

# SEMANTIC SIMILARITY SHAPES HOW AGREEMENT MARKERS SPLIT OVER POSITIONS IN VERB MORPHOLOGY

Borja Herce<sup>1,2 † \*</sup>, Carmen Saldana<sup>1,2 †</sup>, and Balthasar Bickel<sup>1,2</sup>

\*Corresponding Author: borja.hercecallega@uzh.ch

†These authors have contributed equally to this work

<sup>1</sup>Department of Comparative Language Science, University of Zurich, Switzerland

<sup>2</sup>Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Switzerland

Agreement markers that refer to the same feature or argument tend to be found in the same position within inflectional forms (e.g., all subject agreement in suffixes, all object agreement in prefixes; Mansfield, Stoll, & Bickel, 2020); however, little is known about the exceptions to this trend. In this study, we explore the positional properties of subject and object person-number agreement markers in a phylogenetically diverse sample of 227 languages from 97 different stocks—based on AUTOTYP (Bickel et al., 2017) plus additional data collected from the WALS 100-language sample (Dryer & Haspelmath, 2013). The result is 325 person-number paradigms, whose agreement morpheme order was surveyed to explore trends and split patterns. We find that a majority of agreement paradigms only require reference to a single position, thus obeying the principle of CATEGORY CLUSTERING (Mansfield et al., 2020). A sizeable minority (128 paradigms, about 39%), however, show what we call POSITIONAL SPLITS, whereby different person-number bundles are marked in different positions (e.g., prefix or suffix) within the paradigm (as illustrated in Table 1). We ask whether positional splits deviate from category clustering in systematic ways and whether their recurrence is proportional to the amount of shared feature values.

Table 1. Different patterns of syncretism in person-number verbal paradigms.

	NATURAL PATTERN GUMER, <i>open</i> IPFV		L-TYPE PATTERN KOASATI, <i>hear</i> ACT		X-TYPE PATTERN BASQUE, <i>walk</i> PRS	
	SG	PL	SG	PL	SG	PL
1	ə-kəft	ni-kəft-inə	há:lo-l	il-há:l	na-bil	ga-bil-tza
2	ti-kəft	ti-kəft-o	is-há:l	has-há:l	za-bil-tza	za-bil-tza-te
3	yi-kəft	ti-kəft-o	há:l	há:l	da-bil	da-bil-tza

Our survey suggests three patterns: natural, L-type and X-type patterns, which are illustrated with orange, blue and green cells in the examples from Gumer (Ethiopia), Koasati (US) and Basque (Spain) in Table 1. In natural patterns, all forms with agreement affixes in the same position(s) share at least one feature value throughout, e.g., prefixal forms share SG and circumfixal forms share PL in

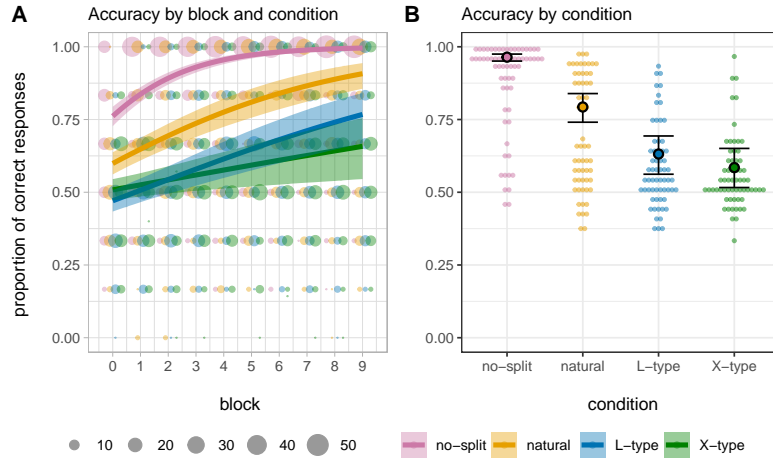


Figure 1. Raw accuracy scores (colour faded) and Bayesian mixed model mean estimates (solid). Error bars and shaded areas show the 90% credible intervals of the mean estimates.

Gumer. In L-type patterns, forms with the same positional properties share values only partially, e.g., two of the prefixal forms in Koasati share PL but differ in person; two share 2, but differ in number. X-type patterns share even less feature values, e.g., circumfixal forms in Basque contain one form that shares no feature value with the other forms. We fitted Bayesian mixed-effects models comparing the occurrences of each type (natural, L-type and X-type) to the occurrences we would expect by chance from all logically possible configurations in person-number  $3 \times 2$  paradigms (with language and family as random slopes). We find that the natural patterns are over-represented in our data, the most unnatural X-patterns are underrepresented, and intermediate-naturalness L-patterns occur with a similar frequency as expected by chance.

These observations suggest a bias towards natural over L over X patterns when languages evolve over time and space. We hypothesize that this bias is grounded in a preference for semantic similarity during the transmission and learning of word forms. To test this hypothesis we conducted an online artificial language learning experiment ( $N=247$ ) where we trained and tested participants on a person-number verbal agreement paradigm with positional splits according to natural vs. L-type vs. X-type patterns (Saldana, Herce, & Bickel, 2022, January 21). We ran a further control condition with no splits (i.e., following CATEGORY CLUSTERING). Results are consistent with the hypothesized learnability gradient *no-split* > *natural* > *L-type* > *X-type* (see Figure 1), thus matching the observed cross-linguistic tendencies. Our findings support the notion that semantic similarity shapes the evolution and transmission of morphological structure (Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Mansfield et al., 2020) and that it does so in a gradient way.

## References

- Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B. (2017). *The AUTOTYP typological databases. Version 0.1. 0*. Retrieved from <https://github.com/autotyp/autotyp-data/tree/0.1.0>.
- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8), 2149–2169. <https://doi.org/10.1111/cogs.12453>.
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from <https://wals.info/>.
- Mansfield, J., Stoll, S., & Bickel, B. (2020). Category clustering: A probabilistic bias in the morphology of verbal agreement marking. *Language*, 96(2), 255–293. <https://doi.org/10.1111/cogs.13107>.
- Saldana, C., Hecce, B., & Bickel, B. (2022, January 21). *Learnability of morphological patterns of syntagmatic splits*. [Preregistration] <https://doi.org/10.17605/OSF.IO/YZCXP>.