Using the Leipzig Corpora Collection for language comparison

Dirk Goldhahn, Uwe Quasthoff

This talk will focus on fully automatic methods for quantifying similarity between languages. Basis for this comparison are lists of the most frequent words or letter N-grams of the languages in question. Unlike other approaches, which use specially designed word lists [Swadesh, 1952] , foundation of this work are lists extracted from text corpora of the Leipzig Corpora Collection [Goldhahn, 2012]. Similarity of these lists is judged by inspecting the distribution of identical elements. In doing so, different parameters like the weightings of the elements are evaluated. Various algorithms capable of comparing the data are examined, among them are rank correlation, cosine similarity or Dice coefficient. The influence of factors such as subject area or size of feature lists is also considered. To overcome the boundaries induced by different writing scripts transliteration is used to extend the analysis.

In a second approach parallel text is utilized. Language similarity is first computed based only on the inter-language distribution of vocabulary, which allows for the identification of translational equivalents on word level. These translations also serve as an input for further analysis. The word pairs are examined for orthographical and phonological closeness to allow for additional investigations of language similarity.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC* (pp. 759-765).

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world′ s languages: a description of the method and preliminary results. *STUF-Language Typology and Universals*, *61*(4), 285-308.