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

Promise Dodzi KPOGLU
CNRS-LLACAN
promisedodzi@gmail.com

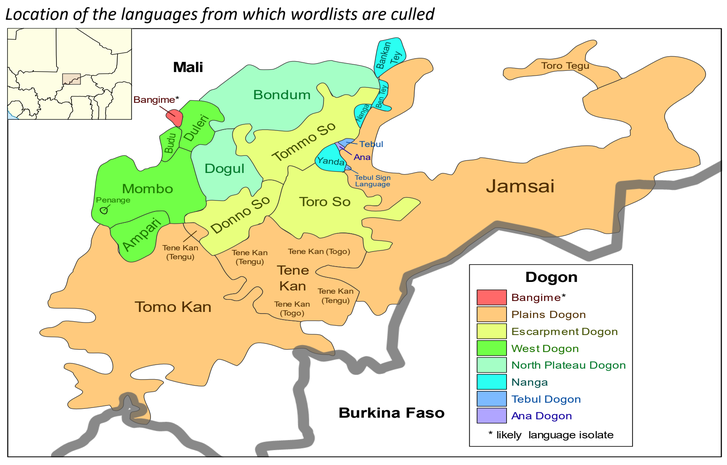
## Introduction

- Cognates are generally understood as words that share a common ancestry.
- Cognates can manifest in various forms depending on the degree of phonetic and semantic similarity (Koch & Hercus, 2013; Meelen & Hill, 2022).
- In spite of the important progress made in automatic cognate detection (cf. Kondrak, 2001; Rama et al., 2018; List, 2012), there is the generally accepted, but often overlooked fact that the quality of input data plays a crucial role in influencing outcomes (cf. List, 2017).
- This paper seeks to address the question: to what extent does morphological segmentation influence the results obtained from automated partial cognate detection?

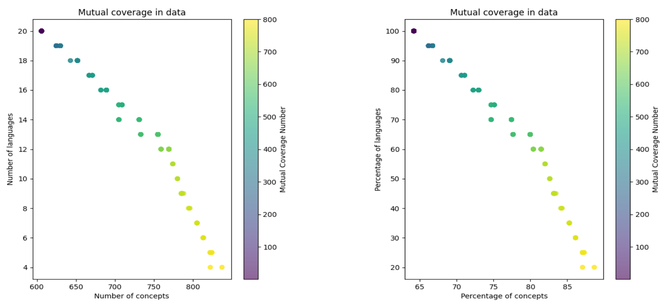
## Materials

- The Dogon language family comprises about 25 languages spoken in central Mali.
- Data – wordlists of core vocabulary, are derived from the Dogon and Bangime Linguistics project (Moran et al., 2016).
- Data is curated and presented in CLDF [Cross-Linguistic Data Format](Forkel et al., 2018).
- Data is publicly available on GitHub as *heathdogon* .

*Location of the languages from which wordlists are culled*




- Data consists of ~600 concepts and 20 languages:
  - Some languages of original dataset do not have enough lexical items
  - Some concepts do not have enough coverage.




- Automatic partial cognate detection on each dataset using LexStat (List, 2012b) - similarity threshold of 0.55, and InfoMap clustering method.
- Cognacy evaluation using two key concepts: cognacy scores (quantitative) and cognacy goodness (qualitative).

## Methodology

Methodology involved the following steps:

- Formatting, and identification of consensus forms of verbs - verbs in Dogon have ATR harmony.
- Division of dataset 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 such as Heath [2018; 2015]), and morpho-phonotactically processed data (which involves a combination of phonetic, morphological and phonotactic segmentation).
- Automatic partial cognate detection on each dataset using LexStat (List, 2012b) - similarity threshold of 0.55, and InfoMap clustering method.
- Cognacy evaluation using two key concepts: cognacy scores (quantitative) and cognacy goodness (qualitative).

## Evaluation metrics

- Results are first visualized using the Edictor tool:

*sample result*

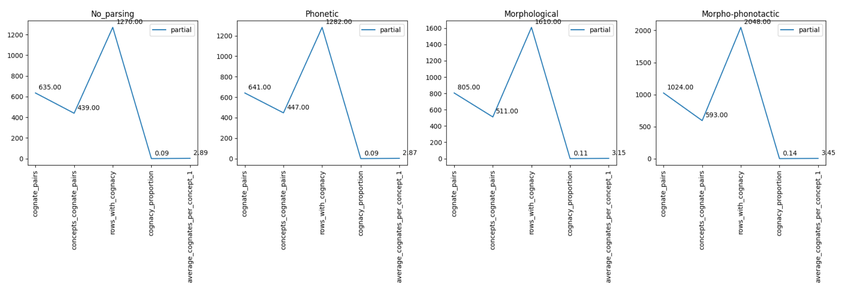
| ID | DOCULECT | CONCEPT | TOKENS | NOTE | COGIDS |
|---|---|---|---|---|---|
| 769 | DogulDomKundialang | (child) be born | n à l + j é | | 347[4] 348[3] |
| 770 | Najamba | (child) be born | n à l + í + j é | | 347[4] 349[2] 348[3] |
| 771 | TommoSoTongoTongo | (child) be born | n à l + í + j é | | 347[4] 349[2] 348[3] |
| 772 | YomoSo | (child) be born | n à r + é é | | 347[4] 350 |

- Cognacy statistics computed:
  - Number of unique cognate pairs.
  - Number of unique concepts involved in cognate pairs
  - Number of rows involved in cognate pairs
  - Proportion of data with cognacy
  - average number of cognate items per concept

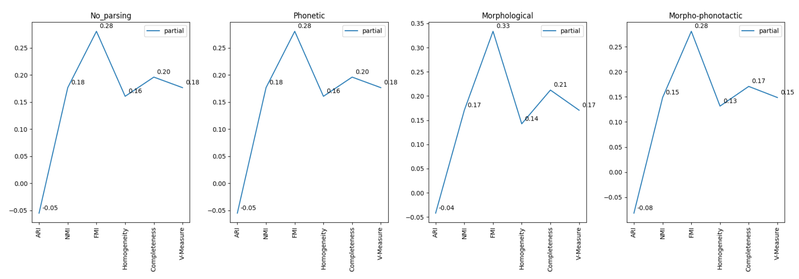
- Cognacy goodness involved cluster comparison with "descriptive-linguist ground truth":
  - Adjusted Rand score
  - Normalized mutual information score
  - Fowlkes-Mallows score
  - Homogeneity score
  - Completeness
  - V-Measure

## Results

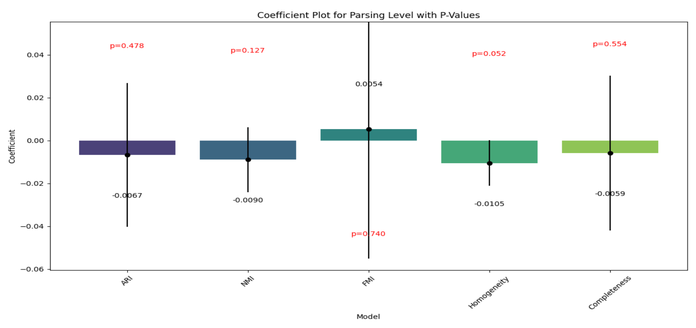
- As the level of linguistic parsing increases, cognacy statistics improves accordingly. For an additional unit of parse there is an average 16.5% increase in the primary cognacy statistics.



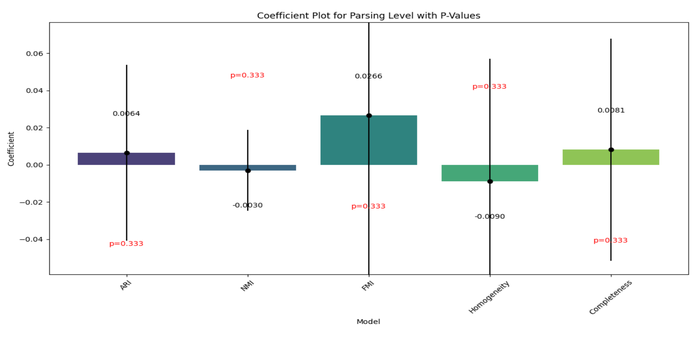

- For cognacy goodness, while phonetic and morphological parsing generally improve it, morpho-phonotactic parsing introduces noise, leading to its decline.




- To further investigate the effect, a linear regression model in which cognacy goodness is the dependent variable, and level of linguistic parsing the independent variable, is constructed.
- The model shows that morpho-phonotactic parsing negatively impacts cognacy goodness, reflected in negative coefficients.




- When morpho-phonotactic parsing is excluded, phonetic and morphological parsing positively affect cognacy goodness, with positive coefficients.

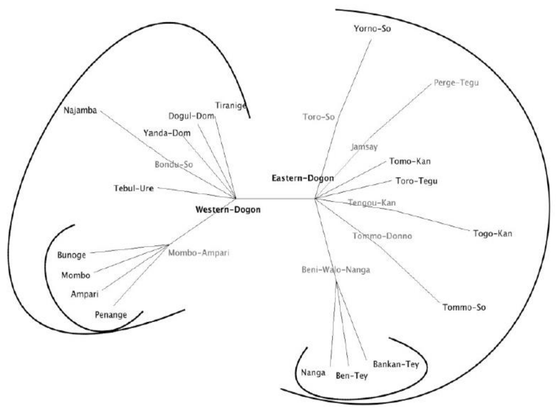



## Conclusions

- This study confirms the hypothesis that while increased linguistic segmentation improves statistical performance in automatic cognate detection, there is a threshold beyond which further parsing introduces noise, reducing the overall quality of the results.
- Relating to partial cognacy, the findings show that while increasing the level of data segmentation can improve the detection of partial cognates, 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.

## Future directions

- Questions of full cognacy and forms involved in "dialexification" remain opened (cf. François & Kalyan, 2023).
- Cognacy goodness is calculated based on "descriptive-linguist ground truth".




Moran & Prokić (2013:12)

- It is important to assess the adequacy of this "ground truth" through simulations.

## References

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

FRANÇOIS A. & KALYAN S. (2023). Dialexification: A tool for studying cross-linguistic patterns of semantic change. In *16th International Cognitive Linguistics Conference*.

HEATH J. (2017). *A grammar of Bunoge (Dogon, Mali).*

HEATH J. (2023). *A grammar of Tebul Ure (Dogon, Mali).*

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.

UMR8135 CNRS INaLCO EPHE
7 rue Guy Môquet 94801 Villejuif Cedex - France

https://llacan.cnrs.fr/
llacan@cnrs.fr

ERC Project BANG
https://bang.huma-num.fr/