

Glottochronology as a heuristic for genealogical language relationships

Søren Wichmann

Max Planck Institute for Evolutionary Anthropology & Leiden University

Eric W. Holman

University of California, Los Angeles

André Müller

University of Leipzig

Viveka Velupillai

Justus-Liebig-Universität Giessen

Johann-Mattis List

Heinrich Heine University Düsseldorf

Oleg Belyaev

Moscow State University

Matthias Urban

Max Planck Institute for Evolutionary Anthropology

Dik Bakker

University of Amsterdam & University of Lancaster

0. Introduction¹

This paper examines whether glottochronological time estimates based on lexical comparisons of a given set of languages are useful for gauging the possibility that these languages are genealogically related. The technique of glottochronology as originally developed by Morris Swadesh (Lees 1953, Swadesh 1955) is based on the idea that the number of cognate lexical items (pertaining to a fixed set of meanings) shared between languages reflects the time that has passed since the languages diverged from one another. In other words, the degree of lexical divergence between related languages

¹ We are grateful to Bernard Comrie, Pamela Brown, and Cecil H. Brown for comments on this paper..

should reflect the amount of elapsed time since the break-up of their shared ancestor into different dialects. In the procedure advocated by Swadesh, cognacy is judged impressionistically, i.e., the items in question are not necessarily linked by regular sound correspondences. In such a procedure some word pairs may be identified as cognate even if they are not. The possibility arises that enough words in unrelated languages are found to be similar that a separation date can be calculated within the range of what is typical for languages that *are* related, albeit distantly. Thus, the ability to calculate an apparently credible glottochronological date is no guarantee that the languages thus dated are, in fact, related.

It is possible also to operate with a version of glottochronology that is not based on cognate identification (Serva and Petroni 2008, Holman et al. 2009). Instead of basing the similarity measure on the number of shared cognates, the similarity between wordlists for different languages may be calculated as the average phonological resemblance holding for pairs of words with the same meaning. The similarity measure used in this paper, as in other recent work within the ‘Automated Similarity Judgment Program’ (ASJP)², derives from a version of the Levenshtein or ‘edit’ distance, which counts the number of substitutions, insertions, and deletions required to transform one of the two compared words into the other. This measure is further modified to take into account variable lengths of the words compared as well as accidental resemblance due to similarities in the phonological inventories of the compared languages (Bakker et al. 2009). It is calculated as follows. We compare each pair of words referring to the same concept on a list of 40 items representing the genealogically most stable items on the 100 item Swadesh list (where stability is defined and measured in Holman et al. 2008). A simplified transcription, described in Brown et al. (2008), is used. For each word pair an automated calculation of the so-called Levenshtein or edit distance (LD) is carried out. This corresponds to the number of substitutions, deletions or insertions which it takes to transform one of the two word forms into the other (the direction of transformation from word A to word B or the other way around does not matter since the LD will be the same in either case). The LD is normalized by dividing it by the length of the two longest

² For papers and other materials relating to the project see <http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm>.

strings compared, to obtain LDN ('LD normalized'). By this operation all LD's are turned into numbers ranging between 0 and 1. A further normalization is applied whereby the average LDN of words referring to the same concepts is divided by the average LDN of words not referring to the same concepts, leading to what we call LDND ('LD normalized divided'). This operation is intended to neutralize the effect of accidental phonological similarities among languages that are not related and thus to enhance the mutual distinctiveness of unrelated languages. Wichmann et al. (2010b) show that the second normalization works as intended: classifications based on LDND tend to more accurately distinguish language families than classifications based on LDN, while the accuracy of within-family classifications are not appreciably different when either LDN or LDND are used. The distance measure used within ASJP, then, is LDND. The corresponding similarity measure, s , is obtained by subtracting LDND from 1. While this similarity measure is not based on an identification of cognacy, it is clearly sensitive to the presence of cognates since related words will tend to exhibit a greater phonological similarity than will unrelated ones. The degree of similarity among unrelated words also contributes to the similarity measure, but only causes small fluctuations in the overall measurement when the languages have a substantial number of cognates. For a very small number of cognates, however, non-cognates may rival or may even surpass cognates with respect to the amount of input they provide into the overall observed similarity. And for unrelated languages that share no cognates at all, this 'noise' is the only contributor to the similarity measure. Random fluctuations in the degree to which unrelated languages are similar in their basic vocabulary by sheer accident may cause some pairs of unrelated languages to look as similar as distantly related languages. Thus, time depths based on similarities among unrelated languages are expected, by mere accident, to sometimes look similar to dates calculated for pairs of languages that have been shown to be related. Also, unrecognized loanwords may contribute to increasing similarities that are not due to common ancestry.

A radical positivist may believe that if something exists it can be measured and vice versa. Glottochronological ages, however, although based on the quite palpable phenomenon of words used on an everyday basis, negate this view since they can sometimes be measured even if they do not exist. This is true of both standard

glottochronology, where the impressionistic approach to the judgment of cognacy may yield incorrect identifications, and of the ASJP method, where the measurement of phonological similarities in basic lexical items is sensitive to random similarities. The approaches are similar, but the latter method has the advantage that the magnitude of the problem can more easily be investigated since it is eminently possible to compare many unrelated languages given that the comparisons are carried out by a computer systematically and fast. In order to determine the frequency with which pseudo-cognates appear in comparisons of unrelated languages a human would need to dedicate months, if not years, to the comparison of wordlists for unrelated languages. Thus, while such exercises have been made (cf. the next section), they have never been carried out for more than a handful of languages. Sporadic investigations are not necessarily very telling, however, since they cannot ultimately determine the probability for a certain amount of apparent cognacy to occur between randomly chosen pairs of unrelated languages, which is what is necessary in order to estimate the utility of glottochronology as a heuristic for establishing distant genealogical language relationships.

We stress at the outset that we do not here address the utility of glottochronology *per se*, only the utility of the glottochronology for the purpose of identifying and establishing deep genealogical relations. Even if glottochronology was not devised with this purpose in mind it makes sense to investigate whether it could be used in that way.

The paper presents the results of measuring lexical similarities for pairs of language families that are normally not assumed to be related and which are also very unlikely to ever be shown to be related. We offer this systematic analogue to the anecdotal examples mentioned in the next section in order to determine what such similarities can tell us about language relationships. We focus on the following three questions: (1) Is there an empirical upper limit to similarities among unrelated languages? (2) How are such similarities for unrelated languages distributed statistically? (3) Which factors contribute to accidental similarities among languages?

Answering these questions will help us to isolate problems arising from using comparisons of basic vocabulary for investigating possible distant relationships and will shed some light on the degree to which very old dates for language families are reliable. We do not intend to show that the comparison of basic vocabulary for the purpose of

identifying distant relationships is futile, but we would like to acquire a better idea of potential pitfalls when the procedure applied is simply one of calculating a date.

The more constructive project of improving methods for vocabulary comparison for the purpose of investigating distant relationships will be addressed in later work.

1. Previous research

A number of studies in the glottochronological literature pertain to the questions of this paper. An early paper is that of Tovar et al. (1961), whose results are relevant even if the authors were apparently in search of real relationships among language families. They compared Basque, Chukchi, Georgian, two North Caucasian languages (Circassian and Avar), and five Afro-Asiatic languages (Rif Berber, Sus Berber, Ancient Egyptian, Coptic, Arabic), and found an average of 5.0% apparent cognates in the Swadesh 100-item list between languages in different families. In a somewhat larger study, Bender (1969) used a modified 100-item list to compare 21 languages, each in a different family. With a strict criterion for cognacy, he found an average of only 0.4% cognates; but with a weakened criterion, of which he said “I believe that it approximates the actual method used in many practical situations,” he found an average of 3.5% cognates. In a small replication of Bender’s study, Campbell (1973) compared Finnish with Quechua and Cakchiquel (a Mayan language). Although he used the same list and criteria as Bender, Campbell reported substantially higher average levels of cognacy: 1.2% and 16.5% with the strict and weakened criteria, respectively. Bender (1976) revisited the question with a comparison of 24 Nilo-Saharan languages and one language from each of three other families. With a criterion described as “the ‘look-alike’ principle modified by the results of a painstaking search for regular phonological correspondences,” he found an average of 3.4% cognates between languages in different families and only 3.8% cognates between languages in different subgroups at the highest level within Nilo-Saharan.

In sum, these studies have produced considerable variation depending on the criteria applied and the linguists applying the criteria.

2. The database

The ASJP database (Wichmann et al. 2010a) consists of wordlists representing the 40 most stable items on the 100-item Swadesh list. Holman et al. (2008) found that, starting from the five most stable items and then gradually increasing the number of items used, there is a steady improvement in how accurate lexicostatistic classifications become compared to the classifications of experts, but that this increase in accuracy eventually wanes such that when 40 or more items are used there is no longer any increase to be observed. This is the reason for using a shorter version of the 100-item Swadesh list. The version of our database on which this paper is based consists of over 3600 languages and dialects, representing close to half of the world's linguistic diversity. It is a convenience sample in the sense that we began the data collection with the idea of sampling all of the world's recorded languages and presently simply find ourselves half way towards this goal. There is, nevertheless, a quite even genealogical spread since we have been focusing on including as many families as possible. The greater part of the languages that are not yet included in the database pertain to large language families such as Niger-Congo and Austronesian.

3. Comparison between some glottochronological dates and ASJP similarities

Since this paper is not only concerned with similarities measured by means of the Levenshtein approach but also claims that the conclusions extend to glottochronology, it is necessary to briefly substantiate the claim that our similarity measures correlate with results from glottochronology. For this purpose we will compare similarities for families that are included in both Swadesh (1959) and the ASJP database. Using the single source of Swadesh (1959) limits the number of possible comparisons, but has the advantage that it can be assumed that the method and the way it is practiced are consistent when only a single paper by one author is the source.

Table 1 presents comparisons between age estimates from Swadesh (1959) and s , which is defined as $100\% - \text{LDND}$, where LDND is the twice-modified Levenshtein distance described in the Introduction. These comparisons yield an overall Spearman rank correlation of $-.59$. We consider this a high correlation given that both methods and data are different (with different wordlists being used and in most cases probably also different samples of languages from the different families).

Thus, statistical findings involving similarities measured by means of the Levenshtein approach should largely be valid also for glottochronological dates. Like cognate percentages in the glottochronological approach our similarity measures can be converted into absolute ages. For the purpose of this paper it is not necessary to present actual dates based on the s values, however. Since this would furthermore require a lengthy discussion of calibration issues, i.e., issues of historical events that are most appropriate for anchoring dates, we restrict ourselves to merely presenting s values.

Table 1. A comparison of some glottochronological age estimates with ASJP similarities

Family	Age estimate (BP) from Swadesh (1959)	s (%) from Holman et al. (2009)
Algonquian	3500	10.55
Boran	1800	16.45
Caddoan	3500	4.08
Chinantecan	1500	27.97
Choco	700	23.80
Eskimo-Aleut	3700	3.31
Ge-Kaingang	4200	4.18
Guaicuruan	4100	13.28
Iroquoian	3400	3.16
Mayan	3800	26.02
Mixtecan	4900	5.10
Muskogean	2800	29.69
Otopamean	5500	9.62
Popolocan	2400	20.71
Salishan	6500	6.85
Subtiaba-Tlapanecan	800	36.48
Totonacan	2600	38.04
Wakashan	2900	16.76
Witoto	5200	13.05
Zaparoan	5500	11.40
Zapotecan	2400	11.74

4. The sample

For the present study we are interested in comparing wordlists for languages generally assumed not to be phylogenetically related. Since it is impossible to prove that two languages are unrelated, we are not using ‘unrelated’ in a definitive sense but in the following specific sense: (1) the languages have not been proven to be related to the satisfaction of the linguistics community broadly speaking, and (2) it is unlikely that it will ever be possible to provide such proof. As an easy way to assure that (1) and (2) nearly always hold, we compare pairs of language families³ where each pertains to respectively the New and the Old World. This set of language family pairs presently only contains one exception to (1) and (2), which is Na-Dene (defined as Tlingit-Eyak-Athapaskan) and Yeniseian. This pair of families is regarded as genealogically related by E. Vajda (most recently, Vajda, forthcoming), a hypothesis that is accepted or regarded as highly probable by many specialists. Since we are presently undertaking a larger statistical exercise we nevertheless include this pair in our comparisons. As we shall see, this will not affect any of our conclusions.

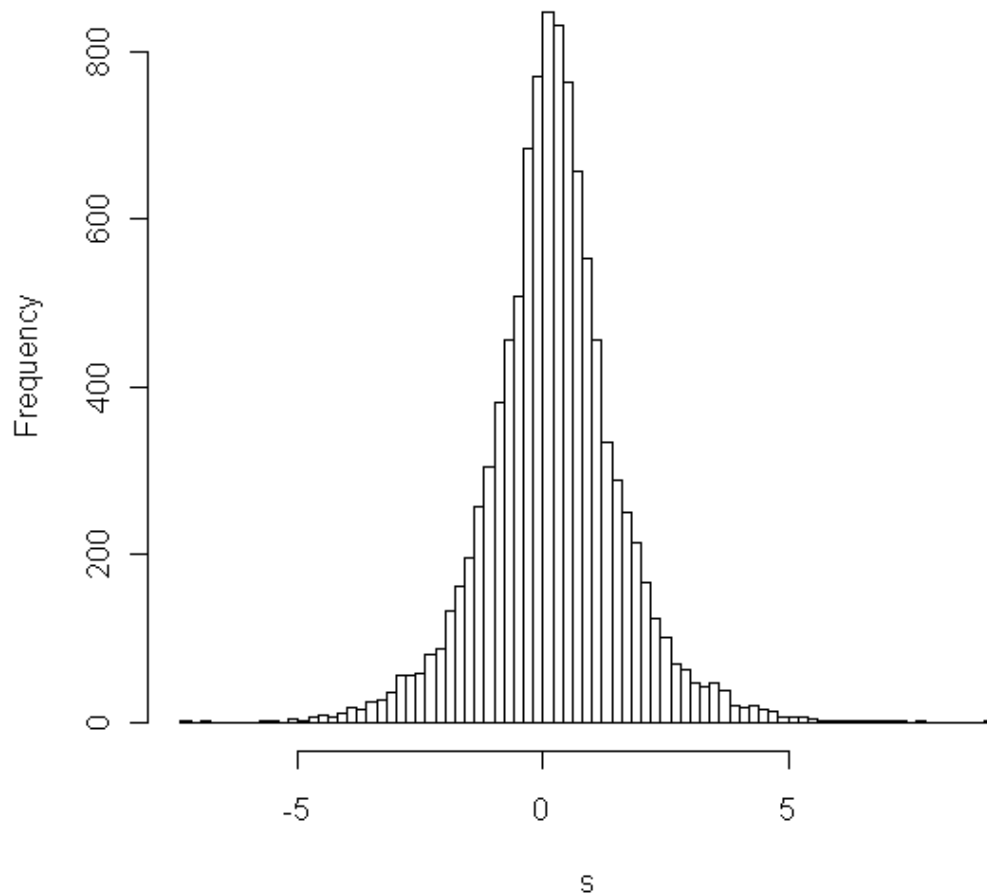
5. Results

Fig. 1 is a histogram of s for 10,356 pairs of unrelated language families, each from one of the two hemispheres. To calculate the s score between two families the similarities are averaged across all language pairs where the members belong to each of the two families. The entire distribution has been divided into 100 bins for the purpose of the plot. The plot shows a normal distribution around a value close to zero. The reason why s can have a

³ In this paper ‘families’ are defined as in Haspelmath et al. (2005), while Holman et al. (2009) use the definitions of Lewis (2009). The difference in choices is entirely due to technical reasons, and choosing one as opposed to the other classification would not have any effect on the observations to be presented in the following.

negative value is that the second modification of the Levenshtein distances may result in a distance (d) greater than 100%. This situation arises when similarities are greater among words not referring to the same concept than among words sharing the same meaning. It is because s is defined as $100\% - d$ that the value of s may be negative.

Figure 1. Histogram showing the frequencies of similarities for unrelated language families, sorted into 100 bins.



It is interesting to note that the average of s is not exactly zero, but more precisely 0.22% (the median is exactly 0.2%). The reason why the value is positive must mainly be due to the tendency for words for certain concepts to contain the same phonemes because of sound symbolism (Wichmann et al. 2010c). In addition, a fraction of the positive value

would be due to a few widespread loanwords, such as Spanish *hueso* ‘bone’, which also occurs in some varieties of Nahuatl, for instance. We investigated whether the average similarity score was affected by removing such loanwords from consideration, however, and still found an average s which rounded up to 0.22%. For the world’s language families as defined in Lewis (2009) only Na-Dene (Tlingit-Eyak-Athapaskan plus Haida) has a similarity score lower than 0.22%.

The right part of the curve shows that there are some outliers among the unrelated families having relatively high similarities. One pair has an s score in the vicinity of Indo-Iranian; 1% of the pairs have a score higher than languages families such as Sino-Tibetan, Caddoan or subgroups such as Ge-Kaingang (Macro-Ge), Chadic and Cushitic (both Afro-Asiatic); 2% have a score higher than families such as Australian, Lakes Plain, Algic, and Iroquoian. As indicated in the previous paragraph, 50% or more have similarity scores that are lower than any of the language families as defined in Lewis (2009), with the exception of Na-Dene (which, with Haida included, is highly controversial).

These results tell us that when interpreting glottochronological ages we should take care not to interpret an age even as low as that of, say, one of the older subgroups of Indo-European, as absolute proof that the languages in question are actually related. However, only a very small minority of unrelated family pairs (1%) have similarities corresponding to ages lower than those of uncontroversial language families (or higher-order subgroups of families) such as Sino-Tibetan, etc. For similarity scores lower than the $s = 0.22\%$ mean for unrelated families one can be certain that a given family is not normally assumed to be related.

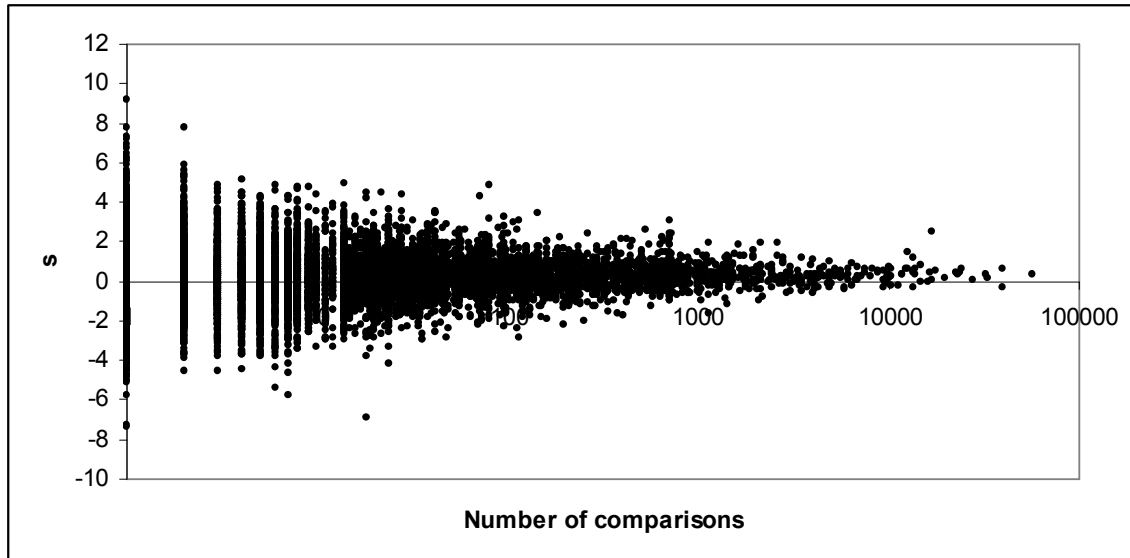
6. Discussion

We may interpret the results presented in the previous section as showing that an s value higher than that of well-established families or subgroups of older families would support a hypothesis of relatedness since such a similarity is almost never found for a pair of unrelated language groups (at least not for a pair containing families pertaining to

different world hemispheres). Lower s values become increasingly less reliable as indicators of language relatedness and, for a similarity score lower than 0.22%, we enter the typical range for unrelated language pairs, and also the range where we can be certain that languages are not broadly accepted as being related, at least given the current state of the art of the comparative method. Time depths based on similarities of this order or lower are meaningless in the sense that the typical range of effects of random fluctuations in accidental similarities has been reached.

The values of s for unrelated families were found by averaging s for language pairs whose members belong to different families. Thus far we have not taken into account the number of language pairs, N , entering into the comparison. Random fluctuations in s due to accidental similarities are expected to be greater for single language pairs than for multiple pairs. Across many language pairs, s is expected to approach the average value of $s = 0.22$. In Figure 2 we show the relationship between s and N . Each dot in the figure refers to a pair of unrelated families. There is, indeed, a clear tendency for fluctuations in s to decrease as N increases. Thus, the probability of encountering extreme values of s (negative as well as positive) is greatest when each of the families compared consists of single members.

Figure 2. The relationship between s for unrelated families and the number of pairwise comparisons, N , on which s is based (to facilitate the inspection of the distribution of small values of N the N -axis has been transformed logarithmically).



Although both extremely positive and extremely negative values of s exhibit the same behavior we are more interested in the cases where highly positive values are reached because these are cases where a scholar may be led to posit a deep genealogical connection. Thus, it may be of interest to look at the list of family pairs exhibiting the most extreme positive values of s . Table 2 shows the 25 highest-scoring family pairs. In all cases we are dealing with either isolates or small families from which it was possible to draw only one or two comparisons.⁴

Table 2. Similarities (s) for languages families (N = number of comparisons feeding into the average)

Family 1	Family 2	N	s
Korean	Warao	1	9.19
West Bougainville	Chapacura-Wanham	2	7.78
Lavukaleve	Tonkawa	1	7.75
Shom Peng	Mura	1	7.28
Burmeso	Taushiro	1	7.18
Hadza	Xincan	1	6.90
Bilua	Timucua	1	6.71
Korean	Chitimacha	1	6.48
Usku	Kunza	1	6.27

⁴ The pair consisting of Na-Dene and Yeniseian does not appear here, and in fact these two families are less similar than are unrelated language families on average, with $s = -0.18$.

Usku	Puinave	1	6.20
Doso	Xincan	1	6.10
Wasi	Urarina	1	5.92
Oksapmin	Katukinan	2	5.91
Mombum	Karok	2	5.64
Korean	Trumai	1	5.57
Burushaski	Washo	1	5.49
Morwap	Mura	1	5.45
Kibiri	Lencan	2	5.44
Bilua	Aymaran	1	5.40
Mombum	Natchez	2	5.36
Korean	Chimúan	1	5.31
Savosavo	Lencan	2	5.29
Karkar-Yuri	Yuchi	2	5.26
Burushaski	Takelma	1	5.26
Burmeso	Waorani	1	5.25

These results strongly indicate that if comparisons of basic vocabulary are to be used for identifying possible deep genealogical language relationships – be it through cognate counts or some measure of phonological distance – it is important to take into account the number of comparisons involved. Such a method may be useful for larger families where consistently high similarity scores across many comparisons of single pairs point to a possible genealogical link, but when only one or two comparisons are involved there is a great danger of picking up similarities that are simply accidental.

7. Conclusions

In this paper we have examined whether glottochronology is useful as a heuristic towards the establishment of genealogical relations among languages. Towards this goal we first argued that an automated parallel to glottochronology is necessary for a systematic investigation of the issue, and we subsequently showed that what holds for this method in a broader statistical perspective is also expected to hold for glottochronology.

We then sampled more than 10,000 pairs of language families that may safely be assumed not to be related, given that members of each pair are spoken respectively in the

Old and New World. We found that the measured similarities have a normal distribution around a positive value of 0.22%, which was attributed to sound symbolism in Wichmann et al. (2010c).

If loanwords are excluded, there is a reasonable certainty that ages lower than what is typically found for well-established families such as Sino-Tibetan, or highest-order subgroups of an old family such as Afro-Asiatic, are real (within a certain margin of error) and, accordingly, that they are due to actual relatedness. For greater age estimates, glottochronology becomes increasingly less reliable as a heuristic for genealogical language relationship.

If measures of similarity in basic vocabulary, whether based on cognacy judgments or the Levenshtein approach, are to be used for investigating the possibility of genealogical relatedness at great time depths, it is vital to take into account the number of language comparisons involved, i.e., the sample size. Moreover, a small, but non-negligible, effect of sound symbolism must be reckoned with (Wichmann et al. 2010c). In the present paper we have to a large extent controlled for the influence of lexical diffusion by looking only at language family pairs whose members belong to different world hemispheres. In an investigation involving languages from the same area or macro-area their geographical distance should also be taken into account, since languages tend to be more similar the closer their geographical proximity. Thus, there are pitfalls associated with the straightforward use of a certain glottochronological date as evidence for a genealogical link. But if (1) sample size, (2) sound symbolism, and (3) expected effects of diffusion are taken into account, measures of lexical similarity in basic vocabulary may still constitute a potential heuristic for evaluating possibilities of distant genealogical language relations. We intend to further substantiate this last observation in future methodological and empirical research.

References

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W.

- Holman. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology* 13: 167-179.
- Bender, Marvin L. 1969. Chance CVC correspondences in unrelated languages. *Language* 45: 519-531.
- Bender, M. Lionel. 1976. Nilo-Saharan overview. In *The Non-Semitic Languages of Ethiopia*, M. Lionel Bender (ed.), 439-483. East Lansing: African Studies Center, Michigan State University.
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 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.
- Campbell, Lyle. 1973. Distant genetic relationship and the Maya-Chipaya hypothesis. *Anthropological Linguistics* 15: 113-135.
- Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Pamela Brown, and Dik Bakker. 2008. Explorations in automated language comparison. *Folia Linguistica* 42: 331-354.
- Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2009. Automated glottochronology: Dating the World's language families. Paper presented at the "Tutorial on Glotto- and Grammachronology", XIth International conference on "Cognitive Modelling in Linguistics" (CML-2009), Constanța, Romania, Sept. 11, 2009.
- Lees, Robert B. 1953. The basis of glottochronology. *Language* 29:113-127.
- Lewis, M. Paul (ed.). 2009. *Ethnologue*. 16th Edition. Dallas: SIL International. <www.ethnologue.com>.
- Serva, Maurizio and Filippo Petroni, Indo-European languages tree by Levenshtein distance. 2008. *Europhysics Letters* 81, paper 68005 (March 2008). [Online journal: <http://www.iop.org/EJ/journal/EPL>].

- Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121-137.
- Swadesh, Morris. 1959. Linguistics as an instrument of prehistory. *Southwestern Journal of Anthropology* 15: 20-35.
- Tovar, Antonio, K. Bouda, R. Lafon, L. Michelena, W. Vycichl, and M. Swadesh. 1961. El método lexico-estadístico y su aplicación a las relaciones del vascuense. *Boletín de la Real Sociedad Vascongada de los Amigos del País* 17: 249-281.
- Vajda, Edward. Forthcoming. A Siberian link with the Na-Dene. *Archeological Papers of the University of Alaska*, New Series, Vol. 6, 75-156.
- Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Matthias Urban, Sebastian Sauppe, Oleg Belyaev, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, and Helen Geyer. 2010a. The ASJP Database (version 12). <<http://email.eva.mpg.de/~wichmann/languages.htm>>
- Wichmann, Søren and Holman, Eric W. and Cecil H. Brown. 2010b. Evaluating linguistic distance measures. *Physica A*. In press; advance online publication: <http://dx.doi.org/10.1016/j.physa.2010.05.011>.
- Wichmann, Søren, Holman, Eric W., and Cecil H. Brown. 2010c. Sound symbolism in basic vocabulary. *Entropy* 12.4: 844-858.