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 word lists, 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.

**References:**  