Mittwoch
08.03.2017
15:45 – 16:30
B4 1, Foyer

A vector-based phonological search for cognates across dictionaries

Max Ionov
Goethe-Universität Frankfurt
max.ionov@gmail.com

Christian Chiarcos
Goethe-Universität Frankfurt
christian.chiarcos@web.de

We present an ongoing work on searching phonologically similar words in related languages in and across dictionaries. This is a part of a larger project devoted to unravelling both synchronic and diachronic lexical connections in related less-resourced languages.

The task of searching for phonologically similar words has two applications: in one language, applied to a corpus, it can cluster inflectional forms of one lexeme, which is extremely useful for low-resourced languages with high inflection and no POS-taggers available. In several related languages, applied across dictionaries or wordlists, it can detect possible cognates – words with the common etymological origin. The latter is a common method for dialectometry (e.g. Heeringa *et al.* 2006), but it was also applied to the field of historical linguistics (List and Moran 2013).

We present an approach that employs PHOIBLE dataset, the universal phonological inventory (Moran *et al.* 2014) and vector-based phoneme representation and compare it with several well-known approaches, starting from simple yet popular minimum edit distance approach (Holman *et al.* 2011 *inter alia*) to more sophisticated approaches like LexStat (List 2012).

Using linguistic insight, we examine the limitations of automatic approaches and propose directions for overcoming them.

CL Poster

References: • Heeringa, W., Kleiweg, P., Gooskens, C., Nerbonne, J. (2006): Evaluation of String Distance Algorithms for Dialectology. Proceedings of the Workshop on Linguistic Distances, 51–62. • Holman, E.W., Brown C.H., Wichmann S., Müller A., Velupillai V., Hammarström H., Sauppe S., Jung H., Bakker D., Brown P., Belyaev O., Urban M., Mailhammer R., List J.-M., Egorov D. (2011): Automated Dating of the World's Language Families Based on Lexical Similarity. Current Anthropology 52(6), 841–875. • List, J.-M. 2012: LexStat. Automatic detection of cognates in multilingual wordlists. Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources, 117–125. • List, J.-M., Moran S. 2013: An Open Source Toolkit for Quantitative Historical Linguistics. Proceedings of the 51st Annual Meeting of the ACL: System Demonstrations, 13–18. • Moran S., McCloy D., Wright R. 2014: PHOIBLE Online, http://phoible.org/.