi:10.18653/v1/P19-1376, https://www.aclweb.org/anthology/P19-1376.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner (2018a), On the complexity and typology of inflectional morphological systems, *arXiv preprint arXiv:1807.02747*.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner (2018b), On the diachronic stability of irregularity in inflectional morphology, *arXiv preprint arXiv:1804.08262*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, *et al.* (2018c), The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection, *arXiv preprint arXiv:1810.07125*.

Ryan COTTERELL, Christo KIROV, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sandra KÜBLER, David YAROWSKY, Jason EISNER, *et al.* (2017), CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages, *arXiv preprint arXiv:1706.09031*.

Ryan COTTERELL, Christo KIROV, John SYLAK-GLASSMAN, David YAROWSKY, Jason EISNER, and Mans HULDEN (2016), The SIGMORPHON 2016 shared task—morphological reinflection, in *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 10–22.

Ryan COTTERELL, Nanyun PENG, and Jason EISNER (2015), Modeling word forms using latent underlying morphs and phonology, *Transactions of the Association of Computational Linguistics*, 3(1).

Ewa DĄBROWSKA (2008), The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology, *Journal of Memory and Language*, 58(4):931–951.

Rick DALE and Gary LUPYAN (2012), Understanding the origins of morphological diversity: The linguistic niche hypothesis, *Advances in Complex Systems*, 15(03n04):1150017.

Nancy C. DORIAN (1978), The fate of morphological complexity in language death: Evidence from East Sutherland Gaelic, *Language*, 54(3):590–609.

Wolfgang U. DRESSLER (2003), Naturalness and morphological change, in Brian D. JOSEPH and Richard D. JANDA, editors, *Handbook of historical linguistics*, pp. 461–471, Blackwell.

Wolfgang U. DRESSLER, Marianne KILANI-SCHOCH, Natalia GAGARINA, Lina PESTAL, and Markus PÖCHTRAGER (2006), On the typology of inflection class systems, *Folia Linguistica*, 40(1-2):51–74.

Markus DREYER and Jason EISNER (2011), Discovering morphological paradigms from plain text using a Dirichlet process mixture model, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 616–627, Association for Computational Linguistics.

Greg DURRETT and John DENERO (2013), Supervised learning of complete morphological paradigms, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1185–1195.

Manaal FARUQUI, Yulia TSVETKOV, Graham NEUBIG, and Chris DYER (2016), Morphological inflection generation using character sequence to sequence learning, in *Proceedings of NAACL-HLT*, pp. 634–643.

Raphael A. Finkel and Gregory T. Stump (2009), Principal parts and degrees of paradigmatic transparency, in James P. Blevins and Juliette Blevins, editors, *Analogy in grammar: Form and acquisition*, pp. 13–53, Oxford University Press.

Bill Forshaw, Lucinda Davidson, Barbara Kelly, Rachel Nordlinger, Gillian Wigglesworth, and Joe Blythe (2017), The acquisition of Murrinhpatha (Northern Australia), in Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pp. 473–494, Oxford University Press.

Stella Frank and Kenny Smith (2018), A model of linguistic accommodation leading to language simplification, *PsyArXiv preprint: https://doi.org/10.31234/osf.io/4ynwu*.

Éric Gaussier (1999), Unsupervised learning of derivational morphology from inflectional lexicons, in Andrew Kehler and Andreas Stolcke, editors, *Unsupervised learning in natural language processing: Proceedings of the workshop*, pp. 24–30, Association for Computational Linguistics.

Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer (1999), Human simulations of vocabulary learning, *Cognition*, 73(2):135–176.

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

Sharon Goldwater, Mark Johnson, and Thomas L Griffiths (2006), Interpolating between types and tokens by estimating power-law generators, in Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *Advances in neural information processing systems 19*, pp. 459–466, Neural Information Processing Systems Foundation.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska (2019), Weird inflects but OK: Making sense of morphological generation errors, in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 140–151, Association for Computational Linguistics, Hong Kong, China, https://www.aclweb.org/anthology/D19-6714.

Joseph H. Greenberg (1960), A quantitative approach to the morphological typology of language, *International Journal of American Linguistics*, 26(3):178–194.

Morris Halle and Alex Marantz (1993), Distributed Morphology and the pieces of inflection, in Kenneth Hale and Samuel Jay Keyser, editors, *The view from building 20*, pp. 111–176, MIT Press.

Morris Halle and Alex Marantz (2008), Clarifying "blur": Paradigms, defaults, and inflectional classes, in Asaf Bachrach and Andrew Nevins, editors, *Inflectional identity*, pp. 55–72, Mouton de Gruyter.

Mary HARE and Jeffrey L. ELMAN (1995), Learning and morphological change, *Cognition*, 56(1):61–98.

Heidi HARLEY and Rolf NOYER (2003), Distributed Morphology, in Lisa CHENG and Rint SYBESMA, editors, *The Second GLOT International state-of-the-article book*, pp. 463–496, de Gruyter Mouton.

Alice C. HARRIS (2008), On the explanation of typologically unusual structures, in Jeff GOOD, editor, *Linguistic universals and language change*, pp. 54–76, Oxford University Press.

Junxian HE, Graham NEUBIG, and Taylor BERG-KIRKPATRICK (2018), Unsupervised Learning of Syntactic Structure with Invertible Neural Projections, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, Association for Computational Linguistics, Brussels, Belgium, `https://www.aclweb.org/anthology/D18-1160`.

Hans Henrich HOCK (1991), *Principles of historical linguistics*, Mouton de Gruyter.

Hans Henrich HOCK and Brian D. JOSEPH (1996), *Language history, language change and language relationship: An introduction to historical and comparative linguistics*, Mouton de Gruyter.

Charles F. HOCKETT (1954), Two models of grammatical description, *Word*, 10:210–234.

Yi Ting HUANG and Steven PINKER (2010), Lexical semantics and irregular inflection, *Language and Cognitive Processes*, 25(10):1411–1461.

Carole Ann JAMIESON (1982), Conflated subsystems marking person and aspect in Chiquihuitlán Mazatec verbs, *International Journal of American Linguistics*, 48(2):139–167.

Lifeng JIN, William SCHULER, Finale DOSHI-VELEZ, Timothy A. MILLER, and Lane SCHWARTZ (2018), Unsupervised grammar induction with depth-bounded PCFG, *Transactions of the Association for Computational Linguistics*, 6:211–224, `https://github.com/lifengjin/db-pcfg`.

Mark JOHNSON, Thomas L. GRIFFITHS, and Sharon GOLDWATER (2006), Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models, in Bernhard SCHÖLKOPF, John C. PLATT, and Thomas HOFFMAN, editors, *Advances in neural information processing systems 19*, pp. 641–648, Neural Information Processing Systems Foundation.

Brian D. JOSEPH (2003), Morphologization from syntax, in Brian D. JOSEPH and Richard D. JANDA, editors, *The handbook of historical linguistics*, pp. 472–492, Blackwell.

Brian D. JOSEPH (2011), Children rule, or do they (as far as innovations are concerned)?, *Bilingualism: Language and Cognition*, 14(2):156–158.

Dunja JUTRONIC (2001), Morphological changes in the urban vernacular of the city of Split, *International Journal of the Sociology of Language*, 147:65–78.

Michael L. KALISH, Thomas L. GRIFFITHS, and Stephan LEWANDOWSKY (2007), Iterated learning: Intergenerational knowledge transmission reveals inductive biases, *Psychonomic Bulletin & Review*, 14(2):288–294.

Herman KAMPER, Shane SETTLE, Gregory SHAKHNAROVICH, and Karen LIVESCU (2017), Visually grounded learning of keyword prediction from untranscribed speech, *arXiv preprint arXiv:1703.08136*.

Katharina KANN, Ryan COTTERELL, and Hinrich SCHÜTZE (2017), Neural multi-source morphological reinflection, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 514–524.

Katharina KANN and Hinrich SCHÜTZE (2016), MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection, in *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 62–70.

Shuan KARIM (2019), Competition between formatives and the diversity of ezafat, Presented at the 24th International Conference on Historical Linguistics (ICHL24).

Kazuya KAWAKAMI, Chris DYER, and Phil BLUNSOM (2018), Unsupervised word discovery with segmental neural language models, *ArXiv e-prints*, arXiv:1811.09353.

Aleksandr E. KIBRIK (1998), Archi, in Andrew SPENCER and Arnold M. ZWICKY, editors, *The handbook of morphology*, pp. 455–476, Blackwell.

David KING, Andrea D. SIMS, and Micha ELSNER (2020), Interpreting sequence-to-sequence models for Russian inflectional morphology, in *Proceedings of the Society for Computation in Linguistics (SCiL)*, Society for Computation in Linguistics, New Orleans, USA.

David KING and Michael WHITE (2018), The OSU realizer for SRST'18: Neural sequence-to-sequence inflection and incremental locality-based linearization, in *Proceedings of the First Workshop on Multilingual Surface Realisation*, pp. 39–48.

Paul KIPARSKY (1968), Linguistic universals and linguistic change, in Emmon BACH and Robert T. HARMS, editors, *Universals in linguistic theory*, pp. 170–202, Holt, Rinehart and Winston.

Simon KIRBY and James HURFORD (1997), Learning, culture and evolution in the origin of linguistic constraints, in *Fourth European conference on artificial life*, pp. 493–502, Citeseer.

Simon KIRBY and James R. HURFORD (2002), The emergence of linguistic structure: An overview of the iterated learning model, in Angelo CANGELOSI and Domenico PARISI, editors, *Simulating the evolution of language*, pp. 121–147, Springer.

Christo KIROV and Ryan COTTERELL (2018), Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate, *arXiv preprint arXiv:1807.04783*.

Christo KIROV, Ryan COTTERELL, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sebastian MIELKE, Arya D. MCCARTHY, Sandra KÜBLER, *et al.* (2018), UniMorph 2.0: Universal morphology, *arXiv preprint arXiv:1810.11101*.

William LABOV (1971), The study of language in its social context, in Joshua A. FISHMAN, editor, *Advances in the sociology of language, vol. 1*, pp. 152–216, Mouton.

William LABOV (2001), *Principles of linguistic change, vol. 2: Social factors*, Blackwell.

William LABOV (2007), Transmission and diffusion, *Language*, 83(2):344–387.

Claire LEFEBVRE (1996), The tense, mood, and aspect system of Haitian Creole and the problem of transmission of grammar in creole genesis, *Journal of Pidgin and Creole Languages*, 11(2):231–311.

Constantine LIGNOS and Charles YANG (2018), Morphology and language acquisition, in Andrew HIPPISLEY and Gregory T. STUMP, editors, *Cambridge handbook of morphology*, pp. 765–791, Cambridge University Press.

Martin MAIDEN (2005), Morphological autonomy and diachrony, in *Yearbook of morphology 2004*, pp. 137–175, Springer.

Robert MALOUF (2017), Abstractive morphological learning with a recurrent neural network, *Morphology*, 27(4):431–458.

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

P.H. MATTHEWS (1972), *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, Cambridge University Press.

Petar MILIN, Dusica FILIPOVIĆ DJURDJEVIĆ, and Fermín MOSCOSO DEL PRADO MARTÍN (2009), The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian, *Journal of Memory and Language*, 60:50–64.

Lesley MILROY (2007), Off the shelf or under the counter? On the social dynamics of sound changes, *Topics in English Linguistics*, 53:149.

Fermín MOSCOSO DEL PRADO MARTÍN, Aleksandar KOSTIĆ, and R. Harald BAAYEN (2004), Putting the bits together: An information theoretical perspective on morphological processing, *Cognition*, 94:1–18.

Gereon MÜLLER (2007), Notes on paradigm economy, *Morphology*, 17(1):1–38.

Thomas MÜLLER, Helmut SCHMID, and Hinrich SCHÜTZE (2013), Efficient higher-order CRFs for morphological tagging, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 322–332.

Garrett NICOLAI, Colin CHERRY, and Grzegorz KONDRAK (2015), Inflection generation as discriminative string transduction, in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 922–931.

Joakim NIVRE, Marie-Catherine DE MARNEFFE, Filip GINTER, Yoav GOLDBERG, Jan HAJIC, Christopher D MANNING, Ryan T MCDONALD, Slav PETROV, Sampo PYYSALO, Natalia SILVEIRA, *et al.* (2016), Universal dependencies v1: A multilingual treebank collection., in *Proceedings of LREC*.

Anna PAPAFRAGOU, Kimberly CASSIDY, and Lila GLEITMAN (2007), When we think about thinking: The acquisition of belief verbs, *Cognition*, 105(1):125–165.

Jeff PARKER (2016), *Inflectional complexity and cognitive processing: An experimental and corpus-based investigation of Russian nouns*, Ph.D. thesis, Department of Slavic and East European Languages and Cultures, The Ohio State University.

Jeff PARKER, Robert REYNOLDS, and Andrea D. SIMS (2019), The role of language-specific network properties in the emergence of inflectional irregularity, in Andrea D. SIMS, Adam USSISHKIN, Jeff PARKER, and Samantha WRAY, editors, *Morphological typology and linguistic cognition*, Cambridge University Press, to appear.

Adam PASZKE, Sam GROSS, Soumith CHINTALA, Gregory CHANAN, Edward YANG, Zachary DEVITO, Zeming LIN, Alban DESMAISON, Luca ANTIGA, and Adam LERER (2017), Automatic differentiation in PyTorch, in *NIPS 2017 Autodiff Workshop*.

Slav PETROV, Leon BARRETT, Romain THIBAUX, and Dan KLEIN (2006), Learning accurate, compact, and interpretable tree annotation, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 433–440, Association for Computational Linguistics.

Steven PINKER and Alan PRINCE (1988), On language and connectionism: Analysis of a parallel distributed processing model of language acquisition, *Cognition*, 28(1-2):73–193.

Vito PIRRELLI, Marcello FERRO, and Claudia MARZI (2015), Computational complexity of abstractive morphology, in Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT, editors, *Understanding and measuring morphological complexity*, pp. 141–166, Oxford University Press.

Brandon PRICKETT, Aaron TRAYLOR, and Joe PATER (2018), Seq2Seq models with dropout can learn generalizable reduplication, in *Proceedings of the*

Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 93–100, Association for Computational Linguistics, Brussels, Belgium, doi:10.18653/v1/W18-5810, https://www.aclweb.org/anthology/W18-5810.

Florencia Reali and Thomas L Griffiths (2009), The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning, *Cognition*, 111(3):317–328.

David E. Rumelhart and James L. McClelland (1986), On learning the past tenses of English verbs, in James L. McClelland, David E. Rumelhart, and PDP Research Group, editors, *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models*, pp. 216–271, MIT Press.

Yoav Seginer (2007), Fast unsupervised incremental parsing, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 384–391.

Miikka Silfverberg, Ling Liu, and Mans Hulden (2018), A computational model for the linguistic notion of morphological paradigm, in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1615–1626.

Andrea D. Sims (2006), *Minding the gaps: Inflectional defectiveness in paradigmatic morphology*, Ph.D. thesis, Department of Linguistics, The Ohio State University.

Andrea D. Sims and Jeff Parker (2016), How inflection class systems work: On the informativity of implicative structure, *Word Structure*, 9(2):215–239.

Royal Skousen (1989), *Analogical modeling of language*, Springer Science & Business Media.

Andrew Spencer (2012), Identifying stems, *Word Structure*, 5(1):88–108.

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

Gregory T. Stump and Raphael A. Finkel (2013), *Morphological typology: From word to paradigm*, Cambridge University Press.

Gregory T. Stump and Raphael A. Finkel (2015), The complexity of inflectional systems, *Linguistics Vanguard*, 1(1):101–117.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014), Sequence to sequence learning with neural networks, in Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in neural information processing systems 27*, pp. 3104–3112, Neural Information Processing Systems Foundation.

Haukur Þorgeirsson (2017), Testing Vocabular Clarity in insular Scandinavian, *Folia Linguistica*, 51(3):505–526.

Peter Tiersma (1982), Local and general markedness, *Language*, 58(4):832–849.

Peter TRUDGILL (2011), *Sociolinguistic typology: Social determinants of linguistic complexity*, Oxford University Press.

Wolfgang U. WURZEL (1989), *Inflectional morphology and naturalness*, Kluwer.

Wolfgang U. WURZEL (2000), Inflectional system and markedness, in Aditi LAHIRI, editor, *Analogy, levelling, markedness: Principles of change in phonology and morphology*, pp. 193–214, Mouton de Gruyter.

Aris XANTHOS, Sabine LAAHA, Steven GILLIS, Ursula STEPHANY, Ayhan AKSU-KOÇ, Anastasia CHRISTOFIDOU, Natalia GAGARINA, Gordana HRZICA, F Nihan KETREZ, Marianne KILANI-SCHOCH, *et al.* (2011), On the role of morphological richness in the early development of noun and verb inflection, *First Language*, 31(4):461–479.

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