

## **Explorations in automated language classification**

Eric W. Holman

*University of California, Los Angeles*

Søren Wichmann

*Max Planck Institute for Evolutionary Anthropology & Leiden University*

Cecil H. Brown

*Northern Illinois University*

Viveka Velupillai

*Justus-Liebig-Universität Giessen*

André Müller

*Leipzig University*

Dik Bakker

*University of Amsterdam & University of Lancaster*

### **0. Introduction**

An earlier paper, to which some authors of the present paper have contributed (Brown et al., 2008), describes a method for automating language classification based on the 100-item referent list of Swadesh (1955). Here we discuss a refinement of the method, involving calculation of relative stabilities of list items and reduction of the list to a shorter one by eliminating least stable items. The result is a 40-item referent list. The method for determining stabilities is explained, as well as a method for comparing the classificatory performance of different-sized reduced lists with that of the full 100-item list. A statistical investigation of the relationship of lexical similarity of languages to their geographical proximity is presented. Finally, we test the possibility that

information involving typological features of languages can be combined with lexical data to enhance classificatory accuracy.

### **1. Summary of Brown et al. (2008)**

The earlier paper describes a procedure for automated comparisons of word lists—henceforth ASJP for ‘automated similarity judgment program’. The approach on the whole is similar to lexicostatistics (Swadesh 1950, 1955), but differs in two fundamental ways: (1) the judgment of similarities is done by a computer program following a consistent set of rules, and (2) graphic branching structures illustrating language relatedness (family trees) are generated through use of standard software and algorithms originally developed for the use of biologists in studying phylogenetic relationships.

A 100-item Swadesh list is assembled for each language to be compared. Words on all lists are transcribed into a standardized orthography — ASJPcode — which employs only symbols of the standard QWERTY keyboard. ASJPcode has 7 different vowel symbols, merging two or more vowels under a single specific symbol when a language has more than 7 vowel qualities. Nasalized vowels are indicated, but vowel length, tone, and stress are not. The orthography also employs 34 consonant symbols. These symbols are used for phonological segments defined by the most common points and manners of articulation. Rarer segments are represented by symbols for the more common segments they most closely resemble in terms of point and manner of articulation.

There are also modifiers to indicate that a single segment is composed of sounds corresponding to two symbols (and, occasionally three), typically in the cases of labialization, aspiration, and palatalization. Other modifiers indicate

glottalization and nasalization. Word-initial glottal stops are not recorded and complex syllabic components involving vowel plus either velar-to-glottal fricatives, glottal stop, or palatal approximant are reduced to vowels.

The ASJP database includes only languages for which at least 70% of the Swadesh items are attested. Items not attested for a given language are treated as missing data and not used in calculations of similarity.

With such lists as input, a computer program identifies similar words in pairwise language comparisons. The computer program can be easily modified and the results of varying instructions can be compared. Our first automated comparison entailed matching rules that allowed for common sound changes, such as  $p > f$  or  $t > d$ . However, better results (judged with respect to how ASJP automated language classification compares to expert classification) were achieved by requiring segments to be identical (after conversion to ASJPcode).

Two words are judged similar if at least two consecutive consonants in the respective words are identical. The two consonants can be either contiguous or interrupted by any vowel symbol or symbols. If there are only two or three symbols in a compared word, vowels are also taken into account, such that identical CV or VC sequences count as a sufficiently good match (where C = consonant and V = vowel); moreover, identity between a vowel in one word and either member of a set of contiguous vowel symbols in the other word is counted toward a match. The modifiers mentioned above allow a few additional matches as described by Brown et al. (2008). The proportion of words with the same meaning judged to be similar for a given pair of languages is the Lexical Similarity Percentage (LSP).

The lexical similarity of words judged similar by ASJP may result from factors other than genetic relationship (shared retentions or shared innovations). For example, they may be found to be similar because of chance resemblance.

This is especially likely for languages having similar phoneme inventories or similar phonotactic preferences. In addition, languages with longer lexical morphemes are more likely to produce spurious matches. To control for these chance factors we calculate the Phonological Similarity Percentage (PSP), which is the average proportion of words with different meanings judged to be similar for a language pair. PSP is then subtracted from the LSP yielding the Subtracted Similarity Percentage (SSP).

Once SSPs are determined for all possible language pairs based on attested words pertaining to the 100-item Swadesh list, a distance matrix is produced by defining the ASJP distance between each language pair as  $100\% - \text{SSP}$ . This matrix serves as a database for the estimation of phylogenetic trees. Embleton (1991) reviews the phylogenetic methods used in early lexicostatical studies. We have used a more recent phylogenetic algorithm called Neighbor Joining (Huson 1998), which has been shown to produce more adequate tree topologies than earlier algorithms (Saitou and Nei 1987).

The database of Brown et al. (2008) consists of 100-item lists transcribed into ASJPcode for 245 languages. With this database, the automated classification method was tested on several non-controversial language families (Mayan, Mixe-Zoque, Otomanguean, Huitotoan-Ocaina, Tacanan, Chocoan, Muskogean, Indo-European, and Austro-Asiatic) and was found to yield classifications that are in substantial agreement with published expert subgrouping of languages within those families.

## **2. Determining item stability**

The Swadesh 100-item referent list can be reduced to a much shorter list that yields equally good classificatory results through ASJP analysis, by restricting the list just to its most stable items. Item stability refers to the degree to which words for an item are retained over time and not replaced by another lexical item from the language itself or a borrowed element: words resistant to replacement are more stable than otherwise. Our method for estimating stability is similar to one used by Wichmann and Holman (n.d., also presented briefly in Holman et al. 2007) to infer stabilities of typological features from data in the *World Atlas of Language Structures (WALS)*: Haspelmath et al. 2005).

The measure of stability is based on the simple idea, which goes back to the early work of Thomas (1960), Kroeber (1963), and Oswalt (1971), that words for more stable items can be identified because they have a greater tendency to yield cognates within groups of closely related languages than words for less stable items. For measuring stability, we have used as groups of closely related languages those languages identified by Dryer (2001, 2005) as belonging to the same genera. Dryer defines the latter as the most inclusive groups believed to have diverged from their common ancestor no more than about 3500 to 4000 years ago.

To correct for chance similarity matches between different items, we determine stability,  $S$ , according to the following formula:

$$(1) \quad S = (R - U)/(1 - U),$$

where  $R$  is a weighted average across genera of the proportion of matches between same items and  $U$  is a weighted average proportion of matches between different items.<sup>1</sup>

Measuring the stability of items on a referent list entails the assumption that words having similar meanings have similar stabilities across languages. To test this assumption, we estimated  $S$  for each of the 100 items, separately for the Eastern Hemisphere, based on 128 languages in 23 families, and the Western Hemisphere, based on 117 languages in 46 families. The Spearman rank correlation of stability across items between hemispheres is 0.37. The observed correlation is within the range of 0.21 to 0.44 reported by Kruskal et al. (1973) and Cavalli-Sforza and Wang (1986) for correlations among stabilities of lexical items estimated by other methods in different language families. Wichmann and Holman (n. d.) found a rank correlation of 0.51 for stability of typological features between hemispheres.

An immediate question is whether the correlation between hemispheres is reduced by geographic variation in the relative stability of items. To answer this

---

<sup>1</sup> The numerator in (1) is the same as the weighted average of SSP, and the denominator ensures that stability is equal to 1 if all items match within genera. This sort of correction for chance is common in deriving similarity indices (Albatineh et al. 2006). Stated symbolically,  $R = \Sigma(m/\sqrt{c})/\Sigma(\sqrt{c})$ , where  $m$  is the number of matches in a genus,  $c$  is the number of comparisons in a genus, and both sums are over genera. To verify that this is a weighted average of proportions, multiply the top and bottom of  $m/\sqrt{c}$  by  $\sqrt{c}$  to get  $\sqrt{c} \times m/c$  and note that  $m/c$  is the proportion of matches. The square-root weighting compensates for the fact that the number of comparisons within a genus increases approximately as the square of the size of the genus, while the amount of data in a genus increases only as the size of the genus.  $U$  is a similarly weighted average of the proportion of matches when the given word is compared to each other word on the list in the other languages in the same genus. The formula for  $U$ , then, is the same except that matches and comparisons are between different items.

question, the 69 families in the sample were randomly split into groups of 23 and 46 families, stabilities were estimated in each group, and the rank correlation calculated between groups. This randomization procedure was repeated 1000 times. The average correlation was 0.39; on 382 of the 1000 trials, the correlation was lower than observed between hemispheres. In other words, the stabilities of Swadesh items are scarcely more different between hemispheres than expected from random sampling error. Wichmann and Holman (n. d.) drew the same conclusion for stabilities of *WALS* features.

Our ranking of the 100 items on the Swadesh list with respect to calculated stability is found in the Appendix.

There has been some discussion in the literature about whether there is a connection between stability of words and their susceptibility to borrowing. According to Wang and Wang (2004), Chen (1996) found words for items on the 100-item Swadesh list to be more resistant to borrowing than those found only on the 200-item list and not on the 100-item list. The same finding was presented by Kessler (2001:104-108) for loans into Albanian, English, French, German, and Turkish. McMahon and McMahon (2005: 94-95) then averaged Kessler's figures to find a borrowing rate of 8.6% on the 100-item list and 15.7% among the items on the 200-item list that are not also on the 100-item list.

A larger sample of languages is available in the preliminary database of 36 languages for the Loanword Typology Project (LWT) of the Linguistics Department at the Max Planck Institute for Evolutionary Anthropology<sup>2</sup>. With this database, borrowing rates can be estimated for 99 of the 100 Swadesh items; the average borrowing rate for these items is 8.5%, almost the same as that found by in Kessler's sample.. We have investigated whether there is a correlation

---

<sup>2</sup> The results of this project have not yet been published, but a description of the project is found here: <http://www.eva.mpg.de/lingua/files/lwt.html> .

between our stability ranking and the rate of borrowability estimated from the LWT data. No such correlation was found: the Spearman rank correlation across 99 items between stability and borrowability is 0.04. There are several possible explanations of this: (a) borrowability may be more variable for given lexical items across geographic areas than stability and, thus, may not be an inherent property of lexical items; (b) borrowability is not a significant contributor to stability, at least as the segment constituted by the Swadesh 100-item list is concerned; (c) there are still far too little data on borrowability to be conclusive (the sample for studying lexical stability consisted of 245 languages, whereas we had only the 36 languages of LWT at our disposal for the study of borrowability).

### 3. Word list size

In investigating the classificatory performance of word lists of different lengths, we examined lists ranging in length from 100 items to 5 items, such that the list of 99 items contained the top 99 most stable items, the list of 98 contained the top 98 items and so on. The ASJP distance matrices produced from the resulting different sized lists were correlated with two broadly accepted language classifications: Dryer's (2005) ranked classification presented in *WALS*, and the unranked trees in *Ethnologue* (Gordon 2005). These quantitative measures were supplemented with comparisons of neighbor-joining trees produced from 100-item lists, 50-item lists, and 40-item lists.

Figure 1 gives an overview of how list length influences the correlations between ASJP distance and taxonomic distance in the *WALS* classification (solid line) and in the *Ethnologue* classification (broken line). The first (solid line) is calculated as the standard Pearson product-moment correlation (across pairs of

languages) with the *WALS* classification, where taxonomic distance is defined as 1 for languages in the same genus, 2 for languages in different genera but the same family, and 3 for languages in different families. The second (broken line) is calculated as the Goodman-Kruskal gamma with the *Ethnologue* classification.<sup>3</sup> We note that the curves are similar in shape. Both correlations increase with list length up to about 40 words but change little for longer lists. So the shortest list providing an optimal fit between the classifications resulting from the ASJP method and both the *WALS* and *Ethnologue* classifications is around 40 items. Visual inspections of trees produced from 100 items and 40 items confirmed that results were similar. Typically small differences either toward greater or smaller accuracy would cancel one another in the two approaches.

Having established the 40 most stable and effective items with respect to language classification, we made some further minor revisions to the list by replacing ‘rain’, ‘kill’, and ‘bark’ with items ranking 41th, 42nd, and 43rd on the

---

<sup>3</sup> The reason why two different correlation measures are used is that the *WALS* classification operates with just three taxonomic levels (the language, the genus, and the family), which are intended to be comparable across the entire classification, whereas the *Ethnologue* classification is more complicated, having variable numbers of taxonomic levels, which are not intended to be comparable across families. Gamma is defined as  $(C-D)/(C+D)$ , where C is the number of concordant comparisons (those ordered in the same direction on both variables), and D is the number of discordant comparisons (those ordered in opposite directions on the two variables). In the present application, one variable is ASJP distance and the other is taxonomic distance in *Ethnologue*. For the latter, if *Ethnologue* classifies two languages in the same group and a third language outside that group, then two comparisons are possible: the distance between the first two languages is less than the distance between the first and third languages, and also less than the distance between the second and third languages. Gamma summarizes the consistency of such comparisons with ASJP distances. Like other correlation coefficients, gamma ranges from -1 to +1 and takes the value 0 if the variables are independent. The fact that gamma is higher than the Pearson correlation merely reflects the fact that gamma ignores ties, which are frequent in taxonomic distances.

list, in order to cope with the fact that in many languages morphemes occurring in words for these three items recur in words for other items on the list, i.e., respectively, ‘water’, ‘die’, and ‘skin’. The starred items in the Appendix are the 40 items on the definitive list to be used in ASJP analysis.

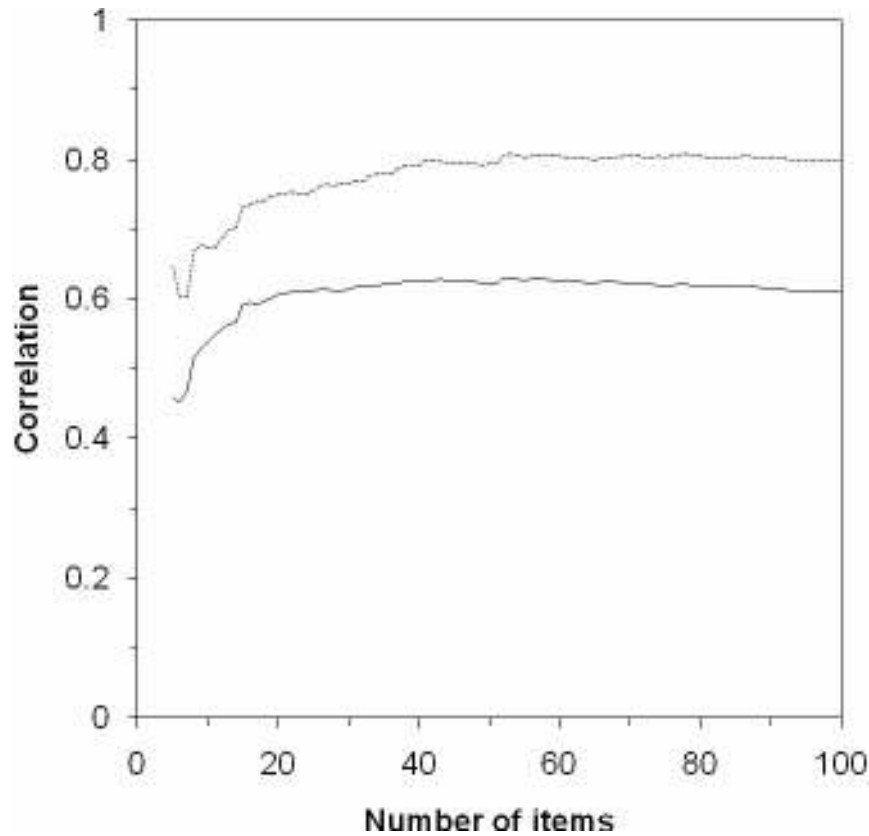


Figure 1. Correlations of ASJP distances with the *Ethnologue* classification (dotted line) and with the *WALS* classification (solid line) as a function of list length.

Other scholars have proposed shorter versions of the Swadesh list, including Yakhontov, to whom a 35-item list is attributed (Starostin 1991: 59-60).

Of these 35 items, 24 are on our 40-item list, 8 are from the rest of the 100-item Swadesh list and 3 are from the rest of the 200-item Swadesh list. When Yakhontov's 32 items selected from the 100-item list are used to produce ASJP distances, the correlations are 0.57 with *WALS* and 0.73 with *Ethnologue*, whereas the 32 most stable items in our list produce correlations of 0.62 and 0.77. The Yakhontov list of 32 items produces about the same correlations as the 15 most stable items shown in the Appendix.

An even shorter list was constructed by Dolgopolsky (1986) consisting of the 23 most stable words in European and Asian languages: 'I/me', 'two/pair', 'thou/thee', 'who/what', 'tongue', 'name', 'eye', 'heart', 'tooth', 'no/not', 'fingernail/toenail', 'louse/nit', 'tear(drop)', 'water', 'dead', 'hand', 'night', 'blood', 'horn', 'full', 'sun', 'ear', 'salt'. There is fairly good agreement with our list inasmuch as 16 items on the Dolgopolsky list are on our 40-item list and 4 are from the rest of the Swadesh 100-item list. ASJP distances based on these 20 items show a correlation of 0.52 with *WALS* and 0.66 with *Ethnologue*, about the same as the 8 most stable items on our list, whereas the 20 most stable items on our list produce correlations of 0.61 and 0.75. The Dolgopolsky list, like the Yakhontov list, would be too short for best results even if it contained the most stable items.

Our findings are consistent with a study of Wang and Wang (2004) showing that Yakhontov's 35-word list or shorter lists do not produce as good results for subgrouping Chinese dialects as the 100-item Swadesh list. Our results differ from the simulation results of Embleton (1986: 92-3) pointing to a greater accuracy of a 200-item list over a 100-item list, and also from the sampling experiments that led Kessler (2001: 67) to conclude: ". . . sample size does matter. All things being equal, it pays to have more words in the sample." Kessler went on to suggest, however, that all things may not be equal, particularly the stability

of lexical items. Our results show that selecting items for stability can compensate for a smaller sample down to about 40 items.

#### **4. Lexical similarity and spatial distance**

While the data used for the investigation of stabilities of the Swadesh items pertained to the same 245 languages investigated in Brown et al. (2008), the database has subsequently been enlarged with 40-item lists for other languages to yield 100-item or 40-item lists for a total of 876 languages, whose locations are displayed in Figure 2. The names and three-letter *Ethnologue* codes of all 876 languages, along with references to the sources of the Swadesh lists, can be found at: <http://lingweb.eva.mpg.de/asjp/index.php/ASJP>. Results reported in this and the next section derive from 859 of the 876 languages, excluding pidgins, creoles, constructed languages, proto-languages, and languages extinct for more than 200 years. In order to eliminate list length as a source of variability, results are based on the 40-item list for all languages, including those for which 100-item lists are available.

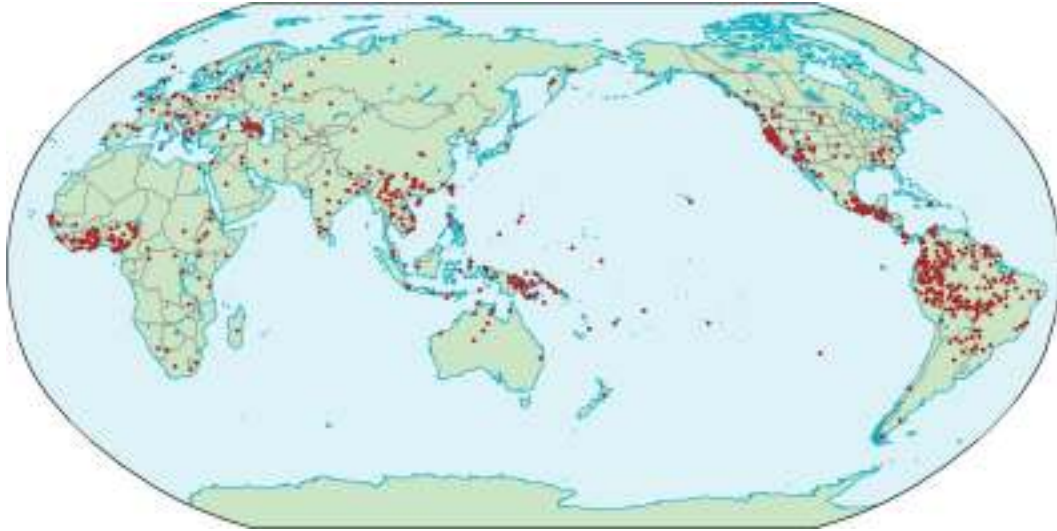


Figure 2. Distribution of the 876 languages currently processed within the ASJP project (produced using Bibiko 2005).

The comparative method of historical linguistics, used to produce expert classifications such as those presented in *WALS* and *Ethnologue*, is designed to distinguish inherited similarity from similarity caused by other factors, notably chance, diffusion, and universal tendencies such as onomatopoeia, sound symbolism, and nursery forms. The definition of SSP makes an effort to correct for chance similarity by subtracting PSP from LSP, but it makes no attempt to distinguish inheritance from diffusion or universal tendencies. The relative influence of these three factors can be estimated empirically, however, by studying SSP as a joint function of taxonomic distance and geographic distance.

The relation between lexical similarity and spatial distance was first studied in dialectometry by Séguy (1971) and Goebel (1984), who found that the lexical similarity of dialects is inversely related to their geographic distance from one another (the more distant, the less similar). Cavalli-Sforza and Wang (1986) found a similar relation in a set of closely related languages, and pointed out that

the decline in lexical similarity with increasing distance could reflect both inheritance through early migration and also diffusion through later contact. To distinguish among these factors in the context of language typology, Holman et al. (2007) plotted typological dissimilarity as a function of spatial distance at two taxonomic levels: languages in the same *WALS* family and languages in different families. Differences in typological dissimilarity between taxonomic levels at the same spatial distance could then be attributed to inheritance, and the effect of distance on dissimilarity between languages in different families could be attributed to diffusion unless controversial inherited relationships among families were invoked instead.

The same analysis is applicable to SSP. To undertake such an 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. Latitudes and longitudes of these centers are provided in *WALS*; for languages not in *WALS*, they were ascertained from maps in Moseley and Asher (1994), or, if necessary, determined from information in *Ethnologue* or in the original sources for lists. Pairs of languages were then sorted according to distance in intervals such as 0-1000 km, 1000-2000 km, etc. Within each distance interval, pairs were grouped into one of the three taxonomic levels in the *WALS* classification, members of the pair being either (1) in the same genus, (2) in different genera but the same family, or (3) in different families. For each group, weighted average SSP and distance were calculated with weights similar to those described in Note 1. Figure 3 plots SSP (on a logarithmic scale) as a function of distance at each taxonomic level. The top two curves do not extend beyond 5000 km because only four families (Austronesian, Indo-European, Niger-Congo, and Sino-Tibetan) contain languages in the sample more than 5000 km apart).

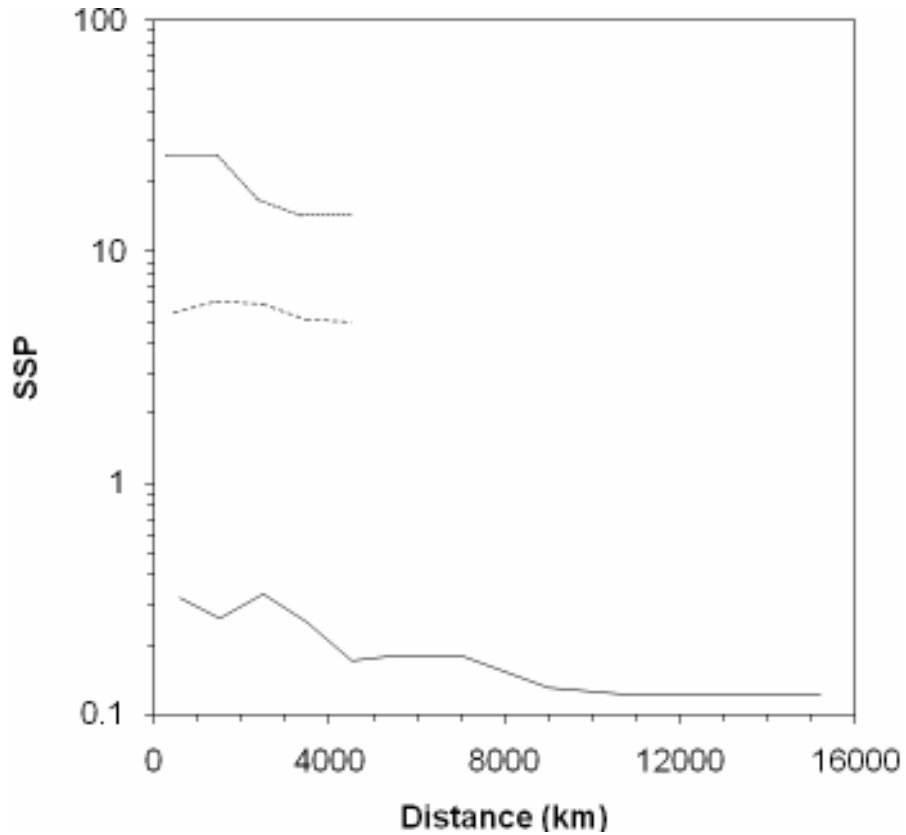


Figure 3. Lexical similarity (SSP) as a function of spatial distance between languages in the same genus (dotted line), different genera in the same family (dashed line), and different families (solid line).

At each distance, SSP is highest for languages in the same genus (dotted line), intermediate for languages in different genera within the same family (dashed line), and lowest for languages in different families (solid line). This effect of taxonomic rank is attributable to inheritance with diffusion held constant. At each rank, SSP tends to decline as distance increases. To test the significance of the decline, the slope of the regression line relating SSP to distance was

calculated separately at each rank for each family with enough data; the slopes are significantly negative at all three ranks by Wilcoxon tests,  $z \geq 2.65$ ,  $p < .01$ .

For languages in different families, the decline of SSP with distance indicates either diffusion or previously unacknowledged phylogenetic relationships, because universal tendencies are expected to be independent of distance. Although numerically small, the decline is statistically significant because most pairs of languages are in different families, producing a very large sample size. Both factors implicated in the decline are evidently very weak, because SSP is only about 0.3% at the shortest distances before declining to about 0.1% at the longest distances. A weak effect of inheritance could mean either that few relationships between families remain to be discovered, or that SSP is insensitive to such remote relationships. A weak effect of diffusion suggests that Swadesh words are rarely borrowed between unrelated languages. It is also possible that some borrowings have gone undetected because changes in sound and meaning in words have caused them to escape ASJP matching criteria, although sound changes in borrowed words generally tend to be small. The asymptote near 0.1% may indicate a small residue of universal tendencies or else diffusion from a few worldwide languages.

For languages in different genera but the same family, the shallow slope of the curve again indicates a weak effect of distance, while a stronger effect is apparent for languages in the same genus. It is not the case, however, that SSP is a generally insensitive measure except for closely related languages. Taxonomic rank has a strong effect on SSP at all degrees of relationship. In particular, SSP is substantially lower for neighboring languages in different families (about 0.3%) than for distant languages in the same family (about 5%), and SSP is also much lower for neighboring languages in different genera of the same family (about 6%) than for distant languages in the same genus (about 14%). In short, one level

of taxonomic distance has much more effect on SSP than 4000 km of geographic distance. It follows that SSP is sensitive mainly to inheritance rather than diffusion. This behavior stands in contrast to that of the typological data in *WALS*, which are sensitive to both inheritance and diffusion (Holman et al. 2007).

## **5. Enhancing results by using typological information**

Given recent uses of typological information to infer genealogical relationships (Dunn et al. 2005, Wichmann and Saunders 2007) it is of interest to investigate which kind of data, typological or lexical, produces the more accurate classifications. Also of interest is whether classification is enhanced when typological and lexical data are combined.

Comparison of lexical and typological data should take into account the amount of data available of each kind. Each ASJP list includes at least 70% or 28 of the 40 items. The languages in *WALS* vary much more widely in their number of attested features. We therefore sorted these languages into nested sets that have respectively a minimum of 20, 40, 60, 80, and 100 attested features out of the 134 nonredundant features in *WALS*. The respective number of languages in each set is 884, 386, 221, 165, and 104. The analyses described below were performed on each of these five sets of languages.

First we compared the classificatory performance of typological data to the performance of lexical data as presented above in Figure 1. Stabilities of typological features were calculated similarly to stabilities of lexical items (see Wichmann and Holman n. d. for a full discussion and Holman et al. 2007 for a brief description of the method). The features were ordered by stability and successive datasets were defined to contain all 134 features, then the 133 most stable, then the 132 most stable, and so on until only the 10 most stable were left.

For each dataset a distance matrix was constructed in which the distance between each pair of languages was defined as the percentage of features with different values among all the features in the dataset attested for both languages. If no features in the dataset were attested for both languages, the distance was treated as missing.

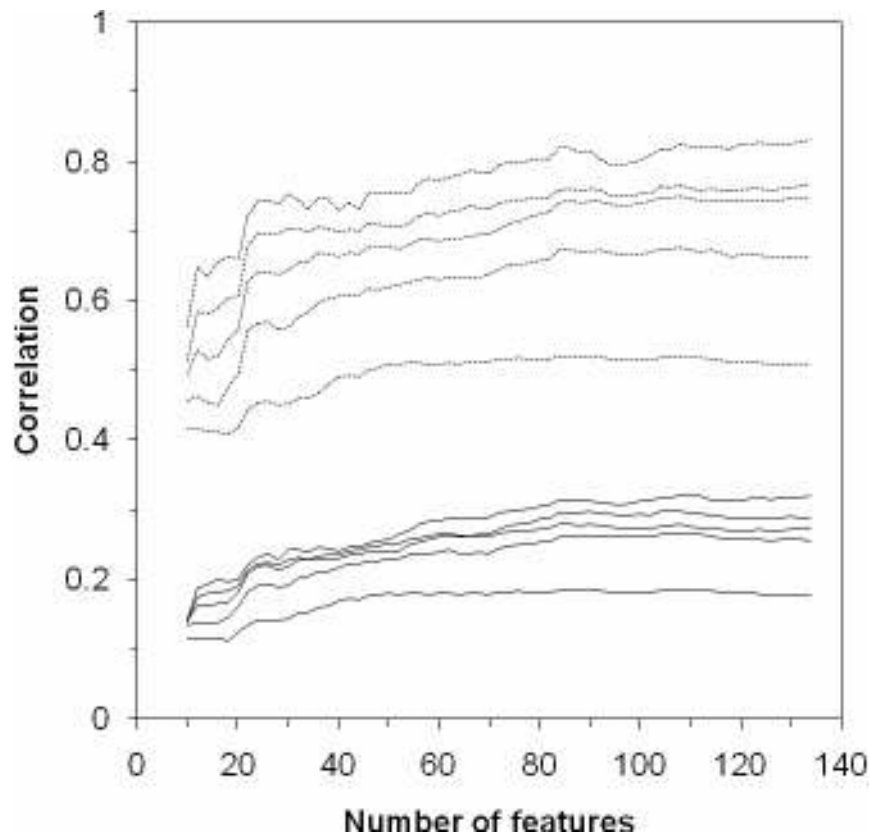


Figure 4. Correlations of *WALS* distances with the *Ethnologue* classification (dotted lines) and with the *WALS* classification (solid lines) as a function of the number of features included. In each group of curves, the lowest (on the left side of the graph) represents the sample of languages with 20 attested features, and

successively higher curves represent languages with 40, 60, 80, and 100 attested features, respectively.

Figure 4 shows correlations between *WALS* distance and taxonomic distance as a function of the number of features in the database. The group of dotted curves describes Goodman-Kruskal gamma with the *Ethnologue* classification and the group of solid curves describes Pearson correlation with the *WALS* classification. Each individual curve refers to languages with a different minimum number of attested features, with 20 features at the bottom of the group and 100 at the top. Not surprisingly, better-attested languages produce higher correlations, no matter how many features are actually used in the database. Each curve describes what happens when successively more features are used in descending order of stability. The curves show that correlations improve as features are added until a maximum is reached with the 85 or so most stable features; correlations change little with more than 85 features. Thus, the ideal is to use the 85 most stable features, and all should be attested for the languages compared. Cysouw et al. (2008, Figure 2) report similar results when features are selected for their compatibility with other features rather than their stability over time.

For the relatively few languages with at least 100 attested features, the maximum correlation with the *Ethnologue* classification is about 0.8, similar to the correlation for 40 lexical items. 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. Clearly the latter approach is to be preferred.

The result for the *WALS* classification brings out a more radical version of this conclusion, since here the maximal correlation is around 0.3, whereas it was

0.6 for the lexical data. It is currently not clear to us why lexical data perform much better than typological data for language classification when the *WALS* classification is the yardstick, but we speculate that it relates to the fact that the *WALS* classification is based on intuitions about similar time depths for the intermediate taxonomic level of genera and is thus more or less directly informed by glottochronology, which in turn is based on the same kind of measurement of distances between lexical data that we use. In any case, it is clear that a large amount of typological data will work no better or possibly worse than a small amount of lexical data for the purpose of classifying languages.

We now turn to the question whether or not a combination of lexical and typological data might work better than either would alone. For this investigation we use the languages attested in *WALS* for which we also have ASJP lists and again construct sets of languages having a minimum of 20, 40, 60, 80 or 100 attested *WALS* features. The sizes of these sets are, respectively, 341, 218, 139, 109, and 79 languages. For each set we compare languages pairwise and construct two distance matrices as described previously: one for lexical (ASJP) distance based on 40-item lists and one for typological (*WALS*) distance based on all 134 features. Now we calculate percentage mixtures of lexical and typological distance such that ASJP data account for anything from 0% to 100% of the total distance and *WALS* data account for the rest.

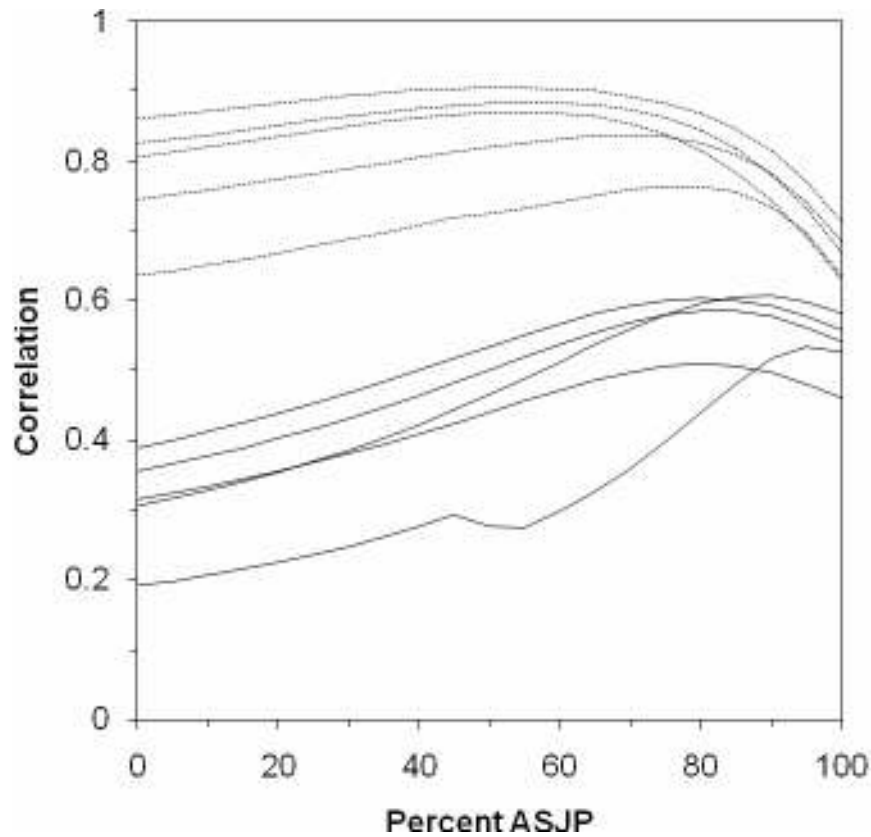


Figure 5. Results of mixing *WALS* and ASJP distances: correlations with the *WALS* classification (solid lines) and the *Ethnologue* classification (dotted lines) as a function of the percentage of ASJP data in the mixture. In each group of curves, the lowest (on the left side of the graph) represents the sample of languages with 20 attested features, and successively higher curves represent languages with 40, 80, 60, and 100 attested features, respectively.

Figure 5 shows correlations between these mixed distances and taxonomic distances as a function of the percentage of ASJP data in the mixture. Moving

from left to right corresponds to increasing the influence of ASJP data and decreasing the influence of *WALS* data correspondingly. As before, dotted curves describe Goodman-Kruskal gamma with the *Ethnologue* classification and solid curves describe Pearson correlation with the *WALS* classification, the curves referring to each of the five sets of languages with a different minimum number of attested features. Since the rightmost points on all the curves are based entirely on ASJP data, differences among these points (for a given correlation) reflect sampling variability among the five sets of languages.

All the curves have intermediate peaks higher than either their beginnings (no ASJP data, only *WALS* data) or their ends (only ASJP data, no *WALS* data). This clearly indicates that a combination of the two types of data leads to better results than either could produce alone. The lowest solid curve shows that in the case where only a minimum of 20 *WALS* features are attested and the *WALS* classification is our yardstick of performance, the gain by using typological features is very small: by letting the distances produced by such features contribute 5% of the combined distance scores we get a small improvement over using lexical distance exclusively (the curve ends at a point slightly lower than the peak).

If the *Ethnologue* classification is the yardstick, a more solid gain is seen with a minimum of 20 attested features. But when a minimum of 40 features are attested, performance increases rather dramatically for the combined distance scores when either *Ethnologue* or *WALS* is the yardstick. Depending on whether *Ethnologue* or *WALS* is the yardstick, performance increases slightly or does not increase with more attested features. Taking into account the great cost of labor involved in producing *WALS*-type data, our results indicate use of around 40 features, which should be attested for all languages. The two groups of curves are slightly at variance with respect to the percentage of lexical and typological data

that produce the best results. But most of the curves are near their peaks within the range of 60-80% lexical data (corresponding to 40-20% typological data).

Language classifications based on distances between pairs of languages yield better results when a combination of lexical and typological features is used than if only lexical features or only typological features are used. Fairly close to optimal results are reached using the 40 most stable lexical features and the 40 most stable typological features for each language, weighted such that lexical features account for three quarters of each distance and typological features account for the remaining quarter.

## **6. Conclusions**

An important result described in this paper is a method for estimating stabilities of lexical items with certain meanings. Based on this method, we propose a selection of 40 Swadesh-list items (meanings) to be used in automated comparative analysis of languages. Since a 40-item list typically can be extracted from lexical sources and transcribed into ASJPcode in about a half hour, we have been able to amass lists for nearly 900 languages in a relatively short period of time. Using this database, we have looked at the relationship between lexical similarity of languages and geographical distances between them, and at how well either lexical or typological data, or a mixture of the two, perform with respect to capturing phylogenetic associations among languages.

The ASJP lists are similar to typological information in that once the data are compiled for individual languages, all pairs of languages can then be compared quickly and objectively. The resulting pairwise matrices can be evaluated relative to the same external standards to address perennial questions about the value of different types of data for inferring phylogenetic relationships.

The short answer from our results is that the lexical data of ASJP are better than the typological data in *WALS* but an appropriately weighted combination is best.

Some more specific reasons for this general conclusion are also available. In particular, the ability to compare all pairs of languages allows the calculation of the baseline similarity observed among languages not generally considered to be related. Such a baseline is not available from the standard comparative method, which assumes some degree of phylogenetic relationship among the languages compared. The very low baseline similarity apparent in Figure 3 between languages in different families indicates that ASJP is very successful at excluding loanwords, the perennial threat to claims of phylogenetic relationship. This success could mean either that loanwords are rare in the 40-item list, or that most loanwords are sufficiently altered in sound or meaning that they are not scored as similar by the ASJP matching rules. In either case, the insensitivity of ASJP scores to diffusion may explain their superiority over typological features for classification.

The low baseline in Figure 3 also has the less welcome implication that ASJP scores are not very sensitive to whatever remote relationships may exist between currently recognized families. Consistent with this implication, Brown et al. (2008) found ASJP to be less successful at finding deeper than shallower branches of phylogenetic trees produced by experts. Typological data may offer more hope here in view of the stability estimates for typological and lexical features, which can be compared because they are based on the same formula (Eq. 1). Wichmann and Holman (n. d.) report an average stability of 35.6% for the 134 *WALS* features with a maximum of 80.8%, while the Appendix below shows an average stability of 23.4% for the 100-item list and 30.5% for the 40-item list, with a maximum of 42.8%. Since the effects of even small differences in stability are amplified with increasing time depth, the slightly higher stability of

typological features may be the ingredient that enhances performance when typological data are mixed with ASJP data.

Another explanation for the superior performance of mixed lexical and typological data is that expert classifications — our standard of comparison — are themselves often based on a combination of lexical and structural features. This explanation is compatible with the previous one because experts can be expected to choose for taxonomic purposes features thought to be relatively stable over time. Although of some methodological interest, the improvement attributable to the added typological data is nevertheless small enough to raise the practical question of whether the benefit justifies the cost in time and effort.

The present results suggest several lines of future research on the use of typological and lexical data for classification. For typological data, a question is whether loans can be distinguished from inherited features well enough to take advantage of the superior stability of some typological features. For lexical data, work is needed on generalizing the ASJP matching rules and on testing data that are encoded in full phonological representations, such as is possible with Unicode. Furthermore, if markedly different typological or lexical databases are developed as suggested, or if language families are sampled more fully for both typological and lexical data, it will then be necessary to compare the performance of the different kinds of data once more before determining the best combination of lexical and typological data for the purpose of classifying languages. These investigations are viable and are currently in preparation, but require the collaboration of many researchers and years of future work.

## **Acknowledgments**

The following colleagues have been helpful in providing data for individual languages: Joseph Atoyebi, Thomas J. Connors, Mark Donohue, David Gil, Zaira Khalilova, Konstantin Krasukhin, Robert Mailhammer, Bill McGregor, Brigitte Pakendorf, Don Stilo, Uri Tadmor, Edward Vajda, Ljuba Veselinova, and Alena Witzlack-Makarevich. We also thank the following scholars for references, comments, discussion, and other input to our project: Barry Alpher, Gene Anderson, Hans-Jörg Bibiko, Robert Blust, Václav Blažek, Pamela Brown, Michael Cahill, Shobhana Chelliah, William Croft, Michael Dunn, Rob Goedemans, Jeff Good, Harald Hammerström, Nick Hopkins, Hagen Jung, Terry Kaufman, Brett Kessler, Maarten Kossmann, David B. Kronenfeld, Frank Landsbergen, Stephen Levinsohn, Sasha Lubotsky, Luisa Maffi, Robert Mailhammer, Steve Marlett, Carolyn Miller, Edith Moravcsik, Maarten Mous, Johanna Nichols, Michael Noonan, Andrew Pawley, Doris Payne, Filippo Petroni, Ger P. Reesink, Keren Rice, Don Ringe, John Roberts, David S. Rood, Malcolm Ross, Fedor Rozhanskiy, Maurizio Serva, Keith Snider, Jae Jung Song, John Stark, Lynn Thomas, Peter Trudgill, Rene van den Berg, Piet van Reenen, James M. Unger, Mary Ruth Wise, and Jan Wohlgemuth. Finally we are grateful to Russell Gray and Martin Haspelmath for permission to use data from the Austronesian Basic Vocabulary Database and the LWT Project, respectively.

## Appendix

A ranking of stability of items on the 100-word Swadesh list with indication (by \*) of members on the reduced 40-item list

Rank	# In list	Meaning	Stability
1	22	*louse	42.8
2	12	*two	39.8
3	75	*water	37.4
4	39	*ear	37.2
5	61	*die	36.3
6	1	*I	35.9
7	53	*liver	35.7
8	40	*eye	35.4
9	48	*hand	34.9
10	58	*hear	33.8
11	23	*tree	33.6
12	19	*fish	33.4
13	100	*name	32.4
14	77	*stone	32.1
15	43	*tooth	30.7
16	51	*breasts	30.7
17	2	*you	30.6
18	85	*path	30.2
19	31	*bone	30.1
20	44	*tongue	30.1

21	28	*skin	29.6
22	92	*night	29.6
23	25	*leaf	29.4
24	76	rain	29.3
25	62	kill	29.2
26	30	*blood	29.0
27	34	*horn	28.8
28	18	*person	28.7
29	47	*knee	28.0
30	11	*one	27.4
31	41	*nose	27.3
32	95	*full	26.9
33	66	*come	26.8
34	74	*star	26.6
35	86	*mountain	26.2
36	82	*fire	25.7
37	3	*we	25.4
38	54	*drink	25.0
39	57	*see	24.7
40	27	bark	24.5
41	96	*new	24.3
42	21	*dog	24.2
43	72	*sun	24.2
44	64	fly	24.1
45	32	grease	23.4
46	73	moon	23.4
47	70	give	23.3

48	52	heart	23.2
49	36	feather	23.1
50	90	white	22.7
51	89	yellow	22.5
52	20	bird	21.8
53	38	head	21.7
54	79	earth	21.7
55	46	foot	21.6
56	91	black	21.6
57	42	mouth	21.5
58	88	green	21.1
59	60	sleep	21.0
60	7	what	20.7
61	26	root	20.5
62	45	claw	20.5
63	56	bite	20.5
64	83	ash	20.3
65	87	red	20.2
66	55	eat	20.0
67	33	egg	19.8
68	6	who	19.0
69	99	dry	18.9
70	37	hair	18.6
71	81	smoke	18.5
72	8	not	18.3
73	4	this	18.2
74	24	seed	18.2

75	16	woman	17.9
76	98	round	17.9
77	14	long	17.4
78	69	stand	17.1
79	97	good	16.9
80	17	man	16.7
81	94	cold	16.6
82	29	flesh	16.4
83	50	neck	16.0
84	71	say	16.0
85	84	burn	15.5
86	35	tail	14.9
87	78	sand	14.9
88	5	that	14.7
89	65	walk	14.4
90	68	sit	14.3
91	10	many	14.2
92	9	all	14.1
93	59	know	14.1
94	80	cloud	13.9
95	63	swim	13.6
96	49	belly	13.5
97	13	big	13.4
98	93	hot	11.6
99	67	lie	11.2
100	15	small	6.3

## References

- Albatineh, Ahmed N., Magdalena Niewiadowska-Bugaj, and Daniel Mihalko. 2006. On similarity indices and correction for chance agreement. *Journal of Classification* 23: 301-313.
- Bibiko, Hans-Jörg. 2005. The Interactive Reference Tool. CD-ROM accompanying Haspelmath et al. (2005).
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Vilupillai. 2008. Automated classification of the world's languages: A description of the method and preliminary results. To appear in *Sprachtypologie und Universalienforschung*.
- Cavalli-Sforza, Luigi L. and William S.-Y. Wang. 1986. Spatial distance and lexical replacement. *Language* 62: 38-55.
- Chen, Baoya. 1996. *Lun yuyanjiechu yu yuyanlianmeng*. Beijing: Yuwen chubanshe.
- Cysouw, Michael, Mihai Albu, and Andreas Dress. 2008. Analyzing feature consistency using similarity matrices. To appear in *Sprachtypologie und Universalienforschung*.
- Dolgopolsky, Aaron B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In: Shevoroshkin, Vitalij V. and Thomas L. Markey (eds.), *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*, pp. 27-50. Ann Arbor: Karoma.
- Dryer, Matthew S. 2005. Genealogical language list. In Haspelmath et al. (eds.), 584-643.

- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072-2075.
- Embleton, Sheila M. 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.
- Embleton, Sheila M. 1991. Mathematical methods of genetic classification. In: Lamb, Sydney M. and E. Douglas Mitchell (eds.), *Sprung from Some Common Source. Investigations into the Prehistory of Languages*, pp. 365-388. Stanford, California: Stanford University Press.
- Goebel, Hans. 1984. *Dialektometrische Studien. Anhand Italo-romanischer und Galloromanischer Sprachmaterialien aus AIS und ALF*. 3 vols. Tübingen: Niemeyer.
- Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue*. 15th Edition. SIL International. <[www.ethnologue.com](http://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 W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11.2: 395-423.
- Huson, Daniel H. (1998). SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14.10: 68–73.
- Kessler, Brett. 2001. *The Significance of Word Lists*. Stanford, Calif.: CSLI Publications.
- Kroeber, Alfred L. 1963. Yokuts dialect survey. *Anthropological Records* 11:177-251.

- Kruskal, Joseph B., Isidore Dyen, and Paul Black. 1973. Some results from the vocabulary method of reconstructing language trees. In: Isidore Dyen (ed.), *Lexicostatistics in Genetic Linguistics*, pp. 30-63. The Hague: Mouton.
- McMahon, April, and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.
- Moseley, Christopher and R. E. Asher (eds.). 1994. *Atlas of the World's Languages*. London: Routledge.
- Oswalt, Robert L. 1971. Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics* 13: 421-434.
- Saitou, Naruya & Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406-425.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335-357.
- Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16: 157-167.
- Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121-137.
- Thomas, David D. 1960. Basic vocabulary in some Mon-Khmer languages. *Anthropological Linguistics* 2.3: 7-11.
- Wang, Feng and Wang, William S.-Y. 2004. Basic words and language evolution. *Language and Linguistics* 5.3: 643-662.
- Wichmann, Søren and Eric W. Holman. N. d. Assessing temporal stability for linguistic typological features. Manuscript under review. Prepublication version:  
<http://email.eva.mpg.de/~wichmann/WichmannHolmanIniSubmit.pdf>.

Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24.2: 373-404.