On the Complexity and Typology of Inflectional Morphological Systems

Ryan Cotterell[ə] and **Christo Kirov**[ə] and **Mans Hulden**[ʔ] and **Jason Eisner**[ə]
[ə]Department of Computer Science, Johns Hopkins University
[ʔ]Department of Linguistics, University of Colorado
{ryan.cotterell,ckirov1,eisner}@jhu.edu, first.last@colorado.edu

Abstract

We quantify the linguistic complexity of different languages' morphological systems. We verify that there is a statistically significant empirical trade-off between paradigm size and irregularity: a language's inflectional paradigms may be either large in size or highly irregular, but never both. We define a new measure of paradigm irregularity based on the conditional entropy of the surface realization of a paradigm—how hard it is to jointly predict all the word forms in a paradigm from the lemma. We estimate irregularity by training a predictive model. Our measurements are taken on large morphological paradigms from 36 typologically diverse languages.

1 Introduction

What makes an inflectional system "complex"? Linguists have sometimes considered measuring this by the *size* of the inflectional paradigms (McWhorter, 2001). The number of distinct inflected forms of each word indicates the number of morphosyntactic distinctions that the language makes on the surface. However, this gives only a partial picture of complexity (Sagot, 2013). Some inflectional systems are more irregular: it is harder to guess how the inflected forms of a word will be spelled or pronounced, given the base form. Ackerman and Malouf (2013) hypothesize that there is a limit to the *irregularity* of an inflectional system. We refine this hypothesis to propose that systems with many forms per paradigm have an even stricter limit on irregularity per distinct form. That is, the two dimensions interact: a system cannot be complex along both axes at once. In short, if a language demands that its speakers use a lot of distinct forms, those forms must be relatively predictable.

In this work, we develop information-theoretic tools to operationalize this hypothesis about the complexity of inflectional systems. We model each inflectional system using a tree-shaped directed graphical model whose factors are neural networks and whose structure (topology) must be learned along with the factors. We explain our approach to quantifying two aspects of inflectional complexity and, in one case, approximate our metric using a simple variational bound. This allows a data-driven approach by which we can measure the morphological complexity of a given language in a clean manner that is more theory-agnostic than previous approaches.

Our study evaluates 36 diverse languages, using collections of paradigms represented orthographically. Thus, we are measuring the complexity of each *written* language. The corresponding *spoken* language would have different complexity, based on the corresponding phonological forms. Importantly, our method does not depend upon a linguistic analysis of words into constituent morphemes, e.g., *hoping* $\mapsto$ *hope*+*ing*. We find support for the complexity trade-off hypothesis. Concretely, we show that the more unique forms an inflectional paradigm has, the more predictable the forms must be from one another—for example, forms in a predictable paradigm might all be related by a simple change of suffix. This intuition has a long history in the linguistics community, as field linguists have often noted that languages with extreme morphological richness, e.g., agglutinative and polysynthetic languages, have virtually no exceptions or irregular forms. Our contribution lies in mathematically formulating this notion of regularity and providing a means to estimate it by fitting a probability model. Using these tools, we provide a quantitative verification of this conjecture on a large set of typologically diverse languages, which is significant with $p < 0.037$.

2 Morphological Complexity

2.1 Word-Based Morphology

We adopt the framework of word-based morphology (Aronoff, 1976; Spencer, 1991). An **inflected lexicon** in this framework is represented as a set of

word types. Each word type is a triple of

- a **lexeme** ℓ (an arbitrary integer or string that indexes the word's core meaning and part of speech)
- a **slot** σ (an arbitrary integer or object that indicates how the word is inflected)
- a **surface form** w (a string over a fixed phonological or orthographic alphabet Σ)

A **paradigm** $\boldsymbol{m}$ is a map from slots to surface forms.[1] We use dot notation to access elements of this map. For example, $\boldsymbol{m}.\textsf{past}$ denotes the past-tense surface form in paradigm $\boldsymbol{m}$.

An inflected lexicon for a language can be regarded as defining a map $\boldsymbol{M}$ from lexemes to their paradigms. Specifically, $\boldsymbol{M}(\ell).\sigma = w$ iff the lexicon contains the triple (ℓ, σ, w).[2] For example, in the case of the English lexicon, if ℓ is the English lexeme $walk_{\text{Verb}}$, then $\boldsymbol{M}(\ell).\textsf{past} = \textit{walked}$. In linguistic terms, we say that in ℓ's paradigm $\boldsymbol{M}(\ell)$, the past-tense slot is filled (or realized) by *walked*.

Nothing in our method requires a Bloomfieldian structuralist analysis that decomposes each word into underlying morphs: rather, this paper is amorphous in the sense of Anderson (1992).

More specifically, we will work within the UniMorph annotation scheme (Sylak-Glassman, 2016). In the simplest case, each slot σ specifies a morphosyntactic **bundle** of inflectional features such as tense, mood, person, number, and gender. For example, the Spanish surface form *pongas* (from the lexeme *poner* 'to put') fills a slot that indicates that this word has the features [TENSE=PRESENT, MOOD=SUBJUNCTIVE, PERSON=2, NUMBER=SG]. We postpone a discussion of the details of UniMorph until §7.1, but it is mostly compatible with other, similar schemes.

2.2 Defining Complexity

Ackerman and Malouf (2013) distinguish two types of morphological complexity, which we elaborate on below. For a more general overview of morphological complexity, see Baerman et al. (2015).

[1] See Baerman (2015, Part II) for a tour of alternative views of inflectional paradigms.

[2] We assume that the lexicon never contains distinct triples of the form (ℓ, σ, w) and (ℓ, σ, w'), so that $\boldsymbol{M}(\ell).\sigma$ has a unique value if it is defined at all.

2.2.1 Enumerative Complexity

The first type, **enumerative complexity** (**e-complexity**), measures the number of surface morphosyntactic distinctions that a language makes within a part of speech.

Given a lexicon, we will measure the e-complexity of the verb system as the average of the verb paradigm size $|\boldsymbol{M}(\ell)|$, where ℓ ranges over all verb lexemes in domain($\boldsymbol{M}$). Importantly, we define the **size** $|\boldsymbol{m}|$ of a paradigm $\boldsymbol{m}$ to be the number of distinct *surface forms* in the paradigm, rather than the number of *slots*. That is, $|\boldsymbol{m}| \stackrel{\text{def}}{=} |\text{range}(\boldsymbol{m})|$ rather than $|\text{domain}(\boldsymbol{m})|$.

Under our definition, nearly all English verb paradigms have size 4 or 5, giving the English verb system an e-complexity between 4 and 5. If $\boldsymbol{m} = \boldsymbol{M}(walk_{\text{Verb}})$, then $|\boldsymbol{m}| = 4$, since $\text{range}(\boldsymbol{m}) = \{walk, walks, walked, walking\}$. The manually constructed lexicon may define separate slots $\sigma_1 = $ [TENSE=PRESENT, PERSON=1, NUMBER=SG] and $\sigma_2 = $ [TENSE=PRESENT, PERSON=2, NUMBER=SG], but in this paradigm, those slots are not distinguished by any morphological marking: $\boldsymbol{m}.\sigma_1 = \boldsymbol{m}.\sigma_2 = walk$. Nor is the past tense *walked* distinguished from the past participle. This phenomenon is known as **syncretism**.

Why might the creator of a lexicon bother to define separate slots σ_1 and σ_2 for English, rather than a single merged slot? A very good reason is the existence of a single English verb, *be*, that does distinguish these slots.[3] Still, the lexicon creator might use a merged slot in general and handle *be* by adding some special slots that are used only with *be*. A second reason is that merged slots may be inelegant to describe using the feature bundle notation: for all English verbs other than *be*, there is a single form shared by the bare infinitive *and* all present tense forms *except* 3rd-person singular, but a single slot for this form could not be easily characterized by a single feature bundle, and so the lexicon creator might reasonably split it for convenience. A third reason might be an attempt at consistency across languages: in principle, an English lexicon is free to use the same slots as Sanskrit and thus list $\textsf{dual}$ and $\textsf{plural}$ forms for every English noun, which just happen to be identical in every case (complete syncretism).

The point is that our e-complexity metric is insensitive to these annotation choices. It focuses

[3] This verb has a paradigm of size 8: {*be*,*am*,*are*,*is*,*was*,*were*,*been*,*being*}.

on observable surface distinctions., and so does not care whether syncretic slots are merged or kept separate. Later, we will construct our i-complexity metric to have the same property.

The notion of e-complexity has a long history in linguistics. The idea was explicitly discussed as early as Sapir (1921). More recently, Sagot (2013) has referred to this concept as **counting complexity**, referencing comparison of the complexity of creoles and non-creoles by McWhorter (2001).

For a given part of speech, e-complexity appears to vary dramatically over the languages of the world. While the regular English verb paradigm has 4–5 slots in our annotation, the Archi verb will have thousands (Kibrik, 1998). However, does this make the Archi system more complex, in the sense of being more difficult to describe or learn? Despite the plethora of forms, it is often the case that one can regularly predict one form from another, indicating that few forms actually have to be memorized for each lexeme.

2.2.2 Integrative Complexity

The second notion of complexity is **integrative complexity** (**i-complexity**), which measures how regular an inflectional system is on the surface. Students of a foreign language will most certainly have encountered the concept of an irregular verb. Pinning down a formal and workable cross-linguistic definition is non-trivial, but the intuition that some inflected forms are regular and others irregular dates back at least to Bloomfield (1933, pp. 273–274), who famously argued that what makes a surface form regular is that it is the output of a deterministic function. For an in-depth dissection of the subject, see Stolz et al. (2012).

Ackerman and Malouf (2013) build their definition of i-complexity on the information-theoretic notion of entropy (Shannon, 1948). Their intuition is that a morphological system should be considered irregular to the extent that its forms are unpredictable. They say, for example, that the nominative singular form is unpredictable in a language if many verbs express it with suffix *-o* while many others use -∅. In §5, we will propose an improvement to their entropy-based measure.

2.3 The Low-Entropy Conjecture

The low-entropy conjecture, as formulated by Ackerman and Malouf (2013, p. 436), "is the hypothesis that enumerative morphological complexity is effectively unrestricted, as long as the average conditional entropy, a measure of integrative complexity, is low." Indeed, Ackerman and Malouf go so far as to say that there need be no upper bound on e-complexity, but the i-complexity must remain sufficiently low (as is the case for Archi, for example). Our hypothesis is subtly different in that we postulate that morphological systems face a trade-off between e-complexity and i-complexity: a system may be complex under either metric, but not under both. The amount of e-complexity permitted is higher when i-complexity is low.

This line of thinking harks back to the equal complexity conjecture of Hockett, who stated: "objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both the morphology and syntax, is about the same as any other" (Hockett, 1958, pp. 180-181). Similar trade-offs have been found in other branches of linguistics (see Oh (2015) for a review). For example, there is a trade-off between rate of speech and syllable complexity (Pellegrino et al., 2011): this means that even though Spanish speakers utter many more syllables per second than Chinese, the overall information rate is quite similar as Chinese syllables carry more information (they contain tone information).

Hockett's *equal* complexity conjecture is controversial: some languages (such as Riau Indonesian) do seem low in complexity across morphology and syntax (Gil, 1994). This is why Ackerman and Malouf instead posit that a linguistic system has *bounded* integrative complexity—it must not be too high, though it can be low, as indeed it is in isolating languages like Chinese and Thai.

3 Paradigm Entropy

3.1 Morphology as a Distribution

Following Dreyer and Eisner (2009) and Cotterell et al. (2015), we identify a language's inflectional *system* with a probability distribution $p(\boldsymbol{M} = \boldsymbol{m})$ over *possible* paradigms.[4] Our measure of i-complexity will be related to the entropy of this distribution.

[4]Formally speaking, we assume a discrete sample space in which each outcome is a possible lexeme ℓ equipped with a paradigm $\boldsymbol{M}(\ell)$. Recall that a random variable is technically defined as a function of the outcome. Thus, $\boldsymbol{M}$ is a paradigm-valued random variable that returns the whole paradigm. $\boldsymbol{M}.\mathsf{past}$ is a string-valued random expression that returns the past slot, so $\pi(\boldsymbol{M}.\mathsf{past} = \textit{ran})$ is a marginal probability that marginalizes over the rest of the paradigm.

For instance, knowing the behavior of the English verb system essentially means knowing a joint distribution over 5-tuples of surface forms such as $(\textit{run}, \textit{runs}, \textit{ran}, \textit{run}, \textit{running})$. More precisely, one knows probabilities such as $p(\boldsymbol{M}.\mathsf{pres} = \textit{run}, \boldsymbol{M}.\mathsf{3s} = \textit{runs}, \boldsymbol{M}.\mathsf{past} = \textit{ran}, \boldsymbol{M}.\mathsf{pastp} = \textit{run}, \boldsymbol{M}.\mathsf{presp} = \textit{running})$.

We do not observe p directly, but each observed paradigm (5-tuple) can help us estimate it. We assume that the paradigms $\boldsymbol{m}$ in the inflected lexicon were drawn IID from p. Any novel verb paradigm in the future would be drawn from p as well. The distribution p represents the inflectional system because it describes what regular paradigms and plausible irregular paradigms *tend to look like*.

The fact that some paradigms are *used* more frequently than others (more tokens in a corpus) does not mean that they have higher probability under the morphological system $p(\boldsymbol{m})$. Rather, their higher usage reflects the higher probability of their lexemes. That is due to unrelated factors—the probability of a lexeme may be modeled separately by a stick-breaking process (Dreyer and Eisner, 2011), or may reflect the semantic meaning associated to that lexeme. The role of $p(\boldsymbol{m})$ in the model is only to serve as the base distribution from which a lexeme type ℓ selects the tuple of strings $\boldsymbol{m} = \boldsymbol{M}(\ell)$ that will be used thereafter to express ℓ.

We expect the system to place low probability on implausible paradigms: e.g., $p(\textit{run}, \textit{~~snur~~}, \textit{~~nar~~}, \textit{run}, \textit{running})$ is close to zero. Moreover, we expect it to assign high conditional probability to the result of applying highly regular processes: e.g., for $p(\boldsymbol{M}.\mathsf{presp} \mid \boldsymbol{M}.\mathsf{3s})$ in English, we have $p(\textit{wugging} \mid \textit{wugs}) \approx p(\textit{running} \mid \textit{runs}) \approx 1$, where *wug* is a novel verb. Nonetheless, our estimate of $p(\boldsymbol{M}.\mathsf{presp} = w \mid \boldsymbol{M}.\mathsf{3s} = \textit{wugs}$ will have support over $w \in \Sigma^* \times \cdots \times \Sigma^*$, due to smoothing. The model is thus capable of evaluating arbitrary wug-formations (Berko, 1958), including irregular ones.

3.2 Paradigm Entropy

The distribution p gives rise to the **paradigm entropy** $H(\boldsymbol{M})$, also written as $H(p)$. This is the expected number of bits needed to represent a paradigm drawn from p, under a code that is optimized for this purpose. Thus, it may be related to the cost of learning paradigms or the cost of storing them in memory, and thus relevant to functional pressures that prevent languages from growing too complex. (There is no guarantee, of course, that human learners actually estimate the distribution p, or that its entropy actually represents the cognitive cost of learning or storing paradigms.)

Our definition of i-complexity in §5 will (roughly speaking) divide $H(\boldsymbol{M})$ by the e-complexity, so that the i-complexity is measured in *bits per distinct surface form*. This approach is inspired by Ackerman and Malouf (2013); we discuss the differences in §6.

3.3 A Variational Upper Bound on Entropy

We now review how to estimate $H(\boldsymbol{M})$ by estimating p by a model q. We do not actually know the true distribution p. Furthermore, even if we knew p, the definition of $H(\boldsymbol{M})$ involves a sum over the infinite set of n-tuples $(\Sigma^*)^n$, which is intractable for most distributions p. Thus, following Brown et al. (1992), we will use a probability model to define a good upper bound for $H(\boldsymbol{M})$ and held-out data to estimate that bound.

For any distribution p, the entropy $H(p)$ is upper-bounded by the cross-entropy $H(p, q)$, where q is any other distribution over the same space:[5]

$$\sum_{\boldsymbol{m}} p(\boldsymbol{m})[-\log p(\boldsymbol{m})] \leq \sum_{\boldsymbol{m}} p(\boldsymbol{m})[-\log q(\boldsymbol{m})] \tag{1}$$

(Throughout this paper, $\log$ denotes $\log_2$.) The gap between the two sides is the Kullback-Leibler divergence $D(p \mid\mid q)$, which is 0 iff $p = q$.

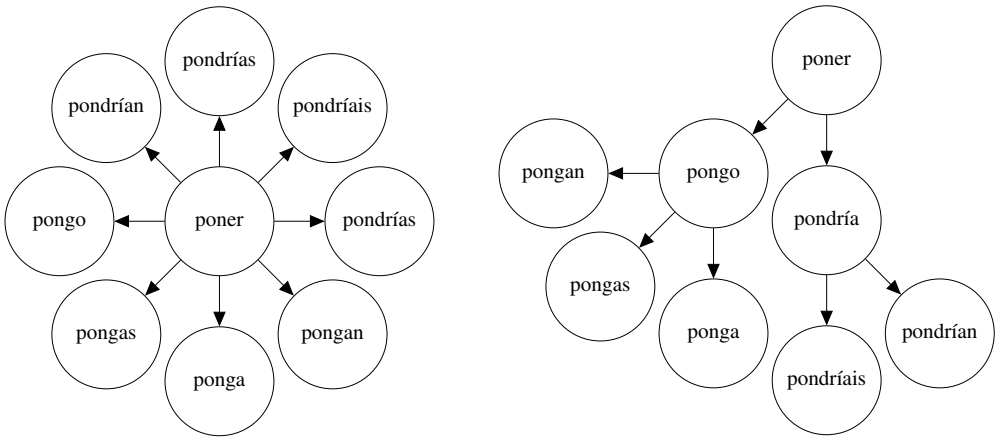
Maximum-likelihood training of a probability model $q \in \mathcal{Q}$ is an attempt to minimize this gap by minimizing the right-hand side. More precisely, it minimizes the sampling-based estimate $\sum_{\boldsymbol{m}} \hat{p}_{\text{train}}(\boldsymbol{m})[-\log q(\boldsymbol{m})]$, where $\hat{p}_{\text{train}}$ is the uniform distribution over a set of training examples that are assumed to be drawn IID from p.

Because the trained q may be overfit to the training examples, we must make our final estimate of $H(p, q)$ using a separate set of held-out test examples, as $\sum_{\boldsymbol{m}} \hat{p}_{\text{test}}(\boldsymbol{m})[-\log q(\boldsymbol{m})]$. We then use this as our (upwardly-biased) estimate of the paradigm entropy $H(p)$. In our setting, both the training and the test examples are paradigms from a given inflected lexicon.

4 A Generative Model of the Paradigm

To fit q given the training set, we need a tractable family $\mathcal{Q}$ of joint distributions over paradigms, with

[5] The same applies for conditional entropies as used in §5.



(a) Lemma paradigm tree

(b) Principal parts paradigm tree

Figure 1: A specific Spanish verb paradigm as it would be generated by two different tree-structured Bayesian networks. The nodes in each network represent the slots of the paradigm class (not labeled). The topology in (a) predicts all forms from the lemma. The topology in (b), on the other hand, makes it easier to predict forms given the others: *pongas* is predicted from *pongo*, with which it shares a stem. Qualitatively, the structure selection algorithm in §4.4 finds trees like (b).

parameters $\boldsymbol{\theta}$. The structure of the model and the number of parameters $\boldsymbol{\theta}$ will be determined automatically from the training set: a language with more slots or more paradigm classes will require more parameters. This means that $\mathcal{Q}$ is technically a semi-parametric family.

4.1 Paradigm Classes

We say that two paradigms $\boldsymbol{m}, \boldsymbol{m}'$ have the same **class** if they define the same slots (that is, $\text{domain}(\boldsymbol{m}) = \text{domain}(\boldsymbol{m}')$) and the same pairs of slots are syncretic in both paradigms (that is, $\boldsymbol{m}.\sigma = \boldsymbol{m}.\sigma'$ iff $\boldsymbol{m}'.\sigma = \boldsymbol{m}'.\sigma'$). Notice that paradigms of the same class must have the same size (but not conversely). Most English verbs fall into 2 classes: 4-form verbs such as regular *sprint* and irregular *stand* where the past participle is syncretic with the past tense, and irregular 5-form verbs such as *eat* where that is not so. There are also a few other English verb classes: for example, *run* has only 4 distinct forms, but in its class, the past participle is syncretic with the *present* tense. The verb *be* is in a class by itself, with 8 distinct forms. The extra slots needed for *be* might be either missing in other classes, or present but syncretic.

Our model $q_{\boldsymbol{\theta}}$ says that the first step in generating a paradigm is to pick its class c. This uses a distribution $q_{\boldsymbol{\theta}}(C = c)$, which we estimate by maximum likelihood from the training set. Thus, c ranges over the set $\mathcal{C}$ of classes that appear in the training set.

4.2 A Tree-Structured Distribution

Next, conditioned on the class c, we follow Cotterell et al. (2017b) and generate all the forms of the paradigm using a tree-structured Bayesian network—a directed graphical model in which the form at each slot is generated conditionally on the form at a single parent slot. Figure 1 illustrates two possible tree structures for Spanish verbs.

Each class c has its own tree structure. If slot σ exists in class c, we denote its parent in class c by $\text{pa}_c(\sigma)$. Then our model is[6]

$$q_{\boldsymbol{\theta}}(\boldsymbol{m} \mid c) = \prod_{\sigma \in c} q_{\boldsymbol{\theta}}(\boldsymbol{m}.\sigma \mid \boldsymbol{m}.\text{pa}_c(\sigma), C{=}c) \tag{2}$$

For the slot σ at root of the tree, $\text{pa}_c(\sigma)$ is defined to be a special slot empty with an empty feature bundle, whose form is fixed to be the empty string. In the product above, σ does not range over empty.

4.3 Neural Sequence-to-Sequence Model

We model all of the conditional probability factors in (2) using a neural sequence-to-sequence model with parameters $\boldsymbol{\theta}$. Specifically, we follow Kann and Schütze (2016) and use an LSTM-based sequence-to-sequence model (Sutskever et al.,

[6] Below, we will define the factors so that the generated $\boldsymbol{m}$ does–usually—fall in class c. We will ensure that if two slots are syncretic in class c, then their forms are in fact equal in $\boldsymbol{m}$. But non-syncretic slots will also have a (tiny) probability of equal forms, so the model $q_{\boldsymbol{\theta}}(\boldsymbol{m} \mid c)$ is *deficient*—it sums to slightly < 1 over the paradigms $\boldsymbol{m}$ in class c.

2014) with attention (Bahdanau et al., 2015).This is the state of the art in morphological reinflection, i.e., the conversion of one inflected form to another (Cotterell et al., 2016).

For example, in German, $q_{\boldsymbol{\theta}}(\boldsymbol{M}.\textsf{nompl} = \textit{Hände} \mid \boldsymbol{M}.\textsf{nomsg} = \textit{Hand}, C = 3)$ is given by the probability that the seq2seq model assigns to the output sequence `H ä n d e` when given the input sequence

```
H a n d C=3 IN=NOM IN=SG OUT=NOM OUT=PL
```

The input sequence indicates the parent slot (nominative singular) and the child slot (nominative plural), by using special characters to specify their feature bundles. This tells the seq2seq model what kind of inflection to do. The input sequence also indicates the paradigm class c. Thus, we are able to use only a single seq2seq model, with parameters $\boldsymbol{\theta}$, to handle all of the conditional distributions in the entire model. Sharing parameters across conditional distributions is a form of multi-task learning and may improve generalization to held-out data.

As a special case, if σ and σ' are syncretic within class c, then we define $q_{\boldsymbol{\theta}}(\boldsymbol{M}.\sigma = w \mid \boldsymbol{M}.\sigma' = w', C = c)$ to be 1 if $w = w'$ and 0 otherwise. The seq2seq model is skipped in such cases: it is only used on non-syncretic parent-child pairs. As a result, if class c has 5 slots that are all syncretic with one another, 4 of these slots can be derived by deterministic copying. As they are completely predictable, they contribute $\log 1 = 0$ bits to the paradigm entropy. The method in the next section will always favor a tree structure that exploits copying. As a result, the extra 4 slots will not increase the i-complexity, just as they do not increase the e-complexity.

We train the parameters $\boldsymbol{\theta}$ on *all* non-syncretic slot pairs in the training set. Thus, a paradigm with n distinct forms contributes n^2 training examples: each form in the paradigm is predicted from each of the $n-1$ other forms, and from the $\textsf{empty}$ form. We use maximum-likelihood training (see §7.2).

4.4 Structure Selection

Given a model $q_{\boldsymbol{\theta}}$, we can decompose its entropy $H(q_{\boldsymbol{\theta}})$ into a weighted sum of conditional entropies

$$H(\boldsymbol{M}) = H(C) + \sum_{c \in \mathcal{C}} p(C{=}c) H(\boldsymbol{M} \mid C{=}c) \tag{3}$$

where

$$H(\boldsymbol{M} \mid C{=}c) = \sum_{\sigma \in c} H(\boldsymbol{M}.\sigma \mid \boldsymbol{M}.\mathrm{pa}_c(\sigma), C{=}c) \tag{4}$$

The cross-entropy $H(p, q_{\boldsymbol{\theta}})$ has a similar decomposition. The only difference is that all of the (conditional) entropies are replaced by (conditional) cross-entropies, meaning that they are estimated using a held-out sample from p rather than $q_{\boldsymbol{\theta}}$. The log-probabilities are still taken from $q_{\boldsymbol{\theta}}$.

It follows that given a fixed $\boldsymbol{\theta}$ (as trained in the previous section), we can minimize $H(p, q_{\boldsymbol{\theta}})$ by choosing the tree for each class c that minimizes the cross-entropy version of (4).

How? For each class c, we select the minimum-weight directed spanning tree over the n_c slots used by that class, as computed by the Chu-Liu-Edmonds algorithm (Edmonds, 1967).[7] The weight of each potential directed edge $\sigma' \to \sigma$ is the conditional cross-entropy $H(\boldsymbol{M}.\sigma \mid \boldsymbol{M}.\sigma', C = c)$ under the seq2seq model trained in the previous section, so equation (4) implies that the weight of a tree is the cross-entropy we would get by selecting that tree.[8] In practice, we estimate the conditional cross-entropy for the non-syncretic slot pairs using a held-out development set (not the test set). For syncretic slot pairs, which are handled by copying, the conditional cross-entropy is always 0, so edges between syncretic slots can be selected free of cost.

After selecting the tree, we could retrain the seq2seq parameters $\boldsymbol{\theta}$ to focus on the conditional distributions we actually use, training on only the slot pairs in each training paradigm that correspond to an edge in the paradigm's class. However, our present experiments do not do this. In fact, training on all n^2 pairs can be seen as a form of multi-task regularization that may improve the model.

[7] Similarly, Chow and Liu (1968) find the best tree-shaped *undirected* graphical model by computing the highest-weighted *undirected* spanning tree. We require a directed model instead because §4.3 provides *conditional* distributions.

[8] Where the weight of the tree is taken to include the weight of the special edge $\textsf{empty} \to \sigma$ to the root node σ. Thus, for each slot σ, the weight of $\textsf{empty} \to \sigma$ is the cost of selecting σ as the root. It is an estimate of $H(\boldsymbol{M}.\sigma \mid C = c)$, the difficulty of predicting the σ form without any parent.

In the implementation, we actually decrement the weight of every edge $\sigma' \to \sigma$ (including when $\sigma' = \textsf{empty}$) by the weight of $\textsf{empty} \to \sigma$. This does not change the optimal tree, because it does not change the relative weights of the possible parents of σ. However, it ensures that every σ now has root cost 0, as required by the Chu-Liu-Edmonds algorithm (which does not consider root costs). Notice that since $H(X) - H(X \mid Y) = I(X; Y)$, the decremented weight is actually an estimate of $-I(\boldsymbol{M}.\sigma; \boldsymbol{M}.\sigma')$. Thus, finding the min-weight tree is equivalent to finding the tree that maximizes the total mutual information on the edges, just like the Chow-Liu algorithm (Chow and Liu, 1968).

5 From Paradigm Entropy to i-Complexity

Having defined a way to approximate paradigm entropy, $H(\boldsymbol{M})$, we finally operationalize our measure of i-complexity for a language.

One Paradigm Class. We start with the simple case where the language has a single paradigm class: $\mathcal{C} = \{c\}$. Our initial idea was to define i-complexity as bits per form, $H(\boldsymbol{M}) \,/\, |c|$, where $|c|$ is the enumerative complexity—the number of distinct forms in the paradigm.

However, $H(\boldsymbol{M})$ reflects not only the language's morphological complexity, but also its "lexical complexity." Some of the bits needed to specify a lexeme's paradigm $\boldsymbol{m}$ are necessary merely to specify the stem. A language whose stems are numerous or highly varied will tend to have higher $H(\boldsymbol{M})$, but we do not wish to regard it as *morphologically* complex simply on that basis. We can decompose $H(\boldsymbol{M})$ into

$$H(\boldsymbol{M}) = \underbrace{H(\boldsymbol{M}.\check{\sigma})}_{\text{lexical entropy}} + \underbrace{H(\boldsymbol{M} \mid \boldsymbol{M}.\check{\sigma})}_{\text{morphological entropy}} \tag{5}$$

where $\check{\sigma}$ here denotes the most predictable slot,

$$\check{\sigma} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\sigma} H(\boldsymbol{M}.\sigma) \tag{6}$$

and we estimate $H(\boldsymbol{M}.\sigma)$ for any σ using the seq2seq distribution $q_{\boldsymbol{\theta}}(\boldsymbol{M}.\sigma = w \mid M.\textsf{empty} = \epsilon)$, which can be regarded as a model for generating forms of slot σ from scratch.

We will refer to $\check{\sigma}$ as the **lemma** since it gives in some sense the simplest form of the lexeme, although it is not necessarily the slot that lexicographers use as the citation form for the lexeme.

We now define i-complexity as the entropy per form when predicting the remaining forms of $\boldsymbol{M}$ from the lemma:

$$\frac{H(\boldsymbol{M} \mid \boldsymbol{M}.\check{\sigma})}{|c| - 1} \tag{7}$$

where the numerator can be obtained by subtraction via equation (5). This is a fairer representation of the morphological irregularity, e.g., the average difficulty in predicting the inflectional ending that is added to a given stem. Notice that if $|c| = 1$ (an isolating language), the morphological complexity is appropriately undefined, since no inflectional endings are ever added to the stem.

If we had allowed the lexical entropy $H(\boldsymbol{M}.\check{\sigma})$ to remain in the numerator, then a language with larger e-complexity $|c|$ would have amortized that term over more forms—meaning that larger e-complexity would have tended to lead to lower i-complexity, other things equal. By removing that term from the numerator, our definition (7) eliminates this as a possible reason for the observed tradeoff between e-complexity and i-complexity.

Multiple Paradigm Classes. Now, we consider the more general case where multiple paradigm classes are allowed: $|\mathcal{C}| \geq 1$. Again we are interested in the entropy per non-lemma form. The i-complexity is

$$\frac{H(C) + \sum_c p(C{=}c) H(\boldsymbol{M} \mid \boldsymbol{M}.\check{\sigma}(c), C{=}c)}{\sum_c p(C{=}c)(|c| - 1)} \tag{8}$$

where

$$\check{\sigma}(c) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\sigma} H(\boldsymbol{M}.\sigma \mid C{=}c) \tag{9}$$

In the case where $|c|$ and $\check{\sigma}(c)$ are constant over all C, this reduces to equation (7). This is because the numerator is esssentially an expanded formula for the conditional entropy in (7)—the only wrinkle is that different parts of it condition on different slots.

To estimate equation (8) using a trained model q and a held-out test set, we follow §3.3 by estimating all $-\log p(\cdots)$ terms in the entropies with our model surprisals $-\log q(\cdots)$, but using the empirical probabilities on the test set for all other $p(\cdots)$ terms including $p(C = c)$. Suppose the test set paradigms are $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ with classes $c_1, \ldots, c_N$ respectively. Then taking $q = q_{\boldsymbol{\theta}}$, our final estimate of the i-complexity (8) works out to

$$\frac{\sum_{i=1}^{N} - \begin{pmatrix} \log q(C{=}c_i) \\ + \log q(\boldsymbol{M} = \boldsymbol{m}_i \mid C{=}c_i) \\ - \log q(\boldsymbol{M}.\check{\sigma}(c_i) = \boldsymbol{m}_i.\check{\sigma}(c_i) \mid C{=}c_i) \end{pmatrix}}{\sum_{i=1}^{N} |c_i| - 1} \tag{10}$$

where we have multiplied both the numerator and denominator by N. In short, the denominator is the total number of non-lemma forms in the test set, and the numerator is the total number of bits that our model needs to predict these forms (including the paradigm shapes c_i) given the lemmas. The numerator of equation (10) is an upper bound on the numerator of equation (8) since it uses (conditional) cross-entropies rather than (conditional) entropies.

	SINGULAR				PLURAL			
CLASS	NOM	GEN	ACC	VOC	NOM	GEN	ACC	VOC
1	*-os*	*-u*	*-on*	*-e*	*-i*	*-on*	*-us*	*-i*
2	*-s*	*-∅*	*-∅*	*-∅*	*-es*	*-on*	*-es*	*-es*
3	*-∅*	*-s*	*-∅*	*-∅*	*-es*	*-on*	*-es*	*-es*
4	*-∅*	*-s*	*-∅*	*-∅*	*-is*	*-on*	*-is*	*-is*
5	*-o*	*-u*	*-o*	*-o*	*-a*	*-on*	*-a*	*-a*
6	*-∅*	*-u*	*-∅*	*-∅*	*-a*	*-on*	*-a*	*-a*
7	*-os*	*-us*	*-os*	*-os*	*-i*	*-on*	*-i*	*-i*
8	*-∅*	*-os*	*-∅*	*-∅*	*-a*	*-on*	*-a*	*-a*

Table 1: Structuralist analysis of Modern Greek nominal inflection classes. (Ralli, 1994, 2002).

6 A Methodological Comparison to Ackerman and Malouf (2013)

Our formulation of the low-entropy principle differs somewhat from Ackerman and Malouf (2013); the differences are highlighted below.

Heuristic Approximation to p. Ackerman and Malouf (2013) first construct what we regard as a heuristic approximation to the joint distribution p over forms in a paradigm. They first provide a structuralist decomposition of words into their constituent morphemes. Then, they consider a distribution $r(\boldsymbol{m}.\sigma \mid \boldsymbol{m}.\sigma')$ that builds new forms by stochastically replacing morphemes. In contrast to our neural sequence-to-sequence approach, this distribution unfortunately does *not* have support over Σ^* and, thus, cannot consider changes other than substitution of morphological exponents.

As a concrete example of r, consider Table 1's (Simplified) Modern Greek example from Ackerman and Malouf (2013). The conditional distribution $r(\boldsymbol{m}.\textsf{gen;sg} \mid \boldsymbol{m}.\textsf{acc;pl} = \ldots\textit{-i})$ over genitive singular forms is peaked since there is exactly one possible transformation: substituting *-us* for *-i*. Other conditional distributions for Modern Greek are less peaked: Ackerman and Malouf (2013) estimated that $r(\boldsymbol{m}.\textsf{nom;sg} \mid \boldsymbol{m}.\textsf{acc;pl} = \ldots\textit{-a})$ swaps *-a* for ∅ with probability $2/3$ and for *-o* with probability $1/3$. We reiterate that no other output has positive probability under their model, e.g., swapping *-a* for *-es* or ablaut of a stem vowel.

Average Conditional Entropy. The second difference is their use of the pairwise conditional entropies between cells. They measure the complexity of the entire paradigm by the average conditional entropy:

$$\frac{1}{n^2 - n} \sum_{\sigma} \sum_{\sigma' \neq \sigma} H(\boldsymbol{M}.\sigma \mid \boldsymbol{M}.\sigma'). \qquad (11)$$

This differs from our tree-based measure, in which an irregular form only needs to be derived from its parent—possibly a similar or even syncretic irregular form—rather than from *all* other forms in the paradigm. So it "only needs to pay once" and it even "shops around for the cheapest deal. Also, in our measure, the lemma does not "pay" at all.

They measure conditional entropies, which are simple to compute because their model q is simple. (Again, it only permits a small number of possible outputs for each input, based on the finite set of allowed morpheme substitutions that they annotated by hand.) In contrast, our estimate uses conditional *cross*-entropies, asking whether our q can predict real held-out forms distributed according to p.

6.1 Critique of Ackerman and Malouf (2013)

Now, we offer a critique of Ackerman and Malouf (2013) on three points: (i) different linguistic theories dictating how words are subdivided into morphemes may offer different results, (ii) certain types of morphological irregularity, particularly suppletion, aren't handled, and (iii) average conditional entropy overestimates the i-complexity in comparison to joint entropy.

Theory-Dependent Complexity. We consider a classic example from English morphophonology that demonstrates the effect of the specific analysis chosen. In regular English plural formation, the speaker has three choices: [z], [s] and [ɨz]. Here are two potential analyses. One could treat this as a case of pure allomorphy with three potential, unrelated suffixes. Under such an analysis, the entropy will reflect the empirical frequency of the three possibilities found in some data set: roughly, $1/4 \log 1/4 + 3/8 \log 3/8 + 3/8 \log 3/8 \approx 1.56127$. On the other hand, if we assume a different model with a unique underlying affix /z/, which is attached and then converted to either [z], [s] or [ɨz] by an application of perfectly regular phonology, this part of the morphological system of English has entropy of 0—one choice. See Kenstowicz (1994, p.72) for a discussion of these alternatives from a theoretical standpoint. Note that our goal is not to advocate for one of these analyses, but merely to suggest that Ackerman and Malouf (2013)'s quantity is analysis-dependent.[9] In contrast, our approach is theory-agnostic in that we jointly learn surface-to-surface transformations, reminiscent of

[9] Other suggested quantifications of morphological complexity have relied on a similar assumption (e.g. Bane, 2008).

a-morphorous morphology (Anderson, 1992), and thus our estimate of paradigm entropy does not suffer this drawback. Indeed, our assumptions are limited—recurrent neural networks are universal approximators. It has been shown that any computable function can be computed by some finite RNN (Siegelmann and Sontag, 1991, 1995). Thus, the only true assumption we make of morphology is mild: we assume it is Turing-computable. That behavior is Turing-computable is a rather fundamental tenet of cognitive science (McCulloch and Pitts, 1943; Sobel and Li, 2013).

In our approach, theory dependence is primarily introduced through the selection of slots in our paradigms, which is a form of bias that would be present in any human-derived set of morphological annotations. A key example of this is the way in which different annotators or annotation standards may choose to limit or expand syncretism — situations where the same string-identical form may fill multiple different paradigm slots. For example, Finnish has two accusative inflections for nouns and adjectives, one always coinciding in form with the nominative and the other coinciding with the genitive. Many grammars therefore omit these two slots in the paradigm entirely, while some include them. Depending on which linguistic choice annotators make, the language could appear to have more or fewer paradigm slots. We have carefully defined our e-complexity and i-complexity metrics so that they are not sensitive to these choices.

As a second example of annotation dependence, different linguistic theories might disagree about which distinctions constitute productive inflectional morphology, and which are derivational or even fixed lexical properties. For example, our dataset for Turkish treats causative verb forms as *derivationally* related lexical items. The number of apparent slots in the Turkish inflectional paradigms is reduced because these forms were excluded.

Morphological Irregularity. A second problem with the model in Ackerman and Malouf (2013) is its inability to treat certain kinds of irregularity, particularly cases of suppletion. As far as we can tell, the model is incapable of evaluating cases of morphological suppletion unless they are explicitly encoded in the model. Consider, again, the case of the English suppletive past tense form *went*— if one's analysis of the English base is effectively a distribution of the choices add [d], add [t] and [ɨd], one will assign probability 0 to *went* as the past tense of *go*. We highlight the importance of this point because suppletive forms are certainly very common in academic English: the plural of *binyan* is *binyanim* and the plural of *lemma* is *lemmata*. It is unlikely that native English speakers possess even a partial model of Hebrew and Greek nominal morphology—a more plausible scenario is simply that these forms are learned by rote. As speakers and hearers are capable of producing and understanding these forms, we should demand the same capacity of our models. Not doing so also ties into the point in the previous section about theory-dependence since it is ultimately the linguist—supported by some theoretical notion—who decides which forms are deemed irregular and hence left out of the analysis. We note that these restrictive assumptions are relatively common in the literature, e.g., Allen and Becker (2015)'s sublexical learner is likewise incapable of placing probability mass on irregulars.[10]

Average Conditional Entropy versus Joint Entropy. Finally, we take issue with the formulation of paradigm entropy as average conditional entropy, as exhibited in equation (11). For one, it does not correspond to the entropy of any actual joint distribution $p(\boldsymbol{M})$, and has no obvious mathematical interpretation. Second, it is Priscian (Robins, 2013) in its analysis in that any form can be generated from any other, which, in practice, will cause it to overestimate the i-complexity of a morphological system. Consider the German dative plural *Händen* (from the German *Hand* "hand"). Predicting this form from the nominative singular *Hand* is difficult, but predicting it from the nominative plural *Hände* is trivial: just add the suffix *-n*. In Ackerman and Malouf (2013)'s formulation, $r(\textit{Händen} \mid \textit{Hand})$ and $r(\textit{Händen} \mid \textit{Hände})$ both contribute to the paradigm's entropy with the former substantially raising the quantity. Our method in §4.4 is able to select the second term and regard *Händen* as predictable once *Hände* is in hand.

7 Experiments

Our experimental design is now fairly straightforward: plot e-complexity versus i-complexity over as many languages as possible, We then devise a numerical test of whether the complexity trade-off conjecture (§1) appears to hold.

[10] In the computer science literature, it is far more common to construct distributions with support over Σ^* (Paz, 2003; Bouchard-Côté et al., 2007; Dreyer et al., 2008; Cotterell et al., 2014), which do not have this problem.

7.1 Data and UniMorph Annotation

At the moment, the largest source of annotated full paradigms is the UniMorph dataset (Sylak-Glassman et al., 2015; Kirov et al., 2018), which contains data that have been extracted from Wiktionary, as well as other morphological lexica and analyzers, and then converted into a universal format. A partial subset of UniMorph has been used in the running of the SIGMORPHON-CoNLL 2017 and 2018 shared tasks on morphological inflection generation (Cotterell et al., 2017a, 2018b).

We use verbal paradigms from 33 typologically diverse languages, and nominal paradigms from 18 typologically diverse languages. We only considered languages that had at least 700 fully annotated verbal or nominal paradigms, as the neural methods we deploy required a large amount of training example to achieve high performance.[11] As the neural methods require a large set of annotated training examples to achieve high performance, it is difficult to use them in a lower-resource scenario.

To estimate a language's e-complexity (§2.2.1), we average over all paradigms in the UniMorph inflected lexicon.

To estimate i-complexity, we first partition those paradigms into training, development and test sets. We identify the paradigm classes from the training set (§4.1). We also use the training set to train the parameters $\boldsymbol{\theta}$ of our conditional distribution (§4.3), then estimate conditional entropies on the development set and use Edmonds's algorithm to select a global model structure for each class (§4.4). Now we evaluate i-complexity on the test set (equation (10)). Using held-out test data gives an unbiased estimate of a model's predictive ability, which is why it is standard practice in statistical NLP, though less common in quantitative linguistics.

7.2 Experimental Details

We experiment separately on nominal and verbal lexicons. For i-complexity, we hold out at random 50 full paradigms for the development set, and 50 other full paradigms for the test set.

For comparability across languages, we tried to ensure a "standard size" for the training set $\mathcal{D}_{\text{train}}$. We sampled it from the remaining data using two different designs, to address the fact that different languages have different-size paradigms.

Equal Number of Paradigms ("purple scheme"). In the first regime, $\mathcal{D}_{\text{train}}$ (for each language) is derived from 600 randomly chosen non-held-out paradigms $\boldsymbol{m}$. We trained the reinflection model in §4.4 on all non-syncretic pairs within these paradigms, as described in §4.3. This disadvantages languages with small paradigms, as they train on fewer pairs.

Equal Number of Pairs ("green scheme"). In the second regime, we trained the reinflection model in §4.4 on 60,000 non-syncretic pairs $(\boldsymbol{m}.\sigma', \boldsymbol{m}.\sigma)$ (where σ' may be empty) sampled without replacement from the non-held-out paradigms.[12] This matches the amount of training data, but may disadvantage languages with large paradigms, since the reinflection model will see fewer examples of any individual mapping between paradigm slots. We call this the "green scheme."

Model and Training Details. We train the seq2seq-with-attention model using the OpenNMT toolkit (Klein et al., 2017). We largely follow the recipe given in Kann and Schütze (2016), the winning submission on the 2016 SIGMORPHON shared task for inflectional morphology. Accordingly, we use a character embedding size of 300, and 100 hidden units in both the encoder and decoder. Our gradient-based optimization method was AdaDelta (Zeiler, 2012) with a minibatch size of 80. We trained for 20 epochs, which yielded 20 models via early stopping. We selected the model that achieved the highest average $\log p(\boldsymbol{m}.\sigma \mid \boldsymbol{m}.\sigma')$ on (σ', σ) pairs from the development set.

8 Results and Analysis

Our results are plotted in Figure 2, where each dot represents a language. We see little difference between the green and the purple training sets, though it was not clear *a priori* that this would be so.

The plots appear to show a clear trade-off between i-complexity and the e-complexity. We now provide quantitative support for this impression, by constructing a statistical significance test. Visually,

[11] Focusing on data-rich languages should also help mitigate sample bias caused by variable-sized dictionaries in our database. In many languages, irregular words are also very frequent and may be more likely to be included in a dictionary first. If that's the case, smaller dictionaries might have lexical statistics skewed toward irregulars more so than larger dictionaries. In general, larger dictionaries should be more representative samples of a language's broader lexicon.

[12] For a few languages, fewer than 60,000 pairs were available, in which case we used all pairs.

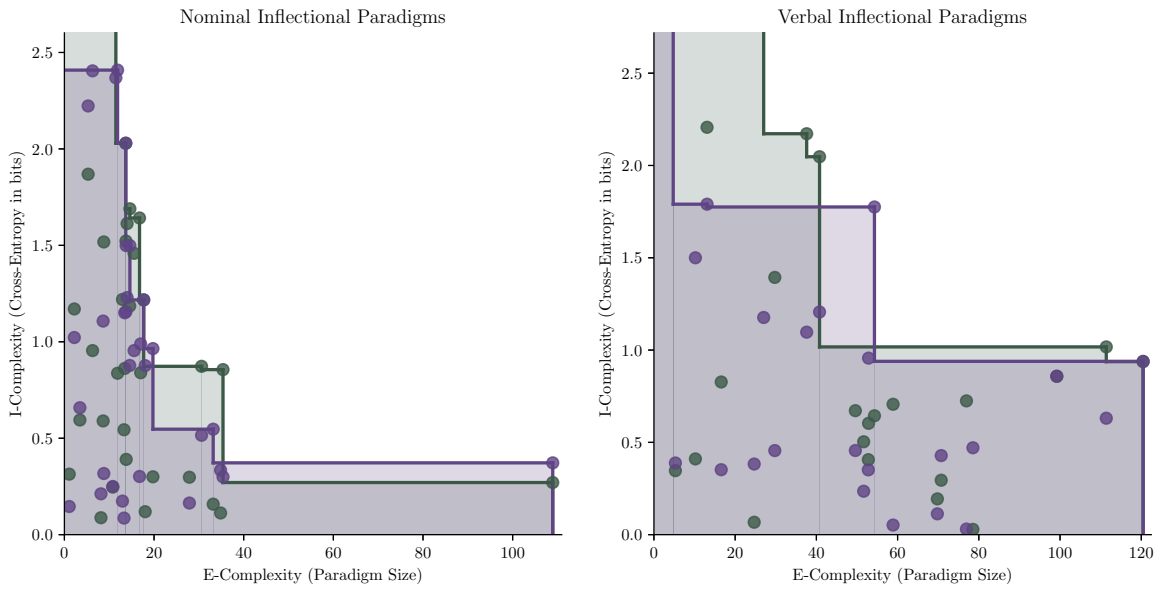


Figure 2: The x-axis is our measure of e-complexity, the average number of distinct forms in a paradigm. The y-axis is our estimate of i-complexity, the average bits per distinct non-lemma form. We overlay purple and green graphs (§7.2): all the purple points are trained on the same number of paradigms, and all the green points are trained on about the same number of slot pairs. The purple curve is the Pareto curve for the purple points, and the area under it is shaded in purple; similarly for green.

our low-entropy trade-off conjecture boils down to the claim that languages cannot exist in the upper right-hand corner of the graph, i.e., they cannot have both high e-complexity and high i-complexity. In other words, the upper-right hand corner of the graph is "emptier" than it would be by chance.

How can we quantify this? The **Pareto curve** for a multi-objective optimization problem shows, for each x, the maximum value y of the second objective that can be achieved while keeping the first objective $\geq x$ (and vice-versa). This is shown in Figure 2 as a step curve, showing the maximum i-complexity y that was actually achieved for each level x of e-complexity. This curve is the tightest non-increasing function that upper-bounds all of the observed points: we have no evidence from our sample of languages that any language can appear above the curve.

We say that the upper right-hand corner is "empty" to the extent that the area under the Pareto curve is small. To ask whether it is indeed emptier than would be expected by chance, we perform a nonparametric permutation test that destroys the claimed correlation between the e-complexity and i-complexity values. From our observed points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, we can stochastically construct a new set of points $\{(x_1, y_{\sigma(1)}), \ldots, (x_m, y_{\sigma(m)})\}$ where σ is a permutation of $1, 2, \ldots, m$ selected uniformly at random. The resulting scatterplot is what we would expect under the null hypothesis of no correlation. Our p-value is the probability that the new scatterplot has an even emptier upper right-hand corner—that is, the probability that the area under the null-hypothesis Pareto curve is $\leq$ the area under the actually observed Pareto curve. We estimate this probability by constructing 10,000 random scatterplots.

In the purple training scheme, we find that the upper right-hand corner is significantly empty, with $p < 0.021$ and $p < 0.037$ for the verbal and nominal paradigms, respectively. In the green training scheme, we find that the upper right-hand corner is significantly empty with $p < 0.032$ and $p < 0.024$ in the verbal and nominal paradigms, respectively.

9 Future Directions

Frequency. Ackerman and Malouf hypothesized that i-complexity is bounded, and we have demonstrated that the bounds are stronger when e-complexity is high. This suggests further investigation as to *where* in the language these bounds apply. Such bounds are motivated by the notion that naturally occurring languages must be learnable. Presumably, languages with large paradigms need to be regular *overall*, because in such a language,

the *average* word type is observed too rarely for a learner to memorize an irregular surface form for it. Yet even in such a language, *some* word types are frequent, because some lexemes and some slots are especially useful. Thus, if learnability of the lexicon is indeed the driving force,[13] then we should make the finer-grained prediction that irregularity may survive in the more frequently observed word types, regardless of paradigm size. Rarer forms are more likely to be predictable—meaning that they are either regular, or else irregular in a way that is predictable from a related frequent irregular (Cotterell et al., 2018a).

Dynamical models. We could even investigate directly whether patterns of morphological irregularity can be explained by the evolution of language through time. Languages may be shaped by natural selection or, more plausibly, by noisy transmission from each generation to the next (Hare and Elman, 1995; Smith et al., 2008), in a natural communication setting where each learner observes some forms more frequently than others. Are naturally occurring inflectional systems more learnable (at least by machine learning algorithms) than would be expected by chance? Do artificial languages with unusual properties (for example, unpredictable rare forms) tend to evolve into languages that are more typologically natural?

We might also want to study whether children's morphological systems increase in i-complexity as they approach the adult system. Interestingly, this definition of i-complexity could also explain certain issues in first language acquisition, where children often overregularize (Pinker and Prince, 1988): they impose the regular pattern on irregular verbs, producing forms like ~~*runned*~~ instead of *ran*. Children may initially posit an inflectional system with lower i-complexity, before converging on the true system, which has higher i-complexity.

Phonology Plus Orthography. A human learner of a written language also has access to phonological information that could affect predictability. One could for example jointly model all the written *and spoken* forms within each paradigm, where the Bayesian network may sometimes predict a spoken slot from a written slot or vice-versa.

Moving Beyond the Forms. The complexity of morphological inflection is only a small bit of the larger question of morphological typology. We have left many bits unexplored. In this paper, we have predicted orthographic forms from morphosyntactic feature bundles. Ideally, we would like to also predict which morphosyntactic bundles are realized as words within a language, and which bundles are syncretic. That is, what paradigm classes are plausible or implausible?

In addition, our current treatment depends upon a paradigmatic treatment of morphology, which is why we have focused on inflectional morphology. In contrast, derivational morphology is often viewed as syntagmatic.[14] Can we devise quantitative formulation of derivational complexity—for example, extending to polysynthetic languages?

10 Conclusions

We have provided clean mathematical formulations of enumerative and integrative complexity of inflectional systems, using tools from generative modeling and deep learning. With an empirical study on noun and verb systems in 36 typologically diverse languages, we have exhibited a Pareto-style tradeoff between the e-complexity and i-complexity of morphological systems. In short, a morphological system can mark a large number of morphosyntactic distinctions, as Finnish, Turkish and other agglutinative and polysynthetic languages do; or it may have a high-level of unpredictability (irregularity); or neither.[15] But it cannot do both.

The NLP community often focuses on e-complexity and views a language as morphologically complex if it has a profusion of unique forms, even if they are very predictable. The reason is probably our habit of working at the word-level, so that all forms not found in the training set are out-of-vocabulary (OOV). Indeed, NLP practitioners often use high OOV rates as a proxy for defining morphological complexity. However, as NLP moves to the character-level, we will need other definitions of morphological richness. A language like Hungarian with almost perfectly predictable morphology may be easier to process than a language like German with an abundance of irregularity.

[13] Rather than, say, description length of the lexicon (Rissanen and Ristad, 1994).

[14] For paradigmatic treatments of derivational morphology, see Cotterell et al. (2017c) for a computational perspective and the references therein for theoretical perspectives.

[15] A language is under no obligation to be morphologically rich—it may have low e-complexity and i-complexity. Carstairs-McCarthy (2010) has pointed out that languages need not have morphology at all, though they must have phonology and syntax.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant No. 1718846. We want to thank Rob Malouf for providing extensive and very helpful feedback, along with the anonymous reviewers.

References

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

Blake Allen and Michael Becker. 2015. Learning alternations from surface forms with sublexical phonology. *Unpublished manuscript, University of British Columbia and Stony Brook University. Available as lingbuzz/002503*.

Stephen R. Anderson. 1992. *A-morphous Morphology*, volume 62. Cambridge University Press.

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. Number 1 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Matthew Baerman. 2015. *The Oxford Handbook of Inflection*. Oxford Handbooks in Linguistic. Part II: Paradigms and their Variants.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2015. Understanding and measuring morphological complexity: An introduction. In Matthew Baerman, Dunstan Brown, and Greville G. Corbett, editors, *Understanding and measuring morphological complexity*. Oxford University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Max Bane. 2008. Quantifying and measuring morphological complexity. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pages 69–76.

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

Leonard Bloomfield. 1933. *Language*. University of Chicago Press. Reprint edition (October 15, 1984).

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Andrew Carstairs-McCarthy. 2010. *The Evolution of Morphology*, volume 14. Oxford University Press.

C. K. Chow and Cong N. Liu. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2018a. On the diachronic stability of irregularity in inflectional morphology. *arXiv preprint arXiv:1804.08262v1*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Gèraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018b. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017a. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, Vancouver, Canada. Association for Computational Linguistics.

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*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2014. Stochastic contextual edit distance and probabilistic FSTs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 625–630, Baltimore, Maryland. Association for Computational Linguistics.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics (TACL)*, 3:433–447.

Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017b. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 759–765, Valencia, Spain. Association for Computational Linguistics.

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017c. Paradigm completion for derivational morphology. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–731, Copenhagen, Denmark. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 101–110, Singapore.

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 (EMNLP)*, pages 616–627, Edinburgh.

Markus Dreyer, Jason Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1080–1089, Honolulu, Hawaii. Association for Computational Linguistics.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.

David Gil. 1994. The structure of Riau Indonesian. *Nordic Journal of Linguistics*, 17(2):179–200.

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

Charles F. Hockett. 1958. *A Course In Modern Linguistics*. The MacMillan Company.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Michael J. Kenstowicz. 1994. *Phonology in Generative Grammar*. Blackwell Oxford.

Aleksandr E. Kibrik. 1998. Archi (Caucasian – Daghestanian). In *The Handbook of Morphology*, pages 455–476. Blackwell Oxford.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of LREC 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Warren S. McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.

John McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5(2):125–66.

Yoon Mi Oh. 2015. *Linguistic Complexity and Information: Quantitative Approaches*. Ph.D. thesis, Université de Lyon, France.

Azaria Paz. 2003. *Probabilistic Automata*. John Wiley and Sons.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87(3):539–558.

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

Angela Ralli. 1994. Feature representations and feature-passing operations in Greek nominal inflection. In *Proceedings of the 8th Symposium on English and Greek Linguistics*, pages 19–46.

Angela Ralli. 2002. The role of morphology in gender determination: evidence from Modern Greek. *Linguistics*, 40(3; ISSU 379):519–552.

Jorma Rissanen and Eric S. Ristad. 1994. Language acquisition in the MDL framework. In Eric S. Ristad, editor, *Language Computation*. American Mathematical Society, Philadelphia.

Robert Henry Robins. 2013. *A Short History of Linguistics*. Routledge.

Benoît Sagot. 2013. Comparing complexity measures. In *Computational Approaches to Morphological Complexity*, Paris, France.

Edward Sapir. 1921. *Language: An Introduction to the Study of Speech*.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27.

Hava T. Siegelmann and Eduardo D. Sontag. 1991. Turing computability with neural nets. *Applied Mathematics Letters*, 4(6):77–80.

Hava T. Siegelmann and Eduardo D. Sontag. 1995. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1):132–150.

Kenny Smith, Michael L. Kalish, Thomas L. Griffiths, and Stephan Lewandowsky. 2008. Cultural transmission and the evolution of human behaviour. *Philosophical Transactions B*.

Carolyn P. Sobel and Paul Li. 2013. *The Cognitive Sciences: An Interdisciplinary Approach*. Sage Publications.

Andrew Spencer. 1991. *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Wiley-Blackwell.

Thomas Stolz, Hitomi Otsuka, Aina Urdze, and Johan van der Auwera. 2012. *Irregularity in Morphology (and beyond)*, volume 11. Walter de Gruyter.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (Unimorph schema). Technical report, Johns Hopkins University.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701v1*.