Vladimir Polyakov

# NEW APPROACHES TO LANGUAGE SIMILARITY MEASURES

**October 13, 2008**

Similarity measures are used in phylogenetic calculations with the aim of establishing genetic relationships between languages. Recently it has been established (Polyakov & Solovyev, Wichmann et al.) that genetic relationships can be reflected when using measures based on typological data. The noise one gets when using WALS, however, and the influence of areal contact on the performance of Jazyki Mira/Languages of the World (JM), appear to have significant impact on the outcome of the calculations. Therefore, an important goal is to create a measure where the influence of areal contact is minimal.

A test set of 48 languages (hereafter "A. A. Kibrik's set") has been proposed by the "Languages of the World" research group at the Institute of Linguistics of the Russian Academy of Sciences in 2005. A method of estimating the quality of a similarity measure has also been proposed by author, based on ranking the languages of each of the main eight families of the set in accordance with the "prototype language" of each family. After the measure is calculated, the languages are sorted eight times based on their closeness to each of the prototype languages. The quality of the measure is estimated as follows:

$$K = (K_1+K_2+K_3+K_4+K_5+K_6+K_7+K_8)/8,$$

where $Ki = Np_i/Ng_i$,

$Np_i$ - the number of related languages which have been placed after the prototype.

$Ng_i$ - the number of related languages in each group.

The tables with measures can be found at www.dblang2008.narod.ru.

While testing the database, a lot of work has been put into choosing the best possible similarity measure and researching the influence of different factors on its quality. Among these factors are: the types of features, their frequencies, their hierarchical positions in the feature set, and also the contribution of various aspects of how the language has been described. It has been found that the best quality values are obtained when using a measure which is simply the number of all conterminous features without any restrictions on their frequency, hierarchical position or the section to which they belong. In this case, full correspondence with the traditional genetic classification is obtained for two grups (Uralic and Turkic) and the quality estimation K is 0.67. For any other combinations of measures, the results were worse. Also, calculating similarities based on only one of the sections of the feature set always gives worse results than using the entire set.

A new set of 39 languages based on overlap of WALS, the Automatic Similarity Judgement Program (ASJP) project, and JM has been offered by Valery Solovyev and specified by Søren Wichmann in 2007. It can be used not

only for estimating the performance of different metrics on JM's data, but also for comparing the genetic trees received from three different sources. Sample trees obtained from the three databases can be ranked as follows: the ASJP tree is the most reliable, the JM tree is second and the WALS tree is the least reliable (the trees can be found at www.dblang2008.narod.ru).

More subtle techinques of estimating the quality of a measure have also been proposed. After calculating the metric the languages are sorted 39 times based on their similarity to each of them. The quality is estimated as follows:

$$K = \sum(K_i)/39,\ i = i \ldots 39$$

$$Ki = Np_i/Ng_i,$$

where:

where $Np_i$ is the number of related languages placed after each language;

and $Ng_i$ is the number of related languages in each group.

Different methods of comparing the resulting trees also exist. In this case the quality of similarity measure is the edit distance.

New heuristics which would allow for improving the quality of similarity measures have been investigated recently. For example, restricting the features based on their frequencies increases the measure's quality up to 0.697 and restricting them based on the sections of the feature set

yields an increase to 0.76. Both results have been obtained on A. A. Kibrik's set.

In the future it is planned to include factors such as stability and genetic markers in the calculation. It appears more preferable to use the second set because it allows for comparison with other resources such as WALS and ASJP.