

## Evaluating linguistic distance measures

Søren Wichmann

Max Planck Institute for Evolutionary Anthropology & Leiden University

Eric W. Holman

University of California, Los Angeles

Dik Bakker

University of Amsterdam & University of Lancaster

Cecil H. Brown

Northern Illinois University

### Abstract

In [1], F. Petroni and M. Serva discuss the use of Levenshtein distances (LD) between words referring to the same concepts as a tool for establishing overall distances among languages which can then subsequently be used to derive phylogenies. The authors modify the raw LD by dividing the LD by the length of the longer of two words compared, to produce what could be called LDN (normalized LD). Other scholars [8-9] have used a further modification, where they divide the LDN by the average LDN among words not referring to the same concept. This produces what could be called LDND. The authors of [1] question whether LDND is a more adequate measure of distance than LDN. Here we show empirically that LDND is the better measure in the situation where the languages compared have not already been shown, by other, more traditional methods of comparative linguistics, to be related. If automated language classification is to be used as a tool independent of traditional methods the further modification is necessary.

Keywords: historical linguistics, phylogenetics, Levenshtein distance, classification, ASJP

## 0. Introduction

While originally introduced by dialectologists [2-4], automated distance measures have recently become popular among scholars classifying forms of speech that are usually not considered dialects of one another, but rather completely distinct languages [5-11]. This has been achieved through the automated pairwise comparison of words referring to the same concept and pertaining to a standardized list of basic, human concepts, such as the so-called Swadesh list [12] or similar lists, representing shorter or longer versions of the Swadesh list or overlapping with it. The comparisons have proceeded in different ways. In [6-7] criteria were developed to determine whether or not words were likely to be cognate (i.e., having a shared origin) and such ‘matching rules’ were implemented in a computer algorithm. The research team behind this approach, which is called ASJP (‘Automated Similarity Judgment Program’), soon favored another approach, however, where a version of the Levenshtein distance (LD) [13] is used to determine phonological distances among words [8-9]. At about the same time other scholars were also using either a version of the LD [5,10] or another string comparison procedure [11] to measure distances among languages which could then be submitted to phylogenetic algorithms for the purpose of generating trees showing relationships among languages.

In this paper we discuss two different versions of the LD. One version consists in a single modification of the original (‘raw’) LD, and the other entails the latter modification and an additional one. In the following, we refer to the once-modified version as LDN (‘Levenshtein Distance Normalized’) and to the twice-modified version as LDND (‘Levenshtein Distance Normalized Divided’). The paper is a reply to [1], where the authors also discuss the two versions, claiming that LDN may be a more adequate measure of linguistic distance than LDND. Our reply seeks to show empirically that LDND is, in fact, the more adequate measure in the default case where languages have not already undergone a prior classification through other methods.

Let us briefly summarize the difference between the two distance measures (for a more extensive and formalized description see [1]). A Levenshtein distance (LD) is defined as the minimum total number of additions, deletions, and substitutions of symbols necessary to transform one word into the other. In [5], a single modification of the LD is used. In this paper, the LD for each pairwise word comparison is normalized by

dividing it by the length of the longer of the two strings compared. Through normalization, the maximal LD for any word comparison is unity and longer words do not influence the overall average of the pairwise comparisons more than shorter words: all comparisons are weighted equally regardless of the lengths of the words compared. This modification, resulting in LDN, is not only meaningful, but indispensable. It is therefore used within ASJP [8-9] as well. In these papers, however, a further modification is introduced: the average LDN for all comparisons of words referring to the same concept is normalized by dividing it by the average LDN (called  $\Gamma$  in [1]) for pairwise comparisons of words that do not refer to the same concept, resulting in LDND.

The motivation for the LDND was not made totally explicit in [8-9] and the empirical proof of its utility was not provided either, perhaps motivating the criticism in [1]. In any case, it is argued in [1] that since LDN and  $\Gamma$  correlate there could be information in the latter which, if lost in the normalization procedure, may render classifications less reliable. Here we take the opportunity to elaborate on the motivation for the second modification and to supply empirical data showing that it improves results when distance measures are compared with the distances among languages contained in classifications provided by experts.

### 1. The motivation for LDND

LDND was designed to take into account the fact that lexical similarity may be influenced by chance resemblances, such as a coincidental overlap in the phoneme inventories or shared phonotactic preferences for the two languages compared. The second modification neutralizes similarities in overall sound structures and serves to avoid the situation where languages are grouped together just because they happen to have such resemblances even if they are unrelated. The premise is that one should not make any prior assumptions about whether the languages compared are related to each other. In fact, a major motivation for automated language classification is precisely that no such assumptions need to be made, such that the enterprise is independent of other methods. The necessity for an entirely objective approach to language classification is shown by the radical differences among existing classifications of the world's languages. As an example of such differences, three recent catalogues [14-16] present exhaustive,

worldwide lists of language groups that are not demonstrably related to each other in the opinion of the authors. The number of such groups is 12 in [14], 241 in [15], and 387 in [16].

In [1], correlations between LDN and  $\Gamma$  are made independently for sets of Austronesian and Indo-European languages. This approach is a mixed one: first languages are sorted into two groups, Austronesian and Indo-European, and only then is the automated approach applied. The first procedure rests entirely on the knowledge accumulated within traditional historical linguistics about the relationship between respectively Austronesian and Indo-European languages. An alternative approach to language classification, however, should have its own, independent means of providing higher-order partitions such as Austronesian and Indo-European. It was assumed by Wichmann, Holman, and collaborators that the second modification of LD would help to prevent unrelated languages from appearing to be related in automated classification. The effects on language classifications of using LDND as opposed to LDN were investigated using the two world-wide languages classifications of [15] and [17] as yardsticks, and it was found that LDND was advantageous. Since these investigations have not been published we report them here, using the most recent version of our linguistic database, which is described in [18]. Just such a comparison of LDN and LDND is recommended towards the end of [1]: “Nevertheless, this conclusion should be empirically supported by comparing the distance matrices corresponding to the two definitions with matrices produced by experts.”

Section 2 describes the database used to generate the two different distance measures that we correlate with expert classifications, and Section 3 presents the results. While sections 2-3 would suffice to close the discussion with [1], we want to go beyond the narrow issue of evaluating linguistic distance measures and inquire into the possible reasons why our procedure for automated language classification shows a differential performance, apparently working better for some families than for others. We devote the rest of the paper to this larger issue, which almost certainly will be at the core of research relating to automated procedures for language classification in years to come.

## 2. Data for evaluating the performances of linguistic distances measures

The ASJP database contains 4169 word lists from languages that are either currently spoken or have gone extinct only within the last three centuries. In addition, it includes a number of pidgins, creoles, mixed languages, constructed languages, earlier stages of currently spoken languages, languages having gone extinct before 1700 CE, and proto-languages. These additional languages, however, are not used in the present study. Among the 4169 word lists, 3065 are considered to represent distinct languages in [15] in the sense of having a unique code in that classification. Thus, the database includes close to one half of the world's spoken languages. Each word list normally contains a minimum of 70% of the items on a 40-item subset of the Swadesh list which was defined in [7].

The languages represented by the 4169 word lists are classified in 211 families as circumscribed in [17], where a family is defined as the most inclusive group proved to the satisfaction of most linguists to have descended from a common ancestral language. The families in [17] are circumscribed slightly more conservatively than those in [15]. The present tests are focused on the 49 families that are represented by at least 10 word lists in our database; these families are represented by a total of 3730 of the 4169 word lists contained in the database.

### 3. Results for the correlation of LDN and LDND with expert classifications

As mentioned in Section 1 above, the motivation for LDND is to keep languages that happen to have similar sound structures from being classified together through automation when distance measures are employed in the construction of phylogenetic trees. We hypothesize (a) that LDND has this preventive effect in the default procedure for language comparison where no prior assumptions are made concerning relatedness of the languages compared. We furthermore hypothesize (b) that the choice of LDN or LDND matters little in the special case where the languages compared are already known to be related, such as when the authors of [1] use prior results by historical linguists to define Austronesian vs. Indo-European groups of languages. The expectation in this case is that the  $\Gamma$ -normalization will only introduce minor fluctuations in the distance measures – fluctuations that are too small to appreciably influence the utility of the distance measure for language classification.

The claimed advantage of using LDND is that it is better suited for distinguishing related languages (i.e., languages in the same family) from unrelated ones (i.e., languages belonging to different families). We test this specific assumption in the following manner. For a given family, let  $d(\text{in})$  be the mean of the distances between languages within the family, let  $d(\text{out})$  be the mean of the distances from each language in the family to each language outside the family, and let  $\text{SD}(\text{out})$  be the standard deviation of the latter distances. Then the distinctiveness of the family (abbreviated Dist) is defined as  $[d(\text{out}) - d(\text{in})]/\text{SD}(\text{out})$ , which is the difference between related and unrelated languages relative to the random variability of unrelated languages.

The two columns under the heading Dist in Table 1 below give the distinctiveness of each family, based respectively on distances measured by LDND and LDN. All but two of the 49 families are more distinctive for distances measured by LDND than by LDN. This is significant by a sign test with families as units of analysis. The average distinctiveness of the 49 families, which appears on the last line of the table, is 28% greater for LDND than for LDN. In summary, LDND is substantially more effective than LDN in distinguishing related from unrelated languages, as predicted by hypothesis (a).

Table 1. Distinctiveness (Dist) and correlations with [17] (RW) and [15] (GE) for LDND and LDN for each family.  $N$  is the number languages and dialects in the database,  $n$  is the total number of languages (as defined by [15]) in the family (as defined by [17]).

AREA			Dist		RW		GE		
	Family	$N$	$n$	LDND	LDN	LDND	LDN	LDND	LDN
AFRICA									
	Khoisan	16	26	4.705	5.401	.7047	.6935	.6899	.7213
	Niger-Congo	558	1532	1.530	1.111	.4335	.4209	.4021	.3540
	Kadugli	11	6	19.144	14.305			.8039	.8039
	Nilo-Saharan	113	199	1.821	1.365	.6209	.6114	.5830	.5572
	Afro-Asiatic	227	374	1.528	0.997	.6444	.6237	.7375	.7443
EURASIA									
	Indo-European	210	439	5.927	4.385	.7529	.7575	.8251	.8004
	Uralic	23	37	6.948	5.534	.5057	.5379	.9742	.9759
	Altaic	75	66	10.847	8.763	.8711	.8722	.9240	.9228
	Nakh-Daghestanian	32	29	7.187	5.632	.6621	.6502	.9397	.9323
	Dravidian	21	85	11.699	8.579	.3612	.3673	.5246	.5117

Sino-Tibetan	141	449	3.077	2.515	.5922	.5835	.6942	.6697
Hmong-Mien	14	38	8.289	6.995			.9333	.9200
Tai-Kadai	56	92	11.948	9.438	.6840	.6959	.7725	.7789
Austro-Asiatic	52	169	5.407	4.514	.5942	.5824	.6459	.6842
Great Andamanese	10	10	12.819	10.349			.1974	.1316
PACIFIC								
Austronesian	833	1257	4.812	3.772	.1589	.1313	.2535	.2362
Border	16	15	6.918	6.125			.7763	.7685
Bosavi	15	8	8.918	7.097			.9369	.9369
Eleman	10	7	9.014	6.982	.9304	.9309	.9574	.9574
Kiwaian	15	7	17.032	13.166			.9441	.9497
Lakes Plain	26	20	4.890	4.207	.4219	.4613	.7100	.7443
Lower Sepik-Ramu	20	32	2.826	2.738	.6054	.6208	.9282	.8873
Marind	32	6	4.573	3.748	.6345	.6507	.9370	.9122
Morehead U. Maro	17	17	3.662	3.105			.6930	.7232
Sepik	26	56	4.665	3.402	.6514	.6685	.8618	.9166
Sko	14	7	4.864	3.640	.8193	.8253	.7450	.6879
Torricelli	31	56	3.924	3.063	.6037	.6090	.8909	.8413
Trans-New Guinea	293	372	0.864	0.829	.5065	.4298	.6748	.6543
West Papuan	34	23	3.547	3.031	.6093	.6202	.7432	.7719
Western Fly	39	4	11.813	8.992			1.0000	.9995
Australian	186	264	1.805	2.234	.3020	.3345	.4463	.4863
AMERICA								
Na-Dene	22	44	11.872	10.118	.6387	.6299	.7728	.7598
Algic	28	44	6.102	5.044	.3969	.4522	.5459	.5813
Uto-Aztecan	81	61	11.584	9.292	.9189	.9174	.7566	.7297
Salishan	12	26	11.139	8.995	.6521	.6601	.8903	.9097
Penutian	21	33	3.315	2.479	.8760	.8683	.8156	.8454
Hokan	24	23	5.424	4.220	.8539	.8531	.5320	.5389
Oto-Manguean	60	177	3.311	2.676	.8507	.8481	.9906	.9877
Totonacan	14	12	25.769	18.151			1.0000	1.0000
Mixe-Zoque	14	17	19.569	16.028			.9803	.9803
Mayan	75	69	17.651	14.553			.8236	.8247
Chibchan	20	21	2.862	2.531	.5796	.5772	.6935	.7677
Tucanoan	19	25	15.807	12.542			.7565	.7447
Panoan	18	28	12.345	9.191			.3802	.3419
Quechuan	18	46	31.385	18.062			.4565	.4565
Arawakan	48	59	3.566	2.927			.4934	.4939
Cariban	19	31	9.761	7.859			.2879	.2867
Tupian	47	76	9.533	6.880	.7594	.7864	.9185	.9176
Macro-Ge	24	29	2.857	2.254	.6887	.6882	.6797	.7294
AVERAGE			8.385	6.526	.6329	.6351	.7331	.7322

We now test hypothesis (b), which says that for the internal classification of groups of languages which have previously been shown to be related (at least to the satisfaction of many linguists), using LDN or LDND matters little. This hypothesis is tested separately relative to the classifications in [17] and [15].

The classification in [17] has three taxonomic ranks. The lowest rank is the individual language. The highest rank is the family. The middle rank is the genus, defined as the most inclusive group thought by linguists to have descended from a common ancestor that was still a single language some 3500 to 4000 years ago. Since these ranks are intended to be comparable across the entire classification, they can be used to define a distance matrix, but the distances between languages within a family take only two possible values: 1 if the languages are in the same genus, and 2 if they are in different genera. The distance matrix for each family can be compared to the matrices of LDN and LDND by means of a point-biserial correlation, which is the special case of Pearson's  $r$  where one of the variables takes only two values. Pearson's  $r$  is also used in [1] and is formally defined in their Eq. (6).

The correlations appear in Table 1 in the two columns under the heading RW (mnemonic for 'Pearson's  $r$  with *The World Atlas of Language Structures*'). Whenever [17] considers a family to be a single genus, the absence of subclassification precludes calculation of RW and the corresponding cells are blank. The average correlations in the 33 remaining families are almost the same, with RW higher by .0022 for LDN than for LDND. There are 17 individual cases where RW has a higher value for LDN and 16 where LDND 'wins', obviously not significant by a sign test.

The classification in [15] is more complicated. The highest and lowest levels are again the family and the language. In between, families may be divided into subgroups, which may be further subdivided, and so on. Levels of subdivision are not intended to be comparable between or even within families, and there are different numbers of levels in different families and in different subgroups of the same family. This structure does not define a complete distance matrix; instead, it defines a partial order on the distances.

For instance, the Indo-European family is divided at the highest level into nine branches. One of these branches, Albanian, is subdivided at the second level into two more branches. All the Albanian languages are so closely related that they are considered

in [14] to be dialects of a single language; thus, the branches of Albanian probably diverged from each other within the last 1000 years or so. Another branch of Indo-European is Indo-Iranian, which is also subdivided at the second level into two branches. These branches are separate genera in [17] and thus probably diverged from each other at least 3500 years ago. The difference in time depth between Albanian and Indo-Iranian is not reflected in the classification, which is not designed for comparing distances among languages in different branches. The classification does predict, however, that the Albanian languages are closer to each other than they are to the Indo-Iranian languages, and likewise that the Indo-Iranian languages are closer to each other than they are to the Albanian languages. In general, languages within a taxonomic group are predicted to be closer to each other than they are to languages outside the group.

This partial order can be compared to the matrices of LDN and LDND by means of Goodman-Kruskal gamma [19]. Given any two sets of distances, gamma is defined as  $(C-D)/(C+D)$ , where C is the number of concordant comparisons (those ordered in the same direction on both distances), and D is the number of discordant comparisons (those ordered in opposite directions on the two distances). To calculate gamma, the computer searches through all sets of three languages; whenever [15] classifies two languages in the same group and the third language outside the group, the partial order implies two comparisons: 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. Each comparison is then counted as concordant or discordant depending upon the relative distances in the matrix of LDN or LDND. Like  $r$ , gamma ranges from  $-1$  to  $+1$  and takes the value 0 if the distances compared are independent of each other.

The observed values of gamma appear in Table 1 in the two columns under the heading GE (mnemonic for ‘Goodman-Kruskal gamma for the correlation with *Ethnologue*). The averages across the 49 families are almost the same, with GE higher by .0009 for LDND than for LDN. LDN ‘wins’ in 20 cases and LDND in 24 cases, again not significant by a sign test. Within families, therefore, LDND and LDN correlate about equally well with the classifications in [17] and [15], as predicted by hypothesis (b).

In addition to the distinctiveness values and the correlations, Table 1 displays the number of word lists held in the database for each family,  $N$ , as well as the number of languages in each family according to [15],  $n$ . These figures will be of importance in the following section. They require a bit of explanation. [17] and [15] differ with respect to the languages they assign to the different families, with [17] generally being more conservative than [15]. For instance, in the case of the Khoisan family [17] excludes the Hadza language while [15] includes it. Since we use the family definitions of [17] but the language statistics of [15], the latter are modified to accommodate the differences between the two classifications. To use the Khoisan example, [15] counts 27 languages as belonging to Khoisan. Since [17] excludes Hadza the total count of languages which we cite is  $n = 27 - 1 = 26$ . There are a few cases where it is not straightforward to combine the family definition of [17] with the language statistics of [15]. For instance, there are two languages within the Macro-Ge family of [15] which are treated as not belonging to Macro-Ge in [17], namely Chiquitano and Jabuti. A third language, Arikapú, is said to be similar to Jabutí in [15], so we can infer that [17] would also exclude Jabutí from Macro-Ge even if Jabutí is not included in the database of [17]. Thus, when [15] lists 32 Macro-Ge languages, we would reduce this to 29 following the family definition of [17]. Nevertheless, there is some uncertainty concerning how [17] would classify some other languages that are also not included in that database and appear as isolated members of Macro-Ge subgroups in [15], such as Botocudo, Kamakan, etc. For the statistical purposes of the language count (see Section 4 below), however, the uncertainty of how a small handful of languages are to be treated following [17] is of little importance.

In a further test of LDN and LDND within families, we generated two actual phylogenetic trees for each of the 49 language families in our sample, one using LDN and one using LDND, and we then compared these trees to the ones in [15] for the same families. The extraction of the classification of [15] for just those languages that pertain to our sample was greatly facilitated by software produced by P. Huff [20]. The trees generated from our data were constructed using the MEGA software [21] by means of Neighbor-Joining [22], which is probably the currently most widely used distance-based phylogenetic algorithm. These two trees were compared to one another and to the *Ethnologue* tree using the Robinson-Foulds distance (a.k.a. Symmetric Difference) [23]

as implemented in software by J. Felsenstein [24]. The results are displayed in Table 2. In the majority of cases LDN and LDND have the same RF distance to the *Ethnologue* tree, and when there are differences, these are small (never exceeding 8). The sum of distances is slightly smaller for LDND, which performs better in 11 individual cases while LDN performs better in 4 cases. The number of ‘wins’ scored by LDND, however, does not reach a  $p < .05$  significance level by a sign test.

Table 2. Robinson-Foulds distances between *Ethnologue* (Ethn) classifications and trees produced from LDN and LDND

Family	LDN tree vs. Ethn	LDND tree vs. Ethn	LDN tree vs. LDND tree	Family	LDN tree vs. Ethn	LDND tree vs. Ethn	LDN tree vs. LDND tree
Khoisan	15	15	0	Sko	4	4	0
Niger-Congo	554	558	200	Torricelli	30	28	10
Kadugli	8	8	0	Trans-New Guinea	261	257	88
Nilo-Saharan	97	95	16	West Papuan	32	32	10
Afro-Asiatic	210	210	68	Western Fly	36	36	14
Indo-European	231	231	42	Australian	153	153	94
Uralic	17	17	2	Na-Dene	24	24	0
Altaic	81	81	4	Algic	28	26	4
Nakh-Daghestanian	18	18	2	Uto-Aztecan	31	31	0
Dravidian	21	19	4	Salishan	7	7	0
Sino-Tibetan	152	152	54	Penutian	11	11	2
Hmong-Mien	10	10	4	Hokan	21	19	8
Tai-Kadai	55	55	10	Oto-Manguean	43	43	24
Austro-Asiatic	153	153	94	Totonacan	6	6	0
Great Andamanese	8	8	0	Mixe-Zoque	5	5	0
Austronesian	876	874	342	Mayan	69	67	2
Border	14	14	4	Chibchan	14	16	8
Bosavi	12	12	2	Tucanoan	12	10	4
Eleman	2	2	0	Panoan	18	18	4
Kiwaian	12	12	0	Quechuan	15	15	0

Lakes Plain	22	22	6	Arawakan	42	42	14
Lower Sepik-Ramu	8	10	4	Cariban	20	20	16
Marind	22	22	2	Tupian	44	44	24
Morehead and Upper Maro Rivers	11	11	6	Macro-Ge	22	24	16
Sepik	21	19	4	<b>Sum</b>	<b>3461</b>	<b>3447</b>	<b>1138</b>

From Tables 1 and 2 we conclude that it matters little whether LDN or LDND is used as the distance measure when languages are sampled according to groupings of related languages as predefined by historical linguists. In this situation, we concur with [1] that LDN may be preferable. The argument, however, would not be that LDN is demonstrably better. The only reason for preferring LDN over against LDND would be that the former is the conceptually more simple measure; additionally it saves computer time. While the correlations within families in Tables 1 and 2 are inconclusive, showing the two methods to be almost equally good, the distinctiveness comparisons in Table 1 clearly support LDND as the more adequate distance measure. For consistency we consider it preferable to use LDND both in the cases where it is demonstrably more adequate and in situations where its performance is similar to that of LDN. A case where we might make an exception is the study of close dialects. But proper studies at the high-resolution level of dialects would, in any case, invite a consideration of somewhat different methods, including a more fine-grained transcription than the very broad phonological transcription system of [6] (or the standard orthographies of [1,5]).

This concludes our response to [1]. In the remainder of the paper we go beyond the discussion and consider the interesting issue, raised by the results in Tables 1 and 2, of why our automated method of language classification (the ASJP approach) apparently works better for some language families than for others. In spite of small differences between different automated methods, it is likely that all the ones currently available, if submitted to a large dataset such as ours, would also show differential performance for different language families. Thus, the issue of differential performance is probably common to the whole enterprise of automated language classification.

#### 4. Differential performance of automated language classification

Four factors are tested for their influence on the degree to which LDND and LDN produce distinctive families and correlate with expert classifications within families: (1) mean % word attestation, (2) % language attestation, (3)  $d(\text{in})$ , and (4) the number of languages in each family according to [15]. ‘Mean % word attestation’ (1) refers to the mean percentage of attested words for the 40 concepts on the standard list used. Since only wordlists that have at least a 70% coverage of the concepts on the list are normally admitted to the ASJP database, there is not much variation in word attestation. ‘% language attestation’ (2) refers to the proportion of languages (as defined in [15]) that are attested in the ASJP database for each family (as defined in [17]). ‘ $d(\text{in})$ ’ (3) is defined above as a component of the distinctiveness of a family; it refers to the mean distance, either LDND or LDN, across all pairs of languages within each family. This quantity is higher in families of distantly related languages and lower in families of closely related languages. ‘Number of languages in each family’ (4) is denoted by  $n$  in Table 1 above. Since the distribution of  $n$  is positively skewed with a long upper tail,  $n$  is transformed logarithmically in the present analysis to produce an approximately normal distribution.

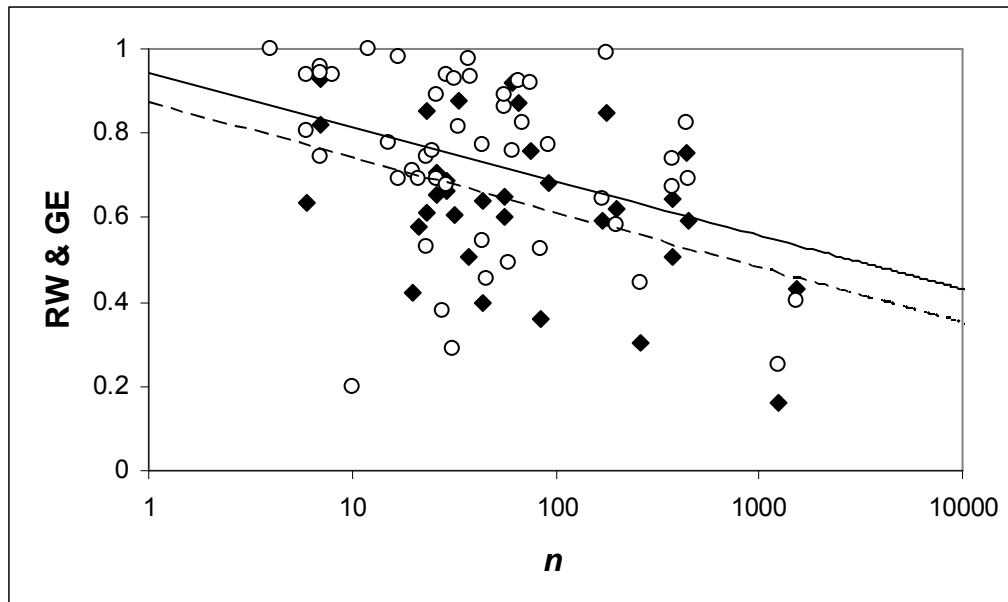
Table 3. Correlations of four factors with Dist, RW, and GE, for LDND and LDN.

Factor	Dist		RW		GE	
	LDND	LDN	LDND	LDN	LDND	LDN
Mean % word attestation	.47	.46	.02	.02	.14	.15
% language attestation	.07	.12	.20	.22	.18	.16
$d(\text{in})$			-.26	-.29	-.11	-.07
log number of languages	-.38	-.42	-.44	-.49	-.37	-.37
Number of families	49	49	33	33	49	49

The first two columns of Table 3 show Pearson’s  $r$  for three of these factors with Dist, which measures the distinctiveness of families in terms of LDND or LDN;  $d(\text{in})$  is omitted because it is by definition a component of Dist. The other four columns show Pearson’s  $r$  for each of the four factors correlated with RW and GE, which measure the fit between LDND or LDN and the expert classifications within families. The

correlations are across the families for which Dist, RW, or GE is defined in Table 1, and the last line of Table 3 shows the numbers of such families. The correlations between mean % attestation are significantly positive for Dist but not for RW or GE. None of the correlations involving % language attestation or  $d(\text{in})$  is significantly different from zero. All of the correlations involving log number of languages are significantly negative. The latter finding indicates clearly that LDND and LDN are less consistent with the expert classifications of [17] and [15] for large families than for smaller ones, in terms both of distinguishing between families and of subgrouping within families. Figure 1 shows how the number of languages per family is related to RW and GE for LDND.

Figure 1. RW (diamonds) and GE (circles) for LDND as functions of  $n$  (on a logarithmic scale) with regression lines showing correlations between  $\log n$  and respectively RW (dashed) and GE (solid)



Since we would not expect a machine to perform worse just because it has more data to deal with, some other explanation of this negative correlation is sought. Arguably, larger families are more likely to come into contact with several unrelated languages than smaller families, leading to an influx of loanwords which could disturb ASJP results since ASJP currently does not have a mechanism for filtering away loanwords. The more qualitatively-based classifications of experts, on the other hand, usually pay attention to

loanwords. An additional possibility – which does not exclude the first one from simultaneously operating – is that experts are less likely to produce accurate genealogical classifications for larger families than for smaller families. The families that score below the average for both RW and GE are Niger-Congo, Nilo-Saharan, Dravidian, Sino-Tibetan, Austro-Asiatic, Austronesian, Lakes Plain, Trans-New Guinea, Australian, and Algic. With the possible exceptions of Dravidian and Algic, historical linguists would probably agree that many relationships among languages in these families still have to be worked out through the application of the traditional criteria of identifying synapomorphies (shared innovations). Moreover, in several cases (in particular, Nilo-Saharan, Trans-New Guinea, Australian) there continue to be controversies over the question of which languages to include or exclude as family members, and these families are among the least distinctive in the database.

Thus, for larger families whose internal phylogenetic configurations still remain to be fully worked out, such as the ones just mentioned, and for world areas where language genetic classification is not well developed, such as New Guinea, automated language classification should be useful as a prognosticator of what may be eventually revealed about genetic affiliation when enough data are accumulated to facilitate accurate expert reckoning.

## 5. Conclusion

The authors of [1] discussed different measures of linguistic distance based on standard comparative wordlists compiled for the ultimate purpose of establishing phylogenetic relationships among languages. They direct criticism at the normalization procedure used within ASJP, whereby, in addition to a normalization of the standard Levenshtein distance by the length of the longer of the strings compared, the mean similarity among words not referring to the same concepts,  $\Gamma$ , is also taken into account. Here we show that this second normalization in fact improves correlations between the distance measures produced by our automated method and taxonomic distances in two standard classifications of the world's languages which seek to summarize the consensus among experts working within individual families. More precisely, correlations improve when the languages analyzed have not undergone a prior sorting into different families with

reference to standard classifications. For the internal classification of predefined language families using or not using the  $\Gamma$ -normalization makes no appreciable difference.

Our defence of the ASJP procedure was followed by a brief report on factors other than the distance measure that might influence automated language classification. The one factor that clearly emerges as significant is size of language families, measured by the number of languages in families of the world as described in the catalogue of [15]. Automated classification seemingly works better for smaller than for larger families. This result may, in part, be due to a greater amount of unrecognized loanwords in larger than in smaller families. As far as this possible factor is concerned, expert classifications should be more reliable than automated ones since experts are prepared to sort loanwords from true cognates. Another explanation for this negative correlation, however, is the expected differential accuracy of classifications produced by experts in a limited period of time when confronted with large as opposed to smaller families. The traditional comparative method whereby reconstructions are worked out and shared innovations identified may often require several person-years of work for even smallish language families of one or two dozen members and substantially greater periods of time for very large families. Large families composed of several hundred languages present exceptional challenges to experts working within the framework of traditional historical linguistics that impede their pace in fleshing out the very most accurate classifications. We speculate, then, that the negative correlation is mainly due to the ability of ASJP to quickly produce reasonable classifications for large families compared to the extended time needed by experts to produce the most valid results for the same families. If this is correct, the correlation is negative not so much because ASJP classification works better for smaller rather than larger families, but rather because experts are more likely to produce, in a limited amount of time, more valid classifications for smaller than for larger families.

Future work should, and will, be dedicated to the direct comparison of the results of automated language classification and those of historical linguists applying traditional methods. A first step in this direction is a series of working papers, posted on the ASJP website [25], which are dedicated precisely to such confrontations for individual

language families. In planned future conferences and publications this area of research will be expanded.

## References

- [1] F. Petroni and M. Serva, *Physica A*, **389**, 2280-2283 (2010).
- [2] B. Kessler, *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics* (Morgan Kaufmann Publishers Inc., San Francisco, 1995), pp. 60-67.
- [3] J. Nerbonne, W. Heeringa and P. Kleiweg, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* ed. by D. Sankoff and J. Kruskal (CSLI, Stanford, 1999), pp. v-xv.
- [4] W. Heeringa, Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. thesis, Rijksuniversiteit Groningen (2004).
- [5] M. Serva and F. Petroni, *Europhysics Letters* **81**, 68005 (2008)
- [6] C. H. Brown, E. W. Holman, S. Wichmann and V. Velupillai, *STUF – Language Typology and Universals* **61**, 285-308 (2008).
- [7] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller and D. Bakker, *Folia Linguistica* **42**, 331-354 (2008)
- [8] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller and D. Bakker, *Quantitative Investigations in Theoretical Linguistics* ed. by A. Arppe, K. Sinnemäki and U. Nikanne (University of Helsinki, Helsinki, 2008), pp. 40-43.
- [9] D. Bakker, A. Müller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant and E. W. Holman, *Linguistic Typology* **13**, 167-179 (2009)
- [10] K. Johnson, *Quantitative Methods in Linguistics* (Blackwell Publishing, Malden, MA, 2008).
- [11] S. S. Downey, B. Hallmark, M. P. Cox, P. Norquest and J. S. Lansing, *Journal of Quantitative Linguistics* **15**, 340-369 (2008).
- [12] M. Swadesh, *Int. J. Am. Ling* **21**, 121-137 (1955).
- [13] V. Levenshtein, *Cybernetics and Control Theory* **10**, 707-710 (1966).

- [14] M. Ruhlen, *A Guide to the World's Languages, Vol. 1: Classification*. (Stanford University Press, Stanford, 1991)
- [15] M. P. Lewis, *Ethnologue: Languages of the World*, 16th ed. (SIL International, Dallas, Tex, 2009) <<http://www.ethnologue.com/>>
- [16] H. Hammarström, *Diachronica* (in press, 2010).
- [17] M. Haspelmath, M. S. Dryer, D. Gil and B. Comrie. The World Atlas of Language Structures Online. (Max Planck Digital Library, Munich, 2008) <<http://wals.info/>>.
- [18] S. Wichmann, A. Müller, V. Velupillai, C. H. Brown, E. W. Holman, P. Brown, M. Urban, S. Sauppe, O. Belyaev, Z. Molochieva, A. Wett, D. Bakker, J.-M. List, D. Egorov, R. Mailhammer and H. Geyer, The ASJP database (version 12) (2010) <<http://email.eva.mpg.de/~wichmann/languages.htm>>.
- [19] L. A. Goodman and W. H. Kruskal, *J. Am. Stat. Ass.* **49**, 732-764 (1954)
- [20] P. Huff, Process\_asjp - a script for generating subsets of ASJP data (version 0.0.1) (2010). <<http://email.eva.mpg.de/~wichmann/software.htm>>.
- [21] K. Tamura, J. Dudley, M. Nei, and S. Kumar. *Mol. Biol. Evol.* **24**, 1596-1599 (2007).
- [22] N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406-425 (1987).
- [23] D. R. Robinson and L. R. Foulds, *Mathematical Biosciences* **53**, 131-147 (1981).
- [24] J. Felsenstein, Phylip (1980+) <<http://evolution.genetics.washington.edu/phylip.html>>
- [25] <<http://email.eva.mpg.de/~wichmann/papers.htm>>.