

An analogical approach to the typology of inflectional complexity

Matías Guzmán Naranjo
University of Freiburg

ABSTRACT

This paper studies the inflectional complexity of nouns, verbs and adjectives in 137 datasets, across 71 languages. I follow Ackerman and Malouf (2013) in distinguishing between E(numerative) complexity and I(ntegrative) complexity. The first one encompasses aspects of inflection, like the number of principal parts, paradigm size, and number of exponents, while the second one captures the implicative relations between paradigm cells (how difficult it is to predict one cell of a paradigm knowing a different cell). I provide a formalism and computational implementation to estimate both I- and E-complexity expressed through Word and Paradigm morphology (Blevins 2006, 2016), which is flexible and powerful enough for typological research. The results show that, as suggested by Ackerman and Malouf (2013), I-complexity is relatively low across the languages in the sample, with only two clear exceptions (Navajo and Yaitepec-Chatino). The results also show that E-complexity can vary considerably cross-linguistically. Finally, I show there is a clear correlation between I- and E-complexity.

Keywords:
inflectional complexity, typology, analogy, Word and Paradigm morphology

The study of morphological complexity has a long history in linguistics and typology (see for example Greenberg 1960, for an early approach), and has seen a renewed interest in recent years (Ackerman and Malouf 2013; Cotterell *et al.* 2019; Bentz *et al.* 2022). However, there is very little unity or agreement regarding how we should measure inflectional complexity, and whether the proposed metrics are cross-linguistically comparable (Igartua and Santazilia 2018; Gutierrez-Vasques and Mijangos 2019; Bentz *et al.* 2022, 2016; Arkadiev and Gardani 2020, among many others). Igartua and Santazilia (2018, p. 439) for example, define morphological complexity as “the extent to which formal differences in inflectional paradigms are semantically or phonologically unmotivated” (i.e. the amount of allomorphy in a morphological system). In contrast, Sinnemäki and Di Garbo (2018, p. 8), following Bickel and Nichols (2007) and Bickel and Nichols (2013), define the inflectional complexity of a verb as: “the number of morphological categories expressed per word in a maximally inflected verb form.”

One key development in the study of inflectional complexity came from Ackerman and Malouf (2013), who propose a distinction between two fundamentally different types of inflectional complexity: Enumerative (E) complexity and Integrative (I) complexity. The first is the complexity in morphosyntactic distinctions and the way languages encode them (be it exponents, morphs, morphemes, etc.), while the second one is the difficulty a paradigm poses to speakers in terms of implicative relations. Ackerman and Malouf (2013, 429) provide the following definitions:

The I-complexity of an inflectional system reflects the difficulty that a paradigmatic system poses for language users (rather than lexicographers) in information-theoretic terms. (Ackerman and Malouf 2013, p. 429)

I-complexity measures how predictable the realisation of a lexeme is, given knowledge about one (or more) cells of its paradigm. This type of complexity measures implicational relations in a paradigm,

and it is not directly dependent on paradigm shape (what the actual realisations are, or how many cells a paradigm has). In contrast, E-complexity is defined as:

E-complexity [is given by] the number of exponents, inflectional classes, and principal parts (Ackerman and Malouf 2013, p. 429)

E-complexity has received considerable attention in the literature (Stump and Finkel 2013; Bentz *et al.* 2022; Finkel and Stump 2007; Baerman *et al.* 2015; Dressler 2011), from several different perspectives, including inflection class systems, paradigm size, principal parts and number of morphs or morphemes. Some of this work, however, faces some practical and theoretical challenges (see Section 2).

At the same time, while there are multiple computational proposals for capturing I-complexity (Bonami and Beniamine 2016; Cotterell *et al.* 2019; Guzmán Naranjo 2020; Ackerman and Malouf 2013; Marzi *et al.* 2019), most studies have looked at a relatively small samples (< 100 datasets) and the emphasis has not been on cross-linguistic comparison (although see Cotterell *et al.* 2019). This means that we still do not have a good picture of how I-complexity varies across languages and systems. For example, one still open question is how verb, noun and adjective paradigms compare cross-linguistically for the sake of consistency in terms of I-complexity.

The objectives of this paper are twofold: First and foremost, it presents a medium-scale typological study of morphological complexity from a Word and Paradigm perspective (Blevins 2006, 2016; Matthews 1972). And second, it presents a new technique for measuring morphological complexity and provides an efficient computational implementation of it. I argue that it is both feasible and desirable to work from a W&P perspective when doing cross-linguistic comparisons of inflectional systems. I also show that some fundamental problems in morphological typology can be completely bypassed when approached from a W&P perspective.

The paper is structured as follows: Section 2 gives a brief overview of the main ideas and approaches to morphological complexity, as well as word and paradigm morphology. Section 3 describes the datasets and the methods used in the paper. Section 4 presents the results, and Section 5 concludes.

This section presents a very brief overview of the main ideas of and trends in the morphological complexity literature from two different perspectives. It also discusses some key differences between morpheme-based and W&P approaches to morphology and argues that beyond theoretical considerations, there are practical reasons why the latter is preferable for doing cross-linguistic studies of the complexity of morphological inflection.

Due to the vast amount of research on the topic of morphological complexity (see for example Baerman *et al.* 2015, 2017; Miestamo *et al.* 2008; Bentz *et al.* 2022, for some overviews and recent takes), a full account of these topics is not feasible within the scope of this article, and I will concentrate on some of the more important works on the topic. Similarly, covering the whole debate between different types of morphological theories is not feasible, and I will only discuss some of the more concrete and practical issues.

2.1

Integrative-complexity

The initial work on I-complexity was approached using information theory, and it focused on measuring the conditional entropy between the cells of the paradigm of a lexeme, often using hand-extracted exponents for each cell (Ackerman and Malouf 2013; Bonami and Beniamine 2016; Blevins 2013; Palancar 2021; Parker and Sims 2020, among many others). More recent papers estimate the conditional entropy of a system using LSTMs¹ (Cotterell *et al.* 2019; Court *et al.* 2022) instead of directly calculating it based on extracted exponents.

I will illustrate I-complexity with two simple toy examples in Tables 1 and 2.² Both examples have three inflection classes with two

¹LSTMs are a type of neural network that performs sequence to sequence predictions. In this context, they are trained to predict fully inflected forms from other fully inflected forms (plus lexeme information). The entropy of the system is calculated on the network itself.

²The elements in each cell are meant to be the suffix (markers) which express the cell content. These are just examples, and the actual cell realisations could be achieved by suffixes, infixes, tones, etc.

and three cells, but the exponent structure is completely different. The system of Table 1, Language 1, only has two markers, *-i* and *-o*, while the system in Table 2, Language 2, contains 9 different markers: *-i*, *-e*, *-a*, *-u*, *-o*, *-∅*, *-ik*, *-ek*, *-æ*. Language 2 has a higher E-complexity both in terms of the number of exponents and paradigm size, however, the situation is reversed for I-complexity.

	Cell 1	Cell 2
class A	-i	-i
class B	-o	-i
class C	-o	-o

Table 1:
I-complexity Language 1

	Cell 1	Cell 2	Cell 3
class A	-i	-e	-a
class B	-u	-o	-∅
class C	-ik	-ek	-æ

Table 2:
I-complexity Language 2

Following Ackerman and Malouf (2013), we can measure the I-complexity of each system using conditional entropy. The entropy of a cell X, $H(X)$, in a paradigm can be calculated as:

$$(1) \quad H(X) = - \sum_i p(x_i) \log_2(p(x_i))$$

Where $p(x_i)$ can be calculated from the frequency of the exponents for a cell across inflection classes, and where, i ranges over contrastive exponents found in a cell. However, for illustration purposes, this example assumes that all inflection classes have the same number of lexemes, meaning we can let i range over inflection classes. For Language 1, the frequency of *-i* for Cell 1 is 1, and the frequency for *-o* is 2, meaning $p(-i) = 1/3$ and $p(-o) = 2/3$, which gives us $H(\text{Cell 1}) = 1/3 \log_2(1/3) + 2/3 \log_2(2/3) = 0.92$. This is a measure of how much information is required to capture Cell 1 for Language 1.

The conditional entropy of a cell X given knowledge of cell Y, $H(X|Y)$, can be calculated as:

$$(2) \quad H(X|Y) = H(X, Y) - H(Y)$$

$$(3) \quad = \sum_i \sum_j p(x_i, y_j) \log_2(p(x_i|y_j))$$

For Language 1, the conditional entropy $H(\text{Cell 1}|\text{Cell 2} = -i) = 1$, and $H(\text{Cell 1}|\text{Cell 2} = -o) = 0$. Then, the average conditional entropy $H(\text{Cell 1}|\text{Cell 2}) = 2/3$ (since *-i* appears in two inflection classes, while *-o* appears in 1). Knowing that for a lexeme Cell 2 has the realisation *-o* provides complete information about what its realisation in Cell 1 must be, namely *-o*. In contrast, knowing that the exponent for Cell 2 is *-i* does not provide information about the realisation of Cell 1 because a lexeme with *-i* for Cell 2, can either *-i* or *-o* in Cell 1. Because Language 1 has a symmetric structure, $H(\text{Cell 2}|\text{Cell 1})$ is also $2/3$, meaning that the average pairwise³ conditional entropy is $2/3$. The results for Language 2 are very different. In this case, every cell provides complete information about every other cell in the paradigm of a lexeme, which means that for all pairwise conditional entropy calculations the results are 0, and the average conditional entropy of Language 2 is 0. This very simplified example illustrates the fact that the average E-complexity of a language (measured in terms of paradigm size or the number of markers) is not necessarily correlated with its I-complexity.

While using conditional entropy is still a relatively popular method to estimate I-complexity, an alternative approach is based on the accuracy of classification, instead of conditional entropy (Guzmán Naranjo 2020; Bonami and Pellegrini 2022). Instead of measuring the amount of information a cell provides about another cell, one can train a classifier⁴ on the content of one cell of a lexeme to predict the realisation of another cell for that lexeme.

As an example of classification, if we are dealing with nominal inflection, we can train a classifier to predict the accusative singular from the nominative singular. The accuracy obtained by that classifier (under cross-validation) is then a measure of the I-complexity of the paradigm. If a classifier has a perfect accuracy of 1 predicting the

³ See Bonami and Beniamine (2016) for a method to calculate the conditional entropy taking multiple cells into account.

⁴ Here classifier is understood as any system which takes some word form as an input and assigns it to a class. The method used could be a rule-based system, logistic regression, neural network, etc. For the purposes of modelling inflection, we usually train classifiers on the phonology and semantics of the forms in question, and predict the inflection class from this information.

inflection class of all lexemes in a language, then we can say that there is effectively no I-complexity to that inflection class. The important point is that, just as with conditional entropy, the I-complexity of a system is mostly independent of the number of inflection classes or exponents in an inflectional system. If there is enough information for the classifier to have perfect accuracy, then the I-complexity of the system will be 0.

Using the previous example, the accuracy for Language 2 will be 1 for all cell pairs, because every cell provides complete information about every other cell in the paradigm of a lexeme, which means that the complexity of Language 2 is also 0. For Language 1, the accuracy of predicting (i.e the number of correct predictions over total number of items) Cell 1 from Cell 2 (and the other way around) is 2/3 (because on average we will be able to correctly predict the realisation 2 out of 3 times). This means that the average complexity of Language 1 is ~ 0.67 .

One advantage of using a predictive technique instead of estimating conditional entropy using LSTMs is that we can easily make use of classifiers that work well even on very small datasets. LSTMs, due to the way they are trained, can struggle with small datasets. Cotterell *et al.* (2019, 336), for example, restrict their study to languages with at least 700 lexemes, because the specific model requires relatively large datasets to achieve acceptable accuracies. As we will see in the results section, these much simpler models perform well on much smaller datasets.

Enumerative complexity

2.2

The initial definition of E-complexity covered the number of principal parts, exponents, and inflection classes. In this section, I will discuss some of the studies that have looked at these, and a few other aspects of E-complexity.

Principal parts

2.2.1

Principal parts are defined as the cells in the paradigm of a lexeme which a speaker needs to know in order to be able to deduce all other cells (Finkel and Stump 2007). For example, it is often proposed that the Latin verb system has 4 principal parts, which a speaker would

need to know in order to be able to produce all other inflected forms of the verb, these are the first person singular present indicative active, active present infinitive, first person singular perfect indicative active, and the passive perfect participle (or future participle) (Bennett 1918). While this is, in principle, a relatively straightforward way of quantifying the complexity of an inflectional system determining the number of principal parts is not straightforward, and will vary depending on the approach one takes to how principal parts should behave within and across paradigms (Finkel and Stump 2007). In this paper, I will not directly consider counting principal parts, but I will come back to the question during the discussion of the results.

2.2.2

Inflection classes

Measuring inflectional complexity in terms of the number of inflection classes is, in theory, straightforward: one simply counts how many inflection classes there are in a system. Although the idea of inflection classes might seem intuitive, the task of counting inflection classes is particularly difficult. Some early work on complexity approached the problem from this perspective (Carstairs 1983; Carstairs-McCarthy 1994), but it has lost favour during the past decade (Sims and Parker 2016). One of the reasons is the move towards questions of I-complexity, but another is that counting inflection classes is anything but simple. For example, Parker and Sims (2020) show how non-trivial it is to count inflection classes for Russian, a very well studied language. A similar conclusion is reached by Beniamine and Guzmán Naranjo (2021), who show that if taken at a surface level, it is difficult to determine the number of inflection classes a language can have (cf. Beniamine Forthcoming).

There are several reasons why counting inflection classes is particularly difficult, but it mainly boils down to the fact that identifying whether two lexemes belong to the same inflection class or not is not easy to operationalize. As a simple example, consider irregular verbs, or partially irregular verbs, or defective verbs. Whatever decision one makes regarding the inflection class they belong to or not, will affect the number of inflection classes.⁵

⁵See Section 2.3 for some further discussions on the challenges of cross-linguistic morphological analysis.

The first approach to examining the complexity in the exponents of an inflectional system comes from Greenberg's work on inflectional complexity (Greenberg 1960). Greenberg proposes a method based on indices of synthesis, agglutination, compounding, derivation, gross-inflection index, prefixation, suffixation, isolation, pure inflection index, and concord. These indices are calculated as the ratio of two formal elements, given their frequencies in a text.⁶ For example, the gross inflection index is the ratio of words to inflectional morphemes in a language corpus (Greenberg 1960, 186). A language in which this ratio is 1 will have one inflectional morpheme per word and thus a very low inflectional complexity, while languages with high inflectional complexity will have ratios much lower than 1. Typological work on different aspects of E-complexity is abundant, I will focus on a few recent examples.

While ideas similar to the inflectional index have remained present in more recent work on inflectional complexity (see below), several recent studies have focused on the number of morphosyntactic distinctions marked through inflection (Lupyan and Dale 2010; Bentz and Winter 2013; Cotterell *et al.* 2019). These studies tend to use typological datasets like the World Atlas of Language Structures (Dryer and Haspelmath 2013) or similar databases. A well-known example is Lupyan and Dale (2010), who use hand-annotated features in WALS like degree of syncretism, the number of morphosyntactic categories expressed by the verb, presence of noun/verb agreement, presence of inflectional evidentiality, presence of inflectional negation, among others, as measures of morphological complexity. The idea is that if a language makes more morphosyntactic distinctions in a paradigm, then it is more complex than a language that makes fewer morphosyntactic distinctions in the same paradigm. A similar approach is also taken by Bentz and Winter (2013) in a more recent study. Effectively, these

⁶Greenberg uses rather short texts of 100 words, which, as he admits, leads to only very preliminary results.

studies use paradigm size as a measure of morphological complexity.⁷

A different set of metrics based on corpora (Gutierrez-Vasques and Mijangos 2018; Oh and Pellegrino 2022) try to estimate exponent complexity indirectly. Perhaps the simplest is the type-token ratio (TTR) (Juola 1998, 2008; Kettunen 2014). The idea behind the TTR is that if there is a 1-to-1 relation between word types and word tokens,⁸ then this means that there is a very high degree of inflection in the language, and thus the language has very high morphological complexity. A TTR closer to 0 indicates lower morphological complexity. In practice, TTR values range between 0.05 and 0.2 or 0.6 (Kettunen 2014).⁹

2.2.4

Other corpus metrics

Another proposed method for measuring morphological complexity is to calculate the perplexity¹⁰ of sublexical units (Gutierrez-Vasques and Mijangos 2018). In a segmented word, one can calculate the conditional entropy or perplexity of the units within a single word. Low conditional entropy means higher predictability, and thus lower morphological complexity. This method relies on morphological segmentations. Gutierrez-Vasques and Mijangos (2018) rely on automatic segmentation produced by Morfessor (Smit *et al.* 2014). Other corpus-based metrics include word entropy¹¹ (Bentz and Alikaniotis 2016), which measures the amount of information carried by a word based

⁷ Arguably, some of the features considered in these approaches, like degree of syncretism, is not directly about paradigm size, but rather paradigm structure. However, most other metrics are proxies for paradigm size.

⁸ Notice this never happens due to Zipfian effects.

⁹ The difference lies in whether one normalises the corpus size or not. Because corpus size can have a sizeable impact on TTR, some authors have suggested taking the moving average of the TTR across a fixed sub-corpus length (Covington and McFall 2008, 2010). Doing this ensures that when comparing the complexity in two different sized corpora, the TTR is measured on sub-corpora of roughly the same size.

¹⁰ Perplexity can be related to entropy as: $P = 2^H$, where H is the entropy.

¹¹ These entropy metrics measure the distribution of words in a corpus and are not to be confused with other entropy measures like those of Ackerman and Malouf (2013), which measure the distribution of inflectional patterns in a lexicon.

on its probability distribution in a corpus; the relative entropy of word structure (Koplenig *et al.* 2017), which is based on a compression algorithm; and word alignment measure (Bentz *et al.* 2016), which assumes that for languages with morphologically complex words, those will be translated into several independent words in morphologically simpler languages.¹²

Although corpus-based metrics have the advantage of not requiring human decisions, they also have a clear downside: they cannot distinguish inflection from derivation and other morphological processes. All current methods based on corpora conflate morphological complexity arising from derivation and morphological complexity which arises from inflection. Moreover, in most implementations, these methods do not separate the complexity of different subsystems within a language. It is possible for a language to have a very high inflectional complexity in the nominal domain, but a very low inflectional complexity in the verbal domain, or the other way around. While this could be explored with tagged corpora, I am not aware of studies which do this.

Complexity correlations and trade-offs

2.2.5

Despite the proliferation of complexity metrics, Bentz *et al.* (2016) argue that most metrics proposed in typology, either based on corpora or hand annotations, are highly correlated with each other. To do this, the authors propose a method to estimate an aggregated metric of inflectional complexity based on WALS features. The process is as follows. First, the authors identify 28 features that they argue to be indicative of the morphological complexity of a language (e.g. number of genders, number of cases, presence of morphological tense marking, etc.). Then, they normalise the values for each feature to be between 0 and 1 in order to make them comparable. Finally, the authors take the mean value of all 28 features for each language. The authors then estimate the correlations of this complexity index with estimates for several corpus-based complexity indices estimated from Bible translations. The fact that Bentz *et al.* (2016) find a relatively

¹²See also Oh and Pellegrino (2022) for a comparison and evaluation of different corpus-based metrics of morphological complexity.

high correlation between all these metrics is taken by the authors as an indication that they indeed capture the same phenomenon.

Finally, another question that has received some attention regarding complexity is whether there are trade-offs between the local complexity of different domains (morphology and syntax). Several studies have found trade-offs between different types of complexity (Koplenig *et al.* 2017; Oh and Pellegrino 2022; Bentz *et al.* 2022). Some work that has looked at E- and I-complexity has proposed that there are trade-offs between the two (Gutierrez-Vasques and Mijangos 2019; Cotterell *et al.* 2019). Gutierrez-Vasques and Mijangos (2019) use the metrics proposed by Bentz *et al.* (2016) for measuring E-complexity, which are based on aggregating 28 morphological features found in WALS. Cotterell *et al.* (2019) use a simpler metric based on the number of cells in a paradigm.¹³ Generally, these studies have found some sort of trade-off between their definition of E-complexity and I-complexity.

2.3

Word and Paradigm morphology for typology

Although intuitive, approaches based on morpheme or morph segmentations face a challenge: segmenting words is difficult and depends on theory and tradition.¹⁴ The key idea here is that it is not always easy to compare segmentations across languages, and even within languages, linguists face what is called the segmentation problem (Spencer 2012), i.e. how to segment words into sublexical units like stems, morphs or morphemes. That is, it is not just that segmenting words into morphemes is difficult, but it can be a problem without a determined solution. Things can be even more complex if one considers that some theories propose zero morphemes, or very complex and abstract morph sequences. In order to compare the complexity of two languages based on metrics that rely on morph or morpheme segmentations, the principles behind the segmentation decisions need to be consistent for all languages, and application needs to be independent of linguistic tradi-

¹³ Recall most E-complexity metrics are correlated with, and a proxy for paradigm size.

¹⁴ For the opposite view the reader can look at Manova *et al.* 2020.

tions associated with the languages in question.¹⁵ As far as I am aware, there are no clear formalisation for how this should be resolved for typological comparison.¹⁶

In several of the approaches to E-complexity mentioned in the previous section, segmentation of words into morphs or morphemes plays a crucial role (e.g. mean number of morphemes per word). However, segmentation-based approaches to morphology from a cross-linguistic perspective have issues which are not easy to overcome. The first issue worth discussing is that of the definition, delimitation and identification of morph and morpheme boundaries. This is a problem without a simple solution. This has been noted before with regards to morphological complexity. Greenberg (1960, p. 188) notes that:

Basic to the synthetic index as well as most of the others is the possibility of segmenting any utterance in a language into a definite number of meaningful sequences which cannot be subject to further division. Such a unit is called a morph. There are clearly divisions which are completely justified and which every analyst would make. For example, everyone would divide English eating into eat-ing and say that there were two units. There are other divisions which are just as clearly unjustified. For example, the analysis of chair into ch-, “wooden object,” and -air, “something to sit on,” would be universally rejected. There is, however, an intermediate area of uncertainty in which opinions differ. Should, for example, English deceive be analyzed into de- and -ceive. (Greenberg 1960, p. 188)

This relates to the segmentation problem (Spencer 2012). The implication of this is that trying to do automatic, or even semi-automatic morpheme identification on large datasets is not feasible.¹⁷ More

¹⁵ By this I mean how linguists analyse sublexical units like phonemes, tones, or discontinuous stems (i.e. roots in Semitic), zero morphemes, so-called subtractive morphology, etc.

¹⁶ Though see below for a computer-aided approach, as well as Sagot and Walther 2011 and Walther and Sagot 2011 for some early approaches in this direction.

¹⁷ While tools like Morfessor can, under some circumstances, do a decent job of approximating human judgements in morpheme segmentation, these are

importantly, segmentation done by linguists is not necessarily objective and will be influenced by different theoretical perspectives, and linguistic traditions (see Bonami and Beniamine 2021 for a discussion on stem segmentation). For example, while it is common to view stems in Semitic languages as discontinuous triconsonantal roots, it is not common to take a similar approach for European languages, instead preferring ideas like stem mutation.

The consequence of these issues for linguistic typology is that cross-linguistic comparison is heavily dependent on the individual decisions made by the individual linguists writing the grammars. Morphological analysis and segmentation is usually taken as a given, and it is not possible to be certain that the guiding principles for morpheme segmentation are consistent across languages.

Although there are some attempts at computational formalisations of morpheme-based approaches (Rathi *et al.* 2022), I am not aware of large-scale validations of these for the purpose of studying inflectional morphology cross-linguistically.

The alternative approach is to take whole, fully inflected words and their relations in a paradigm as a starting point of linguistic comparison. If we define a systematic approach to finding relations between fully inflected words (see the next section), then we can be sure that all languages in our sample are analysed using the exact same principles. If we focus on fully inflected words, the issues related to segmentation and morph(eme) boundaries disappear.

Perhaps the main counterargument one can leverage against W&P morphology is that one needs to provide a solid, cross-linguistically valid, definition of what a word is. It has been argued that such a task is impossible (Haspelmath 2011), and Greenberg himself points out the issues with defining word units (Greenberg 1960). While it is true that identifying words can be challenging, it must be noted that this is also a necessary step in all morpheme-based approaches to inflectional complexity I am aware of. In order to estimate metrics related to paradigm size, one first needs to decide which elements belong to a paradigm and which elements do not. This requires at least a definition of words. If one wants to count whether negation is expressed through

nowhere near good enough for the task at hand, and the quality of the output greatly varies with the type of input provided.

inflection or not, one needs to distinguish between what constitutes one or two words. If one wants to calculate the number of morphemes to words ratio one needs to delimit words. And so on. Even the corpus-based metrics discussed in the previous section require orthographic word segmentation to work.

The difficulty of delimiting words, and having a systematic, cross-linguistically valid definition of what a word is, is not an argument in favour of morpheme-based approaches to inflectional complexity, nor is it an argument against W&P approaches. My solution in this paper is the same as with many typological studies: I trust the grammars (or in this case the datasets). Even if different languages require different criteria for defining and delimiting words, I will assume that the authors of the resources I use (see next section) applied the correct and relevant criteria consistently for each language in question.¹⁸

MATERIALS AND METHODS

3

Datasets

3.1

For this study,¹⁹ I mostly rely on Unimorph data which was available in January 2021 (Kirov *et al.* 2018).²⁰ Additionally, I include the following datasets:

¹⁸The only technique that I am aware of, which can completely ignore the issue of words is based on compression algorithms (Moscoso del Prado 2011; Ehret 2021). This type of complexity is also known as Kolmogorov Complexity. These calculate the compression rate of a corpus for a given language (how much a compression algorithm can compress a corpus), and compare that result with the compression rate of either another language, or a modified (e.g. lemmatized) version of the same corpus. For reasons of space, I will not discuss this approach here.

¹⁹All datasets and code can be found at <https://doi.org/10.5281/zenodo.11147171>.

²⁰I am aware that Unimorph has included some additional datasets since then, but our approach is computationally too intensive for us to keep adding languages indefinitely. With my dataset, it took me around 6 months to process all paradigms.

- Russian nouns (Guzmán Naranjo 2020)
- Kasem nouns (Guzmán Naranjo 2019a)
- Latvian nouns (Beniamine and Guzmán Naranjo 2021)
- Hungarian nouns (Beniamine and Guzmán Naranjo 2021)
- French verbs (Bonami *et al.* 2014)
- Arabic nouns (Beniamine 2018)
- Portuguese verbs (Beniamine *et al.* 2021)
- English verbs (CELEX, Baayen *et al.* 1996)
- Latin nouns (Pellegrini and Passarotti 2018)
- Latin verbs (Pellegrini and Passarotti 2018)
- Navajo verbs (Beniamine 2018)
- Yaitepec verbs (Feist and Palancar 2015)
- Zenzotepec verbs (Feist and Palancar 2015)

In total, this makes for 137 datasets across 71 languages for nouns, adjectives, and verbs. The size of these datasets vary considerably, from some languages having a few hundred lexemes, to others containing over 40,000 lexemes. To be able to better compare results, I created random subsamples of 200, 500, 1000, 2000, and 5000 lexemes for each dataset (when available). Although I am aware of some issues with the Unimorph datasets,²¹ I only performed minimal hand corrections. These datasets are structured in long format with three main columns: lexeme, cell, inflected form. Table 3 shows an example of this structure for the Spanish verb *cantar* ‘sing’. All datasets are in orthographic form, except for those listed above, which were converted to a phonemic representation. No other information is required or provided in these datasets.

A final note about the data is that I included all cells listed in unimorph, including elements separated by spaces. These can be inflected forms with pronouns/clitics (like in Romance), single words made up of two elements but which inflect like a single lexeme (like *high school*), or periphrasis. About 25% of the datasets contain at least one form that

²¹ The Hungarian and Latvian dataset are effectively hand-corrected unimorph datasets, for which Beniamine and Guzmán Naranjo (2021) remove multiple mistakes present in the original data. Similarly, the Arabic nouns dataset was hand-corrected by Beniamine (2018).

Lexeme	Cell	Inflected form	Table 3: Example of basic data structure
cantar	1.sg.pres.ind	canto	
cantar	2.sg.pres.ind	cantas	
cantar	3.sg.pres.ind	canta	
...	

fits this description. For most purposes periphrastic forms behave almost exactly as non periphrastic ones and do not have an impact on the analysis. While it would be possible to exclude all forms containing spaces, leaving them in for the analysis ensures that we are not arbitrarily reducing the complexity of any of the systems in question.

Methods

3.2

In order to estimate the complexity of a morphological system we need a formal model of that system, and from this formal model, we can then estimate the I- and E-complexity of the system. Word-based models of morphology can be divided into two main camps: symbolic and non-symbolic. Under non-symbolic models there are approaches like LSTMs (Cotterell *et al.* 2019; Malouf 2017; Elsner *et al.* 2022; Cardillo *et al.* 2018) or linear discriminative learning (Baayen *et al.* 2019a,b). Non-symbolic models do not require any type of explicit morphological structures, and can predict one cell in the paradigm of a lexeme from another cell or from a meaning without any sort of symbolic manipulating of the strings (see also Elsner *et al.* 2019, for a recent overview). In these types of approaches there are no explicit representations of sublexical units above the grapheme level, instead, they treat words as sequences of individual letters and the cell in the paradigm they realise. LSTMs are trained to predict sequences from sequences. In the case of morphological inflection, they can predict one inflected form from another directly or from its lexeme meaning and cell in the paradigm (depending on the setup).²² In non-symbolic approaches, there are no explicit representations of proportions in the style $Xa \Leftrightarrow Xb$. Despite their impressive performance,

²²See for example Cotterell *et al.* 2019 or Malouf 2017 for more in-detail descriptions of how LSTMs work for morphological reinflection tasks.

non-symbolic models are not appropriate to our objectives for two main reasons. First, existing implementations are too slow to be applicable to large datasets with many languages, or even to languages with many cells. Second, while it is possible to use these systems to estimate I-complexity, I am unaware of any method for estimating E-complexity from the models themselves. Studies that have used LSTMs to explore morphological complexity (Cotterell *et al.* 2019; Marzi *et al.* 2019; Marzi 2020) have explicitly relied on traditional metrics like the number of paradigm cells.

In contrast, symbolic models use explicit representations of the relations between cells. A symbolic model must be comprised of two independent elements: (1) a system of proportions that express the relations between cells, and (2) a method for assigning a lexeme to the correct proportion. Here, (2) is essentially what has been called the classification problem (Guzmán Naranjo 2020), that is, how to determine the inflection class of an inflected form based on its phonology and semantics. There are multiple proposals for solving (1)²³ and (2).²⁴ In this paper I present a new solution for (1), and, for performance reasons, take a very simple approach to (2). These are described in Section 3.2.1 and 3.2.2, respectively.

3.2.1

Analogical proportions

At the core of symbolic W&P approaches to inflection are the analogical proportions between fully inflected forms. Traditionally, these have been expressed informally as $Xa \rightleftharpoons Xo$ (sometimes written as $Xa::Xo$, or some variant thereof), where variables are expressed with upper case letters like X or Y, and segments with lower case letters. This proportion expresses the formal relation between two cells in the paradigm of a lexeme. This example would cover alternations like the following: *ata::ato* ($X = at$), *para::paro* ($X = par$), etc. However, this notation is not well formalised in the sense that it does not readily work

²³See for example Lepage 1998; Stroppa and Yvon 2005; Federici *et al.* 1995a,b; Carstairs 1998, 1990; Albright and Hayes 1999; Albright *et al.* 2001; Beniamine 2017; Lindsay-Smith *et al.* 2024.

²⁴Among others Bybee and Slobin 1982; Guzmán Naranjo 2019a; Albright and Hayes 1999; Albright *et al.* 2001; Arndt-Lappe 2011, 2014; Eddington 2000; Matthews 2005, 2010, 2013; Skousen 1989; Skousen *et al.* 2002; Skousen 1992.

in a computational implementation. The reason is that it is not precise enough to disambiguate cases where there is more than one variable. For example, the alternation $XaY \Leftrightarrow XYo$ is ambiguous for a (toy example) form like *badan* because it is compatible with either *badan::badno* ($X = \text{bad}, Y = n$) and *badan::bdano* ($X = b, Y = \text{dan}$). The reason is that there are no restrictions on how many segments each variable can match, and there is no way of specifying which of the two *-a-* segments should be matched by the infix.

Computationally implemented formalisms of proportional analogies go back several decades and have taken the form of automata (Lepage 1998, 2004; Stroppa and Yvon 2005; Federici *et al.* 1995a,b; Federici and Pirrelli 1997), string unification (Carstairs 1998, 1990), and more recently context rich alternation patterns (Albright and Hayes 1999; Albright *et al.* 2001; Beniamine 2017), and typed-feature structures (Guzmán Naranjo 2019a). Of these, the only formalisation which would be useful for us given the current state of development and tools for automatic induction is that of Beniamine (2017). The idea of context-rich alternation proportions is that they express alternations in the same spirit of the X-notation, but they are stricter, and less flexible, thus producing unambiguous proportions. The general form is $X \Leftrightarrow Y/Z$, meaning that *X* alternates with *Y* in the context of *Z*. For the previous example, the contextual pattern could be written as $a_ \Leftrightarrow _o / \text{bad}_ _$, where the underscores can match single segments, and which would only allow for the match *badan::badno*. While the context-rich pattern approach is certainly an improvement over previous formalisms, it lacks some expressive power and it cannot easily capture more abstract patterns. For example, because this technique does not have anything like named variables, it is not possible to express alternations that rely on reordering (e.g. metathesis) or repeating segments (e.g. reduplication, lengthening). In the formalism by Beniamine (2017), it is not possible to express that a matched segment has to be repeated, or changed to a different position in a string.

In this paper, I propose a modification of context-rich proportions. One key insight of the approach by Beniamine (2017) is that alternations are bounded by one of the edges of the word. While his proposed formalism usually needs to specify a lot of concrete (in terms of specific segments) contextual information, most of the time, all that is

actually needed to avoid ambiguities is to know where from either the right or the left the alternation is taking place. For example, if we have the three pairs: *badan::badno*, *tar::tro* and *kariaban::kariabno* it becomes clear that the alternation targets the vowel between the last two consonants, and that everything before it stays constant.²⁵ It is not actually necessary to specify which consonants are at play, just their positions. Doing so carries an important advantage, namely that we can write more abstract patterns involving any two consonants.

It is important to note here that one of the reasons for using contextual information for Beniamine (2017) is that the context helps disambiguate inflection classes, for example, the context might indicate that the alternation between /a/ and /o/ for some cell pair only happens if the preceding consonant is /n/. This is not important for the present technique because I approach classification as a separate problem which can be solved on its own.

In order to be able to express patterns like metathesis and reduplication, I will rely on named variables. It is important to capture these types of patterns because otherwise the system will need many more individual proportions. For example, in a metathesis situation where the last two segments undergo metathesis: $Xab \Leftrightarrow Xba$, if the system cannot capture this pattern abstractly, we would need specific proportions for every combination of segments that appears across all forms. The same applies to reduplication but see below. The basic notation has the following form:

$$(1) \quad [\langle X1, 2 \rangle a \langle X2, 2 \rangle \Leftrightarrow \langle X1, 2 \rangle o \langle X2, 2 \rangle]$$

Where variables, expressed in angled brackets, are tuples of unique identifiers (X1, X2, X3, ...) and a matching potential, i.e., the number of segments they must match. The matching potential, when expressed with a number, means that the variable must match exactly that many segments. Non-variables are expressed simply as lowercase letters, and \Leftrightarrow separates the two parts of the proportion.

To express that some variables can match arbitrarily many segments, we allow for one named variable in the proportion to use ‘+’ (as in a regular expression) indicating that it can match 1 or more

²⁵One could, of course, characterise this example in terms of syllables, but in this paper I will work exclusively on surface strings due to constraints my data.

segments. However, in order to constrain proportions to specific relative positions within inflected words, proportions need to follow two constraints: (i) in any given pattern, all variables must explicitly state their matching potential (i.e. how many segments they must match), and (ii) only one variable can match arbitrarily many segments. The example in (2) shows what this looks like:

(2) [<X1, + > a <X2, 2 > ⇐ <X1, + > o <X2, 2 >]

The main reason for the restriction on the number of variables which can have + as matching potential is computational. If a proportion contains more than 1 variable with a +, then the proportion can become ambiguous, just like proportions of the form $XaY \Leftrightarrow XYo$ are ambiguous in some cases, like in the case of *badan::badno* and *badan::bdano*. Recall this pattern is ambiguous because it is not clear which *a* should be matched. Fundamentally, any pattern of the form [<X1, + > <X2, + > ...] will be ambiguous because given a 3 segment string <abc>, there is no way to know whether X1 should match 1 or 2 segments, and both matches $X1 = \langle a \rangle$ and $X1 = \langle ab \rangle$ will be valid. Restricting + to apply to maximally one variable removes the potential for ambiguity. This should be emphasised: the main motivation for this constraint is purely computational: to remove potentially ambiguous proportions. Ambiguous proportions lead to mis-inflection, and would defeat the purpose of the system. From a theoretical perspective, this restriction seems to match our expectations for most inflectional systems. Languages in the dataset do not allow for infixation operations in free positions within words, which is what XaY states. To my knowledge, operations are either constrained to an edge (or distance to it), or apply across the whole word systematically (e.g. harmony).

A potential type of counter example would be a language in which morphological alternations are constrained by lexically-specified phonological or prosodic cues, which can occur anywhere within the word, and which are independent of word boundaries. For example, a language in which stressed syllables undergo an alternation as: *'pokolo::'pakolo*, *po'kolo::po'kalo*, *poko'lo::poko'la*, would require patterns with two variables with + as matching potential. In such a case, the proportions would not be ambiguous because the cue would only

allow one match.²⁶ A second type of potential exception are languages which have been described as having free morph order like Chintang (Bickel *et al.* 2007) and Mari (Luutonen 1997, as cited by Bonami and Crysmann 2013). So far, it remains unclear how these languages should be handled from a W&P perspective. As far as I can tell, none of the languages in my sample require these type of proportions with multiple variables with + matching potential.

Throughout this paper I will refer to these proportions as *local inflection classes*, and contrast it with *global inflection classes*. While two lexemes can share local inflection classes for some set of cell pairs, they do not have to share the same global inflection class. I favour the term *local inflection class* over something more traditional like cell realisation, because these proportions are meaningless for individual cells, and only really express the relation between two cells. It is important to note that a pattern like that in (2) fully determines the relation between the two cells in question (here Cell 1 and Cell 2). If we know the realisation of Cell 1 for some lexeme L, we can unambiguously deduce Cell 2, and the other way around, provided that we know the local inflection class for Cell 1 – Cell 2 in L. If we know one cell of the paradigm of a lexeme, we can deduce all other forms in its paradigm if we know all its local inflection classes (i.e. all proportions to all other cells). Effectively, being able to infer the whole paradigm of a lexeme boils down to the classification problem (i.e. how to determine the inflection class of an inflected form based on its phonology).

At some points in this text, I will use ‘.’ as shorthand for any segment or number of segments: [$\langle X1, + \rangle \rightleftharpoons \langle X1, + \rangle$.], in cases referring to abstract proportions and not concrete analogies. Unlike context-rich proportions, these proportions do not need to contain contextual information. For example, (3) is a proportion which includes contextual information of where a change happens. In this example, ‘c’ acts as context because it is part of the non-contrastive material (i.e. is present in both cells in the same position), and could be subsumed by the variable.

²⁶Notice this is not the case when stress is not free to wander across the whole word, but is fixed to some position from an edge, like Spanish; or cases in which stress triggers phonological alternations without morphological contrast like in Russian.

(3) [<X1, + > c a <X2, 2 > \rightleftharpoons <X1, + > c o <X2, 2 >]

Proportions like (3) are unnecessary since the more general pattern (2) already matches this same alternations, and even more cases.

This formalism allows for the following inflectional proportions:

- suffixes [<X1, + > \rightleftharpoons <X1, + > .], [<X1, + > . \rightleftharpoons <X1, + > .], [<X1, + > . \rightleftharpoons <X1, + >]
- prefixes [<X1, + > \rightleftharpoons . <X1, + >], [. <X1, + > \rightleftharpoons . <X1, + >]
- circumfixes [<X1, + > \rightleftharpoons . <X1, + > .]
- metathesis [<X1, + > <X2,1 > <X3,1 > \rightleftharpoons <X1, + > <X3,1 > <X2,1 >]²⁷
- fixed suprasegmentals (e.g. tones marked with numbers): [<X1, + > 1 2 \rightleftharpoons <X1, + > 3 1] or [<X1, + > ' <X2,1 > \rightleftharpoons <X1, + > <X2,1 > ']
- reduplication [<X1, + > <X2,1 > <X2,1 > \rightleftharpoons <X1, + > <X2,1 >]
- any combinations of the previous proportions

Except for reduplication, I implemented automatic induction techniques for these proportions (including any and all combinations between suffixes, prefixes, infixes, fixed suprasegmentals and metathesis). That is, the computational implementation can automatically induce proportions required to capture an inflectional system. This induction technique tries to find the most economical, and the fewest proportions that can express the relations between all pairs of cells in a dataset.²⁸ While expressing reduplication in this formalism is straightforward, induction is not. For this paper, I do not implement the induction of reduplication, mostly because it is not very common

²⁷ Something to point out regarding metathesis is with the current formalism each pattern has a fixed length, and different length metathesis would require different patterns. For example, carabo:caraob and carator:caraort would require two different patterns.

²⁸ For reasons of space I do not discuss the techniques in detail here, but these are provided by the packages `analogyr` (<https://gitlab.com/mguzmann89/analogyR>) and `paradigma` (<https://gitlab.com/mguzmann89/paradigma>).

in in the inflectional systems of my dataset²⁹ and it is too costly for the induction phase.

This formalization is not without drawbacks. I cannot currently capture patterns that require feature structure representations, like more complex supra-segmental structures or voicing alternations, but extending the system to be able to capture these is straightforward. There is nothing special about feature structure representations, and they could be integrated into the formalism without any changes to how proportions are expressed.³⁰ There are two reasons for why I will work with segments in this paper. The main one is that the datasets do not have feature structure representations, and trying to induce phonological representations from orthography is prone to mistakes, without any guarantees that the resulting representation is any better than the orthographic representation. The second reason is that inference of complex feature alternations like downstep or harmony patterns, is much too complex to be viable for this study.

Similarly, this system cannot represent abstractions which are present in some languages, like reference to morphological structure (German *vorspringen-vor-ge-sprungen* (‘jump forward’), where the <ge> occurs between a separable prefix and main verb), or the already mentioned harmony, and voicing alternations. While it would be preferable to have a system which can capture all abstractions of the inflectional system of any and all languages, this is not the aim of this paper. For this paper, we need a system that is capable of producing an inflected cell given another inflected cell in the paradigm of a lexeme. The present formalism is in fact capable of doing this exactly in all cases, even if some of the induced proportions are clumsy or too specific from a human perspective.

²⁹ Arguably, reduplication is the most frequent form of morphology since it is present in languages without affixation. However, it does not play a significant role in our data.

³⁰ The simplest approach would be to allow vectors of phonological features instead of or in addition to the individual segments, this would allow feature structure alternations, for example, given a phonological representation of segments with 2 features (e.g. *high* and *back*, etc.), one could express: [$\langle X1, + \rangle 11 \langle X1, 1 \rangle 1 \Leftrightarrow \langle X1, + \rangle 10 \langle X1, 1 \rangle 0$]. But other alternatives are possible, like including syllable structure with onsets, nucleus and codas, etc.

Additionally, while patterns like harmony are not directly captured by the system, it is unclear that we need to. For example, in Hungarian, stating that there is an abstract marker *-Vk* for first singular, and the *-V-* harmonises with the stem: *lát-látok* ('see') vs. *szeret-szeretek* ('love'), has the same effect as stating that there are two different markers *-ek* and *-ok*, with inflection class restrictions. Since capturing the correct classification of such cases is completely straightforward, it is not evident that modelling systems like Hungarian without a specific harmony mechanism should produce different results in terms of estimating the complexity of the system.

I will not discuss induction in detail in this paper, but the following gives a short overview of how induction works. For every dataset:

1. Extract all cell pairs
2. For each cell pair *Cell_1:Cell_2*, calculate the analogical proportions *Cell_1* \rightarrow *Cell_2* and the proportion *Cell_2* \rightarrow *Cell_1* (i.e. the relations as above)
3. Since for each pair of forms there often are several alternative valid proportions:³¹
 - calculate all 'best' proportions
 - after calculating all proportions for all items in *Cell_1::Cell_2*, rank them by frequency
 - for each form pair keep the most frequent proportion which can apply to it

The result is a system of proportions that fully captures the pairwise relations in the paradigm.

The final issue is how to measure the E-complexity of a paradigm using this system. I will use *fragmentation* as a metric of the relative complexity of a pattern. The fragmentation of a pattern is simply the number of positions with contrastive material between the left and right-hand sides of the proportion, i.e. the number of non-variables. For example, if a pattern like [*<X1,+ > \Rightarrow*

³¹ Strictly speaking there is not need to choose between the many different proportions, since all induced proportions work correctly for the specific lexeme. This filter is useful for the classification step.

$\langle X1, + \rangle$.] (e.g. *sing::sings*) has a fragmentation of 1, while a pattern like $[\langle X1, + \rangle . \rightleftharpoons . \langle X1, + \rangle$.] (e.g. *lachen::gelacht* ‘laugh’ inf::participle) has a fragmentation of 3. This metric is independent of the length and complexity of the actual markers, and their position. Prefixes, infixes and suffix contribute 1 to the the total fragmentation of a pattern. There is a relation between a traditional morph count approach and fragmentation in many situations. In the simplest relation between two cells, syncretism, the fragmentation of the pattern will be 0. If the relation between both cells is that of exclusively affixes or prefixes, then the fragmentation will be 2. A fragmentation of more than 2 means that there are discontinuous inflectional markers, or a prefix-suffix combination.³²

There are several advantages to this technique for measuring E-complexity. First, it completely sidesteps the issue of segmentation. This approach does not need to find morphemes, morphs, stems or any other theoretically motivated sub-lexical unit other than the contrasts between two inflected forms. As a consequence, there is no need to find any sort of optimal multiple alignment of a paradigm, all that is needed are optimal pairwise alignments between cells.

The second advantage is that this method works with relatively small datasets of a few dozen inflected lexemes, at least compared to the types of datasets needed when working with automated morpheme segmentation software, or corpus-based methods. In this approach, we only need paradigms of the lexemes we are interested in, there is no need for large corpora, as is the case with other tools.³³ While in this paper I have tried to include inflectional paradigms as complete as possible, fragmentation could be calculated for just two cells. So even if one has only very sparse, and incomplete information on some

³²This metric is inspired by Bonami and Beniamine (2021), however, in their paper, the definition of the fragmentation of the stem would be equivalent to the number of variables in our proportions, while in this paper I count the number of non-variables in the proportions. In practice, there is very little difference taking one or the other, and additional E-complexity metrics could be developed following similar principles.

³³While tools like Morfessor can be used on similarly small datasets, they will produce better results if trained on larger datasets.

inflectional system, it should be possible to use fragmentation as a measure of its E-complexity.

There are two final caveats regarding fragmentation. The first is that it should be understood as an upper limit of E-complexity. Because the induction method is not perfect, because the data lacks feature structure representation, and because the formalism cannot deal with all types of inflectional patterns found in the languages in question, many of the resulting proportions are more complex than theoretically required. The effect is that the measured fragmentation can be higher than the real fragmentation of the language.

The second one is that fragmentation, and the way it is implemented, assumes that all segment alignments matter, even those that might not correspond to traditionally identified inflectional markers. As an example under the current system, the alternation between the Spanish first person singular and third person singular form in any aspect tense combination will contain an infix: *canto::cantamos* produce $s [\langle X1, + \rangle \langle X2, 1 \rangle \rightleftharpoons \langle X1, + \rangle a m \langle X2, 1 \rangle s]$ because the *o* is not contrastive material. While there might be arguments against this type of full alignment,³⁴ there two in favor. First, it is unclear from a Word and Paradigm perspective why one should allow some but not all segments to align, especially from a crosslinguistic perspective. Second, implementing an algorithm and computational system to produce alignments which match linguistic intuition is remarkably difficult.

Analogical classification

3.2.2

As mentioned in Section 2, I take a classification-based approach to measure I-complexity. Instead of measuring the entropy of the system, we try to predict the local inflection class of each lexeme based on its phonological properties.³⁵ Complexity of the system is then measured in terms of accuracy. That is, if we can successfully predict all local

³⁴ Notice that this is not a unique effect of making pairwise comparison. The same type of alignment would arise in a multiple alignment.

³⁵ There is good evidence that other factors like semantics can also play a role in helping predict the inflection class of a lexeme. However, there is no semantic information for most datasets in our sample. For this reason, I will only focus on phonology.

inflection classes of all lexemes in a morphological system, then the accuracy is 1 and the complexity 0. If we can predict none of the proportions the accuracy is 0 and the complexity is 1.

There are many approaches to analogical classification that have been proposed in the literature, including Skousen’s Analogical Modelling framework (Skousen 1989; Skousen *et al.* 2002; Skousen 1992; Arndt-Lappe 2011, 2014), TiMBL (Daelemans and Van den Bosch 2005; Daelemans *et al.* 1998), Neutral Networks (Guzmán Naranjo 2019a; Matthews 2005, 2010, 2013), Boosting Trees (Guzmán Naranjo and Bonami 2021; Bonami and Pellegrini 2022), and Minimal Generalization Learner (Albright and Hayes 1999; Albright *et al.* 2001), among others. While most of these techniques would likely perform very well on our data (see below for a comparison), they are too slow in most contexts and do not scale very well.³⁶ Additionally, some authors who have pioneered the use of methods like LSTMs (Cotterell *et al.* 2019) suggests very small datasets (< 500 lexemes) might not be adequate for some of these techniques. Since we are predicting all cells in a paradigm from all other cells pairwise, we need a method that can be trained and cross-validated in as little time as possible, but at the same time is as accurate as possible.³⁷

Here, for reasons of computational efficiency and conceptual simplicity, I will use a k -Nearest Neighbours (k -NN) algorithm based on an edge-weighted Levenshtein distance. The k -NN assigns the local inflection class of a word form based on its phonological similarity to its nearest 5 neighbours.³⁸

³⁶Here ‘too slow’ should be understood as too slow for most researchers’ resources. Of course, with unlimited computing power and enough state of the art GPUs, one could fit as many neural network models as needed within some reasonable time limit. However, most researchers (including the author) working on these issue have finite and limited computing power.

³⁷To give a simple example, the dataset for Latin verbs contains 254 cells in total. This means 64,262 models (from every cell to every other possible cell), and that times 10 to account for cross-validation gives 642,620. Assuming 1 minute to train each model (which is rather optimistic for a Neural Network or Boosting Tree), it would take over a year to capture verbal inflection in Latin.

³⁸I arrived at this number as a good choice for N through some previous testing. While it is possible that some systems would be better captured with a different choice for N , trying to optimize each dataset would take too long.

	a	s	a	c
0	1	2	3	4
a	1	0	1	2
s	2	1	0	1
a	3	2	1	0
b	4	3	2	1

Table 4:
Levenshtein distance between *casa* and *basa*

The traditional Levenshtein distance (Levenshtein 1966) calculates the minimum number of operations of insertion, deletion, and substitution needed to convert a string *s* into a different string *t*. Table 4 shows the calculation for the strings *casa* and *basa*.³⁹ In this case, the number of operations necessary to transform *casa* into *basa* is one, namely a substitution of *c* for *b*.⁴⁰ Table 4 shows all possible ways of turning *casa* into *basa* using insertions, deletions and substitutions. An operation is represented as a movement on the matrix. Horizontal movement represents deletion, vertical movement represents insertion, and diagonal movement represents either no operation (when there is no change) or substitution. Each operation has a cost, and the values are the accumulated cost. The smallest number of operations is given on the bottom right corner.

While the Levenshtein distance captures the differences between two strings, it ignores where in the strings these differences take place. However, if we want to emphasise that differences at some edges are more important than differences in the middle of the word, then we need an edge-sensitive metric. We use an edge-sensitive metric instead of a symmetric one for two reasons. First, an edge-sensitive metric will give greater weight to what would traditionally be the segments belonging to either suffixes or prefixes, which have been shown to play a greater role in class assignment than segments that belong to what would be analysed as the stem (Guzmán Naranjo 2020). Second, there is ample research showing that the edges of a word play a greater role in class assignment than the inner segments (Guzmán Naranjo 2019b; Arndt-Lappe 2011; Albright *et al.* 2001).

³⁹ Here I present the reversed strings. This is for clarity in the following examples below.

⁴⁰ It is possible to assign different costs to each operation, but I use a cost of 1 for each for illustration purposes.

Table 5:
Edge weighted Levenshtein
distance between *casa*
and *basa*

	ind		a	s	a	c
ind		0	1	2	3	4
	0	0	1	1.5	1.83	2.08
a	1	1	0	0.5	0.83	1.08
s	2	1.5	1	0	0.33	0.58
a	3	1.83	0.83	0.33	0	0.25
b	4	2.08	1.08	0.58	0.25	0.25

Building an edge-weighted version of the Levenshtein distance is straightforward: we divide the cost of the operation by its relative position to the edge of the word (column ind in Table 5). For example, for the previous pair of *casa* and *basa*, there is one difference in the fourth position from the right edge of the word, meaning that the distance is 0.25. Table 5 shows the corresponding table of operations with accumulated cost. The row and column labelled ‘ind’ show the position of the segment in question from the relevant edge of the word. Notice it is possible to calculate either a right-edge weighted, left-edge weighted, or right-left-edge weighted Levenshtein distance; for the latter we simply average both left-edge and right-edge weighted distances.⁴¹

While most previous approaches to classification have used some form of segmentation into stems and affixes, we can use fully inflected forms as the basis for prediction. The target predictions are the local inflection classes (i.e. the proportions induced as described in the previous section). A simple example will illustrate this. Table 6 shows two cells of the paradigm of 4 Spanish verbs across two inflection classes.

The first step is to build the proportions (already given in the table). These are the local inflection classes we want to predict.

Since we want to measure the I-complexity of the system, we try to predict one cell from the other. Suppose we start with the prediction

⁴¹The reader might find this approach to measuring the distance between words unintuitive. The choice for this metric in this paper was purely practical, and based on initial experiments on a smaller, different datasets, in which it outperformed other Levenshtein-based metrics. It is of course possible that there might be other, better metrics for individual languages, but we were unable to find a better metric that worked consistently better cross-linguistically. See below for some tests.

Gloss	1.SG.PRES.IND	2.SG.PRES.IND	Proportion
touch	toco	tocas	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$
eat	como	comes	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$
sweep	barro	barres	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$
drink	tomo	tomas	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$

Table 6:
Spanish
paradigm
example

from 2.SG.PRES.IND to 1.SG.PRES.IND. The first step is to calculate a distance matrix based on the modified Levenshtein distance discussed before, this is shown in Table 7. For each form, I have highlighted the nearest neighbour.⁴² In this case, the nearest neighbour of *comes* is not *barres* but *tomas*. If we were to do the assignment solely based on this information, we would classify *comes* to the wrong class, namely $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$. However, we can do a filtering step, and remove proportions which are incompatible with the forms we are trying to classify. This step simply means narrowing the search space to those proportions which are real candidates for each lexeme in question. In this case, $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$ is incompatible with *comes* and thus we would correctly classify it as $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$, and the system would produce a perfect accuracy of 1.⁴³

	tocas	comes	barres	tomas
tocas	0	1.3	1.45	0.33
comes	1.03	0	0.95	0.7
barres	1.45	0.95	0	1.45
tomas	0.33	0.7	1.45	0

Table 7:
Distance Spanish
example

There is one thing worth mentioning. In the previous example, we used a right-hand-side weighted distance because we know Spanish is a suffixing language, and we know that the right-hand side of verbs is more important than the left-hand side for inflection class assignments. However, for most systems, we cannot know beforehand which of these three produces the best results for any given cell pair.

⁴²Since we only have 4 items in this example it makes no sense to use 5 nearest neighbours, but the same logic would apply if we consider more neighbours.

⁴³This example is too simple because the filtering is enough to get a perfect accuracy, but it helps illustrate the whole process.

For this reason, for each cell pair, we try all three (right-hand-side, left-hand-side, and average of both) and keep the best one (in terms of accuracy).⁴⁴

After having calculated the accuracy of all cell pairs in both directions, we take the average accuracy of the paradigm as the I-complexity of the paradigm.

Before discussing the results, I present a brief illustration on how edge-weighted Levenshtein distances compare to regular Levenshtein distances in a classification task, and we also compare these to a more general classifier, namely Boosting Trees with XGBoost (Chen and Guestrin 2016). For this comparison I picked 5 language datasets, with two cells for each dataset. The datasets in question are: Hungarian nouns, Latvian nouns, Yaitepec-Chatino verbs, Arabic verbs and Navajo verbs. I chose these datasets somewhat randomly, trying to maximise variety in terms of language families and paradigm structure. For each dataset, I first computed the proportions to go from one cell to the other as described in Section 3.2.1. I then performed the k -NN classification method described in this section using four different distance metrics: Levenshtein Distance (LD), right-hand-side edge-weighted LD (RHS), left-hand-side edge-weighted LD (LHS), and left-right-hand-side edge-weighted LD (LRHS). For all datasets, I computed the accuracy of predicting the inflection class of the pair from each cell.

Additionally, I trained a Boosting Tree classifier using XGBoost. Boosting Trees are a machine learning classification technique which consists of sequentially fitting small classification trees, and aggregating their predictions. Boosting Trees are similar in principle to Random Forest, with the difference that Random Forest fits multiple small classification trees randomly, while Boosting Trees work by sequentially fitting trees which target the errors in the previous tree. In practice, Boosting Trees have been successfully used in several classification tasks (Bonami and Pellegrini 2022; Guzmán Naranjo and

⁴⁴ A single language could use different similarities for different cell pairs. For example, if a cell pair analogy Cell 1 \Rightarrow Cell 2 is [<X1, + > . \Rightarrow . <X1, + >] then doing Cell 1 \rightarrow Cell 2 might work better with right hand side similarity (because it has a suffix) while doing Cell 2 \rightarrow Cell 1 might work better with left hand size similarity because it has a prefix.

Bonami 2021; Bonami *et al.* 2023), and they can perform extremely well. I used slightly different meta-parameters for each dataset, but the basic setup is that when predicting Cell 2 from Cell 1, I take the 5 to 7 final (or initial)⁴⁵ segments of Cell 1, and use them as predictors in the model. For each dataset, I optimised the meta-parameters with grid-search until the model achieved the best accuracy possible. In all cases, with k -NN and Boosting Trees, I performed 10-fold cross-validation.

Table 8 shows a comparison of these models. First, there are the results of k -NN using a simple Levenshtein distance and $k = 5$. Second, the results of k -NN with an edge-weighted Levenshtein distance and $k = 5$, the table shows results for the right-hand-side edge (RHS), left-hand-side edge (LHS), and left- and right-hand-side (LRHS) distances. Finally, it shows the results of a Boosting Tree algorithm trained on the N final (or initial) segments of the source inflected form, and the results of TiMBL fitted to the whole word.

The results show two key points. First, the accuracy of the edge-weighted Levenshtein distance models are systematically higher than the accuracies of the regular Levenshtein distance models, even if only by a small amount in some cases. The implication is that edge-weighting distances produce either equivalent, or better results in these five languages, and cell pairs which were chosen for their diverse structures. In some cases, like Hungarian and Latvian, the difference between regular LD and edge-weighted LD can be as dramatic as 11 percentage points. This performance difference is enough to justify preferring edge-weighted LD for our purpose. Second, and equally as important, the Boosting Tree classifier can outperform the distance-based k -NN classifiers most of the time,⁴⁶ and in some cases, by a very large margin, like in Yaitepec-Chatino or Navajo. This is perhaps not

⁴⁵ I experimented with both sides, and chose the one which produced the best performance

⁴⁶ The cases in which it does not, it reaches a very comparable accuracy. It is unclear why XGBoost sometimes struggles to outperform the k -NN classifiers, but here two factors are likely at play. First, machine learning techniques like XGBoost work better with larger datasets, and some of our datasets in this experiment are not very large (fewer than 1000 lexemes). Second, while I did my best to optimise the hyper-parameters of the models, it is possible that a different parametrization could produce in better results.

Table 8: Accuracy comparison classification methods

Language	POS	N. lexemes	N. classes	Predictor	Predicted	LD	RHS	LHS	LRHS	XGBT	TIMBL
Hungarian	N	11417	44	NOM.SG	ACC.PL	0.8	0.91	0.5	0.83	0.93	0.91
Hungarian	N	11417	44	ACC.PL	NOM.SG	0.97	0.98	0.9	0.94	0.98	0.95
Latvian	N	2515	43	NOM.SG	ACC.PL	0.85	0.91	0.5	0.85	0.93	0.92
Latvian	N	2515	43	ACC.PL	NOM.SG	0.85	0.93	0.6	0.8	0.95	0.95
YC	V	316	105	1CPL	3CPL	0.3	0.34	0.3	0.3	0.34	0.4
YC	V	316	105	3CPL	1CPL	0.4	0.44	0.4	0.4	0.69	0.49
Arabic	V	687	22	IMP.ACT.M/F.2.D	SBJV.ACT.F.3.D	0.97	0.92	0.96	0.97	0.96	0.94
Arabic	V	687	22	SBJV.ACT.F.3.D	IMP.ACT.M/F.2.D	0.96	0.92	0.96	0.95	0.96	0.94
Navajo	V	784	69	IPFV.1:IPA	IPFV.3i:IPA	0.89	0.64	0.93	0.9	0.91	0.85
Navajo	V	784	69	IPFV.3i:IPA	IPFV.1:IPA	0.78	0.63	0.79	0.73	0.95	0.77

surprising, as Boosting Trees can pick up much more complex patterns in the data. The implication is that the results I present in this paper are a complexity baseline, and that it is likely that with more time and computational resources one could fit models which result in higher accuracy and lower complexity than the ones I present here.

Taking stock, these results show that one word edge is clearly more important than the other edge for classification purposes, and that edge-weighted Levenshtein distances outperform regular Levenshtein distances; and also, that more sophisticated classification techniques should be able to produce better results.

RESULTS

4

This section presents the main results of this paper. It is divided into three subsections. First, I discuss the results for I-complexity, then the results of E-complexity, and finally I look at the relations between the two. One crucial fact to keep in mind is that these results should be interpreted as upper bounds on complexity. As I mentioned when discussing the classification method, it is likely that more advanced classifiers would produce lower I-complexity, but the same is true regarding E-complexity. A more sophisticated approach to inducing proportions could be able to reduce the fragmentation of many patterns by finding better and simpler abstractions.

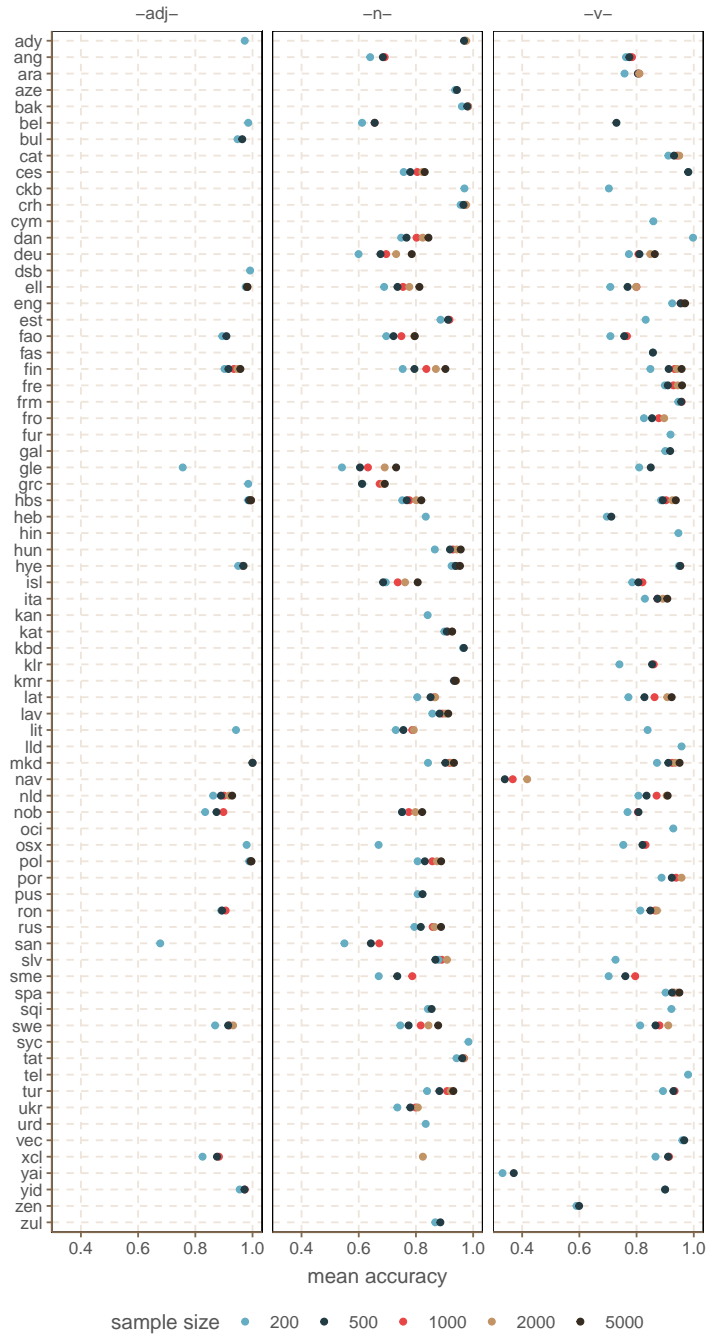
I-complexity

4.1

First, let us analyse the mean I-complexity across the whole paradigm of each language for each sample size.⁴⁷ These results are shown in Figure 1. The accuracy value for each dataset is the mean accuracy across all cell pair predictions. There are several important points worth mentioning here. First, most languages have accuracies

⁴⁷ Recall that for all datasets, we created random subsamples of 200, 500, 1000, 2000, and 5000 lexemes. This allows us to better compare across all languages, for those datasets with very few lexemes.

Figure 1:
Mean accuracy
by language,
part of speech
and sample size



above 0.7 for all sample sizes, for adjectives, nouns, and verbs, and accuracies above 0.8 for sample sizes 2000 and 5000. Even with the very crude approach to classification taken in this paper, and even when only looking at 200 lexemes, we find that most systems have an accuracy above chance level. While these relatively high accuracies might seem unsurprising to linguists familiar with computational work on predicting class assignment of words, these results show that a very simple method for classification can achieve very high accuracies cross-linguistically, for many different types of inflectional systems, only based on phonological distances. It is important to note that these results should be understood as upper complexity limits. More sophisticated classification techniques like LSTMS are likely to be able to produce much better average accuracy scores. If the model reaches a mean accuracy of 96% for some language, this does not mean that the remaining 4% of cell pairs are unpredictable. Rather, it means that given the method and data we were only able to predict 96% of cell pairs. It is very likely that either a more sophisticated method like the ones mentioned in the background section, or more (e.g. simply more lexemes) and better data (e.g. semantic information), would allow us to reach a higher accuracy.

Another key point to remark on is that we are not choosing the best principal part for these results, but rather testing all possible cells and averaging across them. These results are averaged from the worst predictive cells and the best predictive cells. This observation connects to the second point, which is that inflection systems that are usually thought of as needing multiple principal parts, like Latin or Spanish verbs do not actually seem to need principal parts given that the mean accuracy is so high (>0.95 for 5000 lexemes). It is likely that some cells are very bad at predicting some other cells, but more often than not, knowing only one cell is enough to predict a good portion of the remaining cells as can be seen from the results.

If we were to pick the best predicting cell (akin to choosing the principal part), then the accuracy results can go up dramatically. For example, for Spanish verbs, the worst predictive cell in the 5000 sample size is the first singular present of the indicative with 0.86, while the best predictive cell is the infinitive with 0.98. Similarly, for Latin, the worst predictive cell in the 5000 sample size is FIN.IND.PRES.ACT.1.SING with a mean accuracy of 0.84,

while the best predictive cell is FIN.IND.FUT.ACT.3.SING with a mean predictive accuracy of 0.96. This fact further points toward the interpretation that I-complexity seems to be rather low cross-linguistically. Moreover, these results are an alternative way of measuring complexity similar to the principal parts approach, but without the challenges that are related to identifying principal parts already discussed.

A third point worth noting is that there is a very large amount of variation across languages. While some languages like Telugu (tel) have very low complexity in their verbal paradigm, others like Zenaga (zen) have a much larger complexity. There is also variation across domains for the same language. For example, Irish (gle) has a high complexity in the adjectival and nominal system, but lower complexity in the verbal system. These results do not show any clear tendency in terms of I-complexity across domains. While some languages are equally simple in all three domains (e.g. Armenian, hye), others are similarly complex across domains (e.g. Faroese, fao). The only clear trend appears to be that adjectives have lower complexity for this sample (although the sample has fewer adjective paradigms than verb or noun paradigms).

There are two clear exceptions to the high predictability result: Navajo (nav) and Yaitepec Chatino (yai). For languages like Navajo and Yaitepec Chatino, these results suggest that knowing just one cell is clearly not enough, and they raise the question of how many cells we need to know in these languages to be able to deduce the remaining cells. I discuss Navajo in some more detail next.

In the dataset, Navajo verbs can inflect for 7 persons: 1, 2, 3, 3o, 3a (fourth person), 3s (space), and 3i (indefinite);⁴⁸ 3 numbers: singular, dual and plural; and 5 TAM categories: future (FUT), imperfective (IPFV), iterative (ITER), optative (OPT) and perfective (PFV) (see Young 2000, for a more complete description of Navajo verbal inflection). Most verbs in our data have somewhere between 50 cells and 70 cells. Tables 9 and 10 show the inflection table for three verbs: *adika* ('to play cards'), *náháshne* ('to hope around') and *yish'aah* ('to eat').

⁴⁸This is only a small fragment of Navajo verb conjugation, because the dataset only includes subject indices.

The typology of inflectional complexity

Table 9: Conjugation of *adika* ('to play cards'), *náháshne* ('to hope around') and *yish'aah* ('to eat'), part 1

Tense	Person	Number	' <i>adiishk'áq̄h</i>	<i>náháshne</i> '	<i>yish'aah</i>
FUT	1	SG	?atite:ʃk'á:ʃ	nahote:ʃtʃ'á:h	te:ʃ?á:ʃ
FUT	1	DL	?atiti:k'á:ʃ	nahoti:tʃ'á:h	ti:?á:ʃ
FUT	1	PL	tati?ti:k'á:ʃ	ntahoti:tʃ'á:h	tati:?á:ʃ
FUT	2	SG	?atití:k'á:ʃ	nahotí:tʃ'á:h	tí:?á:ʃ
FUT	2	DL	?atito:hk'á:ʃ	nahoto:htʃ'á:h	to:h?á:ʃ
FUT	2	PL	tati?to:hk'á:ʃ	ntahoto:htʃ'á:h	tato:h?á:ʃ
FUT	3	SG	?atito:k'á:ʃ	nahoto:tʃ'á:h	to:?á:ʃ
FUT	3	PL	tati?to:k'á:ʃ	ntahoto:tʃ'á:h	tato:?á:ʃ
FUT	3a	SG	?aʒtito:k'á:ʃ	nahozto:tʃ'á:h	tʃito:?á:ʃ
FUT	3a	PL	tatiʒ?to:k'á:ʃ	ntahozto:tʃ'á:h	taʒto:?á:ʃ
IPFV	1	SG	?ati:ʃk'á:h	nahaʃtʃ'á:h	jiʃ?a:h
IPFV	1	DL	?ati:k'á:h	nahwi:tʃ'á:h	ji:?a:h
IPFV	1	PL	ta?ti:k'á:h	ntahwi:tʃ'á:h	tei:?a:h
IPFV	2	SG	?ati:k'á:h	nahóʃtʃ'á:h	ni?a:h
IPFV	2	DL	?ato:hk'á:h	nahohʃtʃ'á:h	woh?a:h
IPFV	2	PL	ta?to:hk'á:h	ntahohʃtʃ'á:h	ta:h?a:h
IPFV	3	SG	?ati:k'á:h	nahaʃtʃ'á:h	ji?a:h
IPFV	3	PL	ta?ti:k'á:h	natahatʃ'á:h	ta:?a:h
IPFV	3a	SG	?aʒti:k'á:h	nahotʃitʃ'á:h	tʃi?a:h
IPFV	3a	PL	taʒ?ti:k'á:h	ntahotʃitʃ'á:h	taʃi?a:h
ITER	1	SG	n̄?ti:ʃk'á:h	nináháʃtʃ'á:h	náʃ?á:h
ITER	1	DL	n̄?ti:k'á:h	nináhwi:tʃ'á:h	néi:?á:h
ITER	1	PL	n̄ta?ti:k'á:h	ninátahwi:tʃ'á:h	n̄tei:?á:h
ITER	2	SG	n̄?ti:k'á:h	nináhóʃtʃ'á:h	nání?á:h
ITER	2	DL	n̄?to:hk'á:h	nináhóhtʃ'á:h	náh?á:h
ITER	2	PL	n̄ta?to:hk'á:h	ninátahohʃtʃ'á:h	n̄ta:h?á:h
ITER	3	SG	n̄?ti:k'á:h	nináháʃtʃ'á:h	ná?á:h

Table 10:
Conjugation
of *adika'*
(‘to play cards’),
náháshne (‘to
hope around’)
and *yish'aah*
(‘to eat’), part 2

Tense	Person	Number	<i>'adiishk'áq̣h</i>	<i>náháshne'</i>	<i>yish'aah</i>
ITER	3	PL	ńtaʔti:k'á:h	ninátahatʔ'á:h	ńta:ʔá:h
ITER	3a	SG	ńízʔti:k'á:h	nináhoʔʔiʔ'á:h	ńʔʔiʔá:h
ITER	3a	PL	ńtaʔʔti:k'á:h	ninátahoʔʔiʔ'á:h	ńtaʔʔiʔá:h
OPT	1	SG	ʔato:ʔk'á:ʔ	nahóʔʔ'á:h	wóʔʔá:ʔ
OPT	1	DL	ʔato:k'á:ʔ	naho:ʔ'á:h	wo:ʔá:ʔ
OPT	1	PL	taʔto:k'á:ʔ	ntaho:ʔ'á:h	tao:ʔá:ʔ
OPT	2	SG	ʔatoók'á:ʔ	nahó:ʔ'á:h	wó:ʔá:ʔ
OPT	2	DL	ʔato:hk'á:ʔ	naho:hʔ'á:h	wo:hʔá:ʔ
OPT	2	PL	taʔto:hk'á:ʔ	ntaho:hʔ'á:h	tao:hʔá:ʔ
OPT	3	SG	ʔato:k'á:ʔ	nahóʔʔ'á:h	wóʔá:ʔ
OPT	3	PL	taʔto:k'á:ʔ	ntahóʔʔ'á:h	taoʔá:ʔ
OPT	3a	SG	ʔaʔto:k'á:ʔ	nahóʔʔóʔ'á:h	ʔóʔá:ʔ
OPT	3a	PL	taʔʔto:k'á:ʔ	ntahóʔʔóʔ'á:h	taʔʔóʔá:ʔ
PFV	1	SG	ʔati:ʔk'á:ʔ	nahóʔéʔ'á:ʔ	jíʔá
PFV	1	DL	ʔati:hk'á:ʔ	nahóʔi:ʔ'á:ʔ	ji:ʔá
PFV	1	PL	taʔti:ʔk'á:ʔ	ntahóʔi:ʔ'á:ʔ	tei:ʔá
PFV	2	SG	ʔatiniʔk'á:ʔ	nahosiniʔ'á:ʔ	jíniʔá
PFV	2	DL	ʔato:hʔk'á:ʔ	nahóʔo:ʔ'á:ʔ	wo:ʔá
PFV	2	PL	taʔto:hʔk'á:ʔ	ntahóʔo:ʔ'á:ʔ	tao:ʔá
PFV	3	SG	ʔati:ʔk'á:ʔ	nahaʔʔ'á:ʔ	jíʔá
PFV	3	PL	taʔti:ʔk'á:ʔ	ntahaʔʔ'á:ʔ	tá:ʔá
PFV	3a	SG	ʔaʔti:ʔk'á:ʔ	nahóʔʔiʔʔ'á:ʔ	ʔi:ʔá
PFV	3a	PL	taʔʔti:ʔk'á:ʔ	ntahóʔʔiʔʔ'á:ʔ	taʔʔi:ʔá

The main difficulty in Navajo seems to come from predicting across TAM categories. Measuring the predictability within TAM blocks (PFV cells only predicted from other PFV cells, etc.), the mean accuracy of the system is 0.81, which is clearly much better than the 0.42 of the mean accuracy of the whole system. Looking at the best predictors for each TAM block we get the results in Table 11. This means that in Navajo, it is relatively easy to predict all cells of a verb as long as you know one form for each of these blocks. Even taking the worst predictors by TAM in Navajo, as shown in Table 12, the system still has very high inter-predictability.

TAM	predictor	accuracy
FUT	FUT.3.SG	0.947
IPFV	IPFV.3.SG	0.935
ITER	ITER.3.SG	0.952
OPT	OPT.3.SG	0.962
PFV	PFV.3.SG	0.863

Table 11:
Best predictors by TAM category
for Navajo

TAM	predictor	accuracy
FUT	FUT.3I.SG	0.805
IPFV	IPFV.1.PL	0.707
ITER	ITER.1.PL	0.740
OPT	OPT.1.PL	0.731
PFV	PFV.2.SG	0.645

Table 12:
Worst predictors by TAM category
for Navajo

This does not quite mean that Navajo necessarily requires five principal parts. Table 13 shows that for FUT, IPFV, and OPT, the model gets a relatively high accuracy from at least one cell from a different TAM block. These results come from choosing the best predictor found in a different TAM block. For FUT, the accuracy is lower (0.81), and for PFV the accuracy is very low (0.403). This shows that the main difficulty comes from predicting PFV from non-PFV cells.

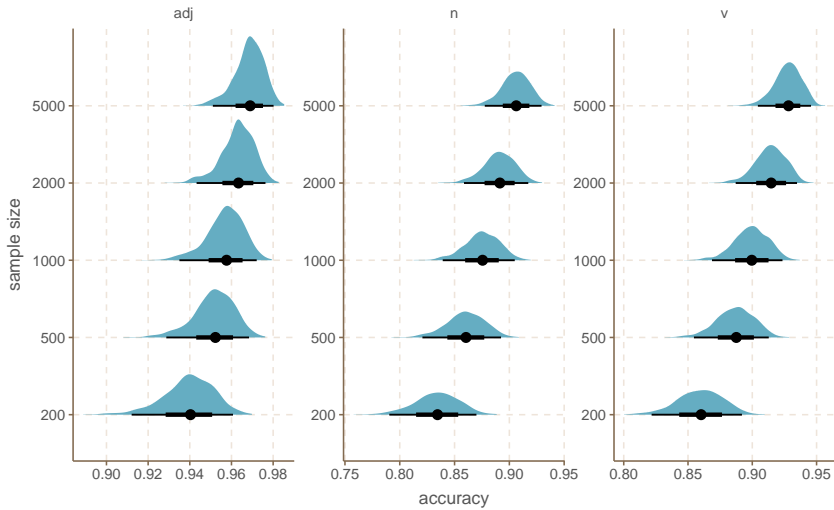
What the Navajo example shows is that even apparently very complex systems like Navajo have only limited I-complexity in the sense that this complexity is mostly restricted to predicting across certain TAM features, and it is not a general property of the whole system.

A different, but equally important question is whether the number of lexemes in a corpus impacts our estimates of I-complexity. I built a

predictor	predicted	accuracy
ITER.3S.SG	FUT.3I.SG	0.816
OPT.3I.SG	IPFV.3S.SG	0.846
FUT.3S.SG	ITER.1.DL	0.710
IPFV.3S.SG	OPT.3O.PL	0.803
OPT.3O.PL	PFV.3.PL	0.403

Table 13:
Best predictors across TAM categories
for Navajo

Figure 2:
Mean accuracy
vs sample size



Bayesian⁴⁹ zero-one inflated Beta regression model where to predict the mean accuracy from the sample size and the part of speech (verb, noun, or adjective) and controls for language by part of speech.⁵⁰ Figure 2 shows the marginal effects of sample size on the mean accuracy. Overall, there is an effect of sample size on mean accuracy, but this effect is relatively small, especially for adjectives. For verbs and nouns, the model does not show any noticeable difference between 2000 and 5000 lexemes, but the difference between 200 and 5000 is more clear. While having larger sample sizes can lead to higher accuracy estimates, it is not clear that relatively small number of lexemes produce bad estimates. Moreover, we can be confident that higher sample sizes lead to higher mean accuracy, meaning that estimates on small sample sizes work well as a lower bound. The consequence is that we can study I-complexity for languages using this method even if we only have access to relatively small datasets. This is a key result. This method allows us to study the I-complexity in languages with

⁴⁹I used Stan (Carpenter *et al.* 2017) with brms Bürkner (2017) for all models in this paper.

⁵⁰The formula in question in brms is `mean-accuracy ~ mo(sample_size) * pos + (1 + mo(sample_size) | language/pos)`, where `mo` is a function to declare monotonic effects.

considerably smaller resources than those needed when using LSTMs (Cotterell *et al.* 2019).

E-complexity

4.2

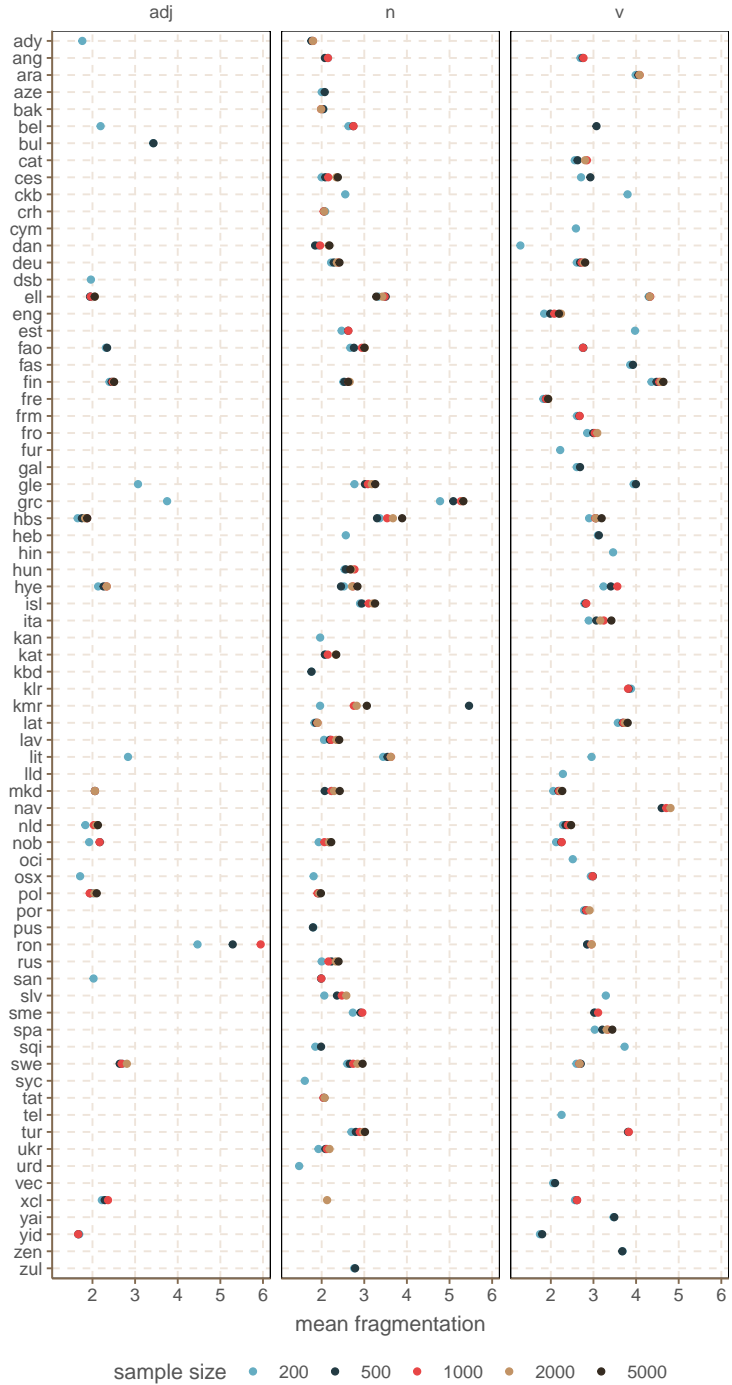
As introduced in Section 3.2.1, we measure E-complexity in terms of the fragmentation of the proportions between cells. Figure 3 shows the mean fragmentation by language and sample size. For the most part, fragmentation stays relatively stable across sample sizes except for Northern Kurdish (kmr).⁵¹

At the same time, the mean fragmentation for most languages is higher than 2. This result shows that inflectional pairs of the form *stem + ending 1:stem + ending 2* are not the most common pattern in our sample, and it shows that it is much more common to have at least two breaks in inflectional pairs. This is even true in European languages which are usually analysed in segmentation-based approaches as being composed of a (mostly invariant) stem and an ending. This does not seem to be the most common situation on average. While these fragmentation values are dependent on the chosen formalization of proportions, they do suggest that for studies of inflectional complexity methods which focus on suffixes and prefixes, and ignore alternations within inflected words, could underestimate E-complexity.

Another important implication of these results is that approaches which follow segmentation based on linguistic traditions, and which are not designed to be language-independent, are likely to overestimate the complexity of some languages, and to underestimate the complexity of other languages. To illustrate this point, we can look at the fragmentation of Arabic (ara), Spanish (spa), and English (eng) verbs. The mean fragmentation of English verbs is of approximately 2.1, for Spanish verbs in the 5000 verb sample is of approximately 3.5, and the mean fragmentation of Arabic verbs is of about 4. However, if one simply follows traditional descriptions of these three languages, Arabic is often characterised as having triconsonantal stems with different affixing schemas, while Spanish and English are characterised

⁵¹ The large difference in fragmentation of Kurdish between the smaller and larger datasets is due to a subset of lexemes with additional periphrastic cells.

Figure 3:
Mean
fragmentation
by language
and sample size



as being a stem + ending type of languages. Our results show that the difference (at least in terms of E-complexity) between Arabic on the one hand, and Spanish and English on the other, is not a categorical one, but rather a gradient one. While Arabic is in fact more complex than English and Spanish, Spanish is much closer to Arabic than it is to English.

Unlike I-complexity, we do find large variation in the E-complexity of different languages, anywhere between 2 and 6 mean fragmentation. This, despite the fact that our method treats all types of *stem changes* in the same way. The result shows that some languages make use of substantially more discontinuous markers (i.e. markers which happen at separate positions) than others.

As with accuracy, we are interested in exploring how sample sizes affect our estimates of fragmentation. I fitted a log-normal model⁵² with the same predictors as for accuracy.⁵³ The result is similar to the others for accuracy, but the effect goes in the opposite direction in terms of complexity, that is, the larger the sample size, the higher the E-complexity. This can be seen in Figure 4. Smaller sample sizes

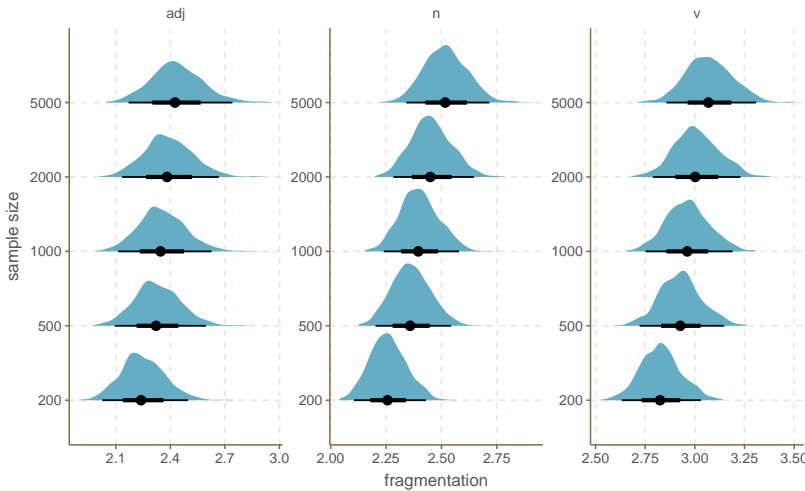


Figure 4:
Mean
fragmentation
vs sample size

⁵²I used a log-normal likelihood because our mean fragmentation values can only be positive.

⁵³As before: $\text{accuracy} \sim \text{mo}(\text{sample_size}) * \text{pos} + (1 + \text{mo}(\text{sample_size}) \mid \text{language/ pos})$.

underestimate the E-complexity of the system, but the effect is very small. Even at only 200 lexemes, the estimates are very close to the estimates with 5000 lexemes. The likely explanation for this effect is that larger samples contain more unique inflection patterns, or suppletive forms which increase the mean fragmentation of the system.

4.3

E- and I-complexity trade-offs

A question that has been asked multiple times in typology is the relation between the complexity of different parts of a grammatical system. With respect to morphology in particular, Cotterell *et al.* (2019) propose a negative correlation between E- and I-complexity. Namely, the authors find that as I-complexity increases, E-complexity decreases, and the other way around. Cotterell *et al.* (2019) use a LSTM approach to estimate the I-complexity of 36 languages, and paradigm size as a measure of E-complexity.⁵⁴ The implication is then that there is effectively a trade-off in terms of complexity, and thus, arguably, a sort of upper level of complexity for any inflectional system.

First, we want to compare the I-complexity results against E-complexity measured in terms of paradigm size. Figure 5 shows the mean accuracy by language and part of speech vs the number of cells in the relevant paradigm.⁵⁵ Unlike in the case of results reported by Cotterell *et al.* (2019), there does not appear to be any type of correlation between I-complexity and the number of cells. There are two possible reasons for this discrepancy in results. One possibility is that our approach to measuring I-complexity just does not show the type of correlation that Cotterell *et al.* (2019) found. While this is possible, it is not possible to test this explanation without direct access to the original dataset used in that paper.⁵⁶ The alternative is that there is

⁵⁴However, Cotterell *et al.* (2019) only count the number of different cell realisations, rather than total number of cells listed.

⁵⁵Since some paradigms have a small amount of variation in the number of cells a lexeme allows depending on the type of lexeme, I take the maximum possible number of cells.

⁵⁶Cotterell *et al.* (2019) also use UniMorph data, but it is not completely clear which version was used, because these datasets have seen changes since the original study was published.

The typology of inflectional complexity

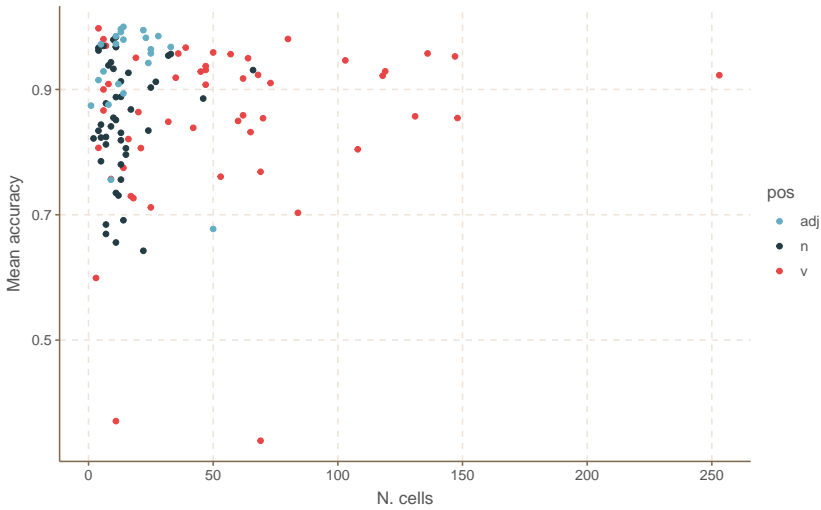


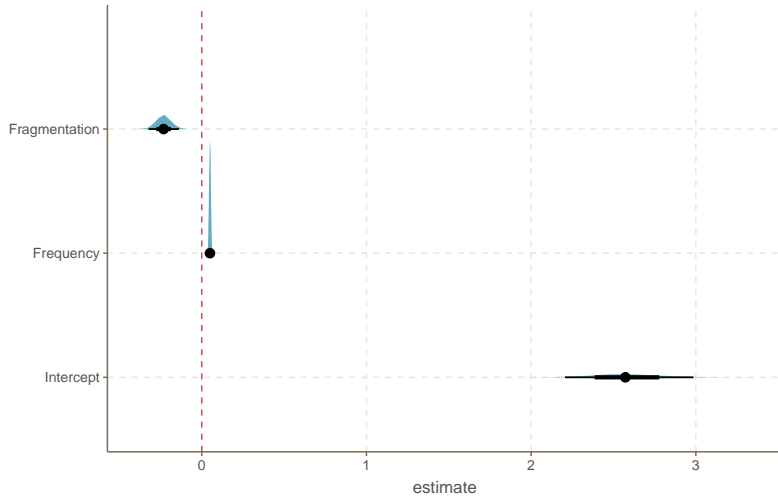
Figure 5:
Mean accuracy
vs number
of cells

bias in the dataset used by Cotterell *et al.* (2019), and that a larger dataset removes any sort of bias of their smaller dataset.

However, this type of approach assumes that the relation between E- and I-complexity happens at the level of the whole inflectional system. There is no *a priori* reason why this should be the case, however. It is more likely that the correlation, if any, should happen at the individual pattern level. For example, it is possible that relatively simple suffixing proportions with low fragmentation like $X \rightleftharpoons Xa$ will also be easier to predict than more complex proportions with higher fragmentation like $uXiY \rightleftharpoons XaYo$. To test this hypothesis, I fitted a binomial model predicting the accuracy of each pattern from its fragmentation and controlled for language. Because there is so much data, and so many proportions, I had to downsample the dataset.⁵⁷ First, I restrict the model to results from the datasets with 1000 lexemes. Additionally, since verbs can have many cells (sometimes in the hundreds), I took a random sample of 500 proportions per language for the verb dataset which left us with around 10,000 proportions instead of 100,000 (about 6000 for nouns and 1800 for adjectives). This leaves

⁵⁷ The issue arises because fitting the group-level effects with a correlation structure is very slow and difficult (i.e. the funnel geometry of the space leads to divergences in the sampling).

Figure 6:
Coefficients
of the model
comparing
pattern accuracy
vs fragmentation



us with a smaller dataset, which should still contain enough information to allow us to estimate any effects in the data.

I fitted a binomial model predicting pattern accuracy from its fragmentation. The question boils down to: is there a relation between the I-complexity of a pattern and its E-complexity? I also controlled for the type frequency of the pattern in the cell,⁵⁸ as well as language and part of speech.⁵⁹ The main coefficients for the model are shown in Figure 6. The results show a negative effect of the proportion's fragmentation on the model's accuracy predicting it, and, as expected, a clear positive effect of frequency on accuracy, meaning that more frequent proportions are easier to predict than less frequent ones. Since accuracy is the opposite of complexity, it means that a higher fragmentation in a pattern generally leads to higher complexity. This result is effectively the opposite of a complexity trade-off. More complex proportions in terms of E-complexity also tend to be harder to predict, while simpler proportions tend to be easier to predict.

⁵⁸A very frequent pattern, i.e. a pattern that applies to many lexemes in a cell, could be easier to predict than a rarer one. The frequency of a pattern by cell could be correlated with its complexity.

⁵⁹The brms model was the following: $\text{correct} \mid \text{trials}(\text{total}) \sim 1 + \text{fragmentation} + \log(\text{total}) + (1 + \text{fragmentation} \mid \text{language/pos})$

Understanding why this effect happens in our data is not completely straightforward, but there are some potential explanations.

If more complex proportions are harder to predict, a reasonable hypothesis is that the number of infixes in the proportions might be driving this effect. To test this, we first look at the mean number of infixes in low and high accuracy proportions. I looked at the results from the datasets with 1000 lexemes. From these, I then extracted the 100 proportions with highest accuracy, and the 100 proportions with lowest accuracy for each language. The results shown in Figure 7. The pattern is clear: the most accurate proportions systematically have the same or fewer number of infixes than the least accurate proportions.

Next, we can approach the question from the opposite direction and only look at the best performing proportions. For this, I further restricted the sample to proportions with a frequency of between 2 and 100 (to control for effects of very high frequent proportions). I also abstracted away all concrete material and matching potential to get basic skeletal patterns: [$\langle X \rangle . \Leftrightarrow \langle X \rangle .$]. Then, I extracted the 10 most frequent proportions among those with an accuracy of 1, those with an accuracy higher than 0.95, and those with an accuracy higher than 0.9, and then compare their relative frequency in those subsamples to their relative frequency in the whole dataset. I did this experiment aggregating across all languages. The results of this comparison are shown in Table 14. By comparing the values in columns ‘acc = 1’, ‘acc > 0.95’, ‘acc > 0.9’ to the values in the baseline column ‘total sample’, one can see the extent to which a pattern is over-represented among the most-accurate proportions, relative to its overall frequency.

Out of the 10 most frequent proportions on the three subsamples, only [$\langle X \rangle . \Leftrightarrow \langle X \rangle .$] shows any clear difference in relative frequency between the subsample and the whole sample, but this difference is considerable. For the subsample on proportions with accuracy of 1, this takes up 0.1 additional total frequency than in the whole sample. What this mean is that the proportion [$\langle X \rangle . \Leftrightarrow \langle X \rangle .$] is very common among easy to predict proportions, while we observe more proportions with more infixes among the harder to predict proportions. This helps create the observed correlation between E- and I-complexity.

Figure 7:
Mean infix
accuracy

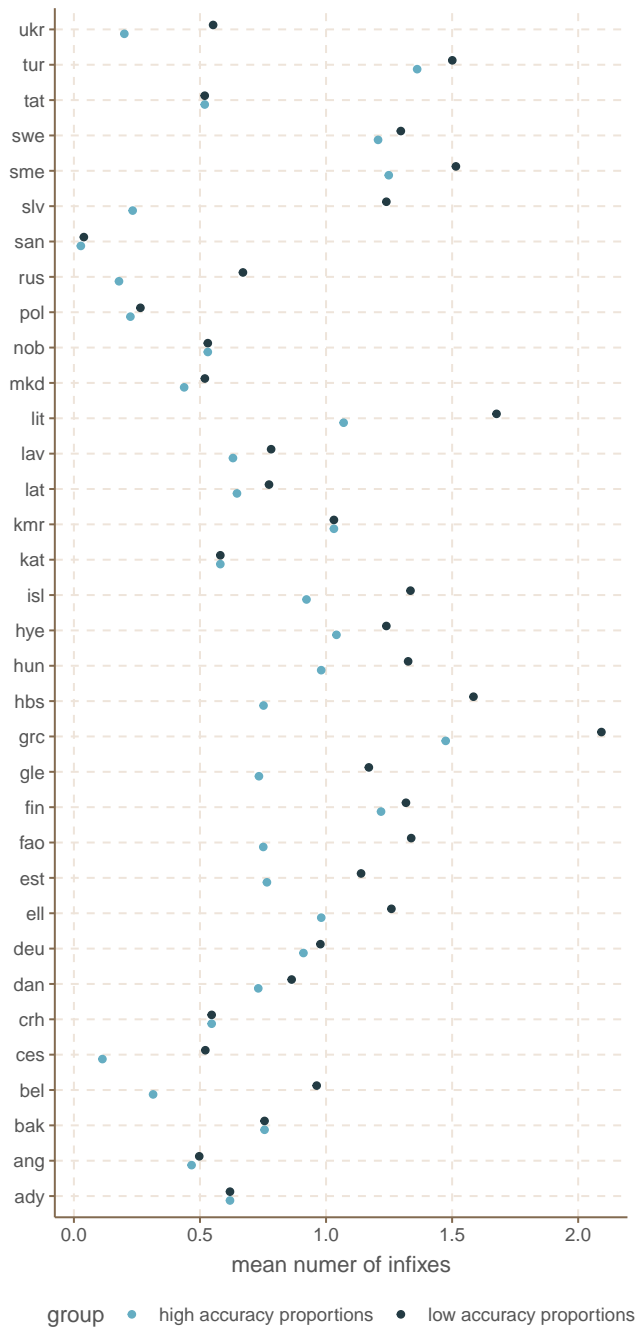


Table 14: Relative frequency of most accurate proportions

Proportion	acc = 1	acc > 0.95	acc > 0.9	Total sample
$\langle X \rangle . \Leftrightarrow \langle X \rangle .$	0.49	0.44	0.43	0.39
$\langle X \rangle . \langle X \rangle . \Leftrightarrow \langle X \rangle . \langle X \rangle .$	0.09	0.10	0.10	0.09
$\langle X \rangle . \langle X \rangle \Leftrightarrow \langle X \rangle . \langle X \rangle$	0.07	0.07	0.07	0.07
$\langle X \rangle . \langle X \rangle \Leftrightarrow \langle X \rangle . \langle X \rangle .$	0.06	0.06	0.06	0.05
$\langle X \rangle . \langle X \rangle . \Leftrightarrow \langle X \rangle . \langle X \rangle$	0.04	0.05	0.05	0.05
$\langle X \rangle . \Leftrightarrow \langle X \rangle$	0.03	0.03	0.03	0.03
$\langle X \rangle \Leftrightarrow \langle X \rangle .$	0.02	0.02	0.03	0.03
$\langle X \rangle . \langle X \rangle . \langle X \rangle \Leftrightarrow \langle X \rangle . \langle X \rangle . \langle X \rangle$	0.02	0.02	0.02	0.02
$\langle X \rangle . \langle X \rangle . \Leftrightarrow \langle X \rangle \langle X \rangle .$	0.01	0.01	0.02	0.02
$\langle X \rangle \langle X \rangle . \Leftrightarrow \langle X \rangle . \langle X \rangle .$	0.01	0.01	0.01	0.02

CONCLUSION

5

In this paper, I have presented an approach to the typology of paradigm complexity in the spirit of Word and Paradigm morphology. I argue that a W&P approach is advantageous for doing cross-linguistic work in inflectional morphology for multiple reasons. First, it gets around the segmentation problem, and second, it allows for relatively simple formalisation in the form of proportional analogies that can be used for efficient automatic induction. I have presented a concrete formalisation of proportional analogies, using named variables with matching potential, restricting morphological patterns to be defined from the word boundary. With this formalisation, I have shown that it is possible to measure both E- and I-complexity in many typologically diverse morphological systems.

The results confirm previous results in the literature (Ackerman and Malouf 2013). The I-complexity of most morphological systems examined were relatively low, and increasing sample sizes leads to a reduction in system complexity. In contrast, E-complexity is less consistent across languages and parts of speech. The results also show that there is a clear correlation between I- and E-complexity of individual patterns: patterns with higher E-complexity lead to higher I-complexity. At the same time, there does not seem to be a clear cor-

relation between I-complexity and paradigm size as has been reported in the literature. The lack of a trade-off between different levels of morphological complexity also point towards the conclusion that, among morphologically complex languages, some are decidedly more complex than others.

There are also some wider implications for the study of morphological typology in general. Using automatic induction has the advantage of being neutral to linguistic tradition, and it allows for systematic and comparable analysis for different languages. The fact that some languages have traditionally been described as using root-and-pattern morphology, or suffixes plus phonological rules for stem alternation, does not play a role in this approach since we analyse everything from a purely surface-based perspective. This is important because it is a fundamental requirement to be able to carry out large scale quantitative studies of morphological systems.

From a methodological perspective, this paper offers two contributions. First, I have shown that computational work in inflectional morphology is feasible with a relatively small number of lexemes. While this is not a completely new insight, it is important to emphasise this point. The fact that data is somewhat limited for many languages does not mean that we need to exclude them in computational approaches to morphology, it just means that we need to use tools capable of coping with small datasets. Second, I provided a new implementation of proportional analogies based on a new formalism. I have shown one potential application of this method to the estimation of inflectional complexity, but other applications are possible, and there is potential for further research on automated morphological analysis. At the same time, while this new formalism can capture a relatively wide range of phenomena, there are still some gaps which we aim to cover in future work, like inducing different types of reduplication and implementing feature structures.

ACKNOWLEDGMENTS

This research was partly funded by the Emmy Noether project Bayesian modelling of spatial typology (grant no. GU 2369/1-1, project number 504155622). I am grateful to Olivier Bonami for his many comments and suggestions.

REFERENCES

- Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: the low conditional entropy conjecture, *Language*, 89(3):429–464, doi:10.1353/lan.2013.0054.
- Adam ALBRIGHT, Argelia ANDRADE, and Bruce HAYES (2001), Segmental environments of Spanish diphthongization, *UCLA Working Papers in Linguistics*, 7(5):117–151.
- Adam ALBRIGHT and Bruce HAYES (1999), An automated learner for phonology and morphology, <https://pdfs.semanticscholar.org/8d74/847ecd575887fcfe42ea022c2d82750fe7d9.pdf>, unpublished manuscript.
- Peter ARKADIEV and Francesco GARDANI (2020), The complexities of morphology, in Peter ARKADIEV and Francesco GARDANI, editors, *The complexities of morphology*, pp. 1–19, Oxford University Press.
- Sabine ARNDT-LAPPE (2011), Towards an exemplar-based model of stress in English noun–noun compounds, *Journal of Linguistics*, 47(3):549–585.
- Sabine ARNDT-LAPPE (2014), Analogy in suffix rivalry: the case of English *-ity* and *-ness*, *English Language and Linguistics*, 18(3):497–548.
- R. Harald BAAYEN, Yu-Ying CHUANG, and Maria HEITMEIER (2019a), WpmWithLdl: implementation of word and paradigm morphology with linear discriminative learning R package version 2.
- R. Harald BAAYEN, Yu-Ying CHUANG, Elnaz SHAFAEI-BAJESTAN, and James P. BLEVINS (2019b), The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning, *Complexity*, 2019:1–39.
- R. Harald BAAYEN, Richard PIEPENBROCK, and Leon GULIKERS (1996), The CELEX lexical database (cd-rom).

Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT, editors (2015), *Understanding and measuring morphological complexity*, Oxford University Press.

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

Sacha BENIAMINE (2017), Un algorithme universel pour l'abstraction automatique d'alternances morphophonologiques, in *24e conférence sur le traitement automatique des langues naturelles (TALN)*, volume 2.

Sacha BENIAMINE (2018), *Classifications flexionnelles: étude quantitative des structures de paradigmes*, Ph.D. thesis, Université Paris Diderot.

Sacha BENIAMINE (Forthcoming), One lexeme, many classes: inflection class systems as lattices, in Berthold CRYSMANN and Manfred SAILER, editors, *One-to-many relations in morphology, syntax and semantics*, Language Science Press.

Sacha BENIAMINE, Olivier BONAMI, and Ana R. LUÍS (2021), The fine implicative structure of European Portuguese conjugation, *Isogloss. Open Journal of Romance Linguistics*, 7:1–35, ISSN 2385-4138, doi:10.5565/rev/isogloss.109, <https://revistes.uab.cat/isogloss/article/view/v7-beniamine-bonami-luis>.

Sacha BENIAMINE and Matías GUZMÁN NARANJO (2021), Multiple alignments of inflectional paradigms, in *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 4, pp. 216–227, doi:10.7275/ymc0-p491.

Charles Edwin BENNETT (1918), *New Latin grammar*, Allyn and Bacon.

Christian BENTZ and Dimitrios ALIKANIOTIS (2016), The word entropy of natural languages, unpublished arXiv manuscript.

Christian BENTZ, Ximena GUTIERREZ-VASQUES, Olga SOZINOVA, and Tanja SAMARDŽIĆ (2022), Complexity trade-offs and equi-complexity in natural languages: a meta-analysis, *Linguistics Vanguard*, doi:10.1515/lingvan-2021-0054.

Christian BENTZ, Tatyana RUZSICS, Alexander KOPLÉNIG, and Tanja SAMARDŽIĆ (2016), A comparison between morphological complexity measures: typological data vs. language corpora, in *Proceedings of the workshop on computational linguistics for linguistic complexity (CLALC)*, pp. 142–153.

Christian BENTZ and Bodo WINTER (2013), Languages with more second language learners tend to lose nominal case, *Language Dynamics and Change*, 3(1):1–27, doi:10.1163/22105832-13030105.

Balthasar BICKEL, Goma BANJADE, Martin GAENZSLE, Elena LIEVEN, Netra Prasad PAUDYAL, Ichchha Purna RAI, Manoj RAI, Novel Kishore RAI, and Sabine STOLL (2007), Free prefix ordering in Chintang, *Language*, pp. 43–73.

- Balthasar BICKEL and Johanna NICHOLS (2007), Inflectional morphology, in Timothy SHOPEN, editor, *Language typology and syntactic description*, volume 3, pp. 169–240, Cambridge University Press, 2 edition.
- Balthasar BICKEL and Johanna NICHOLS (2013), Inflectional synthesis of the verb, in Matthew S. DRYER and Martin HASPELMATH, editors, *The world atlas of language structures online*, Max Planck Digital Library.
- James P. BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):531–573.
- James P. BLEVINS (2013), The information-theoretic turn, *Psihologija*, 46(3):355–375, ISSN 00485705, doi:10.2298/PSI1304355B.
- James P. BLEVINS (2016), *Word and paradigm morphology*, Oxford University Press.
- Olivier BONAMI and Sacha BENIAMINE (2016), Joint predictiveness in inflectional paradigms, *Word Structure*, 9(2):156–182.
- Olivier BONAMI and Sacha BENIAMINE (2021), Leaving the stem by itself, in Marcia HAAG, Sedigheh MORADI, Andrija PETROVIC, and Janie REES-MILLER, editors, *All things morphology*, pp. 82–98, John Benjamins.
- Olivier BONAMI, Gauthier CARON, and Clément PLANCQ (2014), Construction d'un lexique flexionnel phonétisé libre du Français, in *SHS web of conferences*, volume 8, pp. 2583–2596, EDP Sciences, doi:10.1051/shsconf/20140801223.
- Olivier BONAMI and Berthold CRYSMANN (2013), Morphotactics in an information-based model of realisational morphology, in Stefan MÜLLER, editor, *Proceedings of the 20th international conference on Head-Driven Phrase Structure Grammar, Freie Universität Berlin*, pp. 27–47.
- Olivier BONAMI, Lukáš KYJÁNEK, and Marine WAUQUIER (2023), Assessing the featural organisation of paradigms with distributional methods, *Proceedings of the Society for Computation in Linguistics*, 6(1):310–320.
- Olivier BONAMI and Matteo PELLEGRINI (2022), Derivation predicting inflection: a quantitative study of the relation between derivational history and inflectional behavior in Latin, *Studies in Language*, 46(4):753–792, doi:10.1075/sl.21002.bon.
- Joan L. BYBEE and Dan I. SLOBIN (1982), Rules and schemas in the development and use of the English past tense, *Language*, 58(2):265–289.
- Paul-Christian BÜRKNER (2017), Brms: an R package for bayesian multilevel models using stan, *Journal of Statistical Software*, 80(1):1–28, doi:10.18637/jss.v080.i01.
- Franco Alberto CARDILLO, Marcello FERRO, Claudia MARZI, and Vito PIRRELLI (2018), Deep learning of inflection and the cell-filling problem, *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):57–75.

Bob CARPENTER, Andrew GELMAN, Matthew HOFFMAN, Daniel LEE, Ben GOODRICH, Michael BETANCOURT, Marcus BRUBAKER, Jiqiang GUO, Peter LI, and Allen RIDDELL (2017), Stan: a probabilistic programming language, *Journal of Statistical Software, Articles*, 76(1):1–32, ISSN 1548-7660, doi:10.18637/jss.v076.i01.

Andrew CARSTAIRS (1983), Paradigm economy, *Journal of Linguistics*, 19(1):115–128.

Andrew CARSTAIRS (1990), Phonologically conditioned suppletion, in Wolfgang U. DRESSLER, Hans C. LUSCHÜTZKY, Oskar E. PFEIFFER, and John R. RENNISON, editors, *Contemporary morphology*, number 49 in Trends in Linguistics, pp. 17–23, De Gruyter, Berlin.

Andrew CARSTAIRS (1998), Some implications of phonologically conditioned suppletion, in Geert E. BOOIJ and Jaap VAN MARLE, editors, *Yearbook of morphology 1998*, pp. 67–94, Springer.

Andrew CARSTAIRS-MCCARTHY (1994), Inflection classes, gender, and the principle of contrast, *Language*, pp. 737–788.

Tianqi CHEN and Carlos GUESTRIN (2016), Xgboost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

Ryan COTTERELL, Christo KIROV, Mans HULDEN, and Jason EISNER (2019), On the complexity and typology of inflectional morphological systems, *Transactions of the Association for Computational Linguistics*, 7:327–342, doi:10.1162/tacl_a_00271.

Sara COURT, Micha ELSNER, and Andrea D. SIMS (2022), Quantifying factors shaping analogical restructuring of the Maltese nominal system, Talk at the International Morphology Meeting, Budapest.

Michael A. COVINGTON and Joe D. MCFALL (2008), The moving-average type-token ratio, in *Linguistics Society of America*.

Michael A. COVINGTON and Joe D. MCFALL (2010), Cutting the Gordian knot: the moving-average type-token ratio (MATTR), *Journal of Quantitative Linguistics*, 17(2):94–100.

Walter DAELEMANS and Antal VAN DEN BOSCH (2005), *Memory-based language processing*, Cambridge University Press, ISBN 0-521-80890-1.

Walter DAELEMANS, Jakub ZAVREL, Ko VAN DER SLOOT, and Antal VAN DEN BOSCH (1998), TiMBL: Tilburg memory-based learner, Technical report, Universiteit van Tilburg, <https://research.tilburguniversity.edu/en/publications/timbl-tilburg-memory-based-learner-version-10-reference-guide>.

Wolfgang U. DRESSLER (2011), The rise of complexity in inflectional morphology, *Poznań Studies in Contemporary Linguistics*, 47(2):159.

- Matthew S. DRYER and Martin HASPELMATH (2013), *The world atlas of language structures online*, Max Planck Digital Library, <https://wals.info/>.
- David EDDINGTON (2000), Analogy and the dual-route model of morphology, *Lingua*, 110(4):281–298.
- Katharina EHRET (2021), An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data, *Corpus Linguistics and Linguistic Theory*, 17(2):383–410.
- Micha ELSNER, Andrea D. SIMS, Alexander ERDMANN, Antonio HERNANDEZ, Evan JAFFE, Lifeng JIN, Martha Booker JOHNSON, Shuan KARIM, David L. KING, Luana Lamberti NUNES, *et al.* (2019), Modeling morphological learning, typology, and change: what can the neural sequence-to-sequence framework contribute?, *Journal of Language Modelling*, 7(1):53–98.
- Micha ELSNER *et al.* (2022), OSU at SigMorphon 2022: analogical inflection with rule features, in *Proceedings of the 19th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*, pp. 220–225.
- Stefano FEDERICI and Vito PIRRELLI (1997), Analogy, computation, and linguistic theory, in *New methods in language processing*, pp. 16–34, UCL Press London.
- Stefano FEDERICI, Vito PIRRELLI, and François YVON (1995a), Advances in analogy-based learning: false friends and exceptional items in pronunciation by paradigm-driven analogy, in *Proceedings of international joint conference on artificial intelligence (IJCAI'95) workshop on new approaches to learning for natural language processing, Montreal, Canada*, pp. 158–163.
- Stefano FEDERICI, Vito PIRRELLI, and François YVON (1995b), A dynamic approach to paradigm-driven analogy, in Stefan WERMTER, Ellen RILOFF, and Gabriele SCHELER, editors, *IJCAI 1995: connectionist, statistical and symbolic approaches to learning for natural language processing*, volume 1040 of *Lecture Notes in Computer Science*, pp. 385–398, Springer.
- Timothy FEIST and Enrique L. PALANCAR (2015), Oto-Manguean inflectional class database, *University of Surrey*.
- Raphael FINKEL and Gregory STUMP (2007), Principal parts and morphological typology, *Morphology*, 17:39–75.
- Joseph H. GREENBERG (1960), A quantitative approach to the morphological typology of language, *International Journal of American Linguistics*, 26(3):178–194.
- Ximena GUTIERREZ-VASQUES and Victor MIJANGOS (2018), Comparing morphological complexity of Spanish, Otomi and Nahuatl, <https://arxiv.org/abs/1808.04314>, unpublished manuscript.
- Ximena GUTIERREZ-VASQUES and Victor MIJANGOS (2019), Productivity and predictability for measuring morphological complexity, *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(1):48.

Matías GUZMÁN NARANJO (2019a), *Analogical classification in formal grammar*, Empirically Oriented Theoretical Morphology and Syntax, Language Science Press, doi:10.5281/zenodo.3191825.

Matías GUZMÁN NARANJO (2019b), Analogy-based morphology: the Kasem number system, in Stefan MÜLLER and Petya OSENOVA, editors, *Proceedings of the 26th international conference on Head-Driven Phrase Structure Grammar*, University of Bucharest, pp. 26–41, CSLI Publications.

Matías GUZMÁN NARANJO (2020), Analogy, complexity and predictability in the Russian nominal inflection system, *Morphology*, 30:219–262.

Matías GUZMÁN NARANJO and Olivier BONAMI (2021), Overabundance and inflectional classification: quantitative evidence from Czech, *Glossa: a Journal of General Linguistics*, 6(1).

Martin HASPELMATH (2011), The indeterminacy of word segmentation and the nature of morphology and syntax, *Folia Linguistica*, 45(1):31–80.

Iván IGARTUA and Ekaitz SANTAZILIA (2018), How animacy and natural gender constrain morphological complexity: evidence from diachrony, *Open Linguistics*, 4(1):438–452.

Patrick JUOLA (1998), Measuring linguistic complexity: the morphological tier, *Journal of Quantitative Linguistics*, 5(3):206–213.

Patrick JUOLA (2008), Assessing linguistic complexity, in Matti MIESTAMO, Kaius SINNEMÄKI, and Fred KARLSSON, editors, *Language complexity: typology, contact, change*, pp. 89–108, Benjamins, Amsterdam.

Kimmo KETTUNEN (2014), Can type-token ratio be used to show morphological complexity of languages?, *Journal of Quantitative Linguistics*, 21(3):223–245.

Christo KIROV, Ryan COTTERELL, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sebastian J. MIELKE, Arya MCCARTHY, Sandra KÜBLER, *et al.* (2018), UniMorph 2.0: universal morphology, in *Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018)*, European Language Resources Association (ELRA).

Alexander KOPLINIG, Peter MEYER, Sascha WOLFER, and Carolin MÜLLER-SPITZER (2017), The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort, *PLOS ONE*, 12(3):1–25, doi:10.1371/journal.pone.0173614.

Yves LEPAGE (1998), Solving analogies on words: an algorithm, in *COLING 1998 volume 1: the 17th international conference on computational linguistics*, pp. 728–735.

Yves LEPAGE (2004), Analogy and formal languages, *Electronic Notes in Theoretical Computer Science*, 53:180–191.

Vladimir I. LEVENSHTEIN (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, 10(8):707–710.

Emily LINDSAY-SMITH, Matthew BAERMAN, Sacha BENIAMINE, Helen SIMS-WILLIAMS, and Erich R. ROUND (2024), Analogy in inflection, *Annual Review of Linguistics*, 10(1):211–231, ISSN 2333-9683, 2333-9691, doi:10.1146/annurev-linguistics-030521-040935, <https://www.annualreviews.org/doi/10.1146/annurev-linguistics-030521-040935>.

Gary LUPYAN and Rick DALE (2010), Language structure is partly determined by social structure, *Plos One*, 5(1):1–10, doi:10.1371/journal.pone.0008559.

Jorma LUUTONEN (1997), *The variation of morpheme order in Mari declension: suomalais-ugrilaisen seuran toimituksia*, Suomalais-Ugrilainen Seura, Helsinki.

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

Stela MANOVA, Harald HAMMARSTRÖM, Itamar KASTNER, and Yining NIE (2020), What is in a morpheme? theoretical, experimental and computational approaches to the relation of meaning and form in morphology, *Word Structure*, 13(1):1–21.

Claudia MARZI (2020), Modeling word learning and processing with recurrent neural networks, *Information – an International Interdisciplinary Journal*, 11(6):320–334.

Claudia MARZI, Marcello FERRO, and Vito PIRRELLI (2019), A processing-oriented investigation of inflectional complexity, *Frontiers in Communication*, 4(48):1–23.

Clive A. MATTHEWS (2005), French gender attribution on the basis of similarity: a comparison between AM and connectionist models, *Journal of Quantitative Linguistics*, 12:262–296.

Clive A. MATTHEWS (2010), On the nature of phonological cues in the acquisition of French gender categories: evidence from instance-based learning models, *Lingua*, 120(4):879–900.

Clive A. MATTHEWS (2013), On the analogical modelling of the English past-tense: a critical assessment, *Lingua*, 133:360–373, ISSN 0024-3841, doi:10.1016/j.lingua.2013.04.002.

Peter Hugoe MATTHEWS (1972), *Inflectional morphology: a theoretical study based on aspects of Latin verb conjugation*, CUP Archive.

Matti MIESTAMO *et al.* (2008), Grammatical complexity in a cross-linguistic perspective, in Matti MIESTAMO, Kaius SINNEMÄKI, and Fred KARLSSON, editors, *Language complexity: typology, contact, change*, pp. 23–41, Benjamins, Amsterdam.

- Fermin MOSCOSO DEL PRADO (2011), The mirage of morphological complexity, in *Proceedings of the annual meeting of theGG*, 33.
- Yoon Mi OH and François PELLEGRINO (2022), Towards robust complexity indices in linguistic typology: a corpus-based assessment, *Studies in Language*, pp. 1–41.
- Enrique L. PALANCAR (2021), Paradigmatic structure in the tonal inflection of Amuzgo, *Morphology*, 31(1):45–82.
- Jeff PARKER and Andrea SIMS (2020), Irregularity, paradigmatic layers, and the complexity of inflection class systems: a study of Russian nouns, in Peter ARKADIEV and Francesco GARDANI, editors, *The complexities of morphology*, Oxford University Press, Oxford.
- Matteo PELLEGRINI and Marco PASSAROTTI (2018), Latin-flexi: an inflected lexicon of Latin verbs, in Elena CABRIO, Alessandro MAZZEI, and Fabio TAMBURINI, editors, *Proceedings of the fifth Italian conference on computational linguistics (CLiC-it 2018)*, volume 2253, pp. 324–329, Accademia University Press.
- Neil RATHI, Michael HAHN, and Richard FUTRELL (2022), Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff, in *Proceedings of the annual meeting of the Cognitive Science Society*.
- Benoît SAGOT and Géraldine WALTHER (2011), Non-canonical inflection: data, formalisation and complexity measures, *SFCM*, 100:23–45.
- Andrea D. SIMS and Jeff PARKER (2016), How inflection class systems work: on the informativity of implicative structure, *Word Structure*, 9(2):215–239.
- Kaius SINNEMÄKI and Francesca DI GARBO (2018), Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: a typological study of verbal and nominal complexity, *Frontiers in Psychology*, 9, ISSN 1664-1078, doi:10.3389/fpsyg.2018.01141.
- Royal SKOUSEN (1989), *Analogical modeling of language*, Kluwer Academic Publishers.
- Royal SKOUSEN (1992), *Analogy and structure*, Springer.
- Royal SKOUSEN, Deryle LONSDALE, and Dilworth B. PARKINSON (2002), *Analogical modeling: an exemplar-based approach to language*, number 10 in *Cognitive Processing*, John Benjamins.
- Peter SMIT, Sami VIRPIOJA, Stig-Arne GRÖNROOS, and Mikko KURIMO (2014), Morfessor 2.0: toolkit for statistical morphological segmentation, in *The 14th conference of the European chapter of the Association for Computational Linguistics (EACL)*.
- Andrew SPENCER (2012), Identifying stems, *Word Structure*, 5(1):88–108.

The typology of inflectional complexity

Nicolas STROPPA and François YVON (2005), An analogical learner for morphological analysis, in *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pp. 120–127.

Gregory T. STUMP and Rafael FINKEL (2013), *Morphological typology: from word to paradigm*, Cambridge Studies in Linguistics, Cambridge University Press.

Géraldine WALTHER and Benoît SAGOT (2011), Modélisation et implémentation de phénomènes flexionnels non-canoniques, *Traitement Automatique des Langues*, 52(2):91–122.

Robert W. YOUNG (2000), *The Navajo verb system: an overview*, University of New Mexico Press.

Matías Guzmán Naranjo


© 0000-0003-1136-6836
mguzmann89@gmail.com

University of Freiburg

Matías Guzmán Naranjo (2024), An analogical approach to the typology of inflectional complexity, *Journal of Language Modelling*, 12(2):415–475

doi <https://dx.doi.org/10.15398/jlm.v12i2.352>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>