TITLE: Adding typology to lexicostatistics: a combined approach to language classification

CATEGORY: oral

The paper presents a method for automatic language classification based on the word list of Swadesh (1955), and compares this method with a typologically one based on the data in the World Atlas of Language Structures (WALS). The method differs from the original lexicostatistical approach in two ways. Firstly, the comparison between the phonological representations of pairs of potential cognate word forms is done by a computer program (ASJP; automated similarity judgment program), using a refined version of the well-known Levenshtein algorithm (LDND). This provides a distance matrix between languages, expressed as a LDND value per language pair. Secondly, graphic branching structures illustrating language relatedness are generated from this matrix by means of standard software originally developed for the use of biologists in studying phylogenetic relationships. To date, we have collected and transcribed a set of basic word forms for around 2500 languages of the world. We have established that the 40 diachronically most stable elements from the original 100-word Swadesh list perform better than the full list in terms of matching both Dryer's genera as proposed in Haspelmath et al. (2005), the global classification of the *Ethnologue*, and several more specific expert classifications. The paper will briefly discuss the methodological aspects of this operation. Furthermore, we compared the performance of our lexically based method with a typological one. We selected the around 1100 languages that are both in the ASJP database and in the one used for the WALS. We established that LDND performs considerably better than the combined WALS variables, both with respect to the WALS classification in terms of families and genera, and the more fine-grained *Ethnologue* classification. The underperformance of WALS is at least partially due to the very uneven distribution of the 139 variables over the 2560 languages. Therefore, we made a further selection by ranking the WALS variables according to their diachronic stability, applying the same method that we used to establish the stability of the lexical items. This gave a further reduction to the 355 languages in our database for which there are enough WALS variables available. This operation lead to a remarkable improvement of the results. The typological approach now does somewhat better than the lexical one. However, we found that an even better result is attained when both methods are combined. Taking both the 40 most stable Swadesh items and the 85 most stable WALS variables into consideration, assigning equal weights to them, we detected very good fits with all the classifications we used for comparison. This suggests that filling the gaps in the WALS database would provide linguistics with a reliable instrument for language classification, especially when this data is combined with lexical information as available in the ASJP database.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.