On segmented data and partial cognacy: automated cognate detection and input form

Promise Dodzi KPOGLU
Llacan, CNRS, 7 Rue Guy Môquet- Bp 8, 94801, Villejuif, France
promisedodzi@gmail.com

KEYWORDS: partial cognacy, segmentation, cognacy goodness, Dogon languages.

It is generally accepted, but often overlooked, that the quality of input data plays a crucial role in influencing the outcomes of automatic cognate detection. Despite this, there are few in-depth studies that explore how varying levels of data quality impact results. Consequently, apart from some sporadic mentions (cf. List, 2017), the specific details of how input influences the process of cognate detection remain underexplored. This paper seeks to address this gap by focusing on the critical question: to what extent does morphological segmentation influence the results obtained from automated cognate detection?

The data used in this study are derived from the Dogon and Bangime Linguistics project, a language description project focusing on the Dogon family (Moran et al. 2016). The Dogon language family comprises about 25 languages spoken in central Mali. The data is curated and presented in CLDF (Cross-Linguistic Data Format), making it publicly available for further research (Forket et al. 2018). The dataset contains wordlists for core vocabulary items (at least 500) from each of the languages. The first step in the analysis involves formatting the raw data, and also identifying consensus forms of verbs - verbs in Dogon have ATR harmony. After this, the dataset is divided into four distinct sets using hand-coded rules in Python scripts: the raw data (without tones [but cleaned]), phonetically processed data (where phonetic representations are standardized), morphologically processed data (where individual morphemes are identified [by referencing grammatical descriptions of individual languages such as Heath 2018;2015]), and morpho-phonotactically processed data (which involves a combination of phonetic, morphological and phonotactic segmentation). Each of these datasets is then subjected to a detailed automated cognate detection process.

For automated cognate detection, the LexStat algorithm was employed (List, 2012b). LexStat is a powerful algorithm available in LingPy, a python library for historical comparative linguistics work (List et al., 2024). This algorithm is designed to detect full and partial cognates based on phonetic similarity, after classifying sequences into sound classes. In this study, a choice is made to undertake partial cognate detection, with a similarity threshold of 0.55, and the InfoMap clustering method. The LexStat algorithm is then applied

to each of the four datasets. Once partial cognates are identified, the results are evaluated using two key concepts: cognacy scores and cognacy goodness. Cognacy scores involve the quantitative measurement of cognacy results of each dataset. On the other hand, cognacy goodness is concerned with whether the cognates detected by the algorithm on each dataset align with the expectations and judgements of field linguists who have direct experience with Dogon languages.

To asses cognacy goodness, an aggregated distance matrix is constructed for each pair of languages in each dataset. Using the distance matrix as input, average-linkage clustering is applied. The resulting cluster trees, or dendrograms, represent a phylogenetic grouping of the languages, providing a visual representation of how closely related the languages are according to the detected cognates. This phylogenetic grouping, derived from the automated cognate detection, is then compared to an existing phylogenetic classification [which was produced by field linguists employing traditional comparative methods – Moran & Prokić (2013:12)] of the Dogon languages. The comparison is made using several evaluation metrics: Adjusted Rand Score, Normalized Mutual Information Score, Fowlkes-Mallows Score, Homogeneity Score, Completeness Score, and V-Scores.

The results of this comparison reveal several important findings. Firstly, it is observed that, as the level of linguistic parsing increases, cognacy statistics improves accordingly. Specifically, for every additional unit of linguistic parsing – from raw data to phonetic, morphological, and finally morpho-phonotactic processing – there is an average 16.5% increase in the primary cognacy statistics. This suggests that more detailed linguistic segmentation leads to better statistical performance in detecting cognates. However, when it comes to cognacy goodness, the results are more nuanced. While phonetic and morphological parsing generally improve cognacy goodness, morpho-phonotactic parsing introduce noise into the data, leading to a decline in cognacy goodness.

To further investigate this effect, a linear regression model in which cognacy goodness is treated as the dependent variable, and the level of linguistic parsing the independent variable, is constructed. The results of this model confirm the detrimental impact of morpho-phonotactic parsing, with negative coefficients observed when this level of processing is included. By contrast, when morpho-phonotactic parsing is excluded, the coefficients are positive, indicating that phonetic and morphological parsing have a generally positive impact on cognacy goodness. This confirms the hypothesis that while increased linguistic segmentation improves statistical performance, there is a threshold beyond which further parsing introduces too much noise, reducing the overall quality of the results.

The findings of this study suggest that while increasing the level of data segmentation can improve the detection of partial cognates in an automatic cognate detection framework, it does not necessarily lead to better overall results. There is a balance between increasing data complexity through segmentation and maintaining the quality of the results. It can be stated

then that while efforts to enhance cognate detection are important, it is crucial to recognize thresholds where further parsing could negatively affect overall performance, preventing unintended outcomes.

Bibliography

BUIS M. L. (2010). Stata Tip 87: Interpretation of Interactions in Nonlinear Models. *The Stata Journal Promoting Communications On Statistics And Stata*, *10*(2), 305-308. https://doi.org/10.1177/1536867x1001000211.

COXE S., WEST S. G., & AIKEN L. S. (2009). The Analysis of Count Data : A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal Of Personality Assessment*, *91*(2), 121-136. https://doi.org/10.1080/00223890802634175

FORKEL, R., LIST, J., GREENHILL, S. J., RZYMSKI, C., BANK, S., CYSOUW, M., HAMMARSRÖM, H., HASPELMATH, M., KAIPING, G. A., & GRAY, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(1), 1-10. https://doi.org/10.1038/sdata.2018.205

HEATH J. (2015). *A grammar of Togo Kan (Dogon language family, Mali)*. Univ. of Michigan. https://doi.org/10.17617/2.2176494.

Heath, J. (2008). *A Grammar of Jamsay*. Walter de Gruyter.

KONDRAK G. (2000). A new algorithm for the alignment of phonetic sequences. *North American Chapter of the Association for Computational Linguistics*, 288–295. https://www.aclweb.org/anthology/A00-2038.pdf.

KONDRAK G. (2001). Identifying cognates by phonetic and semantic similarity. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.3115/1073336.1073350.

List, J.-M (2012a). Multiple sequence alignment in historical linguistics. A sound class-based approach. *Proceedings of ConSOLE*, *1*, 241–260. https://hcommons.org/deposits/item/hc:30395/.

LIST J.-M. (2012b). LexStat: Automatic Detection of Cognates in Multilingual Wordlists. *Conference of the European Chapter of the Association for Computational Linguistics*, 117–125. https://www.aclweb.org/anthology/W12-0216

LIST J.-M. (2017). Historical language comparison with LingPy and EDICTOR. Jena: Max Planck Institute for the Science of Human History. DOI: 10.5281/zenodo.1042204.

LIST J.-M., LOPEZ P. & BAPTESTE E. (2016). Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 599–605. https://doi.org/10.18653/v1/p16-2097.

LIST J-M., & FORKEL R. (2024). *LingPy. A Python library for historical linguistics*. Version 2.6.13. URL: https://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

MIKHEEV A. (2022). Text segmentation. In R. MITKOV, Ed., *The Oxford handbook of computational linguistics*, chapter 23, p. 549–564. Oxford University Press.

MORAN S., FORKEL R., & HEATH J. (2016). *Dogon and Bangime linguistics*. Max Planck Institute for the Science of Human History.

MORAN S., & PROKIĆ J. (2013). Investigating the relatedness of the endangered Dogon languages. *Literary and Linguistic Computing*, 28(4), 676-691.

RAMA T., LIST J., WAHLE J., & JÄGER G. (2018). Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics? *arXiv Preprint*. https://doi.org/10.18653/v1/n18-2063.