A Computational Typological Analysis of Syntactic Structures in European Languages

Diego Alves, Božo Bekavac, Marko Tadić – Faculty of Humanities and Social Sciences – University of Zagreb

Cleopatra Final Event
2022-04-17/18
Hannover, Germany

Introduction

From 2015 onwards, the usage of deep learning techniques has been predominant in the field of dependency parsing (Otter, D. et al., 2019). These approaches require a large amount of annotated data which can be problematic for some languages considered low-resourced. Linguistic manual annotation of texts can be very costly (Fort, K. et al. 2014), therefore, other solutions for improving PoS-MSD and dependency parsing tagging scores have been proposed in the literature. One way to overcome this issue is to combine data from similar languages according to established typological classifications (e.g.: Smith et al., 2018 and Alzetta, C. et al., 2020). However, these studies do not present a deep analysis of typological features which may play a significant role when corpora are combined, and do not consider statistics concerning possible (or impossible) syntactic constructions inside the available data as a possible typological classification.

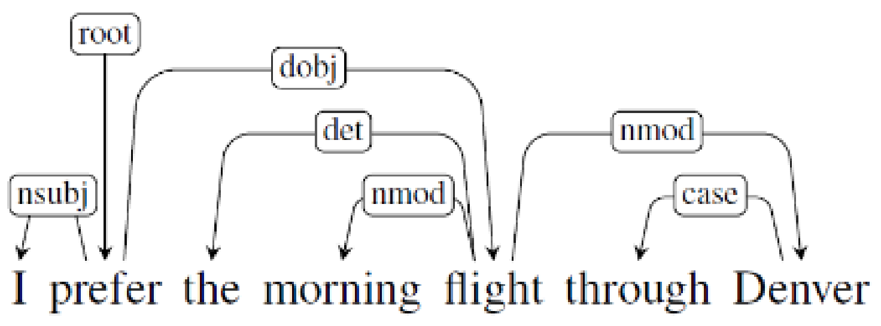

Fig.1. Exemple of dependency analysis (Jurafsky and Martin, 2021)

Aims

- Provide new quantitative methods for the typological classification of EU languages
- Improve the dependency parsing scores of low-resourced EU languages via corpora-combination using these quantitative typological approaches

Languages

24 EU + 10 worldwide languages

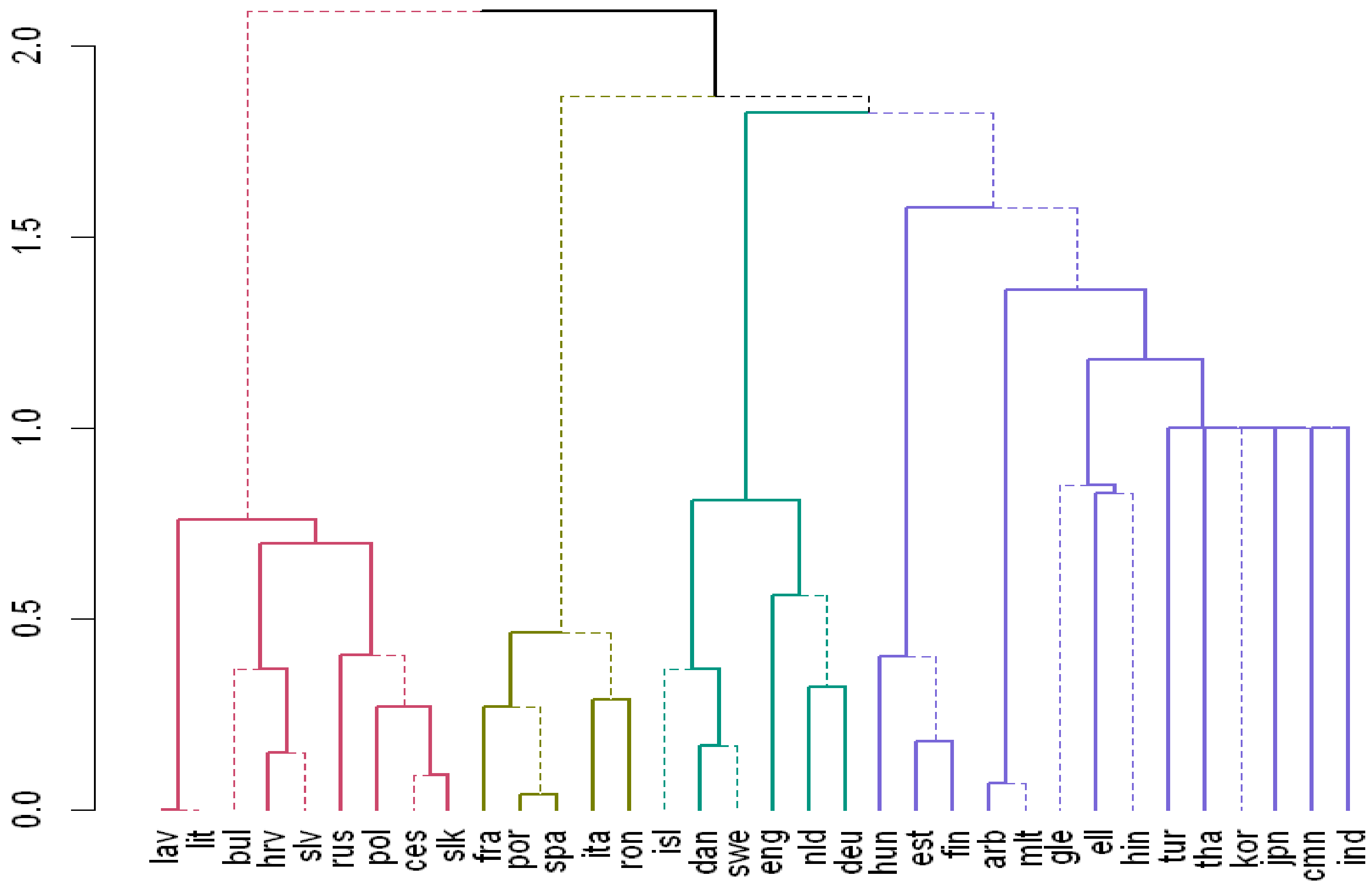

Fig.2. Cosine dendrogram of the lang2vec (Littell et al., 2017) phylogenetic comparison of the 34 selected languages.

Low-resourced EU languages in terms of Universal Dependencies corpus-size and parsing results: Hungarian, Irish, Lithuanian, and Maltese.

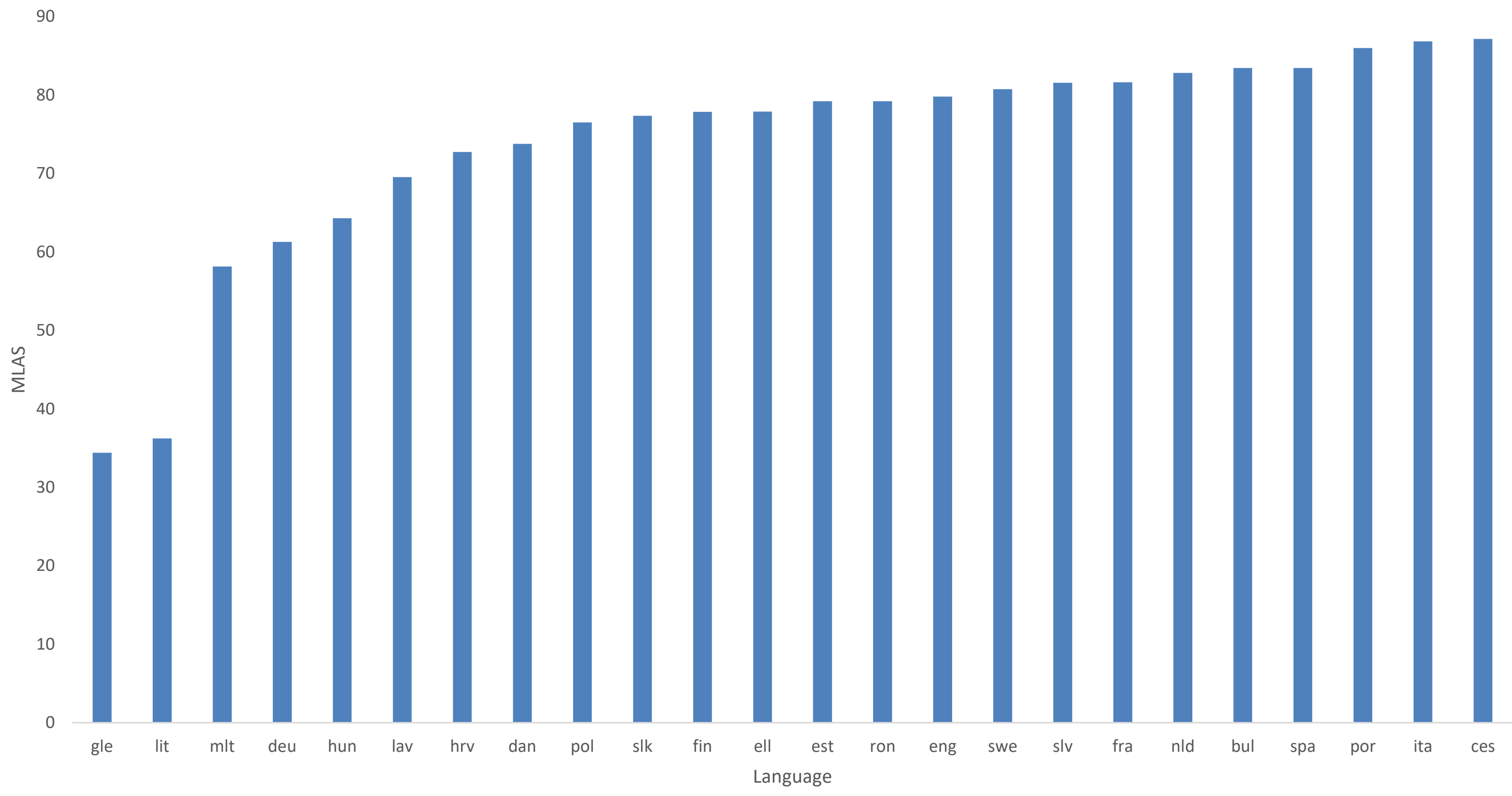

Fig.3. MLAS values for each EU language obtained by Kondratyuk and Straka (2019).

Typological Methods

1) Marsagram

Patterns are identified from context-free grammar extracted from annotated corpora.

a) All patterns → linear, exclude, require, and unicity
b) Only linear

E.g.: *NOUN, precede, DET – det, NOUN - nmod*

VERB, exclude, NOUN – nsubj, PRON - nsubj

2) Head and Dependent relative position

E.g.: *ADV_advmod_precedes_ADJ*

3) Verb and Object relative position

E.g.: *NOUN_obj_follows_VERB*

Steps:

1) Extraction of patterns
2) Comparison of language vectors (Euclidean and cosine)
3) Cluster analysis
4) Correlation with LAS and MLAS results when languages are combined

Results:

Best correlation with dependency parsing synergy results (better than the standard classification obtained with lang2vec syntactic features):

- Marsagram linear patterns (cosine).

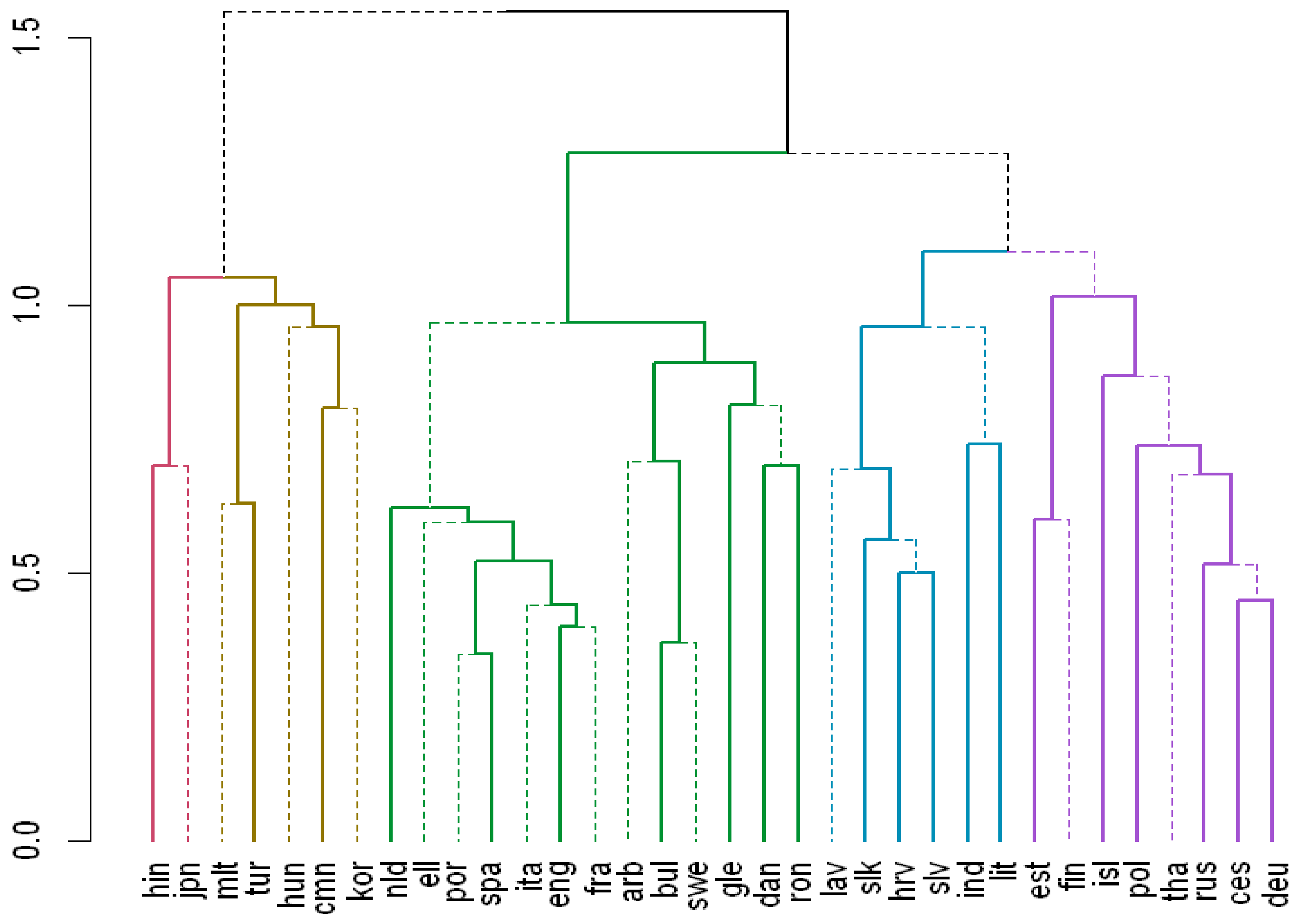

Fig.4. Cosine dendrogram of the lang2vec (Littell et al., 2017) phylogenetic comparison of the 34 selected languages.

Dependency parsing improvement of low-resourced EU languages:

- Hungarian + Dutch → +0.50 (LAS) and +1.27 (MLAS)
- Irish + Portuguese → +0.36 (MLAS)
- Lithuanian + Portuguese → +1.86 (MLAS)
- Maltese + French → +2.51 (LAS) and +4.05 (MLAS)

The best improvement was observed for long sentences (i.e.: more than 50 tokens).

Conclusion

The new proposed typological methods allowed us to classify EU languages with different quantitative syntactic approaches, and, from the comparison with the dependency parsing results, it was possible to identify the best strategy to combine corpora for parsing improvement. The best results were obtained for languages with the smallest training corpora (i.e.: Hungarian and Maltese).

Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997.