Finding proportionality in computational approaches to morphological change

Alexandru Craevschi[*1,2], Sarah Babinski[1,2], and Chundra Cathcart[1,2]

[*]Corresponding Author: alexandru.craevschi@uzh.ch
[1]Department of Comparative Language Science, University of Zurich
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Abstract

Proportional mechanisms are thought to play a major role in morphological change. This paper explores the extent to which simple models of morphological inflection embody proportional behavior. Using models with varying architectures and decoding schemes, we find that errors produced by these models do not often form a valid proportion with forms found in the training data. We discuss the implications of this finding for research seeking to recapitulate diachronic processes using models of this sort.

1. Introduction

Morphological paradigms change over time. Analogical change is a significant factor driving morphological shifts of this kind: when attempting to produce an inflected form, language users draw upon their knowledge of inflectional patterns from other word forms, sometimes resulting in alterations to the intended target.

Morphological paradigms can undergo restructuring through various analogical mechanisms. Traditionally, analogical change mechanisms are categorized into two types: those involving proportional and non-proportional mechanisms (Paul, 1880; Anttila, 1977; Gaeta, 2007; Hock, 2009). While paradigms may occasionally be restructured via non-proportional mechanisms (Haspelmath, 1994; Fertig, 2016; Sims-Williams, 2016), the most commonly cited types of changes affecting paradigms are analogical extension and leveling, which are typically understood to operate proportionally (Hill, 2007; Garrett, 2008). Proportional analogy refers to a phenomenon where a form or pattern is extended or generalized to create new forms or patterns within a language, maintaining consistent relationships between elements. This process involves adhering to linguistic constraints while generating both attested and unattested forms based on established patterns or paradigms. These changes fit within a framework of analogical proportions, exemplified in (1a). A proportion generates both attested and unattested forms, but for it to be considered valid, it must adhere to the linguistic constraints of the language. For

instance, (1b) presents a well-formed and attested analogical proportion in Latin, where the pluralization of the second-declension Latin noun *rīvulus* is patterned after another second-declension noun *fabulus*. Conversely, (1c) demonstrates an invalid proportion where the fourth-declension noun *cēnsus* is incorrectly pluralized as **cēnsī* based on the pattern of *fabulus*; the attested plural form for *cēnsus* would be *cēnsūs* in accordance with its noun class. This disparity illustrates the importance of maintaining linguistic congruence within analogical proportions, as seen in the ill-formed proportion (1d) which attempts to apply a feminine form ending in *-a* to generate the plural of a masculine noun ending in *-us*.

(1) a. A : B :: C : *x*
b. fabulus : fabulī :: rīvulus : **rīvulī**
c. fabulus : fabulī :: cēnsus : ***cēnsī**
d. fābula : fābulae :: rīvulus : ***rīvulae**

The extent to which computational models of morphological change exhibit proportional behavior remains unexplored. Earlier computational work on morphological learning exploits pairwise relationships between inflected forms in order to establish proportional bases for generating inflectional forms (Neuvel & Fulop, 2002). However, the role of proportionality in neural models of inflection, which learn linear and/or nonlinear mappings between semantic features and phonological cues, is not fully understood. Linear discriminative learning (LDL) (Baayen, Chuang, & Blevins, 2018; Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019) is a framework which maps meaning to form and vice versa by learning linear relationships between vector semantic and phonological cues. Its proponents argue that LDL generalizes the standard four-part analogy (1a) beyond set-based conceptions of semantics (e.g., {DOG, SINGULAR}) to vector semantic representations representing the collocational distributions in which a form is found.

In this study, we explore the extent to which proportional behavior emerges in computational models of morphological inflection without the models being explicitly coded to use four-part analogies from (1a) in the process of inference. We apply models of morphological inflection to morphological data sets from different languages, allowing different properties of the models used, namely the architecture and decoding schemes, to vary across model settings. We employ an algorithm to find proportions in the training data that support attested and predicted forms in the test data. Models with a linear regression-based architecture perform consistently better than Long Short-Term Memory (LSTM) models in terms of rates of proportional errors. Proportional support for a test form in the training data is a significant predictor of whether or not a morphological inflection model will generate it accurately, though this can be interpreted as a proxy for type frequency. Analyses of the errors produced by the models show an overwhelmingly low degree of proportionality. Our results suggest that if changes in morphological paradigms

are overwhelmingly proportional, then computational models of morphological learning should be used with care when simulating historical changes.

2. Data

Verbal paradigms were sourced from Unimorph (McCarthy et al., 2020) and converted to a broad IPA transcription using Epitran (Mortensen, Dalmia, & Littell, 2018) for most languages, with a few manual corrections. Phonemic transcriptions for English were taken from the Carnegie Mellon Pronouncing Dictionary (Rudnicky, 2015), available through the Natural Language Toolkit (Bird, Klein, & Loper, 2009). The glosses available in UniMorph, as well as lemmas, were converted to one-hot representations of inflectional features. Models were applied to data from the following languages: Arabic, Dutch, English, Italian, German, Polish, Portuguese, Russian, and Spanish. The main criteria for selection were (1) availability of verbal paradigms in UniMorph; and (2) availability of a grapheme-to-phoneme (g2p) system for obtaining phonological representation of forms. To alleviate the problem of different numbers of lemmas available per language and also to avoid extreme processing times for some of the data sets, we limited ourselves to a sample of 500 lemmas per language or just used all the lemmas, in case a language has fewer than 500 verbal lemmas (e.g., Zulu). Data and code are available at `https://gitlab.uzh.ch/chundra.cathcart/evolang_2024`.

3. Methods

We evaluate the performance of four models varying in the way meaning is mapped onto form and the way predicted sequences are generated. We probe the extent to which errors produced are supported by a proportional basis in the training data, and explore other properties of proportionality with respect to model performance.

Our models vary in the way they map meaning onto form. Following Baayen et al. (2018, 2019), in one set of models we use linear regression to learn linear mappings between inflectional features and trigram phoneme sequences, which can be used to predict phoneme sequences from inflectional features. Linear regression models were fitted using ordinary least squares.

Another set of models utilizes Long Short-Term Memory (LSTM) neural networks to introduce non-linearity. These models follow a standard encoder-decoder architecture commonly employed in sequence-to-sequence tasks. The architecture consists of two embedding layers, one for inflectional features and another for phonemic form. The inflectional features' embedding is fed into the LSTM encoder, with both embedding and hidden layer dimensions set to 128. The output of the encoder is then passed to the decoder for generation of the phonological form. LSTM models were implemented in Keras (Chollet et al., 2015) and trained with the Adam optimizer (Kingma & Ba, 2015) with a categorical cross-entropy loss function. The models were stopped early once overfitting on validation data was observed.

Finally, we vary the models with two different methods for sequence generation of predicted forms at the inference stage: beam search and greedy decoding. Greedy decoding selects the most probable token at each step, in this case selecting the most probable initial trigram/phoneme given some inflectional features before moving on to the following trigram/phoneme. Beam search, on the other hand, maintains a set of top-N candidates, exploring multiple possibilities simultaneously. We considered the top 2 most probable candidates at each generation step, ultimately picking out the sequence with the highest probability overall. For linear regression models decoded using beam search, we follow Baayen et al. (2018) in training a second model that maps trigram phoneme sequences to semantic vectors, choosing the candidate sequence whose predicted semantics correlates most strongly with the input semantic vector.

Models are run separately for each language. To evaluate model performance, we carry out K-fold cross-validation ($K = 10$), randomly holding out 10% of the forms in each data set used as test data. We vary the random number seeds used to sample lemmas and generate folds, using 5 different seeds for each stochastic dimension.

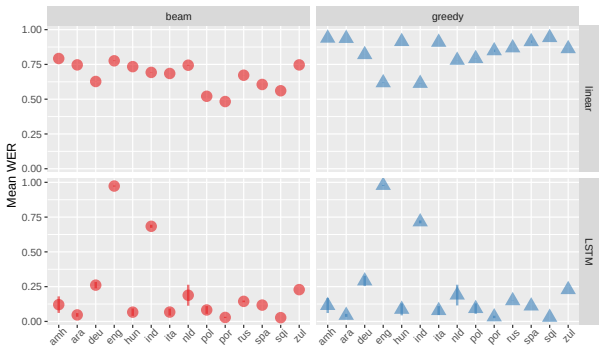

Figure 1. Mean word error rate (WER) by language for each model setting. Error bars (where visible) represent variance across different folds and random number seeds.

4. Model Results

We assess model performance according to the word error rate (WER, the proportion of test items that are produced with at least one error) and the phoneme error rate (PER, the mean normalized Levenshtein distance between each target and each predicted form). WER and PER values are displayed in Figures 1–2.

WER values are relatively high for linear models. Results from LSTM models show considerably lower WER values, with the exception of English and Indone-

Figure 2. Mean phoneme error rate (PER) by language for each model setting. Error bars (where visible) represent variance across different folds and random number seeds.

sian. The poor performance for these languages is striking, and may be due to their generally smaller paradigm size in comparison to the other languages, which are morphologically richer. PER values display a similar trend, although the difference between LSTM and linear regression models is significantly lower in case of PER.

Beam search and greedy decoding schemes do not show consistent differences from each other in terms of performance for these two error metrics. Certain languages show better results in greedy decoding with others benefiting more from beam search. In LSTM models, the differences between beam search and greedy decoding are almost nonexistent.

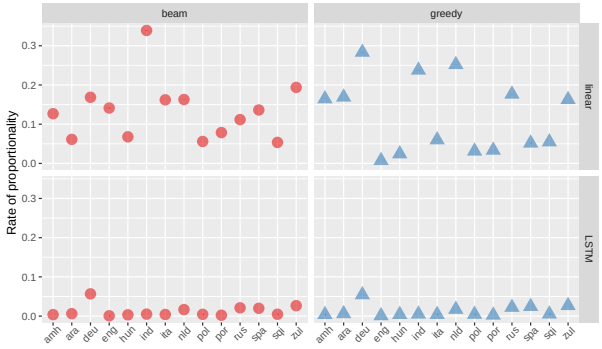

Figure 3. Proportion of model errors with support from at least one proportional basis, by language and for each model setting.

5. Proportionality

We identify proportions in the training data that could give rise to forms in the test data (following Lepage, 1998; Sims-Williams, 2016, 2021). For a given paradigm cell c_i in a given lemma ℓ found in a test split, we iterate through all lemmas $l \in \boldsymbol{\ell}_{\text{training}}$ in the training data, and for each cell $c_j \neq c_i$, generate the proportion $(l, c_j):(l, c_i)::(\ell, c_j):x$. We tabulate the number of proportions in the training data that support each attested target form as well as each predicted form. Using a mixed-effects logistic regression model with word error rate (with values of 1 representing errors) as a response variable and log-transformed proportional strength as a predictor with random intercepts and slopes by language, architecture, and decoding scheme, we find that the proportional support for an attested target form is a significant predictor of accuracy, though the effect is weak ($\beta = -0.0055$ $p < .001$). This may not indicate anything interesting about the effect of proportionality on model accuracy, but may have to do more generally with type frequency. We compute the proportion of errors for each model for which at least one proportion is available in the training data. These values are displayed in Figure 3. Linear architectures generate more proportional errors than LSTM. In many cases, these errors involve regularization, in which case the incorrect prediction will have more proportional support than the attested target form. Greedy decoding and beam search appear to have little influence on the rates of proportionality.

6. Discussion

This paper explored the performance of different models of morphological inflection, with an eye to assessing the extent to which models exhibit proportional behavior. We find that errors produced by these models are unlikely to have support from proportional bases in the training data, with under 35% of errors found across all model settings. Models making use of linear mappings between semantic and phonological cues are found to generate a higher degree of proportional errors.

Our results have implications for research that aims to simulate historical changes using computational models of morphological inflection (e.g., Cotterell, Kirov, Hulden, & Eisner, 2018). If analogical changes that restructure morphological paradigms are in fact overwhelmingly proportional, then care is warranted when choosing models for this particular task. Even for models from the framework of linear discriminative learning, which in a sense incorporates proportionality by learning linear mappings between phonological sequences and semantic variables, the degree of proportional errors produced depends on a range of factors and displays variability across languages. Future work will benefit also from exploring the degree to which models of this sort generate proportions traditionally thought to be invalid, such as *four*:*fork*::*three*:*threek*, or *ear*::*hear*:*eye*::*heye* (Kiparsky, 1968; Deutscher, 2002), in order to probe the extent to which such models can be used to reliably recapitulate processes of diachronic change.

References

Anttila, R. (1977). *Analogy.* The Hague: Mouton.

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230–268.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). 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*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc.

Chollet, F., et al.. (2015). *Keras.* `https://keras.io`.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2018). On the diachronic stability of irregularity in inflectional morphology. *arXiv preprint arXiv:1804.08262*.

Deutscher, G. (2002). On the misuse of the notion of 'abduction' in linguistics. *Journal of Linguistics*, *38*(3), 469–485.

Fertig, D. (2016). Mechanisms of paradigm leveling and the role of universal preferences in morphophonological change. *Diachronica*, *33*(4), 423–460.

Gaeta, L. (2007). Is analogy economic? In G. B. Fabio Montermini & N. Hathout (Eds.), *Selected Proceedings of the 5th Décembrettes: Morphology in Toulouse.* Somerville, MA: Cascadilla Press.

Garrett, A. (2008). Paradigmatic uniformity and markedness. In J. Good (Ed.), *Explaining linguistic universals: Historical convergence and universal grammar* (pp. 124–143). Oxford: Oxford University Press.

Haspelmath, M. (1994). The growth of affixes in morphological reanalysis. In *Yearbook of Morphology 1994* (pp. 1–29). Springer.

Hill, E. (2007). Proportionale Analogie, paradigmatischer Ausgleich und Formerweiterung: ein Beitrag zur Typologie des morphologischen Wandels. *Diachronica*, *24*(1), 81–118. (Publisher: John Benjamins)

Hock, H. H. (2009). *Principles of historical linguistics.* Walter de Gruyter.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *The 3rd International Conference for Learning Representations.* San Diego.

Kiparsky, P. (1968). Linguistic universals and linguistic change. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 170–202). Holt, Rinehart, and Winston.

Lepage, Y. (1998). Solving analogies on words: an algorithm. In *The 17th International Conference on Computational Linguistics (COLING)* (Vol. 1).

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., & Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceed-*

ings of the Twelfth Language Resources and Evaluation Conference (pp. 3922–3931). Marseille, France: European Language Resources Association.

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Neuvel, S., & Fulop, S. A. (2002). Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (Vol. 6, pp. 31–40).

Paul, H. (1880). *Prinzipien der Sprachgeschichte.* Halle: Niemeyer.

Rudnicky, A. (2015). *The Carnegie Mellon Pronouncing Dictionary, Version 0.7b.*

Sims-Williams, H. (2016). *Analogy in morphological change.* Unpublished doctoral dissertation, University of Oxford.

Sims-Williams, H. (2021). Token frequency as a determinant of morphological change. *Journal of Linguistics*, 1–37.