Automated Dating of the World's Language Families based on Lexical Similarity

Eric W. Holman, 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

> The greater the degree of linguistic differentiation within a stock, the greater is the period of time that must be assumed for the development of such differentiations. (Edward Sapir[1])

This paper describes a computerized alternative to glottochronology for estimating time elapsed since parent languages diverged into daughter languages. The method, developed by the ASJP Consortium, is different from glottochronology in four major respects: (1) it is automated and thus is more objective; (2) it applies a uniform analytical approach to a single database of worldwide language coverage; (3) it is based on lexical similarity as determined from Levenshtein (edit) distances, rather than on cognate percentages; and (4) it provides a formula for date calculation that recognizes mathematically the lexical heterogeneity of individual languages, including parent languages just before their breakup into daughter languages. Automated judgments of lexical similarity for groups of related languages are calibrated with historical, epigraphic, and archaeological divergence dates for 52 language groups. The discrepancies between estimated and calibration dates are found to be on average 29% as large as the estimated dates themselves, a figure that does not differ significantly among language families. As a resource for further research that may require dates of known level of accuracy, we offer a list of ASJP time depths for nearly all the world's recognized language families and many subfamilies.

Introduction

Glottochronology, as formulated by Morris Swadesh (1950, 1955), is a method for estimating the amount of time elapsed since phylogenetically related languages diverged from a common ancestral language. This approach involves determining the percentage of words that are cognate in a standard list of basic vocabulary. Working with the assumption that words for items on such a list are replaced in individual languages at a more-or-less constant rate over time, Swadesh devised a formula, using cognate percentage as input, for calculating the length of time since a language divergence occurred.

Glottochronology has had a chequered history since its formulation some sixty years ago. An early review by Hymes (1960) is generally favorable. Later, Embleton (1986) provides a judicious summary of both positive and negative views. More recently, the pros and cons of the method are discussed in numerous chapters of a collection edited by Renfrew, McMahon, and Trask (2000). We do not intend to continue the debate on the theoretical merits and demerits of

glottochronology. Instead, we describe a new approach that infers language divergence from lexical similarity without the protracted linguistic analysis required for cognate identification.

There are several distinct processes that can cause lexical similarity among genetically related languages to diminish with the passage of time. One process is systematic change in sounds whereby commonly inherited words in related languages become phonologically different. Other processes involve replacement of words by totally different words for the same referents. Such replacement may be due to conditions internal to individual languages such as processes of semantic change; or it may be due to the borrowing of a word from one language into another where it is then used as a substitute for a native word. If two languages copy the same word from a third source, then borrowing may actually increase the similarity between the two languages. For lexical similarity to be useful for dating, the net effect of all these processes must be to reduce similarity at an approximately constant rate through time.

The present paper describes a large-scale empirical test of the accuracy of dates produced when assuming a constant rate of decrease for lexical similarity. The test is performed on a database of computer-readable basic vocabulary lists for about one-half of the world's recorded languages. Judgment of lexical similarity is entirely automated and therefore approaches total objectivity. For a set of 52 language groups, lexical similarity determined through automation is calibrated with historical, epigraphic, and archaeological dates of language divergence gathered from published sources. This calibration not only facilitates estimation of dates, but also allows quantitative evaluation of the accuracy of the calculated dates. The observed level of accuracy can serve as the basis for informed decisions on how to use dates calculated by the same method for other groups.

The ASJP Project

The present approach is developed within the Automated Similarity Judgment Program (ASJP),[2] first described by Brown et al. (2008). Brown et al. also review previous research on computerized lexicostatistics, which commenced with Grimes and Agard (1959). A major goal of ASJP is the development of a database of Swadesh lists for all of the world's languages, with all words transcribed into a standard orthography called ASJPcode. Brown et al. (2008:306-7) give a description of this orthography, including IPA equivalents of the ASJPCode symbols. The principal advantage of ASJPcode is that it can be produced with any QWERTY keyboard and thus is highly accessible to transcribers; a disadvantage is that it ignores some features such as tone, vowel length, and suprasegmental traits. A computer program was written to measure the overall lexical similarity of all possible pairs of languages in the database. In Brown et al. (2008) the program was applied to a database consisting of 100-item Swadesh (1955) lists from 245 globally distributed languages, all transcribed into ASJPcode. The automated lexicostatistical classifications of many language families were found to be similar to classifications by expert historical linguists.

Holman et al. (2008) subsequently determined the relative stability of each item on the 100-referent list. A subset of the 40 most stable of the 100 items was found to yield lexicostatistical results (in terms of their correlation with language classifications by specialists)

at least as good as those produced by the full 100-item list. The shorter list facilitated a substantial increase in rate of language list production as did the addition to the project of new transcribers. As a result, the database now consists of lists for 4817 languages and dialects. The worldwide distribution of these is shown on the map of Figure 1. Since some lists are for dialects of the same language, the set of lists represents 3389 of the 6779 spoken languages with different ISO639-3 designations in the 16th edition of *Ethnologue* (Lewis 2009), the most recent worldwide catalogue of languages.

ASJP now employs a different similarity judgment program that produces even better lexicostatistical results compared to language classification by specialists. This program is based on Levenshtein distance (LD), also known as edit distance. LD has previously been applied to language dialects beginning with Kessler (1995), who also reviewed earlier quantitative comparisons of dialects going back to Séguy (1971). To our knowledge, Serva and Petroni (2008) were the first to use LD to calculate language group dates.

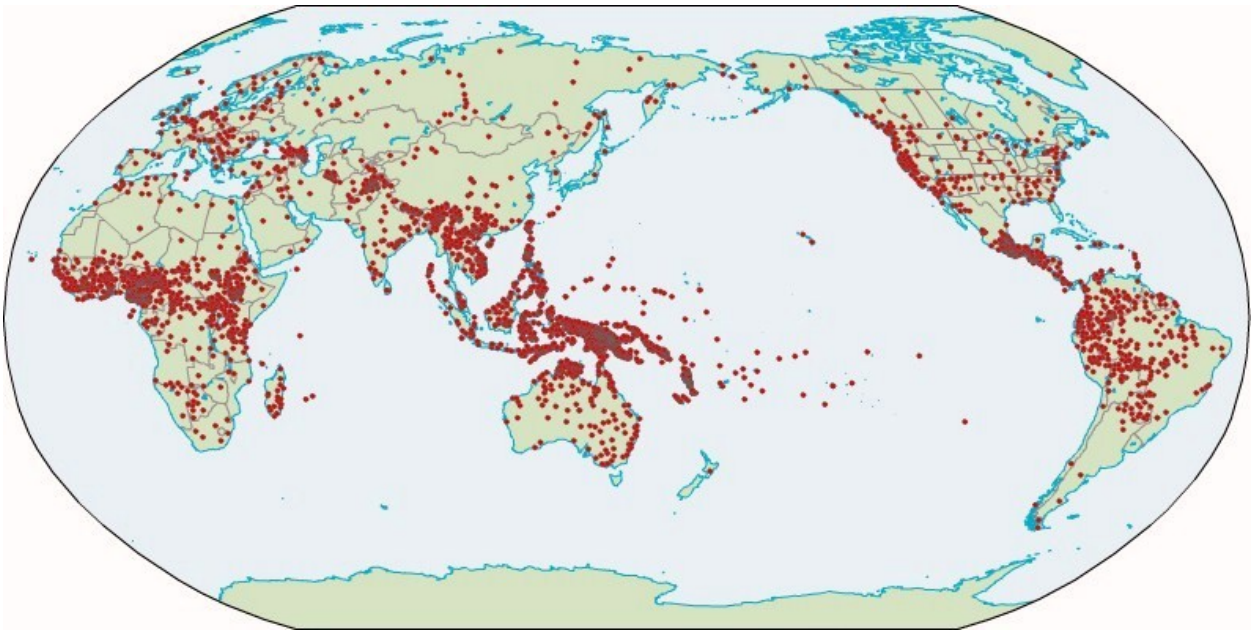

Fig. 1. Distribution of languages and dialects in the ASJP database.

LD is defined as the minimum number of successive changes necessary to convert one word into another, where each change is the insertion, deletion, or substitution of a symbol. For example, in ASJPcode the Spanish word for 'bone' is **weso** and the Italian word is **osso**. In order to convert the Spanish transcription to the Italian one, one insertion, one deletion, and one substitution are required: **s** is added to the Italian word, **w** in the Spanish word is deleted, and **o** is substituted for **e**. Alternatively, to convert the Italian transcription to the Spanish one, **w** is added word-initially to the Italian form, **s** is deleted, and **e** is substituted for the first **o**. Either way, the Spanish and Italian words demonstrate an LD of 3. This symmetry holds in general because deletions and insertions are the inverse of one another and substitutions are not sensitive to the

direction of change. Paired words with smaller LDs are more lexically similar than those with larger LDs.

Levenshtein measurement of similarity treats all changes as equivalent without regard to their phonological plausibility or historical frequency. In comparisons between dialects, Kessler (1995) and Heeringa et al. (2006) explored generalizations of LD in which some changes contribute more to LD than others. These generalizations did not improve the correlations of LD with any of several external criteria. Consistent with these findings, early attempts within the ASJP project to incorporate phonological information into automated similarity judgment did not augment correlations with classifications by specialists, and in some instances even lowered them.

Within the Levenshtein approach, differences in word length can be corrected for by dividing LD by the number of symbols of the longer of the two compared words. This produces normalized LD (LDN), which was used by Serva and Petroni (2008). ASJP includes synonyms on its lists, but no more than two per meaning. For referents represented by two synonyms, LDN is the average LDN of the two. For a given pair of languages, LDN for paired words having the same meaning in the two languages is averaged across all the meanings on the list attested by words in both languages. As a baseline for phonological distance independent of meaning, LDN is also averaged across all pairs with different meanings attested in the two languages. A divided normalized LD (LDND) between the two languages is calculated by dividing the average LDN for all the word pairs involving the same meaning by the average LDN for all the word pairs involving different meanings[3]. As a result, the distance measured by LDND is specifically lexical rather than phonological (Wichmann et al. 2010). Finally, to produce a measure of lexical similarity analogous to the cognate percentages used by Swadesh and others, ASJP similarity (abbreviated *s*) is defined as 1−LDND. Similarity is 100% by definition between identical lists without synonyms, and similarity is near 0% on average between lists from languages that are not at all related by either descent or contact.

Automated use of Levenshtein distance to measure lexical similarity of languages eliminates human judgment of similarity and the ambiguities this entails. Over the years, glottochronology has produced dates for many groups rendered by many different linguists using their individual approaches to cognate identification. In well-studied language families, cognates can be determined rigorously, but this procedure is very labor-intensive and has not been achieved for most families. This lack of uniformity was considered a serious problem in a very early discussion of glottochronology by Swadesh (1955:129; cf., Hymes 1960:18-19). ASJP chronology, in contrast, provides a uniform method for judging lexical similarity and for dating all of the world's phylogenetic language groups, including those for which cognate matches have yet to be worked out.

Lexical Heterogeneity at Time Zero

Swadesh (1950) proposed that if *t* is the time since two related languages diverged from each other and *C* is the proportion of items on a basic vocabulary list that are cognate between the two languages, then *t* can be estimated based on the hypothesis that on average,

(1) $t = \log C / 2 \log r$,

where r is the average proportion of items on the list that are retained after a standard time period (usually 1000 years). This formula can be modified for ASJP chronology by replacing the cognate proportion C with the ASJP similarity s, which has been defined as 1−LDND:

(2) $t = \log s / 2 \log r$.

In other words, s, which is a similarity score derived from a Levenshtein distance, can be used exactly like a proportion of shared cognates in glottochronology to estimate time depth. In (2), r is the average proportion of lexical similarity retained after a standard period of time.

Both formulas (1) and (2) assume that genetically related languages diverged from a single ancestral language that was spoken at $t = 0$ or time zero, the point immediately before the ancestral language began to split into daughter languages. Substitution of $t = 0$ into (1) implies that $\log C = 0$, which in turn implies that $C = 1$, which corresponds to 100% lexical homogeneity for speakers of a single language at time zero. A similar substitution of $t = 0$ into (2) also implies 100% lexical homogeneity at time zero. This is an oversimplification, as Hymes (1960:26-27) recognized. If a time-zero language comprised a chain or network of dialects (Ross 1988:8), then lexical variation almost certainly existed across them. Even if no dialectal diversity were apparent, it is unlikely that any time-zero languages were ever totally lexically homogeneous, since all languages tend to show some variation across speakers, even if distinct dialects are not observed. Consequently, formula (2) should be revised to capture formally the heterogeneity of time-zero languages. For this purpose, we let s_0 represent the average degree of lexical similarity within time-zero ancestral languages. Therefore, the quantity that should start at 1 when $t = 0$ in (2) is not s itself but rather the ratio s/s_0. Since $\log s/s_0 = \log s - \log s_0$, the revised version of (2) is:

(3) $t = (\log s - \log s_0) / 2 \log r$.

Calibration procedure

Once values are established for s, s_0, and r in formula (3), these can be used in the equation, yielding a solution for t, the time depth of a language divergence. The value of s is determined from the data through Levenshtein analysis. The values of s_0, the average degree of lexical similarity within time-zero languages, and of r, the average proportion of lexical similarity retained after a standard period of time, are constants in the formula. Thus, in order to solve for t, the constants s_0 and r must be known. The standard empirical method for determining s_0 and r is linear regression.

Linear regression requires a set of calibration points for groups of genetically related languages with known t and s. The value of t is the date at which the group's ancestral language first began to break apart, as determined from published epigraphic, historical, or archaeological sources. For the few languages with written materials dated near the time of their divergence, the

dates can be determined more or less directly. Otherwise, if speakers of the languages have a recorded history, dates of divergence can be inferred indirectly from the dates of events expected to impair communication between communities, such as migration to places distant from each other, or long-term domination by mutually antagonistic states. For dates before recorded history, archaeology can be used for calibration if words for archaeologically datable objects can be traced to ancestral languages, or if a currently observable association between a language group and a characteristic type of material culture can be extrapolated into the archaeological past. In addition, some of the calibration sources establish dates by correlating loanwords with historically or archaeologically datable periods of contact between languages. As Heggarty (2007) has cogently argued, there are various difficulties in identifying languages with archaeological materials. We nevertheless use archaeological calibrations because they are the only ones available for chronologically deep families.

In addition to the criteria for including calibration points, we also invoke a criterion for excluding candidates. Some sources infer dates for language divergence from archaeological or historical information combined with glottochronology or estimates of similarity between languages. We exclude all potential calibration points of this sort. Our combination of criteria for inclusion and exclusion is intended to identify the most reliable dates that investigators have so far been able to glean from information that is independent of linguistic similarity.

In total, we have assembled 52 published calibration dates which satisfy the criteria described above and for which the ASJP database contains the relevant languages, including those that became extinct after 1900 CE as well as those currently spoken.

By analogy with glottochronology, the similarity score *s* for a language family or subfamily is based on pairwise similarities between the extant (or recently extinct) languages in its highest-level coordinate subgroups. For instance, Indo-Iranian is divided at the highest level into Indic and Iranian, and its similarity score is estimated from the similarities of pairs, each consisting of an Indic language and an Iranian language.

The similarities between such subgroups can be used in different ways to estimate overall similarity for a family or subfamily. One possibility, analogous to the suggestion by Swadesh (1950) for glottochronology, is to use only the smallest similarity score among all the language pairs compared, because the observed lexical similarity for the two least similar languages is least likely to have been influenced by diffusion of words between languages in different subgroups. Other possibilities are the median or the mean of the similarities. ASJP uses the mean for the following reasons. First, Holman et al. (2008, fig. 3) found that, for similarities based on the 40-item list, diffusion has much less effect than phylogenetic relationship. Second, at least one of the pairs of languages with minimum similarity may be a geographic outlier that is atypically different from its sister languages due to removal from them and contact with other languages. Third, the sampling distribution of the mean is less variable than the sampling distributions of the minimum or the median.

The mean similarity calculated for a genetic group such as a language family is directly influenced by the way in which its member languages are sorted into subgroups. In the literature, different classifications are often reported for the same language family, each showing a different

set of highest-level coordinate subgroups. Which one of these classifications is closest to phylogenetic reality is not always obvious, even to those who have specialized knowledge of a particular language family. To minimize the effect of this ambiguity, we require each of our calibration points to be compatible not only with the classification (if any) provided in the source of the calibration date, but also with the classification in the 16th edition of *Ethnologue* (Lewis 2009). For the most part, *Ethnologue* appears to be based on previously published classifications, although the sources of the classifications are not cited.

The classifications in *Ethnologue* of some families are more conservative than those in the calibration sources. For instance, the calibration sources for the Turkic languages distinguish Chuvash from the other surviving Turkic languages, collectively called Common Turkic. *Ethnologue* does not make this distinction, instead listing Chuvash among six coordinate subgroups at the highest level within Turkic. Nevertheless, Common Turkic can be constructed by just combining the five subgroups other than Chuvash. As a general definition, groups from calibration sources are compatible with *Ethnologue* if and only if they can be formed by combining coordinate *Ethnologue* subgroups without moving any languages from one subgroup to another. If a date in a calibration source refers to a group that is not compatible with the *Ethnologue* classification in this sense, we do not use that date.

Classifications from calibration sources and *Ethnologue* are the only guides for language subgroups used for calibration in this study. Through this restrictive approach, ASJP avoids the subjectivity entailed in making choices between competing classifications that could be biased towards a particular result.

Calibration Points

In the following list, the 52 language groups are presented in alphabetical order. A calibration date in years BP is given for each group, followed by an abbreviation indicating whether the date is based on epigraphic (E), historical (H), or archaeological (A) information. The published information itself is described in an immediately following "source" section. When the source gives a range of dates, the middle of the range is used; for convenience, the present time is taken to be the year 2000. Next, the average similarity score is given for the group. This is based on the subgroups listed subsequently, which are named as in *Ethnologue* unless otherwise indicated (each subgroup is followed in parentheses by the number of ASJP lists it contains); the last figure is the number of pairwise language comparisons averaged to produce the score.

Benue-Congo
Date: 6500 (A)
Source: Bostoen and Grégoire (2007:77) link the introduction, during 7000–6000 BP, of new technologies such as macrolithic tools and pottery into the Grassfields region with the break-off of Bantoid from Benue-Congo. They argue that this would fit the hypothesis that the center of dispersion of Benue-Congo is near the confluence of the Niger and Benue rivers, and also mention that pottery-related terminology can be reconstructed to proto-Benue-Congo.
Similarity: 3.58

Comparisons: Akpes (1), Bantoid (258), Cross River (28), Defoid (4), Edoid (27), Idomoid (3), Igboid (5), Jukunoid (2), Kainji (20), Nupoid (4), Oko (1), Plateau (45), Ukaan (6); 44,303 pairs

Brythonic
Date: 1450 (H)
Source: Humphreys (1993:609) concludes from the available historical and linguistic evidence that the distinctiveness of Breton stems from British immigration mainly during the fifth to seventh centuries.
Similarity: 42.60
Comparisons: Breton (1), Welsh (1); 1 pair

Central Southern African Khoisan
Date: 2000 (A)
Source: Güldemann (forthcoming:16) associates the ancestors of the Central Southern African Khoisan (or Khoe-Kwadi, in his terminology) with a cultural sequence starting around 2000 BP, which marks the introduction of food production in the part of Africa where Central Southern African Khoisan speakers are currently located. Support for this hypothesis is provided by the word **gu* 'sheep', which, according to the author, can be reconstructed for the entire language group and which has been borrowed widely into Bantu.
Similarity: 11.67
Comparisons: Nama-Tshu-Khwe (6), Kwadi (1); 6 pairs
Note: Nama-Tshu-Khwe subgroup is from Güldemann (forthcoming).

Cham
Date: 529 (H)
Source: According to Thurgood (1999:44) "[t]here is really no question about the relationship between Western and Phan Rang Cham, as they were the same language until the fall of the southern capital in 1471."
Similarity: 55.06
Comparisons: Eastern Cham (1), Western Cham (1); 1 pair

Chamic
Date: 1550 (H)
Source: Sidwell (2006:198-199) suggests that the breakup of Chamic followed a migration of Chams to Aceh under pressure from Chinese attacks during the 5th century CE.
Similarity: 15.90
Comparisons: Acehnese (1), Coastal-Highlands Chamic (6); 6 pairs
Note: Coastal-Highlands Chamic subgroup is from Sidwell (2006).

Chinese
Date: 2000 (H)
Source: According to Norman (1988:185) the imperial expansion under the Qin and Han dynasties first brought the Chinese language to what are today the Guangdong, Guangxi, Fujian, and southern Jiangxi provinces. This colonization, he argues, was the origin of the differences among modern Chinese languages, particularly the Southern dialect group.
Similarity: 12.97

Comparisons: Hakka (1), Mandarin (2), Min Nan (2), Wu (1), Yue (1); 19 pairs

Cholan
Date: 1600 (E)
Source: Wichmann (2006:283) argues on the basis of epigraphic evidence that Eastern and Western Cholan had split into dialects by 400 CE.
Similarity: 43.26
Comparisons: Chorti (2), Chol-Chontal (3); 6 pairs

Common Turkic (Turkic languages minus Chuvash)
Date: 1419 (H)
Source: In the account of Golden (1998:19-20), "Turkic now became the predominant linguistic element in Mongolia and the steppelands in and around what is now Turkestan and extending into the Pontic zone" in 552 CE, when Bumïn established the Türk Kaghanate. This empire, however, ceased to be united by the end of the rule of Taspar in 581 CE.
Similarity: 37.94
Comparisons: Eastern (2), Northern (7), Southern (30), Western (11); 713 pairs
Note: Common Turkic subgroup is from Golden (1998).

Czech-Slovak
Date: 1050 (E)
Source: Fodor (1962:132) states that "[t]he linguistic unity of the Czech and Slovak languages dissolved in the 10th century."
Similarity: 67.18
Comparisons: Czech (1), Slovak (1); 1 pair

Dardic
Date: 3550 (A)
Source: Parpola (1999:200) correlates the Early Gandhara Grave culture (Ghalegay IV) in Swat (1700–1400 BCE) with Proto-Rgvedic, which he equates with Proto-Dardic.
Similarity: 26.97
Comparisons: Chitral (9), Kashmiri (1), Kohistani (4), Kunar (3), Shina (5); 176 pairs

Eastern Malayo-Polynesian
Date: 3350 (A)
Source: According to Pawley (2009:517), there is a strong association between the first appearance of nucleated villages in the Bismarck Archipelago in 3400-3300 BP and the arrival of Austronesian languages, "specifically with the separation of the large Oceanic branch from its nearest relatives, spoken in the Cenderawasih Bay area at the western end of New Guinea, and in South Halmahera…"
Similarity: 7.56
Comparisons: Oceanic (428), South Halmahera-West New Guinea (43); 18,404 pairs

East Polynesian
Date: 1050 (A)

Source: According to Bellwood and Hiscock (2005:290) radiocarbon dates indicate that the Marquesas, Societies, Cooks, Australs, Tuamotus, Hawaiian Islands, Easter Island, and New Zealand were settled starting around 700 CE and ending several centuries later. Bellwood and Hiscock (2005:292) report general acceptance of the interval between about 700 and 1200 CE as the time when most of central and eastern Polynesia were colonized.
Similarity: 48.34
Comparisons: Rapanui (1), Central (10); 10 pairs

East Slavic
Date: 760 (H)
Source: According to Pugh (2007:10-11), the sacking of Kiev by the Tatars in 1240 CE was soon followed by political fragmentation and linguistic divergence.
Similarity: 39.47
Comparisons: Belarusan (1), Russian (2), Ukrainian (1); 5 pairs

English-Frisian
Date: 1550 (H)
Source: Bremmer (2009:125) states that the Anglo-Saxon conquest of Britain in the fifth century CE implied a separation of what would later become English from the immediate ancestor of Frisian.
Similarity: 30.57
Comparisons: English (2), Frisian (2); 4 pairs
Note: English-Frisian subgroup is from Bremmer (2009).

Ethiopian Semitic
Date: 2450 (E)
Source: According to Ehret (2000:387) the ancestral language of all the members of Ethiopian Semitic is attested in epigraphic records dating to the fifth century BCE at sites in modern-day Eritrea and northern Ethiopia. The relatively wide distribution of epigraphic evidence for the language suggests a geographic dispersal of its speakers and thus the beginning of its breakup.
Similarity: 18.90
Comparisons: North (4), South (14); 56 pairs

Ga-Dangme
Date: 600 (A, H)
Source: Ehret (2000:390-391) states that "the proto-Dangme people of southern Ghana can be tied through both oral tradition and material culture traits to a particular development of town life along the lower Volta River, belonging in the archaeology to the period 1200–1400. Beginning in the fifteenth century, this culture diverged into a set of independent polities, most often consisting of a town and its immediately surrounding rural area."
Similarity: 49.11
Comparisons: Dangme (1), Ga (1); 1 pair

Germanic
Date: 2100 (H)

Source: The emergence of the Cimbri and the Teutones toward the end of the 2nd century BCE (Pohl 2004: 11) was the beginning of the migrations associated with the breakup of Germanic.
Similarity: 29.24
Comparisons: North (7), West (23); 161 pairs

Goidelic
Date: 1050 (E)
Source: According to Jackson (1951:91-92) the Gaelic of Ireland, Scotland, and Man were identical up until the tenth century CE; but from that century onwards, there are indications of divergence between Eastern and Western Gaelic.
Similarity: 27.52
Comparisons: Irish Gaelic (1), Scottish Gaelic-Manx (2); 2 pairs
Note: Scottish Gaelic-Manx subgroup is from Jackson (1951).

Hmong-Mien
Date: 2500 (E)
Source: Sagart, Blench, and Sanchez-Mazas (2005:2-3) date proto-Hmong-Mien to 2500 BP based on the phonological shapes and cultural contents of early loanwords. More specifically, Sagart (1999:208) discusses the Chinese word for 'money', which is among the borrowings into Hmong-Mien. According to Sagart (e-mail, August 4, 2010), this word is first attested in Chinese texts in the 5th century BCE.
Similarity: 5.66
Comparisons: Hmongic (9), Ho Nte (1), Mienic (4); 49 pairs

Indo-Aryan (Indic)
Date: 3900 (A)
Source: Parpola (1999:200) correlates Early Andronovo (Petrovka) (c. 2000–1800 BCE) with proto-Indo-Aryan.
Similarity: 24.79
Comparisons: Central group (1), Central zone (44), Eastern zone (3), Northern zone (1), Northwestern zone (39), Nuristani (2), Sinhalese-Maldivian (2), Southern zone (1); 2586 pairs

Indo-European (minus Anatolian and Tocharian)
Date: 5500 (A)
Source: Anthony (1995:558) argues that proto-Indo-European "existed as a single speech community late enough to experience and create words for wheeled vehicles" and that it cannot have differentiated until after 3500 BCE. Nichols and Warnow (2008:781) then give 5500 BP as a benchmark date for the breakup of Indo-European but mention a slightly earlier divergence of Anatolian, which is not included in the present calibration.
Similarity: 5.29
Comparisons: Albanian (1), Armenian (2), Baltic (2), Celtic (5), Germanic (30), Greek (1), Indo-Iranian (147), Romance (14), Slavic (16); 12,264 pairs

Indo-Iranian
Date: 4400 (A)

Source: Parpola (1999:200) correlates proto-Aryan with the Catacomb Grave and Poltavka cultures (c. 2800–2000 BCE).
Similarity: 8.28
Comparisons: Indic (93), Iranian (54), 5022 pairs

Inuit
Date: 800 (A)
Source: Fig. 2b of Fortescue (1998:27) depicts Neo-Eskimo migration routes with dates and indicates a major dispersal starting around 1200 CE. According to Fortescue (1998:33), this date corresponds to "the first phase of the Thule entry into Greenland" and is based on (recalibrated) Carbon 14 dates.
Similarity: 60.40
Comparisons: North Alaskan Inupiatun (1), Western Canadian Inuktitut (1), Eastern Canadian Inuktitut (1), Greenlandic Inuktitut (1); 6 pairs

Iranian
Date: 3900 (A)
Source: Parpola (1999:200) correlates proto-West-Aryan with the Early Timber Grave and Abashevo cultures (c. 2000-1800 BC).
Similarity: 14.09
Comparisons: Eastern (47), Western (7); 329 pairs

Italo-Western Romance
Date: 1524 (H)
Source: Although Bury (1923:408) argues against a common view that the revolution of 476 CE implied the "Fall of the Western Roman Empire," he does see it as marking the point at which the disintegration of the Empire first extended to Italy.
Similarity: 32.15
Comparisons: Italo-Dalmatian (2), Western (10); 20 pairs

Ket-Yugh
Date: 1300 (H)
Source: Vajda (forthcoming:5) claims that the divergence of Ket and Yugh dates to after the Kirghiz (Turkic) intrusion into the Yenisei region (circa 700 CE).
Similarity: 48.65
Comparisons: Ket (1), Yugh (1); 1 pair

Maa
Date: 600 (H)
Source: Ehret (2000:396) dates proto-Maa to 600 BP. This is an approximate date based on oral traditions of the Maasai and some of their neighbors which indicate that "the breakup of the Proto-Maa society and the emergence of a distinct Maasai society can be dated to not long before the sixteenth century" (Ehret 2000: 385).
Similarity: 67.52
Comparisons: Maasai (1), Samburu (2); 2 pairs
Note: Maa subgroup is from Ehret (2000).

Ma'anyan-Malagasy
Date: 1350 (A)
Source: Adelaar (2006:19) dates the migration of South East Barito speakers to Madagascar to the 7th century CE, after the foundation of Srivijaya.
Similarity: 30.30
Comparisons: Ma'anyan (2), Malagasy (18); 36 pairs
Note: This subgroup is based on Dahl's (1951) identification of Ma'anyan as the language most similar to Malagasy, an identification restated by *Ethnologue*.

Malayo-Chamic
Date: 2400 (A)
Source: According to Sidwell (2006:199) Malayo-Chamic breaks up around 500–300 BCE, when Chamic speakers settle on the mainland and initiate contact with speakers of mainland languages.
Similarity: 24.67
Comparisons: Malayic (23), Chamic (7); 161 pairs
Note: Malayo-Chamic subgroup is from Sidwell (2006). Adelaar (2005) presents evidence that the Bali-Sasak-Sumbawa group diverged from Malayic and Chamic at about the same time in a three-way split, but to be conservative we use only Malayic and Chamic.

Malayo-Polynesian
Date: 4250 (A)
Source: According to Bellwood (2007:40) "[t]he first archaeological appearance to the south of Taiwan of Neolithic communities who used pottery and polished stone adzes, and kept pigs and dogs, occurred in the northern Philippines and western Borneo around 2500–2000 BC."
Similarity: 12.62
Comparisons: Central-Eastern (580), Celebic (61), Chamorro (1), Enggano (3), Greater Barito (57), Javanese (3), Lampungic (24), Land Dayak (3), Malayo-Sumbawan (34), Moklen (1), North Borneo (17), Northwest Sumatra-Barrier Islands (4), Palauan (1), Philippine (151), Rejang (1), South Sulawesi (11); 268,972 pairs

Maltese-Maghreb Arabic
Date: 910 (H)
Source: According to Castillo (2006:29), Arabic domination of Malta lasted from 870 to 1090 CE.
Similarity: 33.57
Comparisons: Maltese (1), Maghreb Arabic (3); 3 pairs

Mississippi Valley Siouan
Date: 2475 (A)
Source: Rankin's (2006:574) Table 41-3 shows proto-Mississippi Valley Siouan breaking up between 2700 and 2250 BP. Rankin (2006:572) infers this date from the observation that the different Mississippi Valley Siouan languages have different words for squash, which became widely cultivated between 500 and 200 BCE.
Similarity: 28.23

Comparisons: Chiwere (1), Dakota (3), Dhegiha (4), Winnebago (1); 27 pairs

Mongolic
Date: 750 (H)
Source: Janhunen (2003:3) describes the ancestral Proto-Mongolic language as the result of intensive linguistic unification under the rule of Chinggiz Khan, and Weiers (2003:248) states that "Moghol developed from the language spoken by the Mongols who during the thirteenth and fourteenth centuries were garrisoned in the west. . . As far as we know, the garrison Mongols who remained in the west never again had any contact with the kinsmen in Mongolia."
Similarity: 20.75
Comparisons: Eastern (7), Western (1); 7 pairs

Northern Roglai-Tsat
Date: 1000 (H)
Source: According to Thurgood (1999:43), "Tsat and Northern Roglai represent a Northern Cham dialect that split into two under the impetus provided by the Vietnamese capture of the northern capital at Indrapura…. As late as around 1000 AD, these two languages probably constituted as single Northern Cham dialect."
Similarity: 25.83
Comparisons: Northern Roglai (1), Tsat (1); 1 pair
Note: Northern Roglai-Tsat subgroup is from Thurgood (1999).

Ongamo-Maa
Date: 1150 (A)
Source: Referring to speakers of the proto-Maa-Ongamo language, Ehret (2000:384-385) states: "Their arrival in central Kenya can be correlated with the appearance in the eighth century of a new pottery, Lanet ware, which has continued to be used by their descendants down to the present…" Moreover, he claims that "the Maa-Ongamo separation took shape by or before 1000 AD, because the Proto-Chaga, a Bantu people of the period 1000–1200, were already by those centuries borrowing Maa-Ongamo words that showed the distinctive phonological features of Ongamo…"
Similarity: 45.17
Comparisons: Maa (3), Ngasa (1); 3 pairs
Note: Maa subgroup is from Ehret (2000).

Oromo
Date: 460 (E)
Source: Ehret (2000:387), based on epigraphic evidence, dates the beginning of the expansion of Oromo people to 1530–1550 CE.
Similarity: 63.46
Comparisons: Orma (1), Borana (2), Eastern (1), West Central (2); 13 pairs

Pama-Nyungan
Date: 4500 (A)
Source: Evans and Jones (1997: 417) link proto-Pama-Nyungan to "new stone and food staple technologies and intensification in the archaeological record," including increased population

density, new art styles, and the extension of long-distance trade networks. They argue that the family would have spread "something like 4000 to 5000 years ago" (Evans and Jones 1997:386).
Similarity: 5.47
Comparisons: Arandic (5), Baagandji (1), Bandjalangic (2), Dyangadi (1), Dyirbalic (4), Galgadungic (2), Gumbaynggiric (2), Guugu-Yimidhirr (1), Iyora (1), Kala Lagaw Ya (1), Karnic (6), Kulinic (7), Maric (9), Muruwaric (1), Paman (21), South-West (23), Tangic (1), Waka-Kabic (4), Wiradhuric (3), Worimi (2), Yalandyic (1), Yanyuwan (1), Yidinic (2), Yotayotic (1), Yugambal (1), Yuin (3), Yuulngu (16); 6693 pairs

Romance
Date: 1729 (H)
Source: Watson (1999:155-156) gives 271 CE as the most likely date for the withdrawal of the last Roman troops to the south of Danube, after which the Latin language persisted north of the river to become Romanian.
Similarity: 28.96
Comparisons: Eastern (2), Italo-Western (12); 24 pairs

Romani
Date: 650 (H)
Source: According to Matras (2002:1) references to 'gypsies' in chronicles allow the reconstruction of "an outwards migration from the Balkans beginning in the fourteenth century, and reaching northern and western Europe in the fifteenth century…"
Similarity: 61.92
Comparisons: Balkan (7), Northern (12), Vlax (7), Dolenjski (1); 243 pairs

Saami
Date: 1750 (A)
Source: Aikio (2006:43) dates the disintegration of proto-Saami to approximately 0–500 CE mainly on the basis of the phonology and distribution of Proto-Scandinavian loanwords.
Similarity: 33.63
Comparisons: Eastern (3), Western (3); 9 pairs

Scandinavian (North Germanic)
Date: 1100 (E)
Source: Haugen (1982:9) states that by the time of the Viking Period (c. 750–1050 CE), a split is observable between East and West Scandinavian.
Similarity: 32.83
Comparisons: East (5), West (2); 10 pairs

Slavic
Date: 1450 (H)
Source: Schenker (1995:9, 15-17) quotes descriptions written in the 6th century CE of the geographic expansion and political anarchy of the Slavs, conditions that initiated the breakup of the common Slavic language.
Similarity: 43.01
Comparisons: East (4), South (6), West (6); 84 pairs

Sorbian (Lusatian)
Date: 450 (E)
Source: Fodor (1962:132) states that "Lower and Upper Lusatian developed from the more or less homogeneous Lusatian in the 16th century, i.e., at the time of the reformation…"
Similarity: 68.78
Comparisons: Lower Sorbian (1), Upper Sorbian (1); 1 pair

Southern Nilotic
Date: 2500 (A)
Source: Ehret (2000:385) correlates the Southern Nilotic languages with the Elmenteitan culture and then states: "In the sixth and fifth centuries BC, a major offshoot of the Elmenteitan moved into the vast Mara and Loita plains south of the western highlands," and he specifically correlates this offshoot with the Tatoga branch of Southern Nilotic.
Similarity: 13.43
Comparisons: Kalenjin (5), Tatoga (6); 30 pairs

Southern Songhai
Date: 550 (H)
Source: According to Moraes Farias (2003:clxxiii), trade diasporas adopted Songhai as a lingua franca, probably during the expansion of the Songhai empire in the fifteenth century CE, and then propagated the language further to the south.
Similarity: 62.85
Comparisons: Dendi (1), Songhay (1), Koyra Chiini Songhay (1), Koyraboro Senni Songhay (1), Zarma (2); 14 pairs

Southwest Tungusic
Date: 236 (H)
Source: Ramsay (1987:216) identifies the Xibe as the descendents of Manchu who were resettled in Xinjiang as border guards in 1764.
Similarity: 53.27
Comparisons: Xibe (1), Manchu (1); 1 pair

Swahili
Date: 1200 (A, H)
Source: Ehret (2000:381) mentions both archaeological and written evidence suggesting that Swahili originated along the Kenya coast, in and around the Lamu archipelago, around 700–900 CE. He further notes that "[a]lready by the close of the eighth and the start of the ninth century, Swahili merchants had planted settlements as far south along the Indian Ocean coast as northern Mozambique and had apparently reached the Comoro Islands."
Similarity: 50.84
Comparisons: Maore (1), Mwani (1), Swahili (8); 17 pairs

Temotu
Date: 3200 (A)

Source: In a paper that describes shared linguistic innovations defining a Temotu subgroup within Austronesian, Ross and Næss (2007:461) cite Green (2003) for an archaeological date of about 3200 BP for the first human occupation of the Reef and Santa Cruz islands, which is ascribed to the Lapita culture (correlated with speakers of Austronesian languages) and is said to be among the earliest examples of this culture outside the Bismarck Archipelago.
Similarity: 7.36
Comparisons: Reefs-Santa Cruz (7), Utupua-Vanikoro (2); 14 pairs

Tupi-Guarani (Coastal)
Date: 1750 (A, H)
Source: Brochado (1984:354) makes reference to ceramic and other archaeological data from Amazonia and adjacent areas as well as ethnohistoric information which together suggest that the ancestors of the Guarani and the ancestors of the Tupinambá evolved independently since 2000–1500 BP.
Similarity: 24.04
Comparisons: Subgroups I-II (7), Subgroup III (3); 21 pairs
Note: Subgroups are from Brochado (1984).

Turkic (Common Turkic and Chuvash):
Date: 2500 (A, H)
Source: Róna-Tas (1991:28) correlates Turkic vocabulary with archaeological and historical information to date the beginning of the Late Ancient Turkic period to the middle of the first millennium BCE, stating: "The beginning of the Late [Ancient] Turkic period was marked by the formation of those Turkic dialects which later became the basis for the various groups and single languages" (Róna-Tas 1991:26).
Similarity: 9.83
Comparisons: Chuvash (1), Common Turkic (50); 50 pairs
Note: Common Turkic subgroup is from Golden (1998).

Wakashan
Date: 2500 (A)
Source: Mitchell (1990:357) infers from archaeology that speakers of the Northern branch of Wakashan expanded into the area around Queen Charlotte Strait, probably from the opposite side of Vancouver Island, in about 500 BCE.
Similarity: 14.80
Comparisons: Northern (2), Southern (3); 6 pairs

Western Turkic (Kipchak)
Date: 900 (H)
Source: The Kipchak empire spread in the 11th and 12th centuries CE (and was destroyed in 1239 CE) according to Troike (1969:191).
Similarity: 45.36
Comparisons: Aralo-Caspian (4), Ponto-Caspian (3), Uralian (4); 40 pairs

This collection of 52 calibration points is substantially larger and more diverse than the 13 points in the calibration of Lees (1953), which has long served as the standard in glottochronology. Lees estimated a constant rate of word replacement by comparing

vocabularies of modern languages to those of older language states attested in textual materials. For example, Catalan, French, Italian, Portuguese, Romanian, and Spanish were compared with Latin. The generality of Lees' calibration is limited by the fact that all but two of the 13 languages (Coptic and Mandarin) are Indo-European. The much larger quantitative test of glottochronology by Blust (2000) involves 224 languages, but all belong to a single family, Austronesian. Among the 52 calibration points employed here, 17 are Indo-European and nine are Austronesian, meaning that one-half refer to groups in other families, including languages of Africa, Australia, and North, Middle, and South America, as well as Europe, Asia, and Oceania. The geographic distribution of points reflects the distribution of available dates, which is thinnest for Australia, New Guinea, and South America.

Testing ASJP chronology

The scatter plot of Figure 2 shows the time depth of each calibration group as a function of the average similarity for the group on a logarithmic scale. The correlation (Pearson's r) between log similarity and time is -0.84. To a good approximation, this strong correlation supports the critical claim that log lexical similarity decreases linearly as time depth increases.

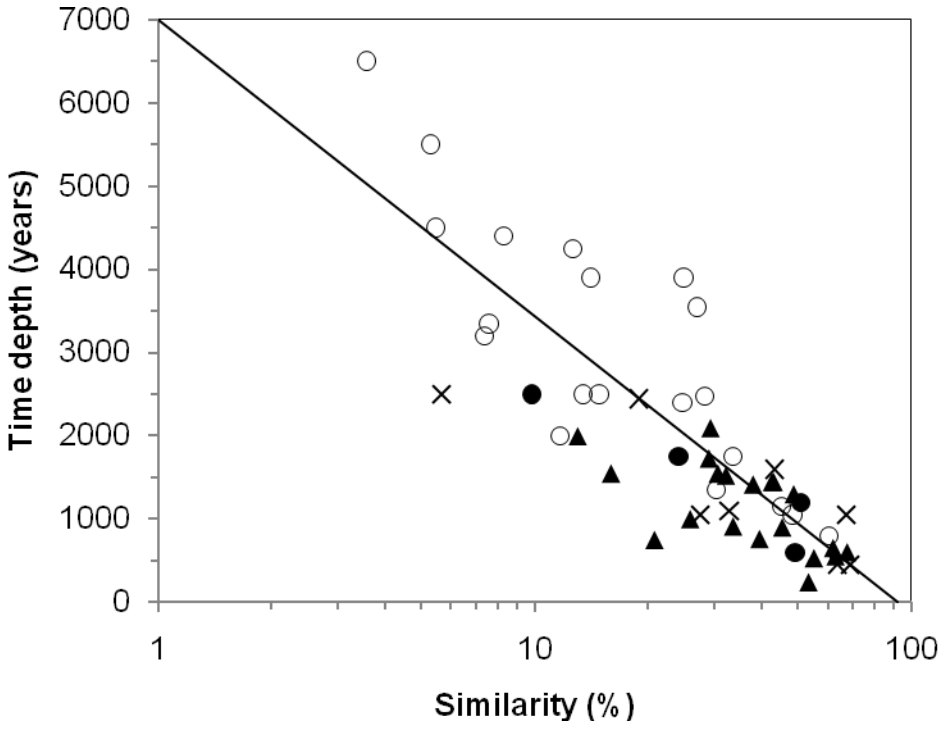

Fig. 2. Time depth (*t*) as a function of similarity (*s*) with regression line; dates are archaeological (open circles), archaeological and historical (filled circles), historical (triangles), and epigraphic (Xs).

The straight line in Figure 2 represents formula (3) with the constants s_0 and r determined by linear regression of t on log s. The line is chosen to make the most accurate possible predictions of time depth from similarity by minimizing the average of the squared distances on the vertical axis (in years) from the values of t to the line, which are the squared distances between the predicted and observed values of t. For this line, s_0 = 92% and r = 0.72 (per 1000 years).[4] Specifically, s_0 is the point where the line crosses the horizontal axis, and 1 / (2 log r) is the slope of the line, which is negative because r is below 1. The fact that regression analysis produces a value of s_0 below 100% is consistent with the usual lexical heterogeneity of languages, including those at time zero.

Table 1 indicates the accuracy of dates based on s_0 = 92% and r = 0.72 for the 52 language groups used for calibration. In the table, language groups are categorized by type of calibration date and rank-ordered within categories by calibration date, from oldest to youngest. The ASJP date is the value of t obtained by substituting the language group similarity score into formula (3) along with s_0 = 92% and r = 0.72. The next column gives the algebraic difference between the calibration date and the ASJP date, and the last column gives this difference as a percentage of the ASJP date.

Table 1. Comparison of calibration dates for 52 language groups with ASJP dates based on s_0 = 0.92 and r = 0.72

Language Group	Calibration Date	ASJP Date	Difference	Percentage
Archaeological				
Benue-Congo	6500	4940	-1560	-32
Indo-European	5500	4348	-1152	-26
Pama-Nyungan	4500	4295	-205	-5
Indo-Iranian	4400	3665	-735	-20
Malayo-Polynesian	4250	3024	-1226	-41
Indo-Aryan (Indic)	3900	1996	-1904	-95
Iranian	3900	2856	-1044	-37
Dardic	3550	1868	-1682	-90
Eastern Malayo-Polynesian	3350	3803	+453	+12
Temotu	3200	3844	+644	+17
Southern Nilotic	2500	2928	+428	+15
Wakashan	2500	2781	+281	+10
Mississippi Valley Siouan	2475	1798	-677	-38
Malayo-Chamic	2400	2003	- 397	-20
Central Southern African Khoisan	2000	3143	+1143	+36
Saami	1750	1532	-218	-14
Ma'anyan-Malagasy	1350	1690	+340	+20
Ongamo-Maa	1150	1083	-67	-6
East Polynesian	1050	979	-71	-7
Inuit	800	640	-160	-25

Archaeological, Historical				
Turkic	2500	3404	+904	+27
Tupi-Guarani (Coastal)	1750	2043	+293	+14
Swahili	1200	903	-297	-33
Ga-Dangme	600	955	+355	+37
Historical				
Germanic	2100	1745	-355	-20
Chinese	2000	2982	+982	+33
Romance	1729	1759	+30	+2
Chamic	1550	2672	+1122	+42
English-Frisian	1550	1677	+127	+8
Italo-Western Romance	1524	1600	+76	+5
Brythonic	1450	1172	-278	-24
Slavic	1450	1157	-293	-25
Common Turkic	1419	1348	-71	-5
Ket-Yugh	1300	970	-330	-34
Northern Roglai-Tsat	1000	1933	+933	+48
Maltese-Maghreb Arabic	910	1534	+624	+41
Western Turkic	900	1076	+176	+16
East Slavic	760	1288	+528	+41
Mongolic	750	2267	+1517	+67
Romani	650	603	-47	-8
Maa	600	471	-129	-27
Southern Songhai	550	580	+30	+5
Cham	529	781	+252	+32
Southwest Tungusic	236	832	+596	+72
Epigraphic				
Hmong-Mien	2500	4243	+1743	+41
Ethiopian Semitic	2450	2408	-42	-2
Cholan	1600	1148	-452	-39
Scandinavian	1100	1569	+469	+30
Czech-Slovak	1050	479	-571	-119
Goidelic	1050	1837	+787	+43
Oromo	460	565	+105	+19
Sorbian	450	443	-7	-2

The algebraic differences between the calibration dates and the ASJP dates include a positive or negative sign and thus show whether the ASJP dates are respectively greater than or less than the calibration dates. As in any linear regression, the algebraic differences have a mean of 0 and a correlation of 0 with the ASJP dates (within rounding error), meaning that the ASJP dates are unbiased. The absolute differences disregard sign and thus show how much the ASJP dates depart from the calibration dates in either direction. Absolute differences have a correlation

of 0.57 with the ASJP dates and tend to be larger for older calibration dates than for younger ones, indicating that older ASJP dates are less accurate than younger ones.

The absolute percentage differences have a correlation of only -0.06 with the ASJP dates, indicating that the discrepancies in ASJP dates are approximately proportional to the dates themselves. The mean absolute percent discrepancy is 29%; of the 52 ASJP dates, five are off by more than 50%, and one is off by more than 100%.

Although the calibration dates older than 2500 BP are all archaeological, Figure 2 shows that younger dates of all four types are about equally close to the regression line. To determine whether there are nevertheless differences among the types of dates, the standard statistical test is a one-way analysis of variance, in which the variance between groups of scores is compared to the variance of scores within groups. The test statistic, F, is a ratio of variances, with degrees of freedom that depend on the number of groups and the number of individual scores. Under the null hypothesis of no differences between groups, the expected value of F is 1, and values of F significantly above 1 indicate significant differences among the groups. A significance criterion of $p < .05$ is used in all tests reported here.

Table 2 gives the results of analyses of variance for type of date and other possibly relevant factors. The second and third columns show the values of F obtained for differences in algebraic and absolute percent discrepancies, respectively, and the fourth column shows the degrees of freedom. If some types of dates are biased low or high relative to others, there would be differences in the algebraic percent discrepancies; and if some types of dates are more accurate than others, there would be differences in the absolute percent discrepancies. The values of F in the first row of the table are not significant, confirming the impression from Figure 2 that the types of date do not differ in bias or accuracy.

Table 2. Analyses of variance on algebraic and absolute percent discrepancies

Factor	F, algebraic	F, absolute	d. f.
Type of date	2.52	0.29	3, 48
Language family	1.12	0.32	16, 35
Geographical area	0.68	0.87	3, 48
Mode of subsistence	0.43	0.53	1, 50

The 52 calibration points pertain to 17 language families. Families whose languages change rapidly would produce ASJP dates older than the calibrations, and families with low rates would produce younger dates. Differences among families in the variability of the rate of lexical change would be reflected in the absolute percent discrepancies. The non-significant F values in the second row of the table suggest no differences among families in the rate or variability of lexical change. To test whether the conditions in different geographical areas influence rates of lexical change, the calibration points are sorted geographically into areas defined as in Tables 3 – 7 below: Africa (11 points), Eurasia (26 points), the Pacific area (10 points), and the Americas (5

points). The third row shows no significant differences between geographical areas. To test the effect of mode of subsistence, Hammarström's (2010) compilation is used to categorize the calibration points according to whether their languages are spoken in predominantly agricultural societies (45 points) or in foraging and pastoral societies (7 points: Central Southern African Khoisan, Inuit, Ket-Yugh, Mississippi Valley Siouan, Pama-Nyungan, Saami, and Wakashan). The last row again shows no significant differences.

In summary, basic vocabulary changes at a sufficiently constant rate to produce a robust correlation of -0.84 between log similarity and calibration date, and the resulting ASJP dates are impervious to all the extraneous factors tested. The observed discrepancies are overestimates if anything because they reflect not only variation in the rate of lexical change but also difficulties in matching dated events with linguistic divergences, as well as uncertainty in the calibration dates themselves, some of which are expressed in the sources as ranges of possible dates. The 29% mean absolute discrepancy thus represents an upper bound on the expected discrepancy between ASJP dates and true dates.

A possible theoretical framework for the present results can be found in Dixon (1997). This author discusses a model of language change involving "punctuated equilibrium," in which languages usually change at a steady rate but occasionally undergo periods of rapid change caused by external events such as natural disasters, material innovations, development of aggressive tendencies, and so on. In Dixon's model, a few languages may undergo more periods of equilibrium and fewer bouts of punctuation (or vice versa) than is typical. As examples of these situations, Bergsland and Vogt (1962) report unusually low rates of lexical change in Icelandic, Georgian, and Armenian, and an unusually high rate of change in East Greenlandic Eskimo. However, for most languages, relative amounts of equilibrium and punctuation are more nearly average, producing roughly similar rates of lexical change as well. This situation is described by Brown (2006:649-650), Ehret (2000:373), Jaxontov (1999:52), and Lohr (2000:219). The rate of lexical change ascertained by ASJP is perhaps best understood as expressing such an average.

The present findings can serve as a baseline for comparison to rates of change in other properties of languages, such as cognates (as in glottochronology), typological features, and also the size (number of languages) and geographical distribution of language groups. Properties found to have sufficiently uniform rates of change could be used to produce alternative dates, or else combined with ASJP lexical similarity for composite chronological estimation.

Worldwide ASJP chronology

The same procedure described for the calibration groups can also be used to calculate ASJP dates for groups that lack alternative information about their time depth. Heggarty (2007) shows how even very rough estimates of linguistic time depth can be used in conjunction with archaeological information to infer sequences of historical events. Compared to the informal dates typically used for this purpose, ASJP dates have the advantages of a uniform definition and a quantitatively known level of accuracy.

Tables 2–6 present ASJP dates (years BP) for nearly all of the world's known language families, calculated by inserting the group average similarity into formula (3), where $s_0 = 92\%$ and $r = 0.72$. The families are all defined in *Ethnologue*, 16th Edition; the subgroups are augmented in three cases from the calibration sources as previously described. Also included are the highest-level subgroups of each family, and in some cases, groups of the next one or two lower taxonomic levels. The choice of which groups to include at lower levels is based on degree of attestation in the database, age (for older families we typically go further down into subgroups), and general interest. Languages in the ASJP database used to generate these dates include those currently spoken and also extinct languages attested by wordlists collected from native speakers after 1700 CE.

The second column in each table shows the number of pairs of lists involved in the calculations. Usually this number is smaller than the number of possible pairs formed from the languages in *Ethnologue*, because some of the languages are not represented by lists in the ASJP database. Occasionally, however, the number of pairs in Column 2 is larger than the number of language pairs, because some languages are represented by several lists for different dialects; this increase in sample size is expected to decrease the variance of the sampling distribution of similarity scores and to leave the mean unchanged.

The third column indicates the number of subgroups across which pairwise similarities are averaged to produce the similarity percentage. The parenthesized numbers in the third column indicate the total number of subgroups according to *Ethnologue*. This may be compared with the number of subgroups used in the calculations to get an idea of the completeness of the data from which a given date was estimated. For instance, in Table 2 Afro-Asiatic has 7 subgroups according to *Ethnologue*. A total of 24,303 pairs were drawn from 6 of these 7 subgroups to produce the date of 6016 BP. When the number of subgroups used differs from the existing number of subgroups, there is the possibility for some eventual improvement in age estimates. The level of individual *Ethnologue* languages, i.e., the level corresponding to a particular ISO-639-3 code, is treated as a taxonomic level in its own right. For instance, in Table 4, the Arai (Left May) subgroup of Arai-Kwomtari is not further sub-classified in *Ethnologue*, so the 4 languages belonging to this subgroup are each treated as coordinate branches under the Arai (Left May) node.

The similarities include a few languages that are missing from *Ethnologue* (being recently extinct or newly described) but are represented in the ASJP database and assigned to subgroups according to the classifications in the sources for the lists. The number of such languages (or their subgroups) is indicated following a + after the number of *Ethnologue* subgroups. For instance, in Table 2, the ASJP dates for Yeniseian are based on two languages included in *Ethnologue* plus four extinct languages in two extinct subgroups. Finally, the groups named in the tables for which there is no information are those that are presented in *Ethnologue* as including only one subgroup or for which the ASJP database includes languages for only one subgroup.

The last two columns show the similarity score and the ASJP date. Three characteristics of the dates are worth noting. First, some of the families reported in *Ethnologue* are controversial and may not be phylogenetically real. If subgroups of a family are not in fact genetically related,

but instead similar only because of contact and diffusion, the lexical similarity score is expected to be relatively small and the date not meaningful. Second, the *Ethnologue* classification of many groups is conservative in the sense that it does not include subgroups identified in other classifications. If the classification of a group is too conservative and separates subgroups that are more closely related than others, the average similarity score for the group will be inflated and the time depth will be underestimated. Third, a few higher-order groups are estimated to be younger than some of their immediate subgroups. As Urban and Wichmann (unpublished manuscript, 2010) point out, these anomalies may reflect random variation in the similarities, mistakes in the classification, or unusually high rates of lexical change in the apparently older subgroups.

Tables 2-6 represent the first attempt known to us to assign dates to most if not all of the world's language families using a uniform method and database. These dates are based on the *Ethnologue* classification because of its comprehensiveness and availability, and *Ethnologue* names of groups are used for consistency. Although Tables 2-6 are restricted to the top few levels of the *Ethnologue* classification, dates for all *Ethnologue* groups represented in the ASJP database are available in Supplement A. It may also be useful to generate ASJP dates based on other classifications identified by specialists as more accurate than those of *Ethnologue*. To facilitate the calculation of such dates, our database and software have been made available online (respectively, Wichmann et al. 2010 and Holman 2010).

Table 3. ASJP dates for language groups of Africa

Group	Pairs	Subgroups	Similarity	Date
Afro-Asiatic	24,303	6 (7)	1.77	6016
Berber	139	4 (4)	29.46	1733
Eastern	2	2 (2)	30.17	1697
Northern	54	3 (4)	43.00	1158
Tamasheq	4	2 (2)	63.83	556
Chadic	2945	4 (4)	3.86	4826
Biu-Mandara	44	2 (3)	4.92	4457
Masa	26	6 (8)	31.13	1649
West	364	2 (3)	6.23	4099
Cushitic	752	4 (4)	4.10	4734
Central	24	4 (4)	30.38	1686
East	891	9 (10)	12.44	3045
South	14	5 (7)	20.20	2308
Omotic	84	2 (2)	3.52	4968
North	141	3 (3)	11.71	3137
South	3	3 (5)	25.34	1963
Semitic	396	2 (2)	10.51	3301
Central	72	2 (2)	16.26	2638
South	72	2 (2)	7.56	3804

Khoisan	31	3 (3)	0.01	14592
Southern African Khoisan	71	3 (3)	2.88	5271
Central	6	2 (3)	11.67	3143
Northern	3	3 (6)	27.36	1846
Southern	4	2 (2)	5.30	4344
Niger-Congo	51,854	4 (4)	1.54	6227
Atlantic-Congo	35,936	3 (3)	1.26	6525
Ijoid	33	2 (2)	16.87	2582
Atlantic	241	3 (3)	1.30	6480
Northern	161	5 (5)	3.32	5055
Southern	9	2 (3)	4.64	4546
Volta-Congo	54,371	5 (5)	2.51	5484
Benue-Congo	46,303	13 (16)	3.58	4940
Dogon	42	8 (14)	21.65	2202
Kru	6	2 (5)	20.08	2317
Kwa	216	2 (2)	5.78	4212
Kordofanian	119	4 (4)	3.77	4861
Heiban	18	2 (2)	17.55	2521
Katla	1	2 (2)	20.72	2269
Talodi	5	2 (2)	4.31	4658
Mande	799	2 (2)	9.74	3417
Eastern	70	2 (2)	26.31	1905
Western	280	2 (2)	12.42	3047
Nilo-Saharan	7676	10 (10)	1.17	6642
Central Sudanic	459	2 (2)	3.20	5114
East	208	4 (4)	8.01	3715
Eastern Sudanic	1091	4 (4)	1.80	5988
Eastern	57	4 (4)	3.23	5103
Nilotic	719	3 (3)	4.76	4508
Western	5	2 (4)	2.32	5601
Kadugli-Krongo	49	6 (6)	41.25	1221
Komuz	18	2 (2)	3.00	5209
Koman	13	4 (5)	17.31	2542
Saharan	3	2 (2)	6.91	3941
Western	2	2 (2)	8.91	3553
Songhai	12	2 (3)	38.31	1333
Northern	1	2 (2)	54.15	807
Southern	14	5 (5)	62.85	580

Table 4. ASJP dates for language groups of Eurasia

Group	Pairs	Subgroups	Similarity	Date
Altaic	1588	3 (3)	1.84	5954

Mongolic	7	2 (2)	20.75	2267
Eastern	16	3 (3)	22.48	2145
Tungusic	99	2 (2)	38.67	1319
Northern	20	3 (3)	44.90	1092
Southern	24	2 (2)	32.26	1595
Turkic	50	2 (2)	9.83	3404
Common	713	4 (7)	37.94	1348
Andamanese	16	2 (2)	4.75	4510
Great Andamanese	12	2 (2)	22.82	2122
South Andamanese	1	2 (3)	42.22	1186
Austro-Asiatic	1843	2 (2)	8.45	3635
Mon-Khmer	3358	8 (9)	9.81	3406
Aslian	29	4 (4)	23.46	2080
Eastern Mon-Khmer	475	4 (4)	18.05	2479
Nicobar	3	3 (5)	11.56	3158
Northern Mon-Khmer	243	4 (4)	10.81	3259
Palyu	1	2 (2)	14.05	2861
Viet-Muong	22	4 (5)	20.45	2289
Munda	60	2 (2)	16.96	2574
North Munda	14	2 (2)	41.58	1209
South Munda	4	2 (2)	17.69	2510
Chukotko-Kamchatkan	6	2 (2)	10.06	3368
Northern Chukotko-Kamchatkan	2	2 (2)	42.04	1192
Dravidian	181	4 (10)	23.84	2055
Central	2	2 (2)	58.29	695
Northern	3	3 (5)	24.24	2030
South-Central	6	2 (2)	18.43	2447
Southern	9	2 (10)	26.50	1894
Hmong-Mien	49	3 (3)	5.66	4243
Hmongic	21	4 (5)	14.84	2777
Indo-European	12,264	9 (9)	5.29	4348
Baltic				
Eastern	1	2 (2)	35.05	1469
Celtic				
Insular	6	2 (2)	7.21	3876
Germanic	161	2 (3)	29.24	1745
North	10	2 (2)	32.83	1569
West	170	4 (4)	36.71	1398
Indo-Iranian	5022	2 (2)	8.28	3665
Indo-Aryan	2586	8 (11)	24.79	1996
Iranian	329	2 (4)	14.09	2856
Italic				
Romance	24	2 (3)	28.96	1759
Slavic	84	3 (3)	43.01	1157
East	5	3 (4)	39.47	1288
South	8	2 (2)	58.44	691

West	11	3 (3)	53.67	820
Japonic	12	2 (2)	32.92	1564
Kartvelian	5	3 (3)	12.82	2999
Zan	1	2 (2)	62.20	596
North Caucasian	160	2 (2)	0.58	7709
East Caucasian	391	7 (7)	7.06	3907
West Caucasian	8	3 (3)	8.37	3649
Sino-Tibetan	1106	2 (2)	2.90	5261
Chinese	19	5 (14)	12.97	2982
Tibeto-Burman	10,352	14 (18)	5.81	4203
Bai	107	3 (3)	34.49	1494
Himalayish	680	2 (3)	11.37	3182
Karen	9	2 (4)	19.71	2345
Kuki-Chin-Naga	65	2 (2)	9.79	3411
Lolo-Burmese	24	3 (4)	9.63	3436
Nungish	2	2 (5)	25.46	1955
Tangut-Qiang	2	2 (2)	4.31	4660
Tai-Kadai	699	3 (3)	10.86	3252
Hlai	2	2 (2)	19.61	2353
Kadai	24	3 (3)	16.53	2613
Kam-Tai	505	3 (3)	19.31	2376
Uralic	240	9 (9)	11.40	3178
Finnic	14	5 (11)	51.75	876
Mordvin	1	2 (2)	54.39	800
Permian	2	2 (2)	49.17	953
Sami	9	2 (3)	33.63	1532
Samoyed	1	2 (5)	14.14	2850
Yeniseian	12	1 (1)+2	16.01	2661
Assan-Kott	1	+2	55.07	781
Awin-Pumpokol	1	+2	14.99	2762
Ket-Yugh	1	2 (2)	48.65	970
Yukaghir	1	2 (2)	24.28	2027

Table 5. ASJP dates for language groups of the Pacific

Group	Pairs	Subgroups	Similarity	Date
Amto-Musan	2	2 (2)	21.84	2189
Arai-Kwomtari	20	2 (2)	0.72	7386
Arai (Left May)	6	4 (4)	13.04	2974
Kwomtari	9	4 (4)	1.82	5968
Australian	10,664	16 (16)	2.84	5296
Bunaban	1	2 (2)	33.50	1538
Daly	95	3 (4)	6.91	3941
Bringen-Wagaydy	25	2 (2)	20.04	2320

Malagmalag	6	2 (2)	31.43	1635
Murrinh-Patha	2	2 (2)	15.14	2747
Djeragan	1	2 (3)	15.10	2750
Giimbiyu	3	3 (3)	70.03	415
Gunwingguan	274	12 (13)	4.73	4517
Burarran	3	3 (4)	8.57	3612
Enindhilyagwa	3	3 (3)	4.07	4746
Gunwinggic	1	2 (2)	13.24	2951
Maran	2	2 (2)	16.02	2661
Rembargic	1	2 (2)	25.97	1925
Yangmanic	1	2 (3)	31.97	1609
Pama-Nyungan	6693	27 (30)	5.47	4295
Arandic	9	4 (6)	26.54	1892
Dyirbalic	6	3 (3)+1	22.60	2137
Galgadungic	1	2 (2)	19.44	2366
Karnic	13	3 (3)+1	14.13	2851
Maric	35	8 (12)	49.98	929
Paman	168	7 (15)	3.64	4918
South-West	228	11 (17)	11.98	3103
Waka-Kabic	5	3 (3)	20.70	2270
Wiradhuric	3	3 (3)	43.82	1129
Worimi	1	2 (2)	18.12	2473
Yidinic	1	2 (2)	40.81	1237
Yuin	3	2 (2) +1	34.27	1503
Yuulngu	76	3 (3)	33.13	1555
West Barkly	3	3 (3)	16.34	2631
Wororan	13	4 (7)	21.93	2183
Yiwaidjan	5	3 (3)	13.85	2882
Yiwaidjic	1	2 (5)	36.50	1407
Austronesian	19,212	10 (11)	8.46	3633
Atayalic	1	2 (2)	15.98	2664
East Formosan	5	3 (3)	19.11	2392
Malayo-Polynesian	268,972	16 (17)	12.62	3024
Celebic	814	4 (4)	28.27	1796
Eastern	43	2 (2)	29.92	1710
Kaili-Pamona	2	2 (2)	45.36	1076
Tomini-Tolitoli	20	2 (2)	35.07	1468
Central-Eastern	51,447	3 (4)	11.92	3111
Central Malayo-Polynesian	4562	9 (10)	18.82	2415
Eastern Malayo-Polynesian	18,404	2 (2)	7.56	3803
Greater Barito	1012	4 (4)	24.23	2031
East	156	3 (3)	26.73	1881
Sama-Bajaw	21	2 (2)	34.59	1489
West	15	2 (2)	45.03	1087
Javanese	2	2 (5)	63.45	566
Lampung	164	3 (3)	54.92	785

Land Dayak	3	3 (5)	34.11	1510
Malayo-Sumbawan	65	3 (3)	27.37	1845
North and East	221	3 (3)	26.44	1898
North Borneo	80	3 (5)	24.46	2016
Melanau-Kajang	1	2 (2)	37.34	1372
North Sarawakan	33	4 (5)	22.08	2172
Sabahan	7	3 (3)	38.32	1333
NW Sumatra-Barrier Islands	5	3 (5)	27.79	1822
Philippine	7587	8 (10)	27.65	1830
Bashiic	9	2 (2)	57.43	717
Bilic	19	3 (4)	31.46	1633
Central Luzon	2	2 (3)	40.42	1252
Greater Central Philippine	1949	8 (8)	38.49	1326
Minahasan	4	2 (2)	61.87	604
Northern Luzon	419	3 (4)	31.72	1621
Sangiric	5	2 (2)	66.95	484
South Sulawesi	39	3 (5)	48.63	970
Bugis	2	2 (3)	51.47	884
Makassar	8	3 (5)	63.78	558
Northern	3	3 (17)	73.36	345
Northwest Formosan	1	2 (2)	21.63	2204
Tsouic	3	3 (3)	20.42	2291
Western Plains	3	2 (2)	16.83	2586
Central Western Plains	2	2 (2)	18.63	2431
Border	64	2 (3)	9.51	3453
Taikat	15	2 (2)	18.96	2404
Waris	24	5 (8)	20.83	2261
Central Solomons	9	4 (4)	8.21	3677
East Bird's Head-Sentani	39	3 (3)	1.19	6615
East Bird's Head	2	2 (2)	8.70	3590
Sentani	20	2 (2)	6.22	4101
East Geelvink Bay	4	2 (10)	6.73	3979
Eastern Trans-Fly	495	4 (4)	10.83	3257
Kaure				
Kaure Proper	1	2 (3)	15.98	2665
Lakes Plain	93	4 (4)	2.87	5279
Rasawa-Saponi	1	2 (2)	12.51	3037
Tariku	149	4 (4)	8.98	3541
Left May	2	2 (2)	15.98	2665
Mairasi	4	2 (3)	41.93	1196
Nimboran	6	2 (5)	23.78	2059
North Bougainville	1	2 (3)	13.46	2925
Pauwasi	12	2 (2)	6.21	4102
Eastern	2	2 (3)	14.22	2842
Western	3	2 (2)	28.69	1774
Piawi	10	2 (2)	11.22	3203

Ramu-Lower Sepik	117	3 (3)	0.96	6942
Lower Sepik	28	4 (4)	9.78	3411
Ramu	28	4 (6)	6.65	4000
Sepik	294	8 (12)	3.86	4827
Ndu	32	6 (12)	41.08	1227
Nukuma	1	2 (3)	28.37	1791
Ram	2	2 (3)	28.35	1791
Sepik Hill	29	4 (4)	9.00	3538
Sko	48	2 (2)	4.85	4478
Krisa	27	3 (4)+4	19.01	2400
Vanimo	14	3 (3)+2	28.24	1798
South Bougainville	2	2 (2)	12.37	3054
Buin	1	2 (3)	29.25	1744
South-Central Papuan	145	4 (4)	1.53	6232
Morehead-Upper Maro	14	3 (3)	2.73	5353
Pahoturi	5	2 (2)	24.02	2044
Yelmek-Maklew	4	2 (2)	35.06	1468
Tor-Kwerba	45	2 (2)	4.99	4435
Greater Kwerba	20	3 (3)	6.18	4109
Kwerba	8	2 (2)	7.32	3852
Orya-Tor	6	2 (3)	8.13	3693
Torricelli	250	6 (7)	2.10	5754
Kombio-Arapesh	12	2 (2)	10.15	3356
Marienberg	34	7 (7)	10.25	3339
Momumbo	1	2 (2)	26.98	1867
Wapei-Palei	7	3 (3)	2.67	5386
Trans-New Guinea	77,005	39 (39)	1.20	6609
Angan				
Nuclear Angan	1	2 (12)	4.71	4523
Asmat-Kamoro	21	4 (5)	21.83	2189
Asmat	6	4 (6)	46.66	1033
Sabakor	1	2 (2)	63.37	567
Binanderean				
Binandere	9	4 (12)	27.42	1842
Bosavi	96	8 (9)	19.66	2349
Chimbu-Wahgi	31	3 (4)	9.41	3470
Chimbu	9	4 (7)	31.42	1635
Hagen	2	2 (2)	34.22	1505
Jimi	1	2 (3)	50.55	912
Duna-Bogaya	1	2 (2)	12.78	3004
East Strickland	18	5 (6)	36.66	1401
Eleman	20	3 (3)	3.80	4851
Nuclear Eleman	8	2 (2)	40.31	1256
Engan	56	3 (3)	14.99	2762
Enga	24	4 (6)	18.94	2406
Angal-Kewa	5	3 (7)	33.12	1555

Finisterre-Huon	70	2 (2)	6.08	4136
Finisterre	9	4 (6)	13.98	2868
Huon	48	2 (3)	12.45	3044
Gogodala-Suki	7	2 (2)	14.36	2827
Gogodala	6	2 (3)	34.47	1494
Inland Gulf	2	2 (2)	13.99	2867
Minanibai	1	2 (6)	21.72	2197
Kainantu-Goroka	140	2 (2)	3.81	4847
Gorokan	76	6 (6)	11.34	3186
Kainantu	20	2 (5)	11.96	3105
Kayagar	5	3 (3)	39.54	1285
Kiwaian	88	7 (7)	35.80	1436
Kolopom	3	3 (3)	13.76	2892
Madang	3025	4 (4)	4.56	4573
Croisilles	1082	7 (7)	6. 19	4107
Rai Coast	378	7 (7)	9.16	3511
South Adelbert Range	66	3 (3)	5.96	4165
Marind	57	3 (3)	6.58	4014
Boazi	12	2 (2)	32.22	1597
Yaqay	2	2 (2)	23.63	2069
Mek	3	2 (2)	38.92	1309
Eastern	3	3 (6)	36.08	1425
Mombum	1	2 (2)	38.82	1313
Ok-Awyu	108	2 (2)	5.56	4272
Awyu-Dumut	27	4 (4)	13.55	2916
Ok	36	2 (5)	17.41	2534
South Bird's Head	69	3 (3)	8.86	3561
South Bird's Head Proper	40	3 (3)	32.52	1583
Southeast Papuan	268	6 (7)	2.85	5286
Kwalean	11	3 (3)	12.55	3032
Goilalan	2	2 (2)	5.70	4233
Koiarian	12	2 (2)	15.70	2691
Mailuan	3	3 (6)	40.79	1238
Manubaran	8	2 (2)	45.70	1065
Teberan	1	2 (2)	20.00	2322
Turama-Kikorian	3	2 (2)	12.58	3028
Turama-Omatian	2	2 (2)	32.59	1580
West	886	5 (5)	3.26	5082
Dani	20	3 (3)	28.53	1782
East Timor	2	2 (2)	26. 13	1916
West Bomberai	2	2 (2)	9.25	3497
West Timor-Alor-Pantar	114	4 (5)	9.04	3531
Wissel Lakes	3	3 (5)	23.77	2060
West Papuan	324	3 (3)	0.24	9083
North Halmahera	97	4 (4)	13.14	2962
West Bird's Head	31	5 (5)	17.66	2512

Yele-West New Britain	1	2 (2)	1.47	6293

Table 6. ASJP dates for language groups of North and Middle America

Group	Pairs	Subgroups	Similarity	Date
Algic	51	3 (3)	2.39	5554
Algonquian	180	4 (7)	10.23	3343
Central	77	7 (8)	15.84	2678
Eastern	28	8 (10)	12.60	3026
Plains	1	2 (3)	3.44	5002
Caddoan	3	2 (2)	3.86	4828
Northern	2	2 (2)	12.52	3035
Chumash	10	5 (7)	28.34	1792
Eskimo-Aleut	8	2 (2)	3.26	5084
Eskimo	16	2 (2)	27.43	1842
Gulf	3	3 (4)	0.53	7859
Hokan	167	3 (3)	3.64	4915
Esselen-Yuman				
Yuman	46	5 (6)	27.01	1865
Northern	47	3 (3)	2.22	5666
Karok-Shasta	4	2 (2)	2.93	5246
Pomo	6	2 (2)	41.12	1226
Iroquoian	6	2 (2)	3.79	4855
Northern Iroquoian	5	2 (3)	11.42	3176
Five Nations	6	2 (2)	30.64	1673
Kiowa-Tanoan	2	2 (2)	9.64	3434
Mayan	1449	5 (5)	21.39	2220
Cholan-Tzeltalan	20	2 (2)	35.91	1432
Cholan	6	2 (2)	43.26	1148
Tzeltalan	6	4 (8)	65.76	511
Huastecan	1	2 (4)	40.29	1257
Kanjobalan-Chujean	15	2 (2)	41.14	1225
Chujean	2	2 (3)	45.89	1058
Kanjobalan	4	2 (2)	54.27	803
Quichean-Mamean	667	2 (2)	31.14	1649
Greater Mamean	100	2 (2)	34.51	1492
Greater Quichean	166	6 (6)	48.30	981
Yucatecan	6	2 (2)	54.76	790
Mopan-Itza	2	2 (2)	51.35	887
Yucatec-Lacandon	1	2 (3)	62.01	601
Misumalpan	3	3 (4)	14.87	2774
Mixe-Zoque[1]	49	2 (2)	36.51	1407
Mixe	14	3 (3)	50.94	900

Zoque	16	3 (3)	54.86	787
Muskogean	8	2 (2)	29.71	1720
Eastern	6	4 (4)	42.16	1188
Western	1	2 (2)	73.34	345
Na-Dene	22	2 (2)	-0.25[5]	
Nuclear Na-Dene	21	2 (2)	0.34	8532
Athapaskan-Eyak	20	2 (2)	5.82	4203
Athapaskan	138	4 (8)	23.74	2062
Oto-Manguean	2108	8 (8)	1.21	6591
Chiapanec-Mangue	1	1 (1)+1	18.46	2445
Chinantecan	6	4 (14)	25.80	1935
Mixtecan	14	2 (2)	4.65	4542
Mixtec-Cuicatec	6	2 (2)	11.69	3140
Trique	1	2 (3)	46.96	1024
Otopamean	10	2 (4)	8.34	3654
Otomian	6	2 (2)	21.48	2214
Popolocan	71	3 (3)	12.52	3036
Chocho-Popolocan	6	2 (2)	21.55	2209
Mazatecan	52	8 (8)	55.29	775
Subtiaba-Tlapanecan	14	5 (5)	49.35	948
Zapotecan	75	2 (2)	11.62	3149
Chatino	3	3 (6)	47.79	997
Zapotec	300	25 (57)	30.60	1676
Penutian	230	7 (8)	2.44	5522
Maiduan	5	3 (4)	41.29	1219
Oregon Penutian	3	2 (3)	0.04	11,886
Coast Oregon	3	3 (3)	3.67	4902
Plateau Penutian	2	2 (2)	6.03	4147
Sahaptin	1	2 (5)	15.35	2725
Yok-Utian	18	2 (2)	5.07	4413
Utian	14	2 (2)	8.29	3663
Miwokan	12	2 (2)	22.54	2141
Salishan	129	5 (5)	7.44	3827
Central Salish	37	5 (6)	18.29	2459
Interior Salish	8	2 (2)	12.98	2980
Siouan	15	2 (2)	1.59	6178
Siouan Proper	56	3 (3)	11.47	3169
Tequistlatecan	1	2 (2)	41.50	1212
Totonacan	4	2 (2)	35.83	1435
Totonac	48	6 (9)+1	65.96	506
Tepehua	3	3 (3)	64.26	546
Uto-Aztecan	781	2 (2)	6.56	4018
Northern Uto-Aztecan	33	4 (4)	16.93	2576
Numic	14	3 (3)	29.40	1737
Southern Uto-Aztecan	754	2 (2)	9.40	3472
Sonoran	61	4 (5)	19.01	2400

Aztecan				
General Aztec	57	2 (2)	34.13	1509
Wakashan	6	2 (2)	14.80	2781
Northern	1	2 (3)	61.78	606
Southern	2	2 (2)	43.11	1154
Yuki	1	2 (2)	17.80	2500

Table 7. ASJP dates for language groups of South America

Group	Pairs	Subgroups	Similarity	Date
Arauan	17	4 (5)	28.88	1764
Arawakan				
Maipuran	796	6 (6)	6.08	4134
Aymaran	2	2 (3)	45.92	1057
Barbacoan	8	3 (4)	12.16	3080
Cayapa-Colorado	1	2 (2)	36.23	1419
Coconucan	1	2 (2)	69.87	419
Cahuapanan	1	2 (2)	42.23	1185
Carib	72	2 (2)	19.50	2362
Northern	38	5 (5)	19.37	2371
Southern	11	3 (3)	18.74	2422
Chapacura-Wanham	1	2 (2)	25.87	1931
Chibchan	210	10 (10))+1	5.11	4400
Aruak	6	3 (3)+1	14.61	2800
Guaymi	2	2 (2)	10.62	3286
Kuna	1	2 (2)	53.68	820
Rama	1	2 (2)	3.19	5117
Talamanca	9	4 (4)	15.30	2731
Choco	7	2 (2)	20.87	2258
Embera	12	2 (2)	51.76	875
Chon	1	2 (2)	14.87	2774
Guahiban	10	5 (5)	39.39	1291
Jivaroan	6	4 (4)	58.94	678
Katukinan	3	3 (3)	25.29	1965
Macro-Ge	245	12 (14)	0.78	7266
Ge-Kaingang	23	3 (3)	3.47	4989
Yabuti	1	2 (2)	32.02	1607
Maku	26	6 (6)	11.81	3124
Mascoian	2	2 (5)	29.76	1718
Mataco-Guaicuru	25	2 (2)	4.19	4701
Guaicuruan	10	5 (5)	13.61	2909
Mataco	10	5 (7)	18.96	2404
Nambiquaran	3	3 (3)	14.55	2807

Panoan	144	7 (8)	27.24	1853
North-Central	5	3 (6)	22.64	2134
Northern	3	3 (5)	44.70	1099
South-Central	5	2 (9)	27.24	1853
Southeastern	2	2 (2)	50.27	920
Quechuan	18	2 (2)	29.77	1717
Quechua II	17	2 (3)	48.51	974
Tacanan	3	2 (2)	32.37	1590
Araona-Tacana	2	2 (2)	40.04	1266
Tucanoan	83	3 (4)	15.62	2699
Eastern Tucanoan	30	2 (3)	40.70	1241
Western Tucanoan	7	3 (3)	22.32	2156
Tupi	570	8 (10)	8.73	3585
Monde	9	4 (5)	29. 87	1712
Munduruku	1	2 (2)	34.79	1480
Tupari	3	3 (5)	27.28	1850
Tupi-Guarani	434	9 (12)	33.22	1550
Yuruna	1	2 (3)	49.26	951
Uru-Chipaya	2	2 (2)	33.89	1520
Witotoan	12	2 (2)	2.49	5491
Boran	2	2 (2)	20.69	2271
Witoto	3	2 (4)	13.66	2903
Yanomam	23	4 (4)	38.66	1319
Zamucoan	2	2 (2)	14.96	2765
Zaparoan	3	3 (7)	11.40	3178

Acknowledgments

We would like to thank the scholars who commented on earlier versions of this work: Gene Anderson, Peter Bellwood, Roger Blench, Robert Blust, Lyle Campbell, W. South Coblin, Michael Coe, Bernard Comrie, Mark Donohue, Chris Ehret, Michael Fortescue, Anthony Grant, Russell Gray, Terrence Kaufman, Victor H. Mair, Johanna Nichols, Andrew Pawley, Robert Rankin, Malcolm Ross, Laurent Sagart, Paul Sidwell, Brian Stross, Edward Vajda, and James Watters. Our gratitude also extends to five anonymous reviewers.

References Cited

Adelaar, Alexander. 2005. Malayo-Sumbawan. *Oceanic Linguistics* 44:357-388.

———. 2006. The Indonesian migrations to Madagascar: making sense of the multidisciplinary evidence. In *Austronesian diaspora and the ethnogenesis of people in Indonesian archipelago: proceedings of the international symposium.* T. Simanjuntak, I.H.E. Pojoh, and M. Hisyam, eds. Pp. 205-232. Jakarta, Indonesia: Indonesian Institute of Sciences, LIPI Press.

Aikio, Ante. 2006. On Germanic-Saami contacts and Saami prehistory. *Suomalais-Ugrilaisen Seuran Aikakauskirja. Journal de la Société Finno-Ougrienne* 91:9-55.

Anthony, David W. 1995. Horse, wagon & chariot: Indo-European languages and archaeology. *Antiquity* 65:554-565.

Bellwood, Peter. 2007. Southeast China and the prehistory of the Austronesians. In *Lost maritime cultures: China and the Pacific*. Tianlong Jiao, ed. Pp. 36-53. Honolulu: Bishop Museum Press.

Bellwood, Peter, and Eusebio Dizon. 2008. Austronesian cultural origins: out of Taiwan, via the Batanes Islands, and onwards to Western Polynesia. In *Past human migration in East Asia: matching archaeology, linguistics, and phylogenetics.* Alicia Sanchez-Mazas, Roger Blench, Malcolm D. Ross, Ilia Peiros, and Marie Lin, eds. Pp. 24-39. London: Routledge.

Bellwood, Peter, and Peter Hiscock. 2005. Australia and the Austronesians. In *The human past: world prehistory and the development of human societies*. Chris Scarre, ed. Pp. 264-305. London: Thames and Hudson.

Bergsland, Knut, and Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3:115-153.

Blust, Robert. 2000. Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics*, vol. 2. Colin Renfrew, April McMahon, and Larry Trask, eds. Pp. 311-331. Cambridge: McDonald Institute for Archaeological Research.

Bostoen, Koen, and Claire Grégoire. 2007. La question bantoue: bilan et perspectives. *Mémoires de la Société de Linguistique de Paris* 15:73-91.

Bremmer, Rolf H., Jr. 2009. *An introduction to Old Frisian. History, grammar, reader, glossary*. Amsterdam, Netherlands: John Benjamins.

Brochado, José Joachim Justiniano Proenza. 1984. *An ecological model of the spread of pottery and agriculture into Eastern South America.* PhD dissertation, University of Illinois, Champaign-Urbana, IL.

Brown, Cecil H. 2006. Glottochronology and the chronology of maize in the Americas. In *Histories of maize*. John Staller, Robert Tykot, and Bruce Benz, eds. Pp. 647-663. Amsterdam, Netherlands: Elsevier.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Vilupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF – Language Typology and Universals* 61:285-308.

Bury, J. B.1923. *A history of the later Roman Empire, from the death of Theodosius I. to the death of Justinian (A.D. 395 to A.D. 565)*. London: Macmillan.

Castillo, Dennis. 2006. *The Maltese cross: a strategic history of Malta.* Westport, CT: Praeger Security International.

Dahl, Otto Chr. 1951. *Malgache et Maanjan: une comparaison linguistique.* Oslo: Egede-Instituttet.

Dixon, R. M. W. 1997. *The rise and fall of languages.* Cambridge: Cambridge University Press.

Ehret, Christopher. 2000. Testing the expectations of glottochronology against the correlations of language and archaeology in Africa. In *Time depth in historical linguistics*, vol. 1. Colin Renfrew, April McMahon, and Larry Trask, eds. Pp. 373-399. Cambridge: McDonald Institute for Archaeological Research.

Embleton, Sheila M. 1986. *Statistics in historical linguistics*. Bochum, Germany: Brockmeyer.

Evans, Nicholas, and Rhys Jones. 1997. The cradle of the Pama-Nyungans: archaeological and linguistic speculations. In *Aboriginal Australia in global perspective: archaeology and linguistics.* Patrick McConvell and Nicholas Evans, eds. Pp. 385-417. Melbourne: Oxford University Press.

Fodor, Istvan. 1962. Comment on 'On the validity of glottochronology'. *Current Anthropology* 3:131-134.

Fortescue, Michael D. 1998. *Language relations across Bering Strait: reappraising the archaeological and linguistic evidence*. London: Cassell.

Golden, Peter B. 1998. The Turkic peoples: a historical sketch. In *The Turkic languages*. Lars Johanson and Éva Á. Csató, eds. Pp. 16-29. London: Routledge.

Green, Roger C. 2003. The Lapita horizon and traditions: signature for one set of Oceanic migrations. In *Pacific archaeology: assessments and prospects*, Christophe Sand, ed. Pp. 95-120. Nouméa: Service des Musées et du Patrimonie de Nouvelle-Calédonie.

Greenhill, S.J., R. Blust, and Russell D. Gray. 2008. The Austronesia Basic Vocabulary Database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4:271-283.

Grimes, Joseph E., and Frederick B. Agard. 1959. Linguistic divergence in Romance. *Language* 35:598-604.

Güldemann, Tom. Forthcoming. Changing profile when encroaching on hunter-gatherer territory: towards a history of the Khoe-Kwadi family in Southern Africa. In *Hunter-gatherers and linguistic history: a global perspective*. Tom Güldemann, Patrick McConvell, and Richard Rhodes, eds. Cambridge: Cambridge University Press.

Hammarström, Harald. 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica* 27:197-213.

Haugen, Einar. 1982. *Scandinavian language structures: a comparative historical survey*. Minneapolis, MN: University of Minnesota Press.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Linguistic Distances.* J. Nerbonne and E. Hinrichs, eds. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney. Pp. 51-62. http://urd.let.rug.nl/nerbonne/papers/heeringa-et-al-coling-2006.pdf

Heggarty, Paul. 2007. Linguistics for archaeologists: principles, methods and the case of the Incas. *Cambridge Archaeological Journal* 17:311-340.

Holman, Eric W. 2010. Program for calculating ASJP dates (version 1.0). http://email.eva.mpg.de/~wichmann/software.htm (accessed August 2010).

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

Humphreys, Humphrey Lloyd. 1993. The Breton language: its present position and historical background. In *The Celtic languages*. Martin J. Ball and James Fife, eds. Pp. 606-643. London: Routledge.

Hymes, Dell H. 1960. Lexicostatistics so far. *Current Anthropology* 1:3-44.

Jackson, Kenneth. 1951. Common Gaelic: the evolution of the Goedelic languages. *Proceedings of the British Academy* 37: 71-97.

Janhunen, Juha. 2003. Proto-Mongolic. In *The Mongolic languages.* Juha Janhunen, ed. Pp. 1-29. London: Routledge.

Jaxontov, S. 1999. Glottochronology: difficulties and perspectives. In *Historical linguistics and lexicostatistics,* V. Shevoroshkin and P. J. Sidwell, eds. Pp. 51-59. Melbourne: Association for the History of Language.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the seventh conference of the European chapter of the Association for Computational Linguistics*, 60-66. San Francisco: Morgan Kaufmann.

Kropp-Dakubu, Mary Esther (ed). 1977-1980.West African language data sheets. Legon, Ghana: West African Linguistic Society.

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

Lewis, M. Paul (ed.). 2009. *Ethnologue*. 16th edition. Dallas, TX: SIL International. http://www.ethnologue.com

Lohr, M. 2000. New approaches to lexicostatistics and glottochronology. In *Time depth in historical linguistics*, vol. 1. Colin Renfrew, April McMahon, and Larry Trask, eds. Pp. 209-222. Cambridge: McDonald Institute for Archaeological Research.

Matras, Yaron. 2002. *Romani: a linguistic introduction*. Cambridge: Cambridge University Press.

Mitchell, Donald. 1990. Prehistory of the coasts of southern British Columbia and northern Washington. In *Handbook of North American indians*, vol. 7. Wayne Suttles, ed. Pp. 340-358. Washington, DC: Smithsonian National Museum of Natural History.

Moraes Farias, P. F. de. 2003. *Arabic medieval inscriptions from the Republic of Mali: epigraphy, chronicles, and Songhay-Tuāreg history*. Oxford: Oxford University Press.

Nichols, Johanna, and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2:760-820.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Parpola, Asko. 1999. The formation of the Aryan branch of Indo-European. In *Archaeology and language, III: artefacts, languages and texts*. Roger Blench and Matthew Spriggs, eds. Pp. 180-207. London: Routledge.

Pawley, Andrew. 2009. The role of the Solomon Islands in the first settlement of Remote Oceania: bringing linguistic evidence to an archaeological debate. In *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*. Alexander Adelaar and Andrew Pawley, eds. Pp. 515–540. Canberra, Australia: Pacific Linguistics.

Pohl, Walter. 2004. *Die Germanen.* Enzyklopädie deutscher Geschichte 57. Munich, Germany: Oldenbourg.

Pugh, Stefan M. 2007. *A new historical grammar of the East Slavic languages,* vol. 1: *introduction and phonology.* Munich: LINCOM Europa.

Ramsay, S. Robert. 1987. *The languages of China.* Princeton: Princeton University Press.

Rankin, Robert L. 2006. Siouan tribal contacts and dispersions evidenced in the terminology for maize and other cultigens. In *Histories of maize*. John Staller, Robert Tykot, and Bruce Benz, eds. Pp. 563-575. Amsterdam, Netherlands: Elsevier.

Renfrew, Colin, April McMahon, and Larry Trask (eds.). 2000. *Time depth in historical linguistics*. Cambridge: McDonald Institute for Archaeological Research.

Róna-Tas, A. 1991. *An introduction to Turkology*. Szeged, Hungary: Universitas Szegediensis de Attila József.

Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra, Australia: Pacific Linguistics.

Ross, Malcolm, and Åshild Næss. 2007. An Oceanic origin for Äiwoo, the language of the Reef Islands? *Oceanic Linguistics* 46:456-498.

Sagart, Laurent. 1999. *The roots of Old Chinese*. Amsterdam: John Benjamins.

Sagart, Laurent, Roger Blench, and Alicia Sanchez-Mazas. 2005. Introduction. In *The peopling of East Asia: putting together archaeology, linguistics and genetics*. Laurent Sagart, Roger Blench, and Alicia Sanchez-Mazas, eds. Pp. 1-14. London: RoutledgeCurzon.

Sapir, Edward. 1916. *Time perspective in aboriginal American culture: a study in method.* Geological Survey of Canada, Memoir 90 ("Anthropological Series," No. 13.) Ottawa.

Schenker, Alexander M. 1995. *The dawn of Slavic: an introduction to Slavic philology*. New Haven: Yale University Press.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35: 335-357.

Serva, Maurizio and Filippo Petroni, Indo-European languages tree by Levenshtein distance. 2008. *Europhysics Letters* 81, paper 68005 (March 2008). http://www.iop.org/EJ/journal/EPL

Sidwell, Paul. 2006. Dating the separation of Acehnese and Chamic by etymological analysis of the Aceh-Chamic lexicon. *Mon-Khmer Studies* 36:187-206.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16:157-167.

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

Thurgood, Graham. 1999. *From Ancient Cham to modern dialects: two thousand years of language contact and change.* Honolulu: University of Hawai'i Press. (Oceanic Linguistics Special Publication No. 28)

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

Vajda, Edward J. Forthcoming. Yeniseic substrates and typological accommodation in Central Siberia. In *Hunter-gatherers and linguistic history: a global perspective*. Tom Güldemann, Patrick McConvell, and Richard Rhodes, eds. Cambridge: Cambridge University Press.

Watson, Alaric. 1999. *Aurelian and the third century*. London: Routledge.

Weiers, Michael. 2003. Moghol. In *The Mongolic languages.* Juha Janhunen, ed. Pp. 248-264. London: Routledge.

Wichmann, Søren. 2006. Mayan historical linguistics and epigraphy: a new synthesis. *Annual Review of Anthropology* 35:279-294.

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

Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck, and Helen Geyer. 2010. The ASJP Database (version 13). http://email.eva.mpg.de/~wichmann/languages.htm (accessed August 2010).

[1] Sapir (1916:76).

[2] Consult http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm for full details on ASJP, including references to sources of data. Especially rich sources are the Austronesian Basic Vocabulary Database (http://language.psy.auckland.ac.nz/austronesian/) described by Greenhill, Blust, and Gray (2008); Kropp-Dakubu

(1977-1980); the Rosetta Project (http://www.rosettaproject.org); and the now defunct online database for South American languages maintained by the late Lincoln Ribeiro (formerly posted as http://paginas.terra.com.br/educacao/GICLI/ListasEnglish.htm). We are particularly grateful to the more than 70 scholars who have contributed original field data.

[3] More formally, let two languages, A and B, be given, and let n be the number of items (out of 40) attested in both languages. Let d_{ij} denote LDN between item i in language A and item j in language B. Then

$$\text{LDND} = [\Sigma_i(d_{ii})/n]/\ [\Sigma_{i\neq j}(d_{ij})/n(n\text{-}1)].$$

[4] Although 92% for s_0 and 0.72 for r are the values that make the best predictions of time depth, they are not the only values consistent with the calibration data. The regression with $s_0 = 92\%$ and $r = 0.72$ is based on the assumption that all the error is in the dates. The alternative assumption that all the error is in the similarities produces $s_0 = 62\%$ and $r = .79$, while intermediate distributions of error produce intermediate values of s_0 and r. These regression analyses imply the testable prediction that if independent estimates of s_0 and r are derived from other data, they should be between 62% and 92% for s_0 and between 0.72 and 0.79 for r.

[5] A negative similarity score indicates that the words not referring to the same concept are more similar on average than words referring to the same concept, which means that the ASJP results do not bring support to this language group as a genealogical unit.