Author(s): Holman, Eric W.[1], Søren Wichmann[2], Cecil H. Brown[3], Viveka Velupillai[4], André Müller[5], and Dik Bakker[6] (ASJP Consortium)
University/Affiliation: University of California, Los Angeles[1], Max Planck Institute for Evolutionary Anthropology & Leiden University[2], Northern Illinois University[3], Justus-Liebig-Universität Giessen[4], Leipzig University[5], University of Antwerp/Lancaster[6]
Email address(es): D.Bakker@uva.nl[6]

Advances in automated language classification

Abstract

The paper presents a method for the automatic reconstruction of language relationships taking the Swadesh (1955) 100-item word list as a point of departure. However, the method differs from the original lexicostatistical approach in two fundamental ways. First, the comparison between word forms is done by a computer program (ASJP; automated similarity judgment program) on the basis of Levenshtein's (1966) algorithm, resulting in a distance matrix between individual languages. And second, graphic branching structures illustrating language relatedness (family trees) are generated from this matrix by the way of standard software and algorithms originally developed for the use of biologists in studying phylogenetic relationships (Huson 1998). To accommodate wordlists originally published in a variety of more or less simplified orthographies, a special alphabet, called ASJPcode, was devised which makes use of the QWERTY keyboard symbols only. It contains just 34 consonant symbols and 7 symbols for vowels. These symbols are used for phonological segments defined by the most common points and manners of articulation. Rarer segments are represented by the symbol they most closely resemble in terms of point and manner of articulation. See Brown et al (to appear 2008) for details.

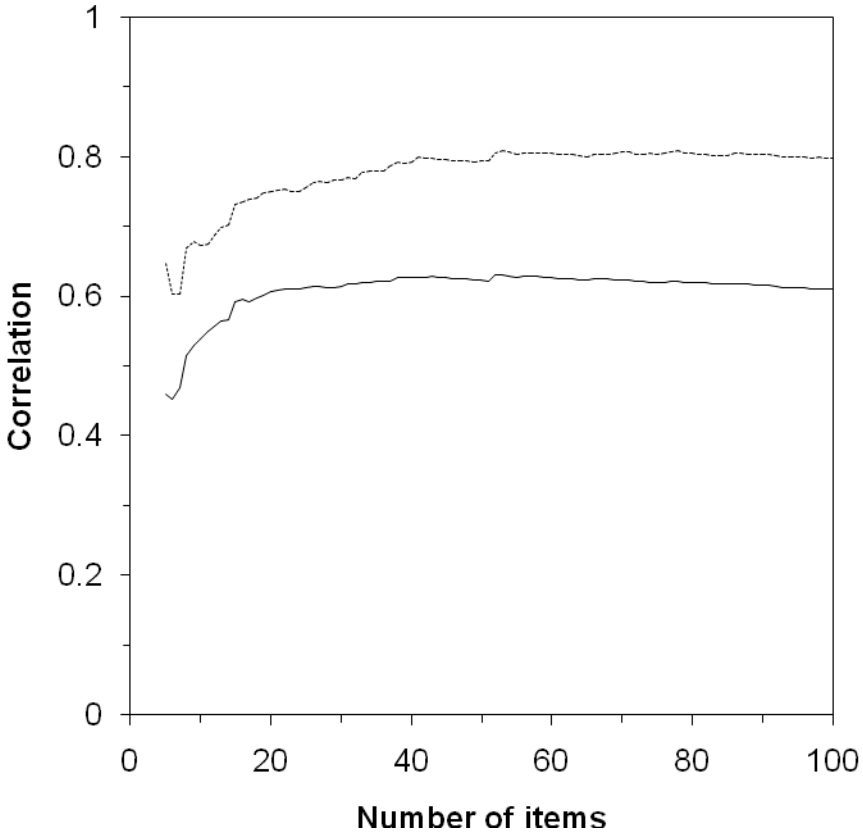
Unlike most other approaches to automatic language classification, such as those described by Oswalt (1971), Atkinson et al. (2005), and Nakleh et al. (2005), the present method automates both the judgments of cognacy and the subsequent inference of phylogeny. We can therefore apply the same objective criteria worldwide to classify an unusually large sample of languages. This facilitates the large scale statistical study of overlaps in lexicons between languages and may reveal previously unknown phylogenetic relationships.

To date, we have collected and transcribed a basic word set for close to 2000 languages of the world. The nearly 2 million language pairs in the database are compared by means of the Levenshtein Distance (LD: see Levenshtein 1966). For any pair of words represented in ASJPcode, LD is defined as the minimum total number of additions, deletions, and substitutions of symbols necessary to transform one word into the other. For any pair of languages L1 and L2, first the LD values are established for each of the N Swadesh words that L1 and L2 share (virtually always the full set that we consider). These LD values are then normalized by dividing each LD by its theoretical maximum giving the normalized LD (LDN). Finally, since lexical similarity may be influenced by chance resemblances, such as an overlap in the phoneme inventories or shared phonotactic preferences for the two languages involved, we correct each LDN by dividing by it the mean LDN of all N(N-1)/2 pairings of words with different meanings,

giving the LDND value for each of the N meaning pairs. The LDND value for the language pair L1 – L2, i.e. their Levenshtein distance, is defined as the mean of the LDND values for the individual word pairs.

Earlier experiments on several hundreds of languages have shown that the 100-item Swadesh list may be reduced to a much shorter one, without loss and even with a gain in classificatory reliability. The subset we selected contains the 40 most stable elements from the original list. Our measure of stability is based on the idea that the more stable items can be identified among a larger set because they have a greater tendency to yield cognates within widely acknowledged groups of closely related languages than words for less stable items. For a comparison of the values in our distance matrix, we have chosen the families and genera as established by Dryer (2005) and the genetic classification of the Ethnologue (Gordon 2005). If we take these classifications as a point of departure, and especially when looking at the more or less firmly and independently established groupings, then the stability factor for the individual lexical items turns out to be consistent across the languages from different hemispheres. Moreover, iterative comparisons lead to a specific subset of 40 items that make better predictions than any smaller subset, and at least as good predictions as any larger subset. The figure below gives an impression of this. A more detailed discussion may be found in Holman et al. (to appear 2008).

The 40-item list contains most of the items in the shorter lists proposed by Yakhontov (see Starostin 1991: 59-60) and Dolgopolsky (1986), and makes better predictions than do the shorter lists.



The ASJPcode was originally introduced for practical reasons: limitations of the keyboard, and problems to represent full IPA code in traditional programming languages.

These two problems have recently been overcome by the project. Full digitalized IPA representations are now automatically converted into equivalent numerical representations that the analyses programs can deal with. Interestingly however, this seems to have no noticeable influence on the results sofar: correlations between the LDN and LDND scores for both IPA and ASJP representations are all significant at the 1% level, and we have noticed no crucial differences between the tree structures produced.

By taking LDND instead of LDN as appoint of departure for further operations, we make an attempt to correct for chance similarities. But no attempt is made to distinguish inheritance from diffusion or universal tendencies. The relative influence of these three factors can be estimated empirically, however, by studying LDND as a joint function of taxonomic distance and geographic distance. For this analysis, geographic distances between languages of the ASJP sample were calculated as the shortest path on the surface of a sphere between the approximate centers of the areas in which the languages are spoken. Comparisons between groups of genetically related and non-related languages show that the amount of similarity declines with distance much more rapidly for the former than the latter groups. This suggests that borrowing of items from the Swadesh list is rather rare between non-related languages, and that most of the weight should be assigned to inheritance. Although there are clear exceptions among the language pairs, our current estimate is that on average not more than 1 or 2 out of the 40 items will be borrowed between non-related languages.

In order to further evaluate the role of lexical comparison we estimated the extent to which acknowledged genetic relationships may be confirmed by other methods. For this purpose we used a subset of the data stemming from the World Atlas of Language Structures (WALS; Haspelmath et al. 2005). Although the WALS project has a purely descriptive goal, and does in no way seek to contribute directly to genetic reconstruction, we think that the wide range and the quality of its database warrants this exercise. So, using the same method as for the Swadesh list to determine the optimal stable subset of the 140 linguistic features of the WALS, we established that for the relatively few languages with at least 100 attested features, the maximum correlations with the Ethnologue and WALS classifications are similar to the correlations for our 40 lexical items in a much larger sample of languages. It follows that equally good results can be achieved either with a high investment of research time in assembling typological features or with a low investment in assembling lexical items.

Furthermore, we studied the behavior of combinations of lexical material and typological features. Our results indicate that fairly close to optimal results are reached using the 40 most stable lexical elements and the 40 most stable typological features for each language, weighted such that lexical elements account for three quarters and typological features account for one quarter of the distance between pairs of languages.

A future goal of the project is to refine the current method of automatically detecting borrowings. It remains to be seen whether for this exercise less abstract representations than the ASJPcode would give better results. An effort will be made, therefore to make full IPA representations available for all languages in the database.

References

Atkinson, Quentin, Geoff Nichols, David Welch, and Russell Gray. 2005. From words to dates: water into wine, mathemagic, or phylogenetic inference? *Transactions of the Philological Society* 103:193-219.

Brown, Cecil et al. (to appear 2008). Automated Classification of the World's languages: A description of the method and prelimary results. *Sprachtypologie und Universalienforschung*.

Dolgopolsky, Aaron B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*, Shevoroshkin, Vitalij V. and Thomas L. Markey (eds.), , 27-50. Ann Arbor: Karoma.

Dryer, Matthew S. 2005. Genealogical language list. In: Haspelmath et al. (eds.), 584-643.

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue.* 15th Edition. SIL International. <www.ethnologue.com>.

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

Holman, Eric et al. (to appear 2008) Explorations in automated language classification. *Folia Linguistica*.

Huson, Daniel H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14.10: 68–73.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707-710.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81:382-420.

Oswalt, Robert L. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3:117-129.

Starostin, Sergei. 1991. *Altajskaja Problema i Proisxozhdenie Japonskogo Jazyka* [The Altaic Problem and the Origin of the Japanese Language]. Moscow: Nauka.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121-137.