Annotating and Inferring Compositional Structures in Numeral Systems Across Languages

Arne Rubehn[1], Christoph Rzymski[2], Luca Ciucci[1], Katja Bocklage[1], Alžběta Kučerová[1], David Snee[1], Abishek Stephen[3], Kellen Parker van Dam[1] & Johann-Mattis List[1]

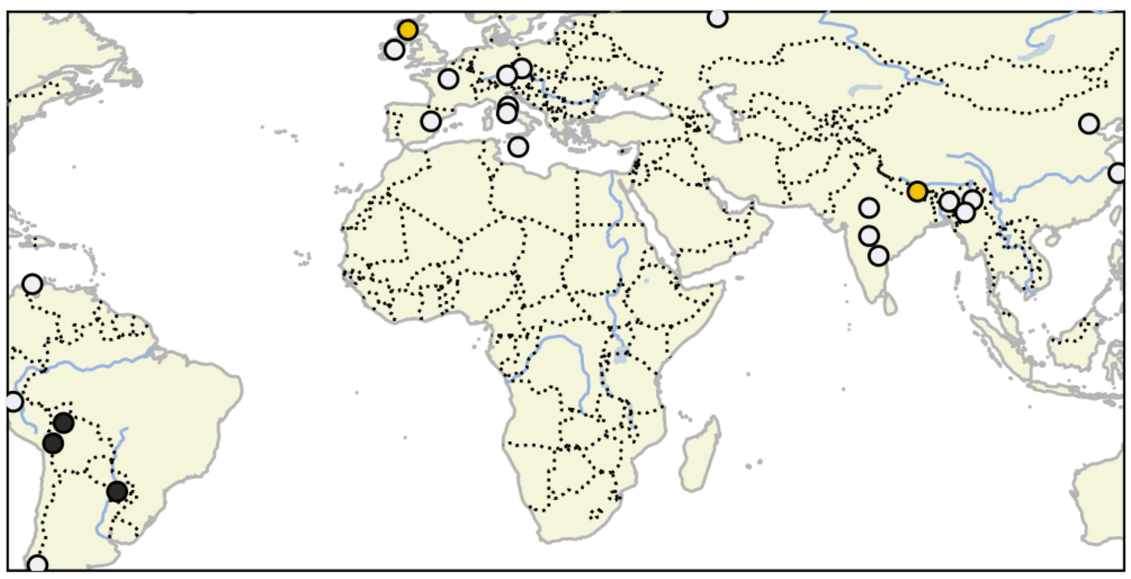
[1] Chair for Multilingual Computational Linguistics, University of Passau, Germany
[2] Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3] Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic

Introduction

We present an annotated sample of 25 typologically diverse numeral system, introducing a simple yet effective annotation scheme, reporting a thorough analysis and experimenting with unsupervised models for morpheme segmentation.

Motivation

- Numeral systems are a showcase example of how linguistic material is recycled to create new forms.
- To enable cross-linguistic comparisons, we need consistent annotation standards that can capture morphological processes.
- With their high degree of compositionality, numeral systems are an ideal candidate for testing how models for morpheme segmentations perform in extremely low-resource scenarios.



Data

Language Sample

- 25 typologically diverse languages from 10 different families of Eurasia and South America.
- Most languages employ a decimal system (*white*), some feature a quinary (*black*) or vigesimal (*orange*) system.
- With their high degree of compositionality, numeral systems are an ideal candidate for testing how models for morpheme segmentations perform in extremely low-resource scenarios.

Annotation

- Morpheme-level annotation for numerals from 1 to 40 using EDICTOR (List et al., 2025).
- Explicit coding for morpheme identity (“cognacy”) using numerical ID’s and morpheme glosses.
- Handling of allophony and allomorphy using inline alignments, mapping different surface forms to one underlying form.

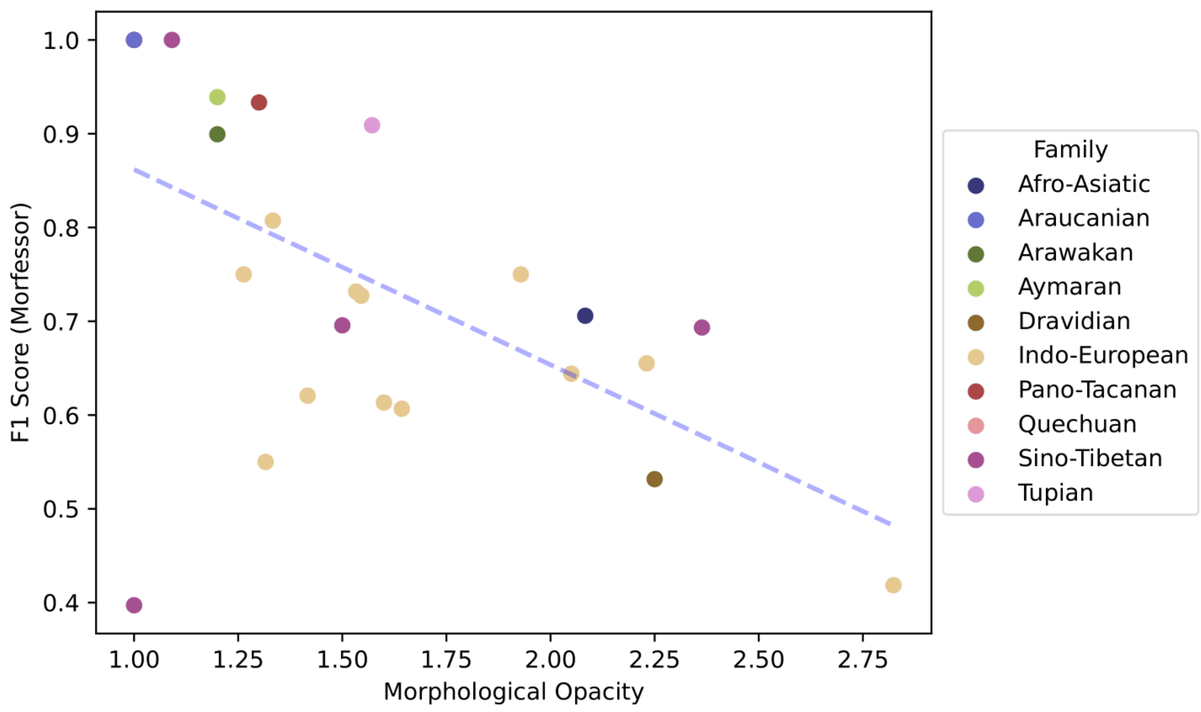
CONCEPT	TOKENS	MORPHEMES	COGIDS
five	kʷ iː ŋ kʷ ɛ	five	3[4]
fifteen	kʷ iː n/ŋ -/kʷ -/ɛ + d ɛ k ɪ̃/ɛ -/m	five + ten	3[4] 8[8]
four	kʷ a tː ʊ ɔ r	four	9[7]
forty	kʷ a dˌtː -ʊ -ɔ r + aː + g ɪ n t aː	four + a-epenthesis + tens_suff	9[7] 14[3] 5[13]
fourteen	kʷ a tː ʊ ɔ r + d ɛ k ɪ̃/ɛ -m	four + ten	9[7] 8[8]

Analysis

We report some simple, quantitative metrics to better understand the different languages’ numeral systems.

- **Number of morphemes:** The number of distinct morphemes employed in a language’s numeral system, both on the surface and the underlying level.
- **Expressivity:** In how many word forms, on average, is the same morpheme used?
- **Opacity:** The ratio between the number of surface morphemes and the number of underlying morphemes.
- **Coding length:** How many morphemes, on average, does a language use to form their numerals?

	Average		Highest		Lowest	
	S	U	S	U	S	U
Morphemes	21.8	13.5	48	20	10	7
Expressivity	5.6	7.9	10.6	15	1.4	3.4
Opacity	1.60		3.18		1	
Code Length	2.53		3.83		1.68	



Experiments

Models for Morpheme Segmentation

- Task: Predict morpheme boundaries in an unsupervised manner.
- We tested Morfessor (Creutz and Lagus, 2002), different variants of Letter Successor Variety/Entropy (Harris, 1955; Hafer and Weiss, 1974), and a simple affix substring matching algorithm (cf. List, 2023).
- Morfessor with the best overall performance, with an average F_1 score of 0.74 on surface forms and 0.88 on underlying forms.
- Performance of models strongly correlates with the opacity of the numeral system.

Subword Tokenization Algorithms

- We tested if popular algorithms for subword tokenization can pick up a genuine morphological signal.
- Poor performance, no generalizable solution for determining a stopping condition.

References: Creutz, M., & Lagus, K. (2002). Unsupervised Discovery of Morphemes. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, 21–30. | Hafer, M. A., & Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10(11–12), 371–385. | Harris, Z. S. (1955). From Phoneme to Morpheme. *Language*, 31(2), 190. | List, J.-M. (2023). Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540), 1–10 | List, J.-M., van Dam, K. P., & Blum, F. (2025). *EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]*. **Acknowledgements:** This study was supported by the ERC Consolidator Grant ProduSemy (Grant No. 101044282).

arne.rubehn.com
arne.rubehn@uni-passau.de