Using Correspondence Patterns to Identify Irregular Words in Cognate Sets Through Leave-One-Out Validation

MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY

Frederic Blum[1,2] Johann-Mattis List[1,2]

[1]Max Planck Institute for Evolutionary Anthropology [2]University of Passau

Abstract

Regular sound correspondences constitute the principal evidence in historical language comparison. Despite the heuristic focus on regularity, it is often more an intuitive judgement than a quantified evaluation, and irregularity is more common than expected from the Neogrammarian model. Given the recent progress of computational methods in historical linguistics and the increased availability of standardized lexical data, we are now able to improve our workflows and provide such a quantitative evaluation. Here, we present the balanced average recurrence of correspondence patterns as a new measure of regularity. We also present a new computational method that uses this measure to identify cognate sets that lack regularity with respect to their correspondence patterns. We validate the method through two experiments, using simulated and real data. In the experiments, we employ leave-one-out validation to measure the regularity of cognate sets in which one word form has been replaced by an irregular one, checking how well our method identifies the forms causing the irregularity. Our method achieves an overall accuracy of 85% with the datasets based on real data. We also show the benefits of working with subsamples of large datasets and how increasing irregularity in the data influences our results. Reflecting on the broader potential of our new regularity measure and the irregular cognate identification method based on it, we conclude that they could play an important role in improving the quality of existing and future datasets in computer-assisted language comparison.

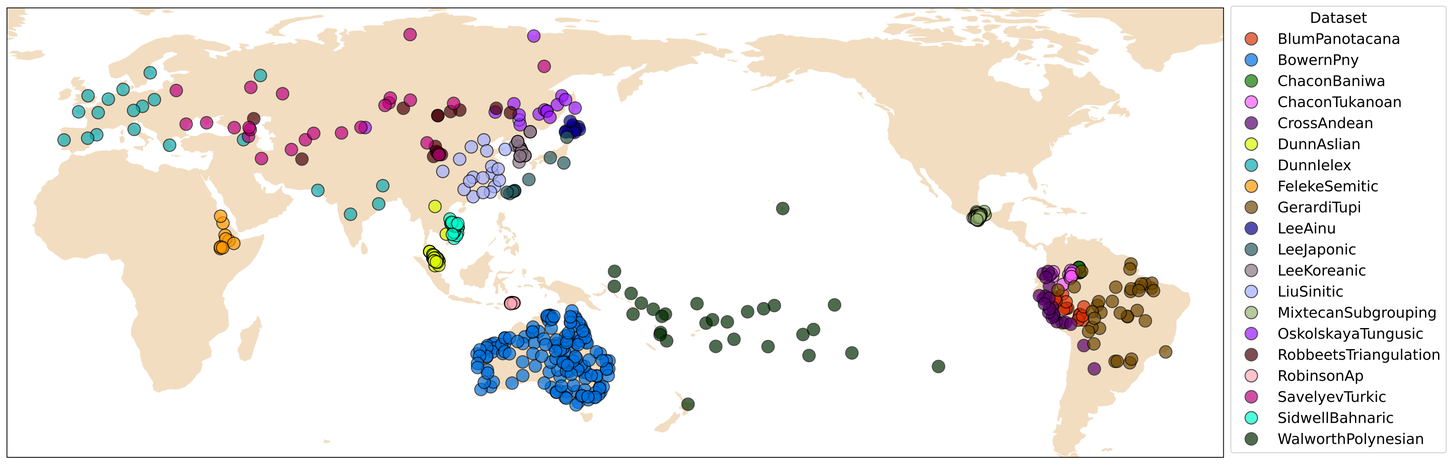

Figure 1: Map of all languages in the sample coloured per dataset.

Figure 2: Data and Code

Background

Regularity of Correspondence Patterns in Historical Linguistics

- Key element of the comparative method [1, 2]
- Alignment site: column of cognate set
- Correspondence pattern: Clustering of compatible alignment sites [3]

	Concept A					Concept B				Concept C				
	I	V	II	V	III	I	V	II	V	III	V	I	V	II
Language 1	k	a	n	o	x	k	e	n	a	x	o	k	e	n
Language 2	k	a	n	o	k	k	i	n	a	k	o	k	i	n
Language 3	k	a	n	o	k	k	i	n	a	k	o	k	i	n
Language 4	k	a	n	o	k	ø	ø	ø	ø	k	o	k	i	n

Figure 3: Artificial example for three cognate sets across four languages with their aligned segments. The columns are numbered and coloured according to their correspondence patterns.

Measuring Regularity in Comparative Wordlists

A Comparable Measure of Regularity

1. Count the recurrence of alignment sites in the data
2. Normalize the count by the total number of sites
3. Log-transform the normalized count
4. Take the exponential of the mean of the log-transformed value

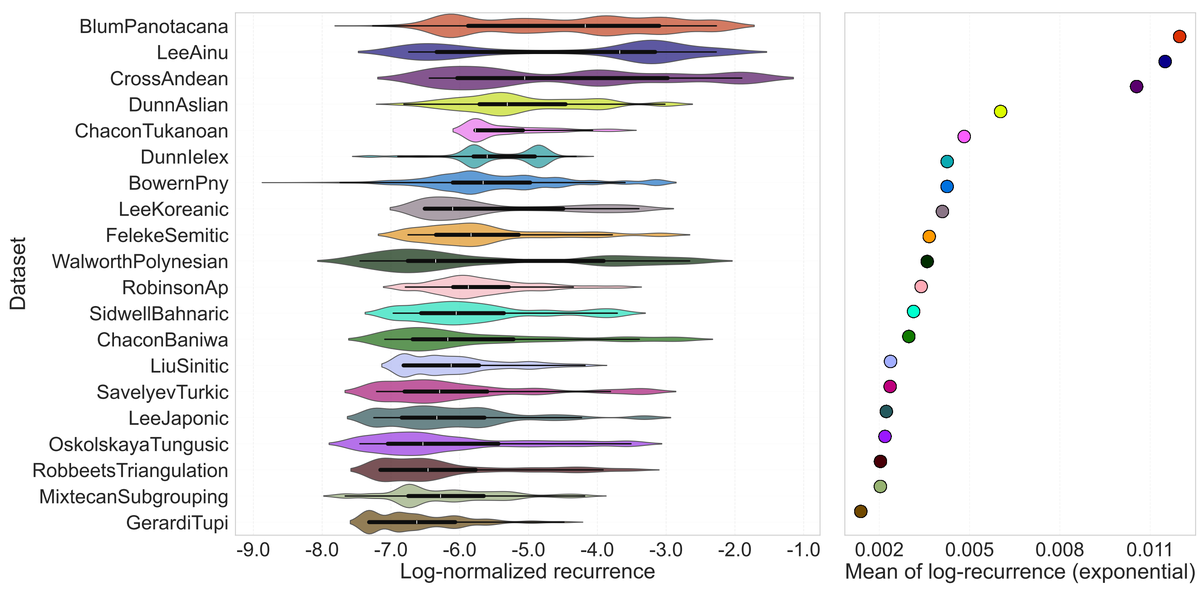

Figure 4: Overall regularity of the 20 datasets in the sample (y-axis) as indicated by measuring the average recurrence of sites in two ways. The left subplot shows the normalized and log-transformed recurrence of each site (x-axis). The right subplot shows the exponential of the mean of that log-transformed recurrence. This score can be interpreted as the balanced average pattern recurrence of an alignment site within each dataset.

Experimental Setup

Leave-one-out experiment for detection of irregular cognate sets

- Randomly replace a word form in 20% of the cognate sets
- Iterating through sites of a cognate set to detect the irregular form

Results

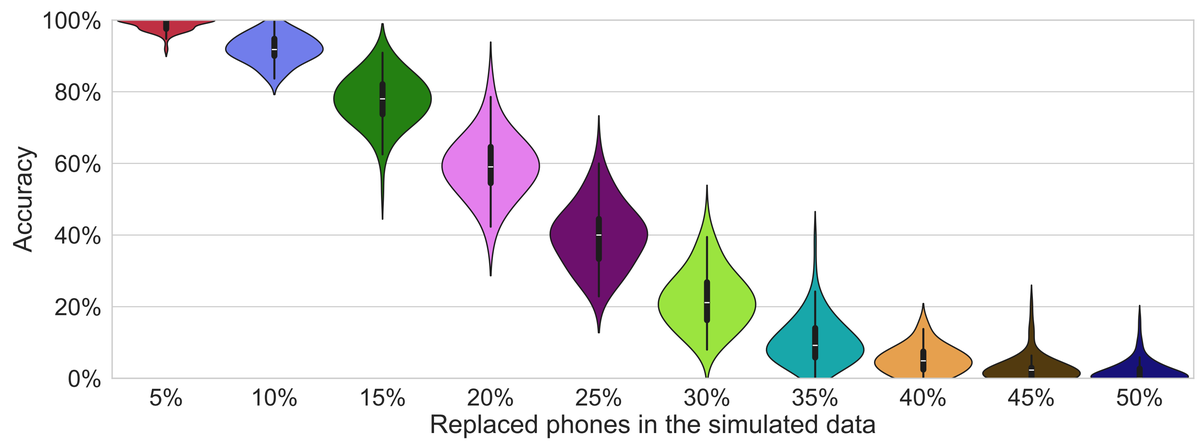

Figure 5: Accuracy for the leave-one-out validation with simulated regular data (y-axis) and random replacement of phones (x-axis) to simulate different levels of irregularity in comparative wordlists.

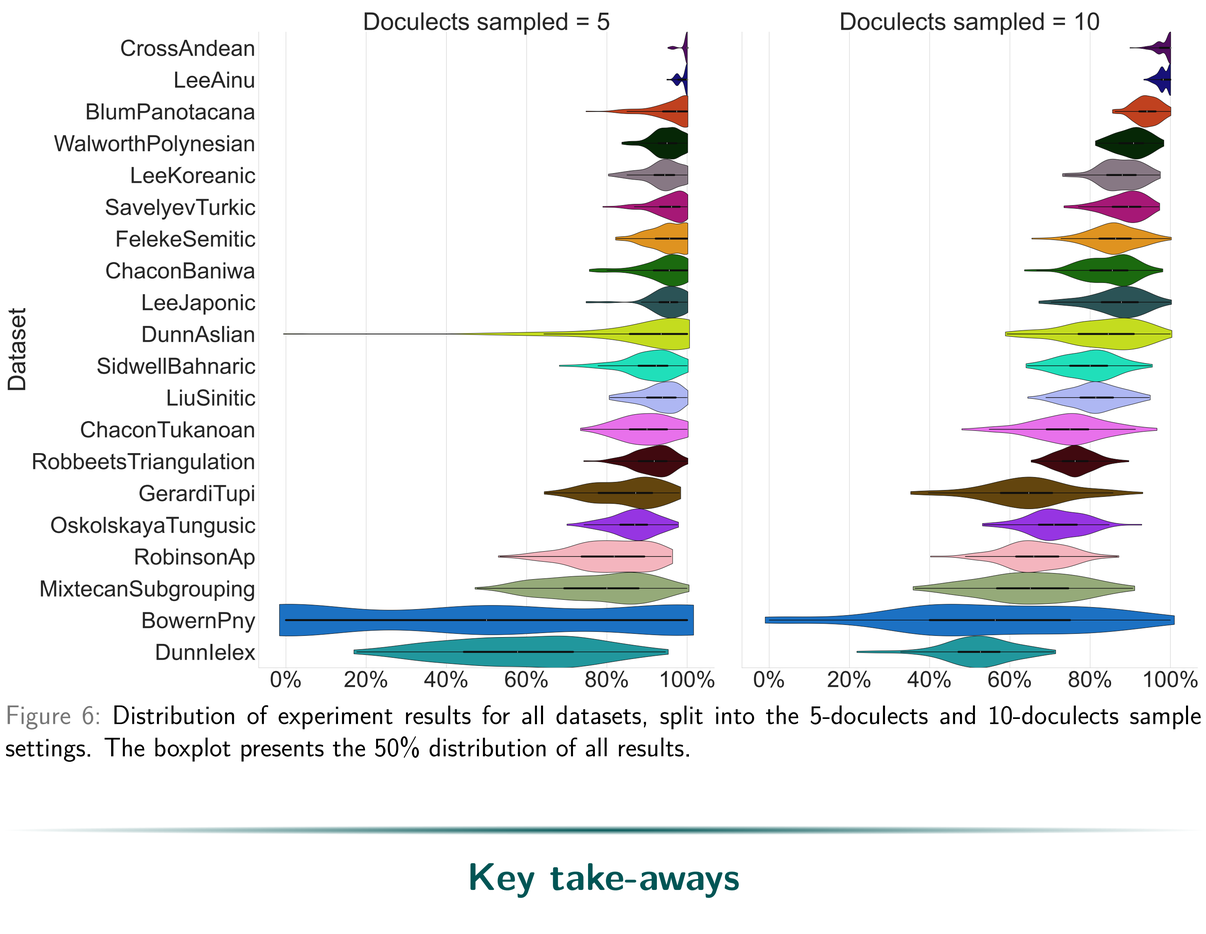

Figure 6: Distribution of experiment results for all datasets, split into the 5-doculects and 10-doculects sample settings. The boxplot presents the 50% distribution of all results.

Key take-aways

1. Regularity is comparable across datasets
2. Regularity can be used to identify irregular forms in cognate sets

References

1. Leskien, A. Die Declination im Slavisch-Litauischen und Germanischen. (S. Hirzel, Leipzig, 1876).
2. Osthoff, H. & Brugmann, K. Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen. (Hirzel, Leipzig, 1878).
3. List, J.-M. Automatic inference of sound correspondence patterns across multiple languages. Computational Linguistics 45, 137–161 (2019).

6th International Workshop on Computational Approaches to Language Change, 28th of March, 2026, Rabat