Languages with longer words have more lexical change

Søren Wichmann and Eric W. Holman

1. Introduction: Aims and data[1]

The findings to be presented in this paper were not anticipated, but came about as an unexpected result of looking at how the application of a version of the Levenshtein distance to word lists compares with cognate counting. We were interested in the degree to which the two correlate. The results of this investigation are intrinsically interesting and will be presented in the following section 2, but even more interesting is our finding that differences between counting cognates and measuring the Levenshtein distances vary as a function of average word lengths in the word lists compared. This observation will occupy the remainder of the paper, with section 3 devoted to establishing the statistical significance of the observation across language families, while section 4 establishes the significance within language groups, and section 5 discusses competing explanations. First we briefly explain the specific version of the Levenshtein distance used and the concept of cognate identification.

In numerous previous papers, beginning in Holman et al. (2008a), the present authors as well as other members of the network of scholars participating in the project known as ASJP (or Automated Similarity Judgment Program) have applied a computer-assisted comparison of word lists in order to derive a measure of differences among languages. Our method consists in comparing pairs of words to determine the Levenshtein distance, LD, which is defined as the number of substitutions, insertions, and deletions necessary to transform one word into another. The LD is divided by the length of the longer of the two words compared such that any distance will come to lie in the range 0%–100%. This normalized measure, called LDN,[2] is averaged over all pairs of words referring to the same concept in lists from two given languages. To enhance discrimination between related and unrelated languages, this average LDN is further divided by the average LDN between words referring to different concepts in the different lists, to obtain what we call LDND ('Levenshtein Distance Normalized Divided'). A similarity measure, here called ASJPsim, is defined by subtracting LDND from 100%.

The higher performance of LDND in comparison to LDN for the purpose of classifying languages is supported in Pompei et al. (2011) and Wichmann et al. (2010a), and Huff and Lonsdale (2011) report similar performances of LDND and the more linguistically informed but also much more computer-intensive ALINE algorithm of Kondrak (2000). Greenhill (2011) reports a low performance for LDN (not looking at LDND), but limits the investigation to the specific case of the Austronesian languages.

LDND and ASJPsim have been put to various uses, such as the dating of proto-languages (Holman et al. 2011), the identification of geographical centers of linguistic diversity for the purpose of identifying homelands (Wichmann et al. 2010c), the estimation of the limitations of word list comparisons for identifying deep genealogical relationships (Wichmann et al. 2010b), and the study of the relationship between population sizes and language change rates (Wichmann and Holman 2009). As objective and easily-obtained measures of the difference and similarity between any given pair of languages, LDND and ASJPsim are potentially useful for the investigation of possible correlations between languages and other kinds of data, such as data pertaining to human culture, prehistory, biology, and ecology.

A different method of measuring similarities between languages is that of counting cognates (related words) on a fixed list of lexical concepts. The percentage of concepts for which the words are cognate in two given languages is here called COGNsim. This method was developed within the framework of lexicostatistics (e.g., Swadesh 1955). In more recent years, cognate identification has been used to establish cognate classes as input to character-based phylogenetic methods, and a variety of issues have been explored using such methods, including dating and classification of language groups (Gray and Jordan 2000; Gray and Atkinson 2003), identification of factors that affect speed of lexical change (Pagel et al. 2007; Atkinson et al. 2008), questions of homelands and language expansions (Gray et al. 2009; Walker and Ribeiro 2011), and relationships between the evolution of different cultural traits (cf. Mace and Jordan 2011 for a review). These studies have mostly been carried out in relation to the three largest groups of languages where word lists coded for cognacy are available: Indo-European (Dyen et al. 1992), Austronesian (Greenhill et al. 2008), and Bantu (Bastin et al. 1999).

Identifying a cognate pair of words is not a trivial task. In the ideal situation a full set of sound correspondences is available which will allow the researcher to match up related words correctly, but the prior identification of regular sound correspondences requires hundreds, if not thousands, of sets of word comparisons. Such information is rarely available. Thus, it is more common to resort to some version of what Gudschinsky (1956: 615) calls the "inspection method", which essentially amounts to educated guesses.

The aim of this paper is to compare ASJPsim to COGNsim. ASJPsim is based on the 40 items identified by Holman et al. (2008b) as the most stable in Swadesh's (1955) 100-item lexicostatistical list. This can be compared to COGNsim at two different levels of resolution, depending on the type of data available. Many published studies present matrices of cognate percentages based on Swadesh's 100-item list, his earlier 200-item list (Swadesh 1952) or a modification of one of these.[3] A minority of studies additionally provide word lists where each lexical item is identified as belonging to a given cognate class, thus allowing for a higher resolution of the comparison in the sense that ASJPsim can be compared to judgments of cognacy at the level of words. These comparisons are based on the items in the 40-item ASJP list that are also included in the lists used in the study. For simplicity, all calculations are based on the first synonym listed if the source includes more than one synonym for a concept. Loanwords identified as such in the source are omitted from the calculations.

Table 1 provides an overview of data and sources. These have been found by a search of pertinent literature. Undoubtedly more could be added, but the sample is sufficiently large and has a sufficient spread in terms of geography and genealogies (language families) that it allows us to test the statistical significance of observations made. Language family designations from *Ethnologue* (Lewis 2009) are followed by the abbreviations that we will use in later tables. We follow the sources in naming the different language groups.[4] See the legend after the table for abbreviations of language group names. These names will be used throughout this paper to identify a dataset from a specific source. For instance, in the context of references to data we use "Austronesian" for a small set of languages whose cognate percentages are given in Dyen (1965) rather than for the family as a whole. The appendix shows how we match languages in the sources with word lists in the ASJP database (Wichmann et al. 2012). We additionally provide a checkmark (√) following the reference when word lists encoded for cognate classes were available,[5] and finally we provide the number of languages (N) within each group for which data were available for both COGNsim and ASJPsim. The total number of languages sampled amounts to around 8% of the world's languages by the definition of Lewis (2009). The sample includes 24 families from all world areas, with no major skewing: Africa (3), Eurasia and SE Asia (7), the Pacific (6), North America (3), South America (3), Middle America (2).

Table 1. Overview of sources and the nature of the data

Family	Abb.	Group	Source	N
Afro-Asiatic	AA	Afras	Militarev (2000) √	18
		Cushitic	Bender (1971)	28
		EthSem	Bender (1971)	13
		Omotic	Bender (1971)	21
Altaic	Alt	Turkic	Troike (1969)	6
Australian	Aus	Daly	Tryon (1974)	13
		Iwaidjan	R. Mailhammer (p.c., 2011)	3
		Mayi	Breen (1981)	5
		Paman	Sommer (1976)	3
		WAustr	O'Grady (1966)	7
		WBarkly	Chadwick (1979)	3
		Worrorran	McGregor & Rumsey (2009)	9
Austro-Asiatic	AuA	MonKhm	Peiros (1998) √	16
Austronesian	An	Austr	Dyen (1965)	10
		Malagasy	Vérin et al. (1969)	18
		Melan	Z'Graggen (1969)	6
		Morob	Hooley (1971) √	55
		NHebr	Tryon (1973)	16
		Philip	Llamzon (1976)	72
		Yapen	Anceaux (1961)	18
Carib	Car	Cariban	Villalon (1991)	10
Dravidian	Dra	Dravidian	Andronov (2001)	10
Hmong-Mien	HM	MiaoY	Peiros (1998) √	6
Indo-European	IE	IndEur	Dyen et al. (1992)	55
Japonic	Jap	Japonic	Hattori (1961)	5
Mayan	May	Mayan	C. H. Brown (p.c., 2011) √	30
Macro-Ge	MGe	Ge	Wilbert (1962)	9
Mixe-Zoque	MZ	MiZo	Cysouw et al. (2006) √	10
Na-Dene	NDe	Athap	Hoijer (1956) √	15
Niger-Congo	NC	Atlantic	Sapir (1971)	21
		Benue-Congo	Bennett & Sterk (1977)	22
		Gur	Swadesh et al. (1966)	20
		Kwa	Heine (1968)	13

Family	Abb.	Group	Source	N
Nilo-Saharan	NS	ESud	Thelwall (1981)	12
		NilSah	Bender (1971)	23
		Southern Luo	Blount & Curley (1970)	5
Quechuan	Que	Quechua	Torero (1970)	9
Salishan	Sal	Salish	Swadesh (1950)	21
Sino-Tibetan	ST	Chinese	Xu (1991)	6
		LoloB	Peiros (1998) √	15
		SinTib	Benedict (1976)	7
Tai-Kadai	TK	Kadai	Peiros (1998) √	11
Torricelli	Tor	Kamas	Sanders & Sanders (1980) √	7
Trans-New Guinea	TNG	Angan	Lloyd (1973)	12
		Awyu	Voorhoeve (1968)	6
		Bosavi	Shaw (1986)	22
		Eleman	Brown (1973)	8
		Finisterre	Claasen & McElhanon (1970)	12
		GVDani	Bromley (1967)	7
		GrMad	Z'Graggen (1969)	50
		Huon	McElhanon (1967) √	14
		Kiwaian	Wurm (1973)	8
		Koiarian	Dutton (1969)	6
		Kolopom	Voorhoeve (1968)	3
		Ok	Voorhoeve (1968)	5
		TurKik	Franklin (1973)	4
Uto-Aztecan	UA	Uto-Aztecan	Miller (1984), Cortina-Borja & Valiñas (1989)	26
West Papuan	WP	NHalm	Chlenov (1986)	8
		Yawa	Jones (1986)	6

Legend: Afras: Afrasian; EthSem: Ethiosemitic; WAustr: West Australian; WBarkly: West Barkly; MoKh: Mon-Khmer; Austr: Austronesian; Melan: Melanesian; Morob: Morobe; NHebr: New Hebrides; Philip: Philippines; MiaoY: Miao-Yao; IndEur: Indo-European; MiZo: Mixe-Zoquean; Athap: Athapaskan; ESud: Eastern Sudanic; NilSah: Nilo-Saharan; LoloB: Lolo-Burmese; SinTib: Sino-Tibetan; Kamas: Kamasau; GVDani: Grand Valley Dani; GrMad: Greater Madang; TurKik: Turama-Kikorian; NHalm: North Halmahera.

2. Comparing ASJPsim and COGNsim

The number of data points for ASJPsim and COGNsim for individual language pairs is so large that it is unwieldy for visual inspection. However, to illustrate an interesting tendency in the data we plot language pairs from five different families in figure 1.

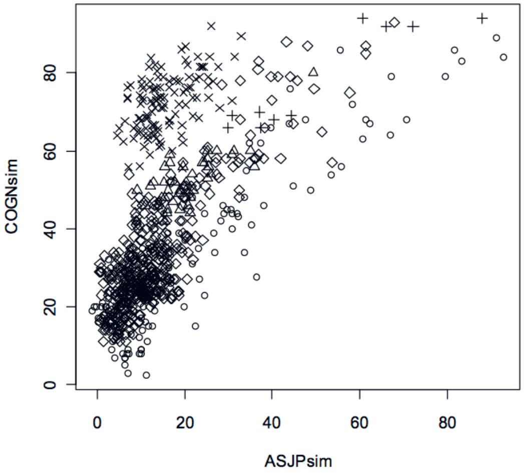

Figure 1. Scatter-plot of COGNsim as a function of ASJPsim for individual language pairs pertaining to five selected families: Australian (o), Uto-Aztecan (◊), Japonic (+), Carib (Δ), Sino-Tibetan (×).

Figure 1 illustrates, for selected data, that different families tend to occupy different regions in a scatter-plot of COGNsim and ASJPsim. For instance, Australian language pairs tend to stay close to the diagonal, whereas Sino-Tibetan language pairs occupy a region where low values for ASJPsim correspond to high values for COGNsim. Language pairs from other families occupy regions in between.

Table 2 provides data on the averages of ASJPsim and COGNsim for all language pairs within each family as well as Pearson's *r* for the correlation between mASJPsim and mCOGNsim across all language pairs belonging to the family.

Reviewing the second and third columns in table 2 we observe that mCOGNsim, the average cognate similarity within families, is always greater than mASJPsim, the corresponding average ASJP similarity. This is because cognates can be less than 100% similar.

Table 2. Data on mean ASJP similarities and mean cognate similarities

Family	mASJPsim	mCOGNsim	r
AA	12.30	23.13	0.856
Alt	57.84	75.57	0.600
An	23.23	31.80	0.726
AuA	13.82	38.66	0.788
Aus	19.36	32.00	0.886
Car	20.42	52.00	0.765
Dra	23.11	33.13	0.869
HM	10.67	72.36	0.715
IE	9.48	24.05	0.921
Jap	50.70	78.00	0.921
May	29.14	47.82	0.862
MGe	29.31	66.56	0.451
MZ	46.42	66.10	0.806
NC	7.54	31.92	0.741
NDe	19.51	51.83	0.739
NS	6.67	12.53	0.934
Que	54.63	82.43	0.325
Sal	11.79	24.56	0.841
ST	14.62	66.85	0.576
TK	21.69	62.81	0.788
TNG	16.95	31.69	0.836
Tor	63.41	77.52	0.929
UA	14.29	46.83	0.837
WP	49.74	75.21	0.648

In figure 2 we plot the relationship between mCOGNsim and mASJPsim. The dotted line, provided as a point of comparison, intercepts at zero and has a slope of 1. The solid line shows the results of a linear regression, where $r = 0.762$ and $p < 0.0001$. Its slope is 0.93, which is so close to 1 that the intercept, at 25.96%, can be interpreted as the percentage that roughly needs to be added to get from mASJPsim to mCOGNsim. The relatively high r and the low p show that ASJPsim and COGNsim are parallel measures of similarities among languages.[6] However, we observe a cone-shaped distribution of the dots in the chart, with a tremendous amount of variation in mCOGNsim for low values of ASJPsim and an increasingly narrower concentration around the regression line for high values of ASJPsim. This reflects the sort of distribution exemplified in figure 1, where language pairs in different families form clouds in different regions of the chart, except that here (in figure 2) we represent each family as a single data point. In the following section we turn to possible explanations for this variability.

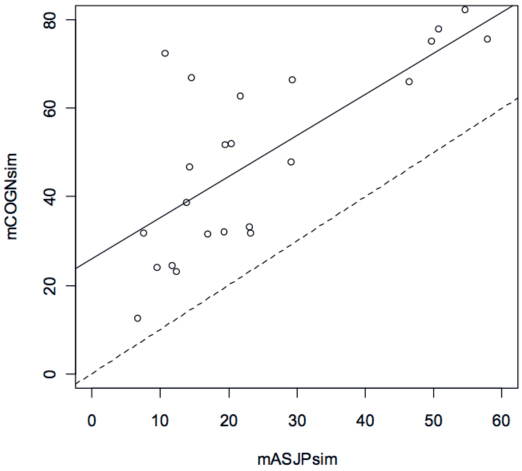

Figure 2. Scatter-plot of mCOGNsim against mASJPsim

3. mCOGNsim vs. mASJPsim in relation to segment inventory size and word length

Given that the relation between ASJPsim and COGNsim is partly a "family matter" we need to somehow capture the relationship across all language pairs pertaining to each family. One way of doing this is simply to take the average of the difference between COGNsim and ASJPsim, which we call mDIFF. Another, more principled approach involves looking at how the two depend on time. Glottochronology (Lees 1953) assumes that the logarithm of COGNsim diminishes in proportion to time, and a similar assumption, supported by evidence from 52 archaeological, historical and epigraphic calibration points, is made for ASJPsim in Holman et al. (2011). Thus, the log of COGNsim should be proportional to the log of ASJPsim. A useful way of characterizing the relation between COGNsim and ASJPsim, then, is to find the average ratio, mRATIO, of log(COGNsim) to log(ASJPsim) across language pairs of a given family.

It could be the case that the size of phonological inventories affects the rate of change in segments. If a language has a relatively large number of segments the phonetic space occupied by each will be relatively small. In this situation a phonetic fluctuation will perhaps more easily cross the phonological boundary of a neighboring segment in the phonetic space, leading to free variation, which may eventually lead to articulatorily driven phonological change. Perceptually driven changes would also seem to occur with a higher probability when the phonetic space is more densely packed, since language users would be more prone to misperceive a sound when phonetically similar sounds constitute part of the inventory. These hypothetical factors translate into the testable prediction that the difference between mASJPsim and mCOGNsim should be positively correlated with the average number of phonological segments found in the group of languages concerned.

While we do not have access to full segment inventories for the languages in our sample, we can use as proxies the number of different segments in the ASJP transcription system (or ASJPcode, cf. Brown et al. 2008) found in the 40-item word lists. In Wichmann et al. (2011) the number of different segments in word lists was used as a proxy for segment inventory sizes in a successful way, inasmuch as we were able to confirm well-known correlations involving segment inventory sizes (Hay and Bauer 2007; Nettle 1995, 1998) using the mean number of segments represented in word lists (mSR).

Another set of hypotheses is that shorter words tend to change faster phonologically than longer ones or that speakers of a language will exchange their words for completely new ones with a relatively high speed when their language contains relatively long words. In both cases, the difference between COGNsim and ASJPsim should be inversely correlated with mean word length, because the rate of loss of COGNsim over time would approach the rate of loss of ASJPsim when words are longer. Now, these are perhaps not the intuitively most plausible hypotheses and we would not have produced them had it not been for the fact that either one or the other is strongly supported by our data. As a measure of word length, we take the first word for each concept in our 40-item lists and average the number of ASJPcode segments (if the translation of the concept is phrasal we still only take the first word). Finally we average the averages within language families to obtain mmWL.

Table 3 contains the data needed to test these different predictions. First we test whether the difference between mASJPsim and mCOGNsim is positively correlated with mSR, and thereby the hypothesis that languages change phonologically faster the more phonemes they have. A linear correlation of mSR and mDIFF indeed shows a positive correlation of *r* = .21, but it is small and non-significant, *p* =.32. As another way of looking at the relationship we can test whether mRATIO, i.e., the mean of the ratios of log(COGNsim) to log(ASJPsim), is negatively correlated with mSR. This is, indeed, the case, but again the correlation is small and non-significant, *r* = –.23, p = .27. Thus, judging from the evidence from language family averages, the hypothesis that languages change faster phonologically the more segments they have is not borne out – in spite of the plausible nature of the hypothesis.

We now go on to test whether differences in word length explain the variability in differences between cognate similarities and ASJP similarities across language families. Again referring to table 3 we first correlate mDIFF and mmWL, which yields a solid *r* = –.50, *p* = .01. Again we alternatively test for mRATIO and find that this property is positively correlated with mmWL, *r* = .53, *p* < .01.

Somewhat unexpectedly we have found that, judging by averages across families, lexical replacement increases as a function of word length or, alternatively, phonological change decreases as a function of word length. In contrast, lexical replacement and phonological change are not significantly affected by the sizes of segment inventories.[7] Later in this paper we discuss the competing explanations for the correlations involving word length. But before that, we would like to establish the findings more firmly by looking at the behavior of individual words within language families.

4. Correlations across items within families

We have seen in the previous section that language families with longer words tend to have fewer cognates relative to their overall lexical similarity than do families with shorter words. Does this correlation apply to the words themselves or only to the families? This question can be addressed in the eleven language groups with checkmarks in table 1, for which the sources of the word lists also indicate which words are cognate. With this information it is possible to calculate mASJPsim and mCOGNsim separately for individual items on the ASJP 40-item list, averaged across language pairs in the group. DIFF and RATIO are then defined for each item as mCOGNsim – mASJPsim and log(mCOGNsim)/log(mASJPsim), respectively. One new property of items, mAsimC, is defined as mASJPsim calculated only for those pairs of words identified as cognate. mAsimC indicates the degree of phonological similarity between words that may have undergone phonological change but have not undergone lexical replacement. To ensure representative samples,

Table 3. Data for correlations with numbers of segments and mean word length

Family	mDIFF	mRATIO	mSR	mmWL
AA	10.83	0.644	27.36	4.06
Alt	17.72	0.545	23.83	3.58
An	8.57	0.795	20.65	4.53
AuA	24.84	0.469	25.25	3.65
Aus	12.64	0.686	18.35	5.21
Car	31.58	0.407	19.30	4.68
Dra	10.02	0.746	19.90	3.96
HM	61.69	0.133	28.33	3.29
IE	14.58	0.570	26.80	3.95
Jap	27.30	0.339	20.40	3.79
May	18.68	0.589	24.96	3.66
MGe	37.25	0.353	23.44	3.82
MZ	19.68	0.577	19.50	3.78
NC	24.37	0.436	24.47	3.73
NDe	32.31	0.302	27.93	3.24
NS	5.87	0.775	25.73	3.83
Que	27.80	0.328	21.56	4.41
Sal	12.77	0.641	32.67	5.06
ST	52.23	0.229	27.52	3.37
TK	41.12	0.284	25.00	2.98
TNG	14.75	0.653	20.49	4.27
Tor	14.12	0.573	21.86	4.17
UA	32.54	0.531	20.42	4.42
WP	25.47	0.467	18.29	4.63

Legend: mDIFF: the difference between mASJPsim and mCOGNsim; mRATIO: the mean ratio of log(COGNsim) to log(ASJPsim); mSR: the mean of the number of different phonological segments in the word lists pertaining to the family; mmWL: the average word length within word lists averaged across the languages the family.

all these quantities are calculated only for items that are attested in at least 70% of the languages in the group. Also, mAsimC is defined for an item only if mCOGNsim is above 0%, because otherwise there are no cognate pairs for the item. RATIO is defined only if mCOGNsim is above 0% and mASJPsim is strictly between 0% and 100%, in order to avoid logarithms of nonpositive numbers or division by 0.

Table 4. Correlations within families

Group	mAs	mCs	DIFF	RATIO	mAsimC
Afras	–.359	–.193	–.369	.415	.180
MoKh	–.370	–.326	–.317	.443	–.269
Morob	–.083	.090	–.135	.209	.010
MiaoY	–.092	–.401	.080	–.080	–.462
Mayan	–.569	–.389	–.495	.609	–.016
MiZo	–.409	–.445	–.092	.288	–.376
Athap	–.342	–.259	–.255	.393	–.166
LoloB	–.063	–.402	.154	.091	–.447
Kadai	.038	–.044	.090	–.076	–.140
Kamas	–.368	–.306	–.160	.157	–.004
Huon	–.267	–.519	.100	–.149	–.119
Mean	–.262	–.290	–.127	.209	–.164
t(10)	4.68	5.12	1.95	2.82	2.65

Legend: mAs: mASJPsim; mCs: mCOGNsim

Cases where a concept is translated by a phrase, i.e., two or more words separated by spaces in the data source, rather than by a single word, are treated differently for different purposes. For the estimates of the mean length of words used in previous sections only the first word in a phrase was counted. This seems appropriate when a word list is used as a random sample of words in a language. However, for comparing the way that specific concepts are expressed across concepts and languages, as is done in the present section, the whole phrase is counted. When translations exceed a single word, the properties mASJPsim and mCOGNsim refer to the entire phrase throughout this paper. The data used are provided as an online appendix.

Mean word length is now correlated across lexical concepts with each of the similarity properties defined in the first paragraph of this section. That is, for each of the 40 items pertaining to the ASJP word lists of each language group we determine their average length as well as mASJPsim, mCOGNsim, etc., and then calculate Pearson's *r*. Table 4 provides the correlations for each of the eleven language groups, ordered as in table 1. The table also gives the mean correlation across groups and the value of Student's *t* (with 10 degrees of freedom) for testing whether the mean correlation differs from 0.

Most of the individual correlations are weak, possibly because of limited variation within language groups. They are collectively quite consistent across groups, however, producing significant effects (*p* < .05) for all but one measure, namely DIFF. Both measures of similarity, mASJPsim and mCOGNsim, show significantly less similarity when items are represented by longer words than when they are represented by shorter ones. Since similarity is calculated across the same pairs of languages for each item, it follows that time depth is the same across items and therefore that items represented by longer words are less stable through time than items represented by shorter ones. The two relative measures, DIFF and RATIO, are consistent with table 3 in showing less lexical similarity relative to phonological similarity for translational equivalents that have longer words, although this effect is significant only for RATIO. Finally, the significantly negative mean correlation for mAsimC implies that longer words undergo more phonological change even if they are not replaced. It follows that the lower lexical similarity of longer words relative to phonological similarity is a consequence of more lexical change rather than less phonological change. In summary, longer words are more likely to be replaced and more likely to change phonologically if they are not replaced.

The significant negative correlation between word length and stability can be extended to the entire ASJP database. For this purpose, mean word length is calculated for each item in each language family and then averaged across families. Stability is defined as in Holman et al. (2008b), except that their similarity measure is replaced by mASJPsim. More specifically, mASJPsim is first determined for each item in each of the language genera established by Dryer (1989, 2011), who defines genera as the most inclusive groups descended from a common ancestral language spoken within the last 3500 to 4000 years. Then stability is equal to the weighted mean of mASJPsim across genera, with each genus weighted by the square root of the number of language pairs in the genus. The correlation between stability and mean word length across the 40 items in the ASJP list proves to be substantial, *r* = –.47, *p* < .01. The other correlations in table 4 cannot be extended in this way because judgments of cognacy are not available for most language groups.

The negative correlation between word length and stability, whether measured by mASJPsim, mCOGNsim, or mAsimC, can be explained by the finding of Pagel et al. (2007) that frequency of use is positively correlated with stability, given that frequent words tend to be shorter than infrequent ones (Zipf 1935). This explanation, however, does not account for the new observation, replicated both across languages and across items, that short words show less lexical change relative to phonological change.

5. Discussion and conclusion

It is a central concern to historical linguistics to identify causes for language change, and many causes are known to exist. External ones include effects of social stratification, lexical adaptation (the creation of new words for new concepts), borrowing and other effects of language contact, imperfect vertical transmission, etc.; among internal ones we can mention the spread throughout the lexicon of sound changes, analogy, grammaticalization, etc. The introduction of glottochronology widened the search for factors that might affect the *rate* of language change. Examples of such factors would be borrowing or word taboo. To date, however, not a single factor, either external or internal to languages, has been identified which *systematically* affects rates of change. Population size is an example of a proposed external factor influencing the rate of language change which has not stood up to a quantitative scrutiny (Wichmann et al. 2008; Wichmann and Holman 2009). The already-mentioned relation between frequency and stability identified by Pagel et al. (2007) is an example of language-internal factors regulating the rate of change, but does not exemplify a factor that systematically predicts that one language changes faster than another.

Thus, our major finding in this paper, namely that longer words tend to be replaced faster than shorter words both within and across languages, is unique. One implication of the finding is that critics of glottochronology for the first time have a weapon other than case studies to attack the idea that lexical change is regular enough to be a useful tool for dating language divergence. The weapon is rather blunt, however, since the effect of average word length is not overwhelming even if statistically significant, and our work on the ASJP dating technique (Holman et al. 2011) still shows a high degree of regularity in the decay of ASJPsim over time. In fact, the present finding that the effect of word length is stronger for COGNsim than for ASJPsim may explain why some studies of glottochronology report less regular results than do Holman et al. Thus, the ‘weapon’ may serve practitioners of lexically-based dating techniques better than their critics since it can potentially be used to improve those techniques.

We have not yet addressed possible explanations for our finding, and cannot hope to do so conclusively at this point. One possibly relevant factor is the differing information provided by long and short words for judgments of cognacy. Maybe false cognates are more likely to be accepted for short words. This sort of inaccuracy is less likely to be important if cognacy is inferred by means of regular sound correspondences rather than judged by inspection. Cognitive biases are even more reduced for ASJPsim, which is normalized by word length and calculated automatically.

The other possible explanatory factor is the process of language change itself. Why should speakers of languages that have longer words in their basic vocabulary replace these words more frequently than speakers of languages that have shorter words? The reason is not, for instance, that speakers want to replace longer words with shorter ones, because we draw our observations from the current state of languages, where the ones that have the longer words have replaced *earlier* words faster.

Our tentative hypothesis is that if a language has rich word-formation strategies at its disposal such that many of the words in a language are formed by derivation and compounding, then the words in the language will tend to be longer and also will tend to be replaced more often. An implication is that the creation of complex lexemes is generally preferred over the creation of simplex ones. A problem for testing this hypothesis is that the ASJP word lists are not based on a consistent definition of what a word is. Generally, the word lists simply reflect whatever is given as a translational equivalent for each concept in a particular linguistic source, with the exception that transcribers have stripped off inflectional elements and class markers when their knowledge of the languages allowed them to do so. This is a minor caveat, since the data can be revisited and adjusted for consistency. A more serious problem is that each of the many lexical items used in this study would ideally have to be tagged for its status as simplex, derived, compounded, or phrasal (and maybe other categories, as well as various combinations) in order to provide more substance to our hypothesis. Thus, the further investigation of the proposed relation between differences in word formation strategies and rates of lexical replacement seems to call for a larger, collective effort and several future case studies.

Online appendix: word lists transcribed in ASJPcode with cognate encoding from the literature

The online appendix is available at
http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/cognatedata.zip

The file contains data used in section 4, i.e., ASJP word lists with information on cognacy, mostly drawn from the literature. The format of the data sheet is explained in the file "description.pdf".

Appendix: matching of languages in the lexicostatistical literature with languages represented in the ASJP database used in this study

The data in tables 2–3 were produced by comparing published cognate percentages with ASJPsim for languages that are represented both in the lexicostatistical literature and in the ASJP database (Wichmann et al. 2012), where the latter has been updated to include as many of the languages in the former as possible and, whenever possible, the actual data on which the cognate counts were based. Language groups for which ASJPsim is calculated from the same dataset which was used for counting cognate percentages are identified by a star following the language group name, and when there are but a handful of exceptions where data from other sources have been added, or when data and cognate judgments are from the same author but in different publications, a star is given in parenthesis (note that ASJPsim is based on the reduced, 40-item version of the Swadesh list whereas published cognate percentages are based on other versions or derivatives, as described in section 1 of this paper, so identity of data sources does not mean complete identity of data). When authors providing cognate counts do not provide the word lists used, alternative data sources are used. All sources for ASJP word lists are found at http:// lingweb.eva.mpg.de/asjp/index.php/ASJP. The doculects represented in our database are uniquely identified by their names. In the lists below we provide the language family names (in bold), language group names (in bold italics), references to sources (in parentheses), names of languages in the sources for cognate counts (in normal font), the ISO 639-3 identifier, and the ASJP designation (in capital letters). Exceptions to these patterns are Iwaidjan (Australian) and Mayan, where cognate judgments were made directly in relation to ASJP word lists; thus, only the ISO-code and the ASJP designation are given in these cases. Language names are joined by + in cases where the source gives separate cognate percentages for varieties with the same ISO-code, and these percentages are averaged in the present calculations. In the ASJP database the language designations make use of the underscore characters instead of spaces (e.g., HARARI_2). In this appendix the underscores are replaced by spaces (e.g., HARARI 2) to allow for more line breaks. The information provided in this appendix is intended to ensure replicability of our results.

Afro-Asiatic: ***Afrasian**** (Militarev 2000): Lebanese Arabic, apc, ARABIC NORTH LEVANTINE; Tigrai, tig, TIGRIGNA; Amharic, amh, AMHARIC 3; Harari, har, HARARI 2; Mehri, gdq, MEHRI 2; Jibbali, shv, SHEHRI; Soqotri, sqt, SOQOTRI 2; Siwa, siz, SIWI; Ghadames, gha, GHADAMES 2; Qabyle, kab, KABYLE; Ahaggar, thv, TAMAHAQ TAHAGGART 2; Zenaga, zen, ZENAGA 2; Hausa, hau, HAUSA 3; Bole, bol, BOLE 2; Beja, bej, BEJA 2; Oromo, hae, EASTERN OROMO 2; Dahalo, dal, DAHALO 2; Kefa, kbr, KEFA 2. ***Cushitic**** (Bender 1971): Beja, bej, BEJA; Bilen, byn, BILIN 2; Qimant, ahg, KEMANT 2; Xamtanga, xan, XAMTANGA 2; Awngi, awn, AWNGI 2; Hadiyya, hdy, HADIYYA 2; Libido, liq, LIBIDO; Kembata, ktb, KAMBAATA 2; Alaba, alw, ALABA; Sidamo, sid, SIDAMO 2; Derasa, drs, GEDEO 2; Burji, bji, BURJI 2; Afar, aar, AFAR 2; Saho, ssy, SAHO 2; Baiso, bsw, BAISO 2; Arbore, arv, ARBORE 2; Dasenech, dsh, DAASANACH 2; Somali, som, SOMALI 2; Rendille, rel, RENDILLE 2; Mecha, gaz, MECHA OROMO; Borena, gax, BORANA OROMO 2; Qottu, hae, EASTERN OROMO; Konso, kxc, KOMSO 2; Gidole, gdl, GIDOLE 2; N Bussa, dox, BUSSA 2; Gawwada + Gobeze + Werize, gwd, GAWWADA 2; Tsamai, tsb, TSAMAI 2; Iraqw, irk, IRAQW 2. ***Ethiosemitic***$^{(*)}$ (Bender 1971): Tigre, tig, TIGRE; Tigrinya, tir, TIGRINYA; Amharic, amh, AMHARIC; Argobba, agj, ARGOBBA; Zway, zwa, ZWAY; Walani, stv, WALANI SILTE; Harari, har, HARARI; Gafat, gft, GAFAT; Soddo, gru, SODDO; Mesmes, mys, MESMES 2; Mesqan, mvz, MESQAN; Chaha + Geto, sgw, GETO; Innemor, ior, INNEMOR. ***Omotic**** (Bender 1971): Dime, dim, DIME; Ari, aiw, ARI; Banna, amf, BANNA; Maji, mdx, MAJI; Sheko, she, SHEKO; Nao, noz, NAO; Southern Mao, myo, SOUTHERN MAO; Shinasha, bwo, SHINASSHA 2; Kefa, kbr, KEFA; Mocha, moy, MOCHA 2; Janjero, jnj, JANJERO; Bencho, bcq, BENCHO; Male, mdy, MALE ETHIOPIA; Basketo, bst, BASKETO; Welamo, wal, WELAMO; Kullo, dwr, KULLO; Dorze, doz, DORZE; Oyda, oyd, OYDA; Kacama, kcx, KACHAMA; Koyra, kqy, KOYRA; Zayse + Zergulla, zay, ZAYSE.

Altaic: ***Turkic**** (Troike 1969): Turkish, tur, TURKISH 2; Azerbaijani (Azeri), azj, AZERBAIJANI NORTH:2; Karachai, krc, KARACHAY BALKAR; Crimean Tatar, tat, CRIMEAN TATAR; Kazan Tatar, tat, KAZAN TATAR; Misher Tatar, tat, MISHER TATAR.

Australian: ***Daly**** (Tryon 1974): Mullukmulluk, mpb, MULLUKMULLUK; Yunggor, zml, YUNGGOR; Kamor, xmu, KAMOR; Marithiel, mfr, MARITHIEL; Marityabin, zmj, MARITYABEN; Maridan, zmd, MARIDAN; Maramanadji, zmm, MARAMANADJI; Marengar, zmt, MARENGAR; Ami, amy, AMI; Manda, zma, MANDA; Pungupungu, wdj, PUNGUPUNGU; Wadyginy, wdj, WADJIGINY; Ngangikurr, nam, NGANGIKURRUNGGURR. ***Iwaidjan**** (Robert Mailhammer p.c., 2011): amg, AMURDAK; ibd, IWAIDJA; mph, MAWNG. ***Mayi**** (Breen 1981): Ngawun, nxn, NGAWUN; Mayi-Kulan, mnt, MAYKULAN; Mayi-Yapi, mnt, MAYI YAPI; Mayi-Thakurti, mnt, MAYI THAKURTI; Mayi-Kutuna, xmy, MAYAGUDUNA. ***Paman*** (Sommer 1976): Bariman Gutinhma, zmv, PARIMANKUTINMA; Umbuygamu, umg, UMBUYKAMU; Lamalama, lby, LAMALAMA COASTAL. ***West Australian*** (O'Grady 1966): Nyungumarda, nna, NYANGUMARTA; Yulbaridja, mpj, YULPARIJA; Warburton Ranges, ntj, NGAANYATJARRA; Pandjima, pnw, PANYTYIMA; Jindjibandi, yij, YINDJIBARNDI;

Ngaluma, nrl, NGALOOMA; Wadjeri, wbv, WAJARRI. ***West Barkly*** (Chadwick 1979): Wambaya, wmb, WAMBAYA; Djingili, jig, DJINGILI; Gudandji, nji, GUDANJI. ***Worrorran***$^{(*)}$ (McGregor and 2009): Wunambal, wub, WUNAMBAL; Gunin Kwini, gww, GUNIN/KWINI; Ngarinyin, ung, NGARINYIN; Unggumi, unp, UNGGUMI; Bunuba, bck, BUNABA; Gooniyandi, gni, GOONIYANDI; Kija, gia, KITJA; Miriwoong, mep, MIRIWUNG; Walmajarri, wmt, WALMAJARRI.

Austro-Asiatic: ***Mon-Khmer**** (Peiros 1998): Jeh, jeh, JEH; Bahnar, bdq, BAHNAR; Chrau, crw, CHRAU; Kui, kdt, KUI THAILAND; Khmer, khm, KHMER; Semai, sea, SEMAI; Mon, mnw, MON; Nyakur, cbn, NYAKUR; Vietnamese, vie, VIETNAMESE; Ruc, scb, RUC; Wa, wbm, WA; Deang, pce, DEANG; Khmu, kjg, KHMU; Ksinmul, puo, KSINMUL; Khasi, kha, KHASI; Mundari, unr, MUNDARI.

Austronesian: ***Austronesian*** (Dyen 1965): Banoni, bcm, BANONI; Saposa, sps, TAIOF; Iai, iai, IAAI; Tongan, ton, TONGAN; Tanna, tnn, NORTH TANNA; Fiji, fij, FIJIAN; Zabana, kji, ZABANA; Roviana, rug, ROVIANA; Acira, adz, ADZERA; Yapese, yap, YAPESE. ***Malagasy**** (Vérin et al. 1969): Betsileo Ambositra, plt, MALAGASY AMBOSITRA; Antaisaka, bjq, MALAGASY ANTAISAKA; Antambahoaka, bjq, MALAGASY ANTAMBAHOAKA; Antankarana, xmw, MALAGASY ANTANKARANA; Bara, bhr, MALAGASY BARA; Betsimisaraka, bmm, MALAGASY BETSIMISARAKA; Betsileo Fianarantsoa, bjq, MALAGASY FIANARANTSOA; Mahafaly, tdx, MALAGASY MAHAFALY; Merina, plt, MALAGASY MERINA; Sakalava 1, skg, MALAGASY SAKALAVA 1; Sakalava 2, skg, MALAGASY SAKALAVA 2; Sihanaka, plt, MALAGASY SIHANAKA; Taimoro, plt, MALAGASY TAIMORO; Antandroy 1, tdx, MALAGASY TANDROY 1; Antandroy 2, tdx, MALAGASY TANDROY 2; Tsimihety, xmw, MALAGASY TSIMIHETY; Vezo, skg, MALAGASY VEZO; Zafisoro, bjq, MALAGASY ZAFISORO. ***Melanesian**** (Z'Graggen 1969): Manam, mva, MANAM; Sepa, spb, SEPA; Gedaged, gdd, GEDAGED; Bilbil, brz, BILBIL; Takia, tbc, TAKIA; Matukar, mjk, MATUKAR. ***Morobe**** (Hooley 1971): Wagao, bzh, WAGAU; Mapos, bzh, MAPOS; Manga, kby, MANGA; Patep, ptp, PATEP; Kumaru, ksl, KUMARU; Zenag, zeg, ZENAG; Towangara, goc, TOWANGARA; Sambio, tbx, SAMBIO; Dambi, dac, DAMBI; Piu, pix, PIU; Buasi, val, BUASI; Latep, zeg, LATEP; Dunguntung, mpl, DUNGUNTUNG; Dangal, mcy, DANGAL; Silisili, mpl, SILISILI; Bubwaf, mpl, BUBWAF; Dagin, lbq, DAGIN; Azera, adz, AZERA; Wampar, lbq, WAMPAR; Sirak, srf, SIRAK; Guwot, gve, GUWOT; Duwet, gve, DUWET; Musom, msu, MUSOM; Sukurum, zsu, SUKURUM; Sirasira, zsa, SIRASIRA; Maralango, mcy, MARALANGO; Wampur, waz, WAMPUR; Mari, hob, MARI; Onank, una, ONANK; Yaros, adz, YAROS; Amari, adz, AMARI; Labu, lbu, LABU; Bukaua, buk, BUKAUA; Kela, kcl, KELA; Kaiwa, kbm, KAIWA; Sipoma, sij, SIPOMA; Hote, hot, HOTE; Yamap, ymp, YAMAP; Jabem, jae, JABEM; Tami, tmy, TAMI; Malasanga, mqz, MALASANGA; Gitua, ggt, GITUA; Lukep, apr, LUKEP; Mangap, mna, MANGAP; Barim, bbv, BARIM; Mutu, tuc, MUTU; Tuam, tuc, TUAM; Sio, xsi, SIO; Nengaya, met, NENGAYA; Roinji, roe, ROINJI; Arawe, aaw, ARAWE; Maleu, mgl, MALEU; Nakana, nak, NAKANA; Halia, hla, HALIA; Gedaged, gdd, GEDAGED. ***New Hebrides***(*) (Tryon 1973): Toga (Torres), lht, TOGA; Mosina (Banks), msn, MOSINA; Peterara (Maewo), mwo, CENTRAL MAEWO; Nduindui (Aoba), nnd, WEST AMBAE; Sakau (Santo), sku, SAKAO; Malo (Santo), mla, NORTH MALO; Fortsenal (Santo), frt, FORTSENAL; Raga (Pentecost), lml, RAGA; Sa (Pentecost), sax, SA; Dakaka (Ambrym), bpa, DAKAKA BAIAP; Aulua (Malekula), aul, AULUA; Big Nambas (Mal.), nmb, BIG NAMBAS UNMET; Lewo (Epi), lww, LEWO FILAKARA; Nguna (Efate), llp, NORTH EFATE NGUNA; Sie (Erromanga), erg, SIE; Lenakel (Tanna), tnl, LENAKEL LENAUKAS. ***Philippines*** (Llamzon and Martin 1976): Agta, agt, AGTA; Atta, att, ATTA PAMPLONA; Balangaw, blw, BALANGAW; Batak, bya, BATAK PALAWAN; Bilaan Koronadal, bpr, BILAAN KORONADAL; Bilaan Sarangani, bps, BILAAN SARANGANI; Binukid, bkd, BINUKID; Bontoc, bnc, CENTRAL BONTOC; Dumagat, dgc, DUMAGAT CASIGURAN; Gaddang, gdg, GADDANG; Amganad Ifugao, ifa, IFUGAO AMGANAD; Batad Ifugao, ifb, IFUGAO BATAD; Bayninan Ifugao, ify, IFUGAO BAYNINAN; Ilonggot, ilk, ILONGOT KAKIDUGEN; Inibaloi, ibl, INIBALOI; Isneg, isd, ISNEG; Itbayaten, ivv, ITBAYATEN BATANES ISLANDS; Itneg, itb, ITNEG BINONGAN; Ivatan, ivv, IVATAN BATANES ISLANDS; Kalagan, klg, KALAGAN; Kalinga, knb, KALINGA GUINAANG; Kallahan Kayapa, kak, KALLAHAN KAYAPA PROPER; Kallahan Keleyqiq, ify, KALLAHAN KELEYQIQ IFUGAO; Kankanay, xnn, KANKANAY NORTHERN; Mamanua, mmn, MAMANWA; Ata Manobo, atd, MANOBO ATA; Dibabawon Manobo, mbd, MANOBO DIBABAWON; Ilianen Manobo, mbi, MANOBO ILIANEN; Kalamsig Manobo, mta, MANOBO KALAMANSIG COTABATO; Sarangani Manobo, mbs, MANOBO SARANGANI; Tigwa Manobo, mbt, MANOBO TIGWA; Western Bukidnon Manobo, mbb, MANOBO WESTERN BUKIDNON; Mansaka, msk, MANSAKA; Siasi, sml, SAMAL; Sambal, sbl, SAMBAL BOTOLAN; Sangil, snl, SANGIL SARANGANI ISLANDS; Sangir, sxn, SANGIR; Sindangan Subanon, syb, SUBANUN SINDANGAN; Siocon Subanon, suc, SUBANON SIOCON; Tboli, tbl, TBOLI TAGABILI; Aborlan Tagbanwa, tbw, TAGBANWA ABORLAN; Kalamian tagbanwa, tbk, TAGBANWA KALAMIAN; Tausug, tsg, TAUSUG; Tagalog, tgl, TAGALOG; Cebuano, ceb, CEBUANO; Hiligaynon, hil, HILIGAYNON; Waray, war, WARAY WARAY; Ilocano, ilo, ILOKANO; Bicol, bcl, CENTRAL BICOLANO; Pampango, pam, KAPAMPANGAN; Pangasinan, pag, PANGASINA; Tagakaolo, klg, KALAGAN TAGAKAOLO; Yakan, yka, YAKAN; Sibutu, ssb, SIBUTU SOUTHERN SAMA; Kapul, abx, INABAKNON; Palun Mapun, sjm, MAPUN; Maranao, mrw, MARANAO; Tasaday, mdh, MAGUINDANAO; Kiniray'a, krj, KINARAY-A; Masbateño, msb, MASBATENYO; Sorsogonon, bks, NORTHERN SORSOGON; Butuanon, btw, BUTUANON; Hanunoo, hnn, HANUNOO; Itawes, itv, ITAWIT; Ibanag, ibg, IBANAG; Yogad, yog, YOGAD;

Aklanon, akl, AKLANON; Capiznon, cps, CAPIZNON; Cagayanzillo, cgc, KAGAYANEN; Romblonon, rol, ROMBLOMANON; Tiruray, tiy, TIRURAY; Mandaya, mry, MANDAYAN CARAGA. ***Yapen*** (Anceaux 1961): Woi, wbw, WOI; Pom, pmo, POM; Marau, alu, MARAU; Ansus, and, ANSUS; Papuma, ppm, PAPUMA; Munggui, mth, MUNGGUI; Serui Laut, seu, SERUI-LAUT; Ambai, amk, AMBAI; Wadapi-Laut, amk, WADAPI LAUT; Wabo, wbb, WABO; Kurudu, kjr, KURUDU; Wandamen, wad, WANDAMEN; Dusner, dsn, DUSNER; Ron, rnn, RON; Biak, bhw, BIAK; Waropen, wrp, WAROPEN; Mor, mhz, MOR; Irarutu, irh, IRARUTU.

Carib: ***Cariban*** (Villalon 1991): Yabarana, yar, YABARANA; Panare, pbh, ENAPA WOROMAIPU; Pemon + Kamarakoto + Taurepan, aoc, PEMON; Makushi, mbc, MACUSHI; Oayana, way, WAYANA; Carib, car, KALINA; Yukpa, yup, YUKPA; Bakairi, bkq, BAKAIRI; Makiritare, mch, MAQUIRITARI; Hianacoto-Umaua, cbd, CARIJONA.

Dravidian: ***Dravidian***(*) (Andronov 2001): Tamil, tam, TAMIL; Malayalam, mal, MALAYALAM; Kannada, kan, KANNADA; Telugu, tel, TELUGU; Kolami, kfb, NORTHWESTERN KOLAMI; Parji, pci, PARJI; Gondi, ggo, ADILABAD GONDI; Kurukh, kru, KURUKH; Malto, mjt, SAURIA PAHARIA; Brahui, brh, BRAHUI.

Hmong-Mien: ***Miao-Yao**** (Peiros 1998): Hmu, hea, HMU; Xiangxi Hmong (or Xx), mmr, XIANGXI HMONG; Hmong Njua, hnj, HMONG NJUA; Bunu, bwx, BUNU; She, shx, SHE CHINA; yao, ium, YAO.

Indo-European: ***Indo-European*** (Dyen et al. 1992): Irish, gle, IRISH GAELIC; Welsh, cym, WELSH; Breton, bre, BRETON; Rumanian, ron, ROMANIAN 2; Vlach, rup, VLACH; Italian, ita, ITALIAN; French, fra, FRENCH; Provençal, frp, ARPITAN; Spanish, spa, SPANISH; Portuguese, por, PORTUGUESE; Catalan, cat, CATALAN; German, deu, STANDARD GERMAN; Dutch, nld, DUTCH; Afrikaans, afr, AFRIKAANS; Flemish, vls, WESTVLAAMS; Frisian, fry, FRISIAN WESTERN; English, eng, ENGLISH; Takitaki, srn, SRANAN TONGO; Swedish, swe, SWEDISH; Danish, dan, DANISH; Riksmal, nob, NORWEGIAN BOKMAAL; Icelandic, isl, ICELANDIC; Faroese, fao, FAROESE; Lithuanian, lit, LITHUANIAN; Latvian, lav, LATVIAN; Lusatian L, dsb, LOWER SORBIAN 2; Lusatian U, hsb, UPPER SORBIAN; Czech, ces, CZECH; Slovak, slk, SLOVAK; Polish, pol, POLISH; Ukrainian, ukr, UKRAINIAN; Byelorusssian, bel, BELARUSIAN; Russian, rus, RUSSIAN; Bulgaria, bul, BULGARIAN; Slovenian, slv, SLOVENIAN; Serbocroatian, srp, SERBOCROATIAN; Singhalese, sin, SINHALA; Kashmiri, kas, KASHMIRI; Lahnda, pnb, WESTERN PANJABI SHAHPUR; Marathi, mar, MARATHI; Gujarati, guj, GUJARATI; Panjabi, pan, PUNJABI MAJHI; Hindi, hin, HINDI; Bengali, ben, BENGALI; Nepali, nep, NEPALI; Ossetic, oss, DIGOR OSSETIAN; Afghan, pbu, NORTHERN PASHTO;

Waziri, pst, BANNU PASHTO; Wakhi, wbl, CENTRAL GOJAL WAKHI; Persian, pes, PERSIAN; Tadzik, tgk, TAJIK; Greek, ell, GREEK; Armenian, hye, WESTERN ARMENIAN; Albanian T, als, ALBANIAN TOSK.

Japonic: ***Japonic**** (Hattori 1961): Tokyo, jpn, TOKYO JAPANESE; Kyoto, jpn, JAPANESE KYOTO; Naha, ryu, NAHA; Shuri, ryu, SHURI; Yonamine, xug, YONAMINE.

Mayan: ***Mayan**** (Cecil H. Brown p.c., 2011): CHICOMUCELTEC, cob; SOUTHERN CAKCHIQUEL SAN ANDRES ITZAPA, ckf; jai, JACALTEC; qum, SIPAKAPENSE; kjb, QANJOBAL SANTA EULALIA; toj, TOJOLABAL; usp, USPANTEKO; ctu, CHOL TUMBALA; mhc, MOCHO; poa, POCOMAM EASTERN; quc, CENTRAL QUICHE; kek, EASTERN KEKCHI CAHABON; ixj, IXIL CHAJUL; mop, MOPAN; quv, SACAPULTECO SACAPULAS CENTRO; ttc, TECO TECTITAN; tzj, TZUTUJIL SAN JUAN LA LAGUNA; knj, ACATECO SAN MIGUEL ACATAN; caa, CHORTI; mam, MAM NORTHERN; pob, POQOMCHI WESTERN; tzb, TZELTAL BACHAJON; agu, AGUACATEC; chf, CHONTAL TABASCO; cnm, CHUJ; lac, LACANDON; tzz, ZINACANTAN TZOTZIL; hva, HUASTEC; itz, ITZAJ; yua, MAYA YUCATAN.

Macro-Ge: ***Ge*** (Wilbert 1962): Apinaye, apn, APINAYE; Creye, xre, KREYE; Canela, ram, APANIEKRA; Craho, xra, KRAHO; Pucobye, gvp, PYKOBJE; Suya, suy, SUYA; Cayapo, txu, KAYAPO; Shavante, xav, XAVANTE; Sherente, xer, XERENTE.

Mixe-Zoque: ***Mixe-Zoque**** (Cysouw et al. 2006): North Highland Mixe, mto, NORTH HIGHLAND MIXE; South Highland Mixe, mxp, SOUTH HIGHLAND MIXE; Lowland Mixe, mco, LOWLAND MIXE; Sayula Popoluca, pos, SAYULA POPOLUCA; Oluta Popoluca, plo, OLUTA POPOLUCA; Texistepec Zoque, poq, TEXISTEPEC ZOQUE; Soteapan Zoque, poi, SOTEAPAN ZOQUE; Santa Maria Chimalapa Zoque, zoh, MARIA CHIMALAPA; San Miguel Chimalapa Zoque, zoh, MIGUEL CHIMALAPA; Chiapas Zoque, zoc, CHIAPAS ZOQUE;

Na-Dene: ***Athapaskan**** (Hoijer 1956): Hare, scs, HARE; Chipewyan, chp, CHIPEWYAN; Beaver, bea, BEAVER; Carrier, crx, CARRIER; Kutchin, gwi, KUTCHIN; Sarcee, srs, SARCEE; Galice, gce, GALICE; Kato, ktw, KATO; Mattole, mvb, MATTOLE; Hupa, hup, HUPA 2; Navaho, nav, NAVAHO; San Carlos, apw, SAN CARLOS; Chircahua, apm, CHIRICAHUA; Jicarilla, apj, JICARILLA; Lipan, apl, LIPAN.

Niger-Congo: ***Atlantic*** (Sapir 1971): Fula, fuc, FULA; Wolof, wol, WOLOF; Serer, srr, SERER SINE; Lehar, cae, LEHAR; Safen, sav, SAFEN; Non, snf, NON; Ndut, ndv, NDUT FALOR; Fogny, dyo, JOLA; Manjaku, mfv, MANJACA CHURO;

Papel, pbo, PAPEL; Balanta, ble, BALANTA; Biafada, bif, BIAFADA; Pajade, pbp, PAJADE; Nalu, naj, NALU; Bijago, bjg, BIJOGO; Temne, tem, TEMNE; Mmani, buy, MMANI; Sherbro, bun, SHERBRO; Krim, krm, KRIM; Kisi, kqs, KISSI; Gola, gol, GOLA. ***Benue-Congo*** (Bennett and Sterk 1977): Nupe, nup, NUPE; Gade, ged, GADE; Igbira, igb, IGBIRRA; Idoma, idu, IDOMA; Eloyi, afo, ELOYI; Igbo, ibo, IGBO ONITSHA; Igala, igl, IGALA; Yoruba, yor, YORUBA; Ora, ema, EMAI; Bini, bin, EDO; Urhobo, urh, URHOBO; Isoko, iso, ISOKO; Degema, deg, DEGEMA 2; Aten, etx, ITEN; Mambila, mzk, MAMBILA; Tiv, tiv, TIV 2; Tunen, baz, TUNEN; Jarawa, jar, BANKALA; Bobangi, bni, BOBANGI; Nyanja, nya, NYANJA; Kikuyu, kik, GIKUYU; Kwanyama, kua, KWANYAMA. ***Gur*** (Swadesh et al. 1966): Basal, bud, BASSARI; Konkomba, xon, KONKOMBA 2; Gurma, gux, GOURMANCHEMA; Pilapila, pil, YOM; Naudem, nmz, NAWDM; Buli, bwu, BULI GHANA; Dagbani, dag, DAGBANI; Mampruli, maw, MAMPRULI; Kusal, kus, KUSAL; Hanga, hag, HANGA; Frafra, gur, NINKARE; Moore, mos, MOORE; Dagaari, dga, DAGAARE; Vagala, vag, VAGALA; Sisala, sil, SISAALA TUMULUNG; Kasem, xsm, KASEM; Lamba, las, LAMA; Kabre, kbp, KABIYE; Mambar, myk, MAMARA SENOUFO; Pantera + Fantera, nfr, NAFAARA. ***Kwa*** (Heine 1968): Twi, aka, TWI ASANTE; Logba, lgq, IKPANA; Adele, ade, ADELE; Lipke, lip, LIKPE; Santroko, snw, SELE; Akpufu, akp, AKPAFU; Lelemi, lef, BUEM LELEMI; Avatime, avn, SIDEME; Nyangbo, nyb, TUTRUGBU; Bowili, bov, TUWULI; Ahlo, ahl, AHLO; Animere, anf, ANIMERE; Ewe, ewe, EWE ADANGBE.

Nilo-Saharan: ***Eastern Sudanic*** (Thelwall 1981): Meidob, mei, MEIDOB NUBIAN; Debri, dil, DILLING; Dongolawi, kzh, NUBIAN OF DONGOLA; Nobiin, fia, NOBIIN; Gaam, tbi, INGASSANA; Liguri, liu, LOGORIK; Shatt, shj, SHATT; Nyala + Lagowa, daj, NYALA; Sila, dau, SILA; Temein, teq, TEMEIN; Dinka, dik, REK; Shilluk, shk, SHILLUK; ***Nilo-Saharan***$^{(*)}$ (Bender 1971): Nuer, nus, NUER; Anyuak, anu, ANYUAK; Shilluk, shk, SHILLUK; Jumjum, jum, JUMJUM; Mabaan, mfz, MABAAN; Burun, bdi, BURUN; Inyangatom, nnj, NYANGATOM 2; Tirma, suq, TIRMA; Mursi, muz, MURSI; Meen, mym, MEEN; Kwegu, xwg, KWEGU; Zilmamu, koe, BAALE; Murle, mur, MURLE; Mesengo, mpe, MESENGO; Nara, nrb, NARA; Ingassana, tbi, INGASSANA; Kunama, kun, KUNAMA; Wetawit, wti, WETAWIT; Uduk, udu, UDUK; C. Koma, xom, CENTRAL KOMA; Langa, lgn, LANGA; N. Koma, kmq, GWAMA; Gumuz, guk, GUMUZ; ***Southern Luo**** (Blount and Curley 1970): Lango, laj, LANGO; Acholi, ach, ACHOLI 2; Alur, alz, ALUR; Luo, luo, LUO; Shilluk, shk, SHILLUK.

Quechuan: ***Quechua*** (Torero 1970): Corongo, qwa, YANAC; Caras, qwh, QUECHUA HUAYLAS ANCASH; Tarma, qvn, QUECHUA NORTH JININ; Ferreñafe, quf, INKAWASI; Cajamarca, qvc, CHETILLA; Chachapoyas, quk, QUECHUA CHACHAPOYAS; Ayachuco, quy, QUECHUA AYACUCHO; Cuzco, quz, QUECHUA DE CUSCO; Potosí + Chuquisaca, quh, MARAGUA.

Salishan: ***Salish*** (Swadesh 1950): Bella Coola, blc, BELLA COOLA; Comox, coo, SLIAMMON; Seshelt, sec, SECHELT; Fraser + Nanaimo, hur, COWICHAN; Squamish, squ, SQUAMISH; Lkungen + Lummi, str, SALISH STRAITS; Clallam, clm, CLALLAM; Nootsak, nok, NOOKSACK; Twana, twa, TWANA; Cowlitz, cow, COWLITZ; Chehalis + Satsop, cjh, CHEHALIS UPPER; Quinault, qun, QUINAULT; Tillamook, til, TILLAMOOK; Lillooet, lil, LILLOOET; Thompson, thp, THOMPSON; Shuswap, shs, SHUSWAP; Okanagan, oka, OKANAGAN COLVILLE; Spokane, spo, SPOKANE; Kalispel + Pend d’Oreille, fla, KALISPEL-PEND DOREILLE; Columbia, col, COLUMBIA WENATCHI; Coeur d’Alène, crd, COEUR DALENE.

Sino-Tibetan: ***Chinese*** (Xu 1991): Xiamen, nan, AMOY MINNAN CHINESE; Meixian, hak, HAKKA; Guangzhou, yue, CANTONESE; Changsha, hsn, XIANG; Suzhou, wuu, SUZHOU WU; Beijing, cmn, MANDARIN 2. ***Lolo-Burmese**** (Peiros 1998): Burmese, mya, BURMESE; Zaiwa, atb, ZAIWA; Achang, can, ACHANG; Nusu, nuf, NUSU; Akha, ahk, AKHA; Biyue, byo, BIYUE; Lahu, lhu, LAHU; Jino, jiu, JINO; Mpi, mpz, MPI; Bisu, bzi, BISU; Xide, iii, XIDE; Dafang, yig, DAFANG; Nanjiang, ywt, NANJIANG; Lisu, lis, LISU; Naxi, nbf, NAXI. ***Sino-Tibetan*** (Benedict 1976): Burmese, mya, BURMESE; Tibetan, bod, TIBETAN LHASA; Lushai, lus, LUSHAI; Kachin, kac, JINGPHO; Garo, grt, GARO; Mandarin, cmn, MANDARIN 2.

Tai-Kadai: ***Kadai**** (Peiros 1998): Siamese, tha, SIAMESE; Longzhou, zzj, ZHUANG SOUTHERN; Zhuang, zyb, ZHUANG NORTHERN; Saek, skb, SAEK; Ong Be, onb, ONG BE; Lakkja, lbc, LAKKJA; Mulao, mlm, MULAO; Kam, kmc, SOUTHERN DONG; Maonan, mmd, MAONAN; Sui, swi, SUI.

Torricelli: ***Kamasau**** (Sanders and Sanders 1980): Tring, kms, TRING; Wau, kms, WAU; Kamasau, kms, KAMASAU; Yibab, kms, YIBAB; Wandomi, kms, WANDOMI; Kenyari, kms, KENYARI; Paruwa, kms, PARUWA; Samap, kms, SAMAP.

Trans-New Guinea: ***Angan**** (Lloyd 1973): Angaataha, agm, ANGAATAHA; Ankave, aak, ANKAVE; Ampale, apz, AMPALE; Baruya, byr, BARUYA 2; Ivori, ago, IVORI; Kamasa, klp, KAMASA; Kapau, hmt, KAPAU; Kawacha, kcb, KAWACHA; Lohiki, miw, LOHIKI; Menya, mcr, MENYA 2; Simbari, smb, SIMBARI; Yagwoia, ygw, YAGWOIA. ***Awyu***$^{(*)}$ (Voorhoeve 1968): Aghu, ahh, AGHU; Kaeti, bwp, KAETI; Pisa, psa, PISA; Syiagha, aws, SIAGHA; Yenimu, awy, YENIMU; Wambon, wms, WAMBON. ***Bosavi*** (Shaw 1986): Duna, duc, DUNA; Bimin, bhl, BIMIN; Bogaia, boq, BOGAYA; Pare, ppt, PARE; Agala, agl, AGALA; Kubo, jko, KUBO; Samo, smq, SAMO; Bibo, goi, GEBUSI; Honibo, goi, HONIBO; Oibae, goi, OIBAE; Kalamo, kkc, ODOODEE; Bedamini, beo, BEDAMINI; Etoro, etr, ETORO; Onabasulu, onn, ONABASULU; Kaluli, bco, KALULI 2; Sunia, siq, SUNIA; Kasua, khs, KASUA 2; Aimele, ail, AIMELE;

Kamula, xla, KAMULA; Bainapi, dby, DIBIYASO; Namumi, faa, NAMUMI; Bamu, bcf, BAMU 2. ***Eleman**** (Brown 1973): Aheave, xeu, AHEAVE; Kaipi, oro, KAIPI; Keuru, xeu, KEURU; Opao, opo, OPAO; Orokolo, oro, OROKOLO 2; Sepoe, tqo, SEPOE; Toaripi, tqo, TOARIPI 2; Uaripi, uar, UARIPI. ***Finisterre*** (Claasen and McElhanon 1970): Nankina, nnk, NANKINA; Awara, awx, AWARA; Wantoat, wnc, WANTOAT; Nek, nif, NEK; Yabong, ybo, YABONG; Saep, spd, SAEP; Ganglau, ggl, GANGLAU; Kolom, klm, KOLOM; Suroi, ssd, SUROI; Lemio, lei, LEMIO; Usino, urw, USINO; Sinsauru, snz, SINSAURU. ***Grand Valley Dani**** (Bromley 1967): Upper Pyramid, dni, UPPER PYRAMID DANI; Pyramid-Wodo, wlw, PYRAMID WODO; Mid-Grand Valley, dnt, MID GRAND VALLEY DANI; Lower Valley Hitigama, dni, HITIGIMA DANI; Lower Valley Tangma, dni, TANGMA DANI; Jalimo Angguruk, yli, ANGGURUK YALI; Kiniageima Amo, wul, KINIAGEIMA. ***Greater Madang**** (Z'Graggen 1969): Isebe, igo, ISEBE; Bau, bbd, BAU; Amele, aey, AMELE; Garus, gyb, GARUS; Yoidik, ydk, YOIDIK; Rempi, rmp, REMPI; Garuh, gaw, GARUH; Foran, fad, KAMBA; Mawan, mcz, MAWAN; Utu, utu, UTU; Saruga, sra, SARUGA; Kare, kmf, KARE; Usino, urw, USINO; Sumau, six, SUMAU; Urigina, urg, URIGINA; Korak, koz, KORAK; Waskia, wsk, WASKIA; Malas, mkr, MALAS; Bunabun, buq, BUNABUN; Dimir, dmc, DIMIR; Pay, ped, PAY; Pila, sks, PILA; Saki, sks, SAKI; Tani, pla, TANI; Ulingan, mhl, ULINGAN; Bepour, bie, BEPOUR; Mawak, mjj, MAWAK; Musar, mmi, MUSAR; Wanambre, wnb, WANAMBRE; Wanuma, wnu, WANUMA; Yaben, ybm, YABEN; Parawen, prw, PARAWEN; Amaimon, ali, AMAIMON; Moresada, msx, MORESADA; Ikundun, imi, IKUNDUN; Pondoma, pda, PONDOMA; Wanambre, wnb, WANAMBRE; Katiati, kqa, KATIATI; Osum, omo, OSUM; Atemple, ate, ATEMPLE; Angaua, anh, ANGAUA; Emerum, ena, EMERUM; Musak, mmq, MUSAK; Paynamar, pmr, PAYNAMAR; Kaian, kct, KAIAN; Gamei, gai, GAMEI; Mikarew, msy, MIKAREW MAKARUB; Anor, anj, ANOR; Rao, rao, RAO; Banaro, byz, BANARO. ***Huon**** (McElhanon 1967): Kâte, kmg, KATE; Dedua, ded, DEDUA; Mape, mlh, MAPE 2; Hube, kgf, HUBE; Tobo, tbv, TOBO; Kosorong, ksr, KOSORONG; Mindik, bmu, MINDIK; Burum, bmu, BURUM; Ono, ons, ONO; Komba, kpf, KOMBA; Selepet, spl, SELEPET; Timbe, tim, TIMBE; Nabak, naf, NABAK; Momolili, mci, MOMOLILI. ***Kiwaian*** (Wurm 1973): Wabuda, kmx, WABUDA; Middle Bamu Kiwai, bcf, BAMU; Morigi, mdb, MORIGI; Kerewo, kxz, KEREWO; Urama, kiw, URAMA; Gope, kiw, GOPE; Gibaio, kiw, GIBAIO; Arigibi / Anigibo / Anigibi / Ani, kiw, ANIGIBI. ***Koiarian*** (Dutton 1969): Koita, kqi, KOITA; Koiari, kbk, KOIARI 2; MtnKoiari, kpx, MOUNTAIN KOIARI; Aomie, aom, AOMIE; Barai, bbb, BARAI; Managalasi, mcq, ESE MANAGALASI. ***Kolopom*** (Voorhoeve 1968): Kimaghana, kig, KIMAGHAMA; Riantana, ran, RIANTANA; Ndom, nqm, NDOM. ***Ok*** (Voorhoeve 1968): Asmat, cns, ASMAT CENTRAL; Telefol, tlf, TELEFOL; Kati, yon, NORTH KATI; Aghu, ahh, AGHU; Mombum, mso, MOMBUN; ***Turama-Kikorian*** (Franklin 1973): Omati, mgx, OMATI; Ikobi, meb, IKOBI; Mena, meb, MENA; Kairi, klq, RUMU.

Uto-Aztecan: ***Uto-Aztecan*** (Miller 1984, Cortina-Borja and Valiñas 1989): Northern Paiute, pao, NORTHERN PAIUTE; Panamint, par, PANAMINT; Shoshoni, shh, SHOSHONI; Comanche, com, COMANCHE; Kawaiisu, xaw, KAWAIISU; Southern Paiute, ute, SOUTHERN PAIUTE; Ute, ute, UTE 2; Tübatulabal, tub, TUBATULABAL; Cahuilla, chl, CAHUILLA; Cupeno, cup, CUPENO; Luiseno, lui, LUISENO; Hopi, hop, HOPI; Papago, ood, TOHONO OODHAM; Nevome, ood, UPPER PIMA; Northern Tepehuan, ntp, NORTHERN TEPEHUAN; Guarijio, var, WARIHIO; Tarahumara, tar, CENTRAL TARAHUMARA; Opata + Eudeve, opt, OPATA; Mayo, mfy, MAYO; Yaqui, yaq, YAQUI; Tubar, tbu, TUBAR; Huichol, hch, HUICHOL; Cora, crn, EL NAYAR CORA; Tetelcingo Nahuatl, nhg, TETELCINGO NAHUATL; Zacapoaxtla Nahuatl, azz, HIGHLAND PUEBLA NAHUATL; Pipil, ppl, PIPIL.

West Papuan: ***North Halmahera*** (Chlenov 1986): Loda, loa, LODA; Galela, gbi, GALELA; Tobelo, tlb, TOBELO; Tabaru / Tobaru, tby, TABARU; Pagu / Isam, pgu, PAGU; Madole / Modole, mqo, MADOLE; Sahu, saj, SAHU; Tidore, tvo, TIDORE. ***Yawa**** (Jones 1986): Tindaret, yva, TINDARET; Ambaidiru, yva, AMBADAIRU; Ariepi, yva, ARIEPI; Sarawandori, yva, SARAWANDORI; Konti Unai, yva, KONTI UNAI; Mariadei, yva, MARIADEI.

Notes

1. We thank Cecil H. Brown and Robert Mailhammer for providing cognate judgments for Mayan and Iwaidjan, respectively. An earlier version of this paper was presented at the ICHL in Osaka, July 2011. We are grateful to Claire Bowern for highly useful comments on that occasion.
2. The normalization leading to LDN is argued by Serva and Petroni (2008) to be absolutely necessary for arriving at good results for language classification. To our knowledge, other types of normalization have not been tested for string comparisons involving languages that are not closely related, analyses having been limited to the field of dialectology (Heeringa 2005). This is a potentially interesting item for future research.
3. This variation in lists would be expected if anything to increase the variability of COGNsim relative to ASJPim, which is always based on the same list. The additional variability would tend to weaken the observed correlations, thus rendering our tests conservative.
4. An exception is the name ‘Greater Madang’, which we use as a collective term for the following Trans-New Guinea groups included in Z’Graggen (1969): Madang, Isumrud, Kaukombaran, Mawamuan, Pihom, Josephstaal, and Wanang.
5. An exception where we provide cognate judgments ourselves is the Torricelli group Kamasau. This is a set of dialects for which it requires no special expertise

to distinguish cognates from non-cognates. In the case of Mixe-Zoque we used the judgments of Cysouw et al. (2006), but corrected for some typos in that paper.

6. Not surprisingly, the magnitude of the correlation between COGNsim and ASJPsim is greater when both measures are based on the same data set. In the appendix we indicate for each language group whether the data sets for the two measures are the same, almost so, or not. Families containing language groups where the data sets are all the same or almost so include AA, Alt, AuA, Dra, HM, Jap, May, MZ, NDe, TK, and Tor. The average *r* for these families is .81. Families containing groups where the data sets are all different include Car, IE, MGe, NC, Que, Sal, and UA. The average *r* for these is .70. The families Aus, NS, ST, TNG, and WP are represented by language groups for which the sources are mixed, some data sets being the same or almost so and some not. Average *r* for these families is .77.
7. It was observed by a referee that if mmWL correlates significantly with both mDIFF and mRATIO, while mSR does not, then mSR and mmWL perhaps do not correlate significantly, which would run counter to previous observations about an inverse relationship between word length and segment inventory sizes (Nettle 1995, 1998; Wichmann et al. 2011). The data used in the present paper are limited, so differences from the results of Wichmann et al. (2011) with regard to word length and segment inventory size are entirely expected and not particularly telling. The correlation in the present data is $r = -.38$, $p = .07$. These figures depend highly on the small number of data points, as can be seen by an increase in the correlation to $r = -.74$, $p < .0001$ with the removal of a single outlier, Salishan.

References

Anceaux, Johannes Cornelis
1961 *The Linguistic Situation in the Islands of Yapen, Kurudu, Nau and Miosnum, New Guinea*. Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde. 'S-Gravenhage: Martinus Nijhoff.

Andronov, Mikhail
2001 *Dravidian Historical Linguistics*. München: Lincom Europa.

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill, and Mark Pagel
2008 Languages evolve in punctuational bursts. *Science* 319: 588.

Bastin, Yvonne, A. Coupez, and Michael Mann
1999 *Continuity and Divergence in the Bantu Languages. Perspectives from a Lexicostatistic Study.* Tervuren: Musée Royal de l'Afrique Centrale.

Bender, M. L.
1971 The Languages of Ethiopia. *Anthropological Linguistics* 13: 165–288.

Benedict, Paul K.
1976 Sino-Tibetan: Another look. *Journal of the American Oriental Society* 96: 167–197.

Bennett, Patrick R. and Jan P. Sterk
1977 South Central Niger-Congo: A reclassification. *Studies in African Linguistics* 8: 241–273.

Blount, Ben and Richard T. Curley
1970 The Southern Luo languages: A glottochronological reconstruction. *Journal of African Languages* 9: 1–18.

Breen, Gavan
1981 *The Mayi Languages of the Queensland Gulf Country*. Canberra: Australian Institute of Aboriginal Studies.

Bromley, H. Myron
1967 The linguistic relationships of Grand Valley Dani: A lexico-statistical classification. *Oceania* 37: 286–308.

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: 285–308.

Brown, Herbert A.
1973 The Eleman language family. In *The Linguistic Situation in the Gulf District and Adjacent Areas, Papua New Guinea*, Franklin, Karl J. (ed.), 281–376. (Pacific Linguistics Series C – N. 26.) Canberra: Australian National University.

Chadwick, Neil
1979 The West Barkly languages: an outline sketch. In *Australian Linguistic Studies*, Stephen A. Wurm (ed.), 653–711. (Pacific Linguistics, Series C, No. 54.) Canberra: Australian National University.

Chlenov, M. A.
1986 North Halmahera languages: A problem of internal classification. *Papers in New Guinea Linguistics* 24. Pacific Linguistics, A-70, 39–44. Canberra: Australian National University.

Claassen, Oren R. and Kenneth A. McElhanon
1970 Languages of the Finisterre Range, New Guinea. Pacific Linguistics A-23: 45–78. Canberra: Australian National University.

Cortina-Borja, Mario and Leopoldo Valiñas C.
1989 Some remarks on Uto-Aztecan classification. *International Journal of American Linguistics* 55: 214–239.

Cysouw, Michael, Søren Wichmann, and David Kamholz
2006 A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean. *Journal of Quantitative Linguistics* 13: 225–264.

Dryer, Matthew S.
1992 The Greenbergian word order correlations. *Language* 68: 81–138.

Dryer, Matthew S.
2011 Genealogical language list. In *World Atlas of Language Structures Online*, ed. Matthew S. Dryer, and Martin Haspelmath. Munich: Max Planck Digital Library, chapter iv. Available online at http://wals.info/chapter/iv.

Dutton, Thomas Edward
1969 *The Peopling of Central Papua. Some Preliminary Observations*. Pacific Linguistics, Series B – Monographs, No 9. Canberra: Australian National University.

Dyen, Isidore
1965 A *Lexicostatistical Classification of the Austronesian Languages*. Supplement to International Journal of American Linguistics, Vol. 31, No. 1. Baltimore: Waverly Press.

Dyen, Isidore, Joseph Kruskal, and Paul Black
1992 An Indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82.5.

Franklin, Karl J.
1973 Other language groups in the Gulf district and adjacent areas. In *The Linguistic Situation in the Gulf District and Adjacent Areas, Papua New Guinea*, Karl J. Franklin (ed.), 263–277. (Pacific Linguistics Series C – N. 26.) Canberra: Australian National University.

Gray, Russell D. and Quentin Atkinson
2003 Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature* 426: 435–439.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill
2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323: 479–483.

Gray, Russell D. and Fiona M. Jordan
2000 Language trees support the express-train sequence of Austronesian expansion. *Nature* 405: 1052–1055.

Greenhill, Simon J.
2011 Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*. doi: 10.1162/COLI_a_00073.

Greenhill, Simon J., Robert Blust, and Russell D. Gray
2008 The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4: 271–283

Gudschinsky, Sarah
1956 The ABCs of lexicostatistics (glottochronology). *Word* 12: 175–210.

Hattori, Shiro
1961 A glottochronological study on three Okinawan dialects. *International Journal of American Linguistics* 27: 52–62.

Hay, Jennifer and Laurie Bauer
2007 Phoneme inventory size and population size. *Language* 83: 388–400.

Heeringa, Wilbert
2004 Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. diss., Rijksuniversiteit Groningen.

Heine, Bernd
1968 *Die Verbreitung und Gliederung der Togorestsprachen*. Berlin: Dietrich Reimer.

Hoijer, Harry
1956 The chronology of the Athapaskan languages. *International Journal of American Linguistics* 22: 219–232.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov
2011 Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52: 841–875.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker
2008a Advances in automated language classification. In *Quantitative Investigations in Theoretical Linguistics*, Antti Arppe, Kaius Sinnemäki, and Urpu Nikanne (eds.), 40–43. Helsinki: University of Helsinki.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Pamela Brown, and Dik Bakker
2008b Explorations in automated language classification. *Folia Linguistica* 42: 331–354.

Hooley, Bruce A.
1971 Austronesian languages of the Morobe District, Papua New Guinea. *Oceanic Linguistics* 10: 79–151.

Huff, Paul and Deryle Lonsdale
2011 Positing language relationships using ALINE. *Language Dyna-mics and Change* 1: 128–162.

Jones, Larry B.
1986 The dialects of Yawa. *Papers in New Guinea Linguistics* 25. Pacific Linguistics, A-74, 31–68. Canberra: Australian National University.

Kondrak, Grzegorz
2000 A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, 288–295.

Lees, Robert B.
1953 The basis of glottochronology. *Language* 29: 113–127.

Lewis, M. Paul (ed.)
2009 *Ethnologue. Languages of the World*. 16th ed. Dallas: SIL International. Online version: http://www.ethnologue.com.

Llamzon, Teodoro and Teresita Martin
1976 A subgrouping of 100 Philippine languages. In *South-East Asian Linguistic Studies*, Vol. 2, Nguyen Dang Liem (ed.), 141–172. (Pacific Linguistics, Series C, No. 42.) Canberra: Australian National University.

Lloyd, Richard G.
1973 The Angan language family. In: *The Linguistic Situation in the Gulf District and Adjacent Areas, Papua New Guinea*, Karl J. Franklin (ed.), 31–110. (Pacific Linguistics Series C – N. 26.) Canberra: Australian National University.

Mace, Ruth and Fiona M. Jordan
2011 Macro-evolutionary studies of cultural diversity: A review of empirical studies of cultural transmission and cultural adaptation. *Philosophical Transactions of the Royal Society B* 366: 402–411.

McElhanon, Kenneth A.
1967 Preliminary observations on Huon Peninsula languages. *Oceanic Linguistics* 6: 1–45.

McGregor, William B. and Alan Rumsey
2009 *Worrorran Revisited: The Case for Genetic Relations among Languages of the Northern Kimberley Region of Western Australia*. (Pacific Linguistics.) Canberra: Australian National University.

Militarev, Alexander
2000 Towards the chronology of Afrasian (Afroasiatic) and its daughter families. In *Time Depth in Historical Linguistics*, Vol. 1, Colin Renfrew, April McMahon, and Larry Trask (eds.), 267–307. Cambridge: McDonald Institute for Archaeological Research.

Miller, Wick R.
1984 The classification of the Uto-Aztecan languages based on lexical evidence. *International Journal of American Linguistics* 50: 1–24.

Nettle, Daniel
1995 Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33: 359–367.

Nettle, Daniel
1998 Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics* 5: 240–245.

O’Grady, Geoffrey N.
1966 Proto-Ngayarda phonology. *Oceanic Linguistics* 5: 71–130.

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade
2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449: 717–720.

Peiros, Ilya

1998 *Comparative Linguistics in Southeast Asia*. (Pacific Linguistics Series C, 142.) Canberra: Australian National University.

Pompei, Simone, Vittorio Loreto, and Francesca Tria

2011 On the accuracy of language trees. *PLoS One* 6.6, e20109.

Sanders, Joy and Arden G. Sanders

1980 Dialect Survey of the Kamasau Language. *Papers in New Guinea Linguistics*, No. 20. (Pacific Linguistics Series A, No. 56.) Canberra: Australian National University.

Sapir, J. David

1971 West Atlantic: An inventory of the languages, their noun class systems and consonant alternation. In *Current Trends in Linguistics*, Vol. 7: *Linguistics in Sub-Saharan Africa*, Thomas A. Sebeok (ed.), 45–112. The Hague: Mouton.

Serva, Maurizio and Filippo Petroni

2008 Indo-European languages tree by Levenshtein distance. *EuroPhysics Letters* 8: 68005.

Shaw, R. Daniel

1986 The Bosavi language family. *Papers in New Guinea Linguistics* 24: 45–76. (Pacific Linguistics A-70). Canberra: Australian National University.

Sommer, Bruce A.

1976 Umbuygamu: The classification of a Cape York Peninsular language. *Papers in Australian Linguistics* 10: 13–31. (Pacific Linguistics Series A, No. 47.) Canberra: Australian National University.

Swadesh, Mauricio and Evangelina Arana, with John T. Bendor-Samuel and W. A. A. Wilson

1966 A preliminary glottochronology of Gur languages. *Journal of West African Languages* 3: 27–65.

Swadesh, Morris

1950 Salish internal relationships. *International Journal of American Linguistics* 16: 157–167.

Swadesh, Morris

1952 Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96: 452–463.

Swadesh, Morris

1955 Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121–137.

Thelwall, Robin

1981 Lexicostatistical subgrouping and lexical reconstruction of the Daju group. In *Nilo-Saharan: Proceedings of the First Nilo-Saharan Linguistics Colloquium, Leiden, September 8–10, 1980*, Thilo C. Schadeberg and M. Lionel Bender (eds.), 167–184. Dordrecht: Foris Publications.

Torero, Alfredo
1970 Lingüística e historia de la sociedad andina. *Anales Científicos de la Universidad Agraria* 8: 231–264.

Troike, Rudolph C.
1969 The glottochronology of six Turkic languages. *International Journal of American Linguistics* 35: 183–191.

Tryon, Darrell T.
1973 Linguistic subgrouping in the New Hebrides: A preliminary approach. *Oceanic Linguistics* 12: 303–351.

Tryon, Darrell T.
1974 *Daly Family Languages, Australia*. (Pacific Linguistics Series C, No. 32.) Canberra: Australian National University.

Vérin, Pierre, Conrad P. Kottak, and Peter Gorlin
1969 The glottochronology of Malagasy speech communities. *Oceanic Linguistics* 8: 26–83.

Villalón, María Eugenia
1991 A spatial model of lexical relationships among fourteen Cariban varieties. In *Language Change in South American Indian Languages*, Mary Ritchie Key (ed.), 54–94. Philadelphia: University of Pennsylvania Press.

Voorhoeve, C. L.
1968 The Central and South New Guinea Phylum. In *Papers in New Guinea Linguistics*, No. 8, C. L. Voorhoeve, Karl J. Franklin, and G. Scott (eds.), 1–17. (Pacific Linguistics, Series A – No. 16.) Canberra: Australian National University.

Walker, Robert S. and Lincoln A. Ribeiro
2011 Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B*. doi: 10.1098/rspb.2010 .2579.

Wichmann, Søren and Eric W. Holman
2009 Population size and rates of language change. *Human Biology* 81: 259–274.

Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown
2010a Evaluating linguistic distance measures. *Physica A* 389: 3632–3639.

Wichmann, Søren, Eric W. Holman, André Müller, Viveka Velupillai, Johann-Mattis List, Oleg Belyaev, Matthias Urban, and Dik Bakker
2010b Glottochronology as a heuristic for genealogical language relationships. *Journal of Quantitative Linguistics* 17: 303–316.

Wichmann, Søren, André Müller, and Viveka Velupillai
2010c Homelands of the world's language families: A quantitative approach. *Diachronica* 27 (2): 247–276.

Wichmann, Søren, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Julia Bishoffberger, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Harald Hammarström, Agustina Carrizo, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, and Anthony Grant
2012 The ASJP database (version 15). http://email.eva.mpg.de/ ~wichmann/languages.htm.

Wichmann, Søren, Taraka Rama, and Eric W. Holman
2011 Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15: 177–197.

Wichmann, Søren, Dietrich Stauffer, Christian Schulze, and Eric W. Holman
2008 Do language change rates depend on population size? *Advances in Complex Systems* 11: 357–369.

Wilbert, Johannes
1962 *Material lingüístico ye*. Caracas: Editorial Sucre.

Wurm, Stephen A.
1973 The Kiwaian language family. In *The Linguistic Situation in the Gulf District and Adjacent Areas, Papua New Guinea*, Karl J. Franklin (ed.), 217–260. (Pacific Linguistics Series C – N. 26.) Canberra: Australian National University.

Xu, Tongjiang (Hsu tong chiang)
1991 *Lishi yuyanxue* [Historical linguistics]. Beijing: Shangwu Yingshuguan.

Z'Graggen, John Anton
1969 Classificatory and typological studies in languages of the Western Madang district, New Guinea. Ph.D. Diss., Australian National University.

Zipf, George K.
1935 *Psycho-Biology of Language*. Boston, MA.: Houghton Mifflin.