Computational Approaches for Romance Related Words Discrimination

Abstract

Natural languages are living eco-systems, they are constantly in contact and, by consequence, they change continuously. Traditionally, the main problems in historical linguistics ("How are languages related?", "How do languages change across space and time?") have been investigated with comparative linguistics instruments. The main idea of the comparative method is to perform a property-based comparison of multiple sister languages in order to infer properties of their common ancestor. It is a time-consuming manual process that required a large amount of intensive work.

The identification of cognates is a fundamental process in historical linguistics, on which any further research is based. On the other hand, discriminating between lexical borrowings and inherited words is considered one of the most difficult and important tasks in HL (Jäger, 2019), for which "the computerised approach" is regarded as the appropriate solution even by classical linguists (Heggarty, 2012). We propose here computer-assisted methods for identifying cognates, for discriminating between cognates and borrowings, and for discriminating between inherited and borrowed Latin words.

Firstly, we introduce a method to automatically determine if a pair of words (*u*,*v*) are cognates or not, and we use it on a large database comprising the main Romance languages (Romanian, Italian, French, Spanish and Portuguese), applying it as well in subsequent tasks. Given an input pair of words, the initial task is to automatically determine if they are cognates or not. We developed a machine learning method for automatically producing the answer based on sequence alignment. To align pairs of words, we employed the Needleman Wunsch global alignment algorithm, which has been successfully used in natural language processing and computational biology. We used words as input sequences and a basic substitution matrix, which gives equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *é* were matched). For the machine learning part, we used an ensemble of methods. We applied our method to multiple data sets, showing that our approach improves on previous results, also having the advantage of requiring less input data, which is essential in historical linguistics, where resources are generally scarce. In the process of discriminating between cognate and borrowing, we tried to answer the following question: given a pair of words, are they cognates, borrowings, or neither? For the automatic discrimination between inherited and borrowed Latin words, the best results were obtained by a system based on SVM using features extracted from the word-etymon pairs. We apply our method on both graphic and phonetic forms of the words.

Keywords

Romance languages, cognates, borrowings, inherited words.

Acknowledgements

This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108, CoToHiLi, within PNCDI III.

References

[1] Ciobanu, Alina Maria and Liviu P. Dinu. 2019. Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, vol. 45, No. 4, pages 667–704.

[2] Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai and Ana Sabina Uban, 2021. Automatic Discrimination between Inherited and Borrowed Latin Words in Romance Languages In: *Proc. EMNLP 2021(Findings)*, Punta Cana, 2021

[3] Alina Maria Ciobanu and Liviu P Dinu, 2014. Automatic detection of cognates using orthographic alignment. In *Proc. ACL 2014*, June 22-27, 2014, Baltimore, MD, USA

[4] Alina Ciobanu and Liviu P Dinu, 2015. Automatic discrimination between cognates and borrowings. In Proc. ACL 2015, July 26-31, 2015, Beijing, China

[5] Gerhard Jäger. 2019. Computational Historical Linguistics. *Theoretical Linguistics* | Volume 45: Issue 3-4, 45: Issue 3–4.

[5] Paul Heggarty. 2012. Beyond Lexicostatistics: How to Get More out of "Word List" Comparisons. In *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*, pages 113–137. Benjamins