Lexical databases for computational analyses: A linguistic perspective

Lexical databases for computational analyses: A linguistic perspective

Robert Malouf[1], **Farrell Ackerman**[2], **and Arturs Semenuks**[3]
[1]Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University
rmalouf@sdsu.edu
[2]Department of Linguistics, University of California, San Diego
fackerman@ucsd.edu
[3]Department of Cognitive Science, University of California, San Diego
asemenuk@ucsd.edu

Abstract

Large typological databases have permitted new ways of studying cross-linguistic morphological variation. Recently, computational modelers with typological interests have begun to turn to broad multilingual text databases. In this paper, we will focus particularly on the UniMorph database, a collection of morphological paradigms, mostly gathered automatically from the crowd-sourced multi-lingual dictionary Wiktionary. It was designed to make the large quantity of data contained in Wiktionary available for NLP researchers by standardizing the data and putting it into a form that is easy to access. For typological studies, however, the requirements for a linguistically informed view of morphological variation are quite different. They involve using a morphological database as a scientific instrument to both formulate and test hypotheses about the nature and organization of language systems. The requirements are, accordingly, much higher. In this paper, we survey some of the methodological challenges and pitfalls involved in using corpora for typological research, and we end with a proposal for best practices and directions for further research.

1 Introduction

The availability of large typological databases (e.g., Dryer and Haspelmath 2013; Bickel and Nichols 2002) has made it possible to both model and hypothesize about the nature of cross-linguistic morphological variation. Recently, computational modelers with typological interests have begun to turn to broad multilingual text databases (e.g., Key and Comrie 2015; Dellert and Jäger 2017; McCarthy et al. 2018). While working from raw linguistic data opens up the possibility for new kinds of discoveries, it also poses significant challenges for the analyst, both with respect to the appropriateness of the selected data for explicitly specified goals and for identifying how these goals relate to alternatives that appeal to similar sorts of data.

Since Greenberg's (1963) pioneering work, we can roughly divide research in morphological typology into three strands. The first, and (arguably) most productive so far, has involved the careful construction of language samples designed by the author(s) of the study for answering specific questions. For example, Baerman et al. (2002, 2005) provide a cross-linguistic study of patterns of syncretism based on a database of all syncretic forms found in 30 genetically diverse languages and a larger database of person syncretisms in 111 languages, and Cysouw (2003) used a database of 102 types of person-marking system found in 309 languages.

This methodology has the advantage that both sample selection and coding is controlled by the researcher and can be designed specifically for the task at hand. However, while a few of the database entries may be based on the typologists' personal linguistic knowledge, for the most part information in the database is derived from dictionaries and grammatical descriptions, which necessarily reflect the analytic choices made by other linguists.

The second strand of typological research leverages the effort put into creating more general-purpose typological databases crafted to address multiple questions, but adaptable to address unanticipated and novel issues. For example, Bentz and Winter (2013) use the information about the case inventories of 261 languages in Iggesen (2013), which in turn is derived from Iggesen's (2005) detailed cross-linguistic study of case marking. Using existing resources in this way allows hypotheses about correlations among typolgical variables to be tested relatively easily, without months or years of work collecting language data. However, it is necessarily limited in the kinds of phenomena

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 297-307.
New Orleans, Louisiana, January 2-5, 2020

that can be examined, and is self-evidently dependent on the analytic choices made by the typologist who assembled the database and the linguists who wrote the grammars that the entries are based on.

Finally, a recent and very promising direction for morphological typology is the direct use of lexicons and corpora to extract cross-linguistics patterns (e.g., Wälchli and Cysouw 2012; Levshina 2016). This 'primary-data typology' has been made possible by the availability of large quantities of text in a diverse range of languages coupled with powerful statistical and computational methods. These methods allow us to investigate typological questions that cannot be addressed via grammatical descriptions. And, while all linguistic data is dependent (explicitly or implicitly) on an underlying analysis, working directly with texts makes us less dependent on the analytic choices made by other linguists. However, just as the other methodologies discussed above, this strand of typological research poses some significant challenges that researchers need to recognize and develop strategies to address.

In this paper, we will focus particularly on the UniMorph database (Kirov et al., 2016, 2018) and use it as a case study to highlight what types of obstacles 'primary-data typology' needs to take into account. UniMorph is a collection of morphological paradigms, mostly collected automatically from the crowd-sourced multi-lingual dictionary Wiktionary (`wiktionary.org`). It was designed to make the large quantity of data contained in Wiktionary available for NLP researchers by standardizing the data and encoding it in a form that is easy to access.

UniMorph has been broadly adapted as a testbed for evaluating morphological processors (e.g., Aharoni and Goldberg 2017; Shearing et al. 2018). Its main advantage is that it is larger and simpler to use than any existing competitors. While it is plausibly preferable to use broader typological samples as a measure of progress, one can make the argument that, all databases are flawed in some way, and evaluating systems on a variety of languages, however restricted, is certainly preferable to testing on only English data. There is a danger of 'overfitting' to standard datasets as a research community, but this can be minimized by continuing to expand and improve available test sets (Kyle Gorman and Markowska, 2019).

Another promising use for resources like UniMorph is for evaluating claims about morphological systems in general separate from the tools we use to process them. For example, a number of recent papers (e.g., Cotterell et al. 2019; Pimentel et al. 2019; Wu et al. 2019) have used UniMorph to offer answers to some basic questions about the structure of morphological systems. But, in contrast to the the engineering applications of UniMorph, the requirements for engaging in such a linguistically informed view of morphological variation are quite different. They involve using a morphological database as a scientific instrument to both formulate and test hypotheses about the nature and organization of language systems. The requirements (and the stakes) are, accordingly, much higher. In linguistics, as in any other field, analysis of an inappropriate data sample can lead to misplaced confidence in unsupported conclusions and unlicensed general inferences about e.g., morphological organization.

It seems likely that the UniMorph project can form the basis of a database suitable for use in typological research, if suitably modified. Forms in the UniMorph database are annotated with features from the UniMorph Schema (Sylak-Glassman, 2016), and considerable effort was put into designing these feature representations to allow cross-linguistic comparison of categories. But, in contrast to this care, the selection of languages in the sample was made opportunistically determined by what was available in Wiktionary, rather than being selected to explore different strategies of morphological organization and related questions concerning the learnability of attested systems. These are core linguistic concerns in relation to the typological sampling of empirical phenomena.

In the following sections, we will survey some of the methodological challenges and pitfalls involved in using corpora for typological generalizing, and we will end with a proposal for best practices and directions for further research.

2 Representativeness

Any database that purports to develop generalizations about language in general has to be representative of the range of possible human languages. UniMorph[1] includes data from 106 languages, including noun paradigms for 74 and verb paradigms

[1]The version of UniMorph we use for this paper consists of all repos with three letter names containing a datafile with a three letter name in the `https://github.com/unimorph` organization, downloaded on 27 July 2019.

for 87. These languages represent 16 families (e.g., Indo-European, Uralic) and 30 genera (e.g., Celtic, Finnic). This is a very small fraction of the world's languages. By comparison, the World Atlas of Linguistic Structures (Dryer and Haspelmath, 2013) includes data for 2,679 languages representing 256 families and 544 genera in total. Or, since WALS does not include values for every feature for every language, the median feature in WALS is specified for 257 languages in 96 families and 177.5 genera.

A small sample, correctly constructed, can support cross-linquistic inferences. However, the languages in UniMorph are not representative of the diversity of human language. Almost half (47 out of 106) of the languages in UniMorph are from just three genera (Romance, Germanic, and Slavic). While the problem of individuating and enumerating languages is a difficult one with no clear solution, some of the 'languages' in UniMorph are arguably not different languages and would normally be considered dialects of a common language (e.g., German, Low German, Middle High German, and Middle Low German). Sometimes the same language is given different names and treated as if it were multiple languages for political or historical reasons.

In addition, 98 of the languages in UniMorph are spoken in Eurasia (i.e., the landmass comprising Europe and Asia) with only three languages in North America, two languages in each of South America and Africa, and only one language in Australia (see Figure 1). As Dryer (1989) demonstrated, Eurasian languages are not generally representative of languages throughout the world. This reinforces the observation that any representative sample needs to include languages with wide geographic and phylogenetic dispersion.

In addition to genetic and geographic homogeneity the data lack varietal representativeness with respect to word structure. The languages in the sample are overwhelmingly of a familiar morphological type, organized around stems and affixes. The African languages in the sample are both Bantu languages (Swahili and Zulu), which are broadly similar to Eurasian languages with respect to displaying a concatenatively affixal strategy for morphotactic organization. The four Semitic languages in the sample show one kind of templatic morphology, but no languages in the sample use tones, reduplication, vowel length patterns, or many other types of morphological expression.

By its nature, Wiktionary only includes languages with a written form and those mostly using their practical orthography, in contrast to phonologized lexicons such as Flexique (Bonami et al., 2013). This raises several potential problems. Of particular note, orthographic systems vary widely in phonological transparency, and many orthographies neglect important distinguishing morphophonological details such as tone, segment length, and stress placement (e.g., see Parker and Sims in press): this creates problems with respect to identifying the correct inventory of forms that need to be compared. For example, the Estonian orthography underrepresents "gradation" in all but the stop consonants and, thereby, misrepresents the actual variety of contrasting forms in Estonian paradigms. Roughly speaking, Estonian consonants and vowels display a three-way contrast (short, long, and overlong) which is not represented in the orthography. This leads to the following differences in the orthographic representations versus the phonological reality for the noun *keel* 'language' (Mürk, 1997, 107):

	Orth.	**Phon.**
NOM SG	keel	ke::l
GEN SG	keele	ke:le
PART SG	keelt	ke:lt::
ILL SHORT SG	keelde	ke::le ~ ke::lde
GEN PL	keelte	ke:lte

Finally, different scripts may pose different modeling challenges, making it difficult to directly compare a model-based metric across languages written using various alphabets, abjads, syllabaries, etc.

A sample of 106 *closely related* or overlapping written languages provides a lot less information about the space of possible languages than a sample of 106 *unrelated* languages would. This is not a flaw in UniMorph per se and it does not reduce its value as a test-bed for developing morphological processors, particularly for the constrained class of variation it models. Given the limited range of morphological variation represented in UniMorph, any results concerning morphological organization beyond that sample can only support modest claims to greater generality, which themselves need to articulated into testable hypotheses. This is, of course, the same standard appropriately posed for linguistic theories that seek to motivate wide ranges of morphological organization exhibiting extraordinarily divergent strategies of surface

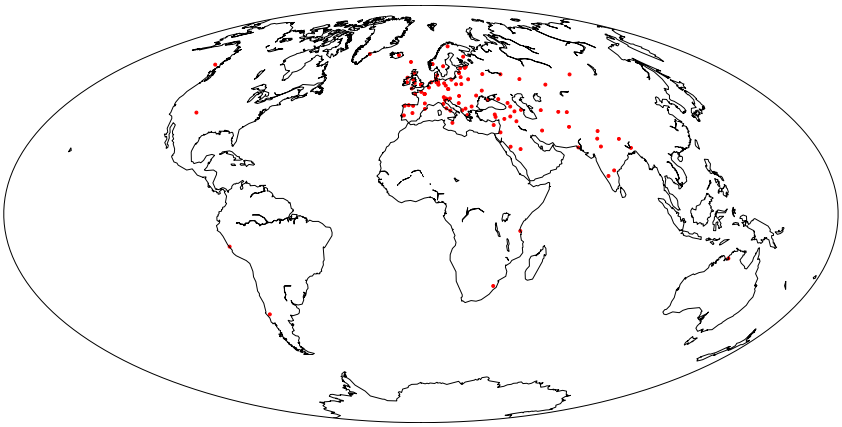

Figure 1: Geographic distribution of languages in UniMorph (languages locations from (Dryer and Haspelmath, 2013))

encoding: their credibility too is dependent on the empirical scope and reliability of the data they analyze.

3 A case study

As a concrete example, we will consider the relationship between paradigm size and predictability in morphological paradigms. Ackerman and Malouf (2013) distinguish two dimensions of morphological complexity: E-complexity (the number of affixes, allomorphs, inflection classes, etc.) and I-complexity (the interpredictability of forms in a paradigm). Ackerman and Malouf (2013) conjecture that I-complexity is what is relevant for language learnability, and that across languages E-complexity can vary widely so long as I-complexity is low enough. More recent work (Cotterell et al., 2019; Semenuks, 2019) suggests that E-complexity and I-complexity may be interrelated, and that the threshold for 'low enough' I-complexity may decrease as E-complexity increases. In what follows, we will consider some of the methodological choices that need to be made in order to properly test this claim.

For the sake of discussion, we will measure E-complexity as paradigm size, or the number of distinct feature values encoded in the database. For example, if a nominal paradigm encodes 7 cases and 2 numbers, the size of the paradigm is 14. If the paradigm size varies between lexemes, we use the most common value (i.e., the mode). To estimate I-complexity or predictability, we train a model to map a citation form and feature set to a surface form (SIGMORPHON 2016 task 1; Cotterell et al. 2016). Specifically, we use a neural encoder-decoder architecture (Kann and Schütze, 2016; Silfverberg and Hulden, 2018) implemented using OpenNMT-tf (Klein et al., 2017). Using the model, we then calculate the average per-form negative log likelihood ($-L$) of held out data.[2] The closer this value is to zero, the better the model is able to predict the correct forms. Note that we are not claiming that this is the correct way to estimate either E- or I-complexity: we have chosen it mostly because it is easy to calculate in a reproducible way. Our goal is to focus on methodological issues, not the viability of any specific linguistic analysis.

3.1 Lexicon size

One issue that immediately arises is that the performance of neural models can be highly dependent on the quantity of training data. Since there are large differences in lexicon sizes across languages in UniMorph, difference in model prediction (reflected in $-L$) may be due to training issues and not to structural differences between languages. This, of course, is important to know, since otherwise our results might be comparing incomparable phenomena.

[2] See https://github.com/rmalouf/SCiL2020 for implementation details.

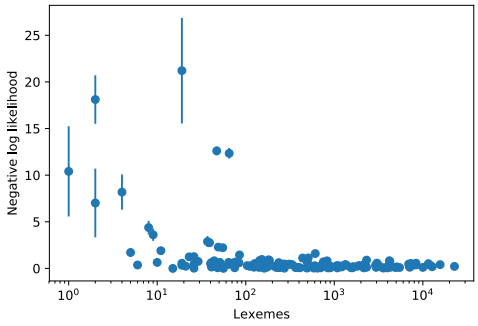

Figure 2: Negative log likelihood ($-L$) vs. lexicon size

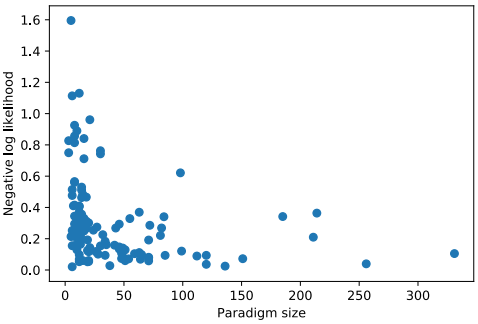

Figure 3: Negative log likelihood ($-L$) vs. paradigm size, for paradigms with >100 lexemes

To test this, we performed five-fold cross-validation to estimate $-L$ and its standard error (the standard deviation divided by $\sqrt{5}$). The results for the 87 verb paradigms and 73 noun paradigms (we exclude Tajik nouns, which list only one inflected form per lexeme) are given in Figure 2. For languages with small lexicons (≤100 lexemes), we see both poor model performance (i.e., high $-L$) and high variability across train/test splits. For languages with more than 100 lexemes, however, performance looks much more consistent.

If we exclude paradigms with fewer than 100 lexemes, we are left with 55 noun paradigms and 61 verb paradigms over a total of 77 languages. The results are shown in Figure 3. At first glance, this appears to support the claim that languages can have higher I-complexity if they have low E-complexity. But, this is only true if high $-L$ is due to structural properties of the language being tested. In the following sections, we will look at a number of factors that can increase $-L$ for particular languages without any increase in I-complexity.

3.2 Overabundance

One issue that arises in examining the UniMorph data is that many (sub)paradigms permit more than a single form in a cell for a given lexeme: particular combinations of feature values can be realized by more than one exponent. For example, the past tense of English *dive* can be either *dived* or *dove*. There are several causes for this. Some examples are simply data processing errors: two distinct forms have been erroneously assigned the same feature values in extracting the data from Wiktionary. In other cases, the forms do share the same features but are not interchangeable for other reasons.

For example, the Spanish lexicon lists both *sentir* and *sentirse* as infinitive forms of the verb *sentir* 'to feel', even though the second of the two forms is (arguably) the infinitive of a different lexeme. Similarly, the Zulu verb lexicon lists both *ngiyadla* and *ngidla* as the 1st person singular present tense positive absolute form of the verb *ukudla* 'to eat'. But, these forms are not completely synonymous. The exact nature of the difference between these forms is unclear (see, e.g., Buell 2006), but they should be distinguished somehow.

The majority of cases, however, are due to genuine **overabundance**: multiple forms are listed because multiple forms are possible (Thornton, 2011, 2019). Wiktionary lists *troféen* or *trofeen* or *troféet* or *trofeet* as alternate definite singular forms of *trofé* 'trophy' in Norwegian Nynorsk, with no difference in meaning. This creates a problem for any metric which assumes that every paradigm cell has exactly one realization. This includes models evaluated using accuracy or, in our case, negative log likelihood. Using our metric, paradigms exhibiting overabundance will show higher negative log likelihood than ones that do not, for reasons that have no connection to how predictable or systematic the morphological system is.

Overall, although many languages left in the sample don't have any lexemes with multiple forms filling in a paradigm cell, it is also not rare: 18 out of 55 languages with noun paradigms and 19 out of 61 languages with verb paradigms exhibit this pattern, out of which 16 (for nouns) and 14 (for verbs) have more than multiple forms in cell for more than 10% of the lexemes. Regardless of whether the reason for this pattern is genuine overabundance or data processing errors, it nevertheless introduces difficulties into further analyses.

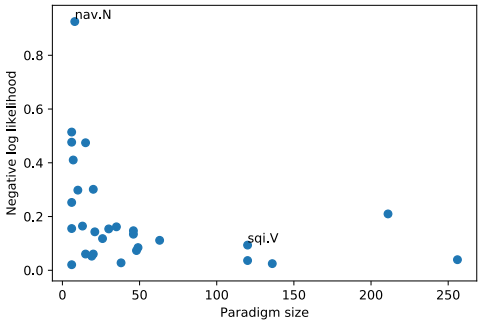

Figure 4: Negative log likelihood vs. paradigm size, for reduced sample

3.3 Defectiveness

Paradigms in the UniMorph database display many missing forms. In many cases this is due simply to incompleteness: the forms exist, but for whatever reason are not included in Wiktionary or were not extracted. However, missing forms can also be due to paradigm **defectiveness**. This is the converse of overabundance: these are paradigm cells for which there is no valid realization.

Like overabundance, missing data raises problems for any metric which assumes that every paradigm cell has exactly one realization. Forms which are missing due to incompleteness may have the effect of hurting model performance (and raising $-L$) without an underlying difference in predictability. If forms are missing due to true paradigm defectiveness, then the fact that the form is missing is something that the model needs to learn. As argued by Sims (2015), the absence of a form is as much a part of the morphological system as its presence.

3.4 Complications

To avoid modeling problems raised by overabundance and defectiveness, we can remove from the sample any paradigms with any overabundant forms and more than ten defective paradigms. This leaves 17 verb paradigms and 12 noun paradigms from 28 languages. The results for this reduced sample are shown in Figure 4 and Table 1.

The outlier in the upper left (nav.N) is Navajo nouns. The high $-L$ value for Navajo nouns is surprising, as Navajo nominal morphology is fairly straightforward. Examination of the data shows a number of inaccuracies or infelicities in the data that lead to poor model performance.

Some of the errors were introduced in the process of extracting forms from Wiktionary. The paradigm for *éé'* 'clothes' is shifted up one row: the 1p singular possessed form is listed as *singular* rather the correct *she'éé'*, the 2p singular is listed as *she'éé'* rather than *ne'éé'*, and so on.

Most of the problems with the Navajo nominal data, however, are consequences of the decisions made by the designers of the Navajo wiktionary. First, a brief summary of Navajo nominal morphology: nouns in Navajo form a fairly small, closed class. Inalienably possessed nouns (mostly kin relations and body parts) appear in an indefinitely possessed form (*átáá'* 'someone's forehead') or with a possessive prefix (*shítáá'* 'my forehead'). Alienably possessed nouns may appear as a bare stem (*sǫ'* 'star'), as possessed form (*azǫ'* 'someone's star'), or as a possessed form with a possessive prefix (*shizǫ'* 'my star'). The possessive prefixes show relatively little allomorphy, but the possessed form and the bare stem sometimes differ in arbitrary ways. Most Navajo nouns are unmarked for number, but a few personal nouns take a plural suffix *-ké* or *-yóó*.

The Navajo noun paradigms in Wiktionary list only the possessed forms. For alienably-possessed nouns, the bare stem (e.g., *sǫ'*) is the citation form for the lexeme but is not included in the paradigm. For inalienably-possessed nouns, the indefinite possessed form is the citation form. This inconsistency makes the two noun classes look more different than they actually are. More problematic is the fact many nouns have separate dictionary entries for possessed forms: *ké* 'foot' is also listed under *bikee'*, *hakee'*, and *akee'*, the 3rd person, 4th person, and indefinite possessed forms. From the model's perspective, this looks like four separate lexemes (with four different citation forms) that happen to share the same inflected forms.

Three other high $-L$ paradigms in Table 1 are Pashto nouns, Urdu nouns, and Yiddish verbs. Like all the language samples, these paradigms are written using the practical orthography of the language. In the case of Urdu and Pashto, the writing system (based on Arabic by way of Persian) is an **abjad**: consonants are included, but many vowels are left unspecified when they should be clear to the reader from context. The Yiddish alphabet is adapted from Hebrew and is a full alphabet, but the mapping between Yiddish letters and

Language	pos	features	$-L$	s.e.	macroarea	family	genus
Albanian	V	120	0.094	0.002	Eurasia	Indo-European	Albanian
Ancient Greek	N	15	0.475	0.018	Eurasia	Indo-European	Greek
Bulgarian	V	20	0.060	0.012	Eurasia	Indo-European	Slavic
Catalan	V	48	0.073	0.002	Eurasia	Indo-European	Romance
Classical Syriac	N	13	0.164	0.112	Eurasia	Afro-Asiatic	Semitic
Crimean Tatar	N	6	0.155	0.021	Eurasia	Altaic	Turkic
Danish	V	6	0.021	0.018	Eurasia	Indo-European	Germanic
Dutch	V	15	0.060	0.006	Eurasia	Indo-European	Germanic
Estonian	N	30	0.154	0.014	Eurasia	Uralic	Finnic
Friulian	V	46	0.147	0.023	Eurasia	Indo-European	Romance
Georgian	N	19	0.052	0.006	Eurasia	Kartvelian	Kartvelian
Hebrew	N	26	0.118	0.027	Eurasia	Afro-Asiatic	Semitic
Hindi	V	211	0.210	0.116	Eurasia	Indo-European	Indic
Irish	V	63	0.111	0.010	Eurasia	Indo-European	Celtic
Lithuanian	V	49	0.084	0.010	Eurasia	Indo-European	Baltic
Lower Sorbian	V	21	0.143	0.058	Eurasia	Indo-European	Slavic
Navajo	N	8	0.925	0.317	North America	Na-Dene	Athapaskan
Occitan	V	46	0.134	0.013	Eurasia	Indo-European	Romance
Pashto	N	6	0.477	0.125	Eurasia	Indo-European	Iranian
Persian	V	136	0.025	0.006	Eurasia	Indo-European	Iranian
Quechua	N	256	0.039	0.023	South America	Quechua	Quechua
Quechua	V	38	0.028	0.016	South America	Quechua	Quechua
Romanian	V	35	0.162	0.026	Eurasia	Indo-European	Romance
Slovenian	V	20	0.301	0.042	Eurasia	Indo-European	Slavic
Tatar	N	6	0.252	0.024	Eurasia	Altaic	Turkic
Turkish	V	120	0.036	0.006	Eurasia	Altaic	Turkic
Urdu	N	6	0.514	0.107	Eurasia	Indo-European	Indic
Yiddish	V	7	0.410	0.177	Eurasia	Indo-European	Germanic

Table 1: Results for reduced sample

Unicode characters is not one-to-one. It is possible that these orthographic differences might make estimates of $-L$ difficult to compare across languages with different writing systems.

Ancient Greek nouns also have a high $-L$, but likely not for orthographic reasons. Rather, these paradigms encode overabundance using punctuation rather than multiply filled paradigm cells. For example, the genitive singular of κοῦρος 'youth' is given as "κούρου / κουροῖο / κούροιο / κουρόο / κούροο". This is presumably meant to reflect five variant forms, but the model would count that as one long (and hard to guess) form.

Another outlier, this time in the number of features, is Albanian verbs (sqi.V). According to UniMorph (and Wiktionary), each Albanian verb has 120 distinct forms. However, this number includes periphrastic tenses formed by combining an inflected verb with a particle and/or an auxiliary verb. This is a bit like counting *will have been being seen* as a distinct form of the verb 'see' in English.

The design choices embodied in Wiktionary are not necessarily incorrect. It is helpful for Navajo learners to have separate dictionary entries for prefixed forms. And, a strong argument can be made that periphrastic forms should be included as part of the paradigm in both Albanian and in English (e.g., Ackerman and Webelhuth 1998; Ackerman and Stump 2004; Bonami 2015). But, if one's goal is to use UniMorph data for cross-linguistic comparison, then these kinds of choices need to be made in a standardized way and clearly articulated. The issue is not whether data choices are right or wrong, but whether those choices are transparent and appropriate for a particular use.

3.5 Galton's problem

Even excluding Navajo nouns and the other outliers, the pattern of languages shown in Figure 4 suggests that languages in the sample with large paradigms show low $-L$. Without Navajo nouns, there are 17 verb paradigms and 11 noun paradigms from 27 languages in the sample. Is this enough to draw any conclusions about language in general?

So far, in our discussion we have used quantitative but not statistical methods. The difficulty with applying standard hypothesis testing methods to the problem is that languages that are genetically and/or areally related cannot be treated as independent observations. Of the 23 languages in the remaining sample, 16 are Indo-European and 21 are Eurasian. If the data is not analyzed using methods taking these phylogenetic and geographic proximities between the data points into account, the analyses could produce spurious correlations (Roberts and Winters, 2013). This is what Naroll (1965) calls **Galton's Problem**: the problem of making inferences based on auto-correlated observations.

Early work in quantitative typology addressed this problem through careful sample construction (Bybee, 1985; Dryer, 1988; Perkins, 1989). More recent efforts have applied hierarchical modeling techniques to control for genetic and areal affects. A survey of these techniques is beyond the scope of this paper, but see Bakker (2011) and (Bickel, 2015) for some proposals.

3.6 Construct validity

Based on the results so far, there is suggestive evidence for a relationship between the number of cells in a paradigm and $-L$ as predicted by an encoder-decoder model. The final step in any typological study has to be to show that these metrics applied in this way to this dataset connect to a relevant linguistic notion. In this case, a crucial question is whether $-L$, a measure of how well a model predicts forms, is a reasonable measure of the I-complexity of a paradigm, or how predictable forms are. This is the question of **construct validity**: does the test measure what it claims to measure?

As we said above, our goal in this paper is to highlight some of the methodological issues that come in using text databases (such as UniMorph) for typology. Our use of $-L$ is only for the sake of demonstration and we make no particular claims about its linguistic relevance. But, if this were a paper making a typological claim, then it would be essential to justify our confidence in the particular metric being used. Readers need to keep this requirement in mind when assessing and interpreting the linguistic value of results based on computational analyses of natural language data.

4 Conclusions

Large text databases open up exciting prospects for typological research, but they also create new challenges for cross-disciplinary collaboration: linguistic morphologists and typologists are practiced curators of the types of data that are most profitably investigated by new computational techniques. The previous section presented a hypothet-

ical typological investigation using UniMorph in order to highlight some of the difficulties in carrying out such an investigation. Any work applying computational models to primary linguistic data (e.g., information-theoretic investigations of UniMorph along the lines of Cotterell et al. 2019; Pimentel et al. 2019; Wu et al. 2019) need to be carried out and evaluated with these challenges in mind. As an emergent interdisciplinary community, we should develop a set of best practices for using the resources we have and in developing a collaboratively determined direction for improving those resources.

As a start, we propose some basic requirements:

- Use UniMorph (Kirov et al., 2016, 2018) as a resource for building databases, not as a database itself: text databases should be seen as a guide for formulating directions of inquiry and identifying the types and nature of data required for systematic inquiry. The data established for this purpose must be reliable and representative for the task at hand.

- Document all choices: In order to achieve maximum transparency and replicability, all choices concerning data selection, preprocessing, representation, parsing, and modeling should be clearly specified, along with their rationales.

- Intended claims and hypotheses associated with analysis and results should be clearly articulated in order to identify their importance in the context of similar research within relevant linguistic approaches to morphological analysis. This is crucial in order to evaluate the research results from both a linguistic and computational perspective: if such results are novel, in what ways do they contribute to our understanding of natural language morphology and to the computational analysis of morphological phenomena.

- Given the cross-disciplinary nature of the relevant contributions, the vetting process for the evaluation of submissions should be distributed among linguists and computational modelers, in order to ensure research that reflects the most accurate and critical assessments from contributing fields.

References

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

Farrell Ackerman and Gregory Stump. 2004. Paradigms and periphrasis: A study in realization-based lexicalism. In Louisa Sadler and Andrew Spencer, editors, *Projecting Morphology*, pages 111–157. CSLI Publications, Stanford.

Farrell Ackerman and Gert Webelhuth. 1998. *A Theory of Predicates*. CSLI Publications, Stanford.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2002. The Surrey syncretisms database. `http://www.smg.surrey.ac.uk/syncretism/index.aspx`.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2005. *The Syntax-Morphology Interface: A Study of Syncretism*. Cambridge University Press.

Dik Bakker. 2011. Language sampling. In Jae Jung Song, editor, *The Oxford Handbook of Typology*. Oxford University Press.

Christian Bentz and Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, pages 1–27.

Balthasar Bickel. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Balthasar Bickel, Bernd Heine, and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis (2 ed.)*. Oxford University Press.

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas*.

Olivier Bonami. 2015. Periphrasis as collocation. *Morphology*, 25:63–110.

Olivier Bonami, Gauthier Caron, and Clément Plancq. 2013. Flexique: an inflectional lexicon for spoken French.

Leston Buell. 2006. The Zulu conjoint/disjoint verb alternation: focus or constituency? In Laura J. Downing, Lutz Marten, and Sabine Zerbian, editors, *Papers in Bantu grammar and description*, pages 9–30. Zentrum für Allgemeine Sprachwissenschaft, Sprachtypologie und Universalienforschung, Berlin.

Joan L. Bybee. 1985. *Morphology: A study of the relation between meaning and form*. Benjamins, Philadelphia.

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.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin. Association for Computational Linguistics.

Michael Cysouw. 2003. *The Paradigmatic Structure of Person Marking*. Oxford University Press.

Johannes Dellert and Gerhard Jäger, editors. 2017. *NorthEuraLex (version 0.9)*.

Matthew S. Dryer. 1988. Object-verb order and adjective-noun order: Dispelling a myth. *Lingua*, pages 185–217.

Matthew S. Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language*, 13(2):257–292.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Grammar*, pages 73–113. MIT Press, Cambridge.

Oliver A. Iggesen. 2005. *Case-asymmetry: A World-Wide Typological Study on Lexeme-class-dependent Deviations in Morphological Case Inventories*. Lincom Europa, Muenchen.

Oliver A. Iggesen. 2013. Number of cases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Mary Ritchie Key and Bernard Comrie, editors. 2015. *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126.

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

Ryan Cotterell Ekaterina Vylomova Miikka Silfverberg Kyle Gorman, Arya D. McCarthy and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *CoNLL 2019*.

Natalia Levshina. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica*, 50(2):507–542.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Harri William Mürk. 1997. *A Handbook of Estonian: Nouns, Adjectives and Verbs*. Indiana University Uralic and Altaic Series, v. 163. Indiana University, Bloomington.

Raoul Naroll. 1965. Galton's problem: The logic of cross-cultural research. *Social Research*, 32:428–451.

Jeff Parker and Andrea Sims. in press. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. In P. Arkadiev and F. Gardani, editors, *The Complexities of Morphology*. Oxford University Press.

Revere D. Perkins. 1989. Statistical techniques for determining language sample size. *Studies in Language*, 13:293–315.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Seán Roberts and James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS ONE*, 8(8):1–13.

Arturs Semenuks. 2019. Investigating relationship between i-complexity and population size. Poster presented at the Workshop on Interaction and the Evolution of Linguistic Complexity, Edinburgh.

Steven Shearing, Christo Kirov, Huda Khayrallah, and David Yarowsky. 2018. Improving low resource machine translation using morphological glosses (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 132–139, Boston, MA. Association for Machine Translation in the Americas.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Andrea D. Sims. 2015. *Inflectional Defectiveness*. Cambridge University Press, Cambridge.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (UniMorph schema). working draft, v. 2.

Anna M. Thornton. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In Martin Maiden, John Charles Smith, Maria Goldbach, and Marc-Olivier Hinzelin, editors, *Morphological Autonomy: Perspectives From Romance Inflectional Morphology*. Oxford University Press.

Anna M. Thornton. 2019. Overabundance in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.