INFERRING THE WORLD TREE OF LANGUAGES FROM WORD LISTS

GERHARD JÄGER

Institute of Linguistics, Tübingen University
Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

SØREN WICHMANN

Leiden University Center of Linguistics, Leiden University
Leiden, The Netherlands
and
Laboratory of Quantitative Linguistics, Kazan Federal University
Kazan, Russia
wichmannsoeren@gmail.com

Since its launch in 2007, the *Automated Similarity Judgment Program* has collected basic vocabulary lists from more than 6,000 languages and dialects, covering close to two thirds of the world's languages. Using these data and phylogenetic techniques from computational biology, such as weighted sequence alignment and distance-based phylogenetic inference, we computed a phylogenetic language tree covering all continents and language families. Our method relies on word lists in phonetic transcription only, i.e. it does not rely on expert cognacy judgments. This decision enabled us to perform inference across the boundaries of language families. The world tree of languages thus obtained largely recaptures the established classification of languages into families and their sub-groupings. Additionally it reveals intriguing large-scale patterns pointing at a statistical signal from deep time.

1. Introduction

Hardly any element of human culture is as highly susceptible to vertical transmission and preserved as faithfully over time as language (Holman, Wichmann, Brown, & Eff, 2015). For these reasons language relations offer a unique framework for the study of cultural processes: by projecting such processes onto linguistic phylogenies ancestral states and horizontal transmission events can be inferred (Mace & Holden, 2005). Existing linguistic phylogenies, however, impose a limitation on such exercises. Both the traditional comparative method of historical linguistics and character-based methods from the modern biological toolkit offer tools for classifying languages, but they are only applicable to groups of languages that have already been demonstrated to be related. In this paper we

present cutting-edge distance-based methods which do not rely on this assumption and therefore allow for determining a relationship between any pair of languages drawn from the pool of the entire global linguistic diversity. More specifically, using the so-called ASJP database (Wichmann et al., 2013), we employ pairwise sequence alignment and distance-based phylogenetic inference to infer a tree of c. 6,000 languages and dialects, covering all continents and language families. This tree correctly identifies most established language families to a very good approximation and recovers their assumed internal structure with high accuracy. Additionally, it reveals a signal of common descent or contact beyond the level of established families. A quantification of linguistic distances such as the one which is tested here through the inference of a world language tree, promises to bridge gaps between historical linguistics and other disciplines within the social sciences. [a]

2. The Automated Similarity Judgment Program

Since its launch in 2007, the collaborative project known as the Automated Similatity Judgment Program (ASJP) has achieved the compilation of a database of close to two thirds of the world's languages consisting of 40-item lists of universally stable lexical concepts (Holman et al., 2008). Publications drawing upon these data have mostly employed a modification of the Levenshtein distance called LDND (Wichmann, Holman, Bakker, & Brown, 2010) in order to compute a linguistic distance between the doculects of the database. The distance measure employed in the present paper demonstrably represents an improvement over LDND. Among the resources published on the ASJP site is (different versions of) an 'ASJP World Language Tree of Lexical Similarity' (see http://asjp.clld.org/download), similar in spirit to the tree discussed in the present paper, but based on LDND and vanilla Neighbor-Joining. The trees made available on the ASJP site were never intended as real publications, only as specimens providing some potentially useful insights into the data. In contrast, the tree that we present here is the result of extensive research towards developing an optimal distance measure and finding the most adequate algorithm for inferring a phylogeny based on the distances computed.

[a] In recent years, linguistic distances computed from the ASJP database have become a tool widely used by economists for studying how linguistic differences influence investment, trade, tourism, migration preferences and the L2 proficiency and general success of migrants. A paper by Isphording and Otten (2011) seems to have initiated this trend; cf. (Melitz & Toubal, 2014) for one of many recent examples. For comparative anthropology ASJP distances have also proven useful (Walker, Wichmann, Mailund, & Atkisson, 2014). Additionally, current research to which the present authors have contributed suggests that paleoanthropology and genetics can also profit by introducing an ASJP-derived distance measure for the purpose of correlational studies.

3. Distance measures

We defined two pairwise distance measures between doculects. The first one (taken from Jäger, 2013; see also Jäger, 2015) — called *PMI distance* as it is built on the notion of *Pointwise Mutual Information* between sound strings —, quantifies the lexical similarity between lists using sequence alignment. It aggregates information both about sound changes and the gain/loss of cognate classes. PMI distances are determined via sequence alignment, using differential weights for different symbol pairings. These weights are determined in a data-oriented way via unsupervised learning from the ASJP data.

To estimate the likelihood of sound correspondences, a corpus of *probable cognate pairs* was compiled from the ASJP data using two heuristics. First, a similarity measure between word lists related to the above-mentioned LDND distances was defined and the 1% of all ASJP doculect pairs with highest similarity were kept as *probably related*. (This notion is rather strict; English, for instance, turns out to be "probably related" to all and only the other Germanic doculects. In total, 99.9% of all doculect pairs defined that way belong to the same language family.) Second, the normalized Levenshtein distance was computed for all translation pairs from probably related doculects. Those with a distance below a certain threshold were considered as *probably cognate*. These probable cognate pairs were used to estimate PMI scores. Subsequently, all translation pairs were aligned using the PMI scores from the previous step as weights. This resulted in a measure of string similarity, and all pairs above a certain similarity threshold were treated as probable cognates in the next step. This procedure was repeated ten times. In the last step, approximately 1.3 million probable cognate pairs were used to estimate the final PMI scores.

Again, the similarity threshold being used is rather strict. For illustration, the only probable cognates pair between English and German that were kept during the last iteration are `fiS/fiS` 'fish', `laus/laus` 'louse', `bl3d/blut` 'blood', `horn/horn` 'horn', `brest/brust` 'breast', `liv3r/leb3r` 'liver', `star/StErn` 'star', `wat3r/vas3r` 'water', and `ful/fol` 'full'. To determine the distance between two word lists, all string similarities in the Cartesian product of the two lists are calculated. The distance between the word lists is a measure of how much the similarities between synonymous words (which are candidates for cognate pairs) exceed the similarity of non-synonymous pairs (i.e, random pairs of words). For more details, see (Jäger, 2013). The full PMI distance matrix is available online at http://www.evolaemp.uni-tuebingen.de/details.html.

To calculate the secondary distance measure we represented each doculect as a binary vector representing the presence/absence of bigrams of the 41 ASJP sound classes in the corresponding word lists. The *bigram inventory distance* between two doculects is then defined as the Jaccard distance between the corresponding vectors.

4. Phylogenetic inference

Based on the PMI distance and the bigram inventory distance, a phylogenetic tree was inferred using the *Minimum Variance Reduction* (MVR) algorithm (Gascuel, 2000) as implemented in the R package *ape* (Paradis, Claude, & Strimmer, 2004). Phylogenetic inference proceeds in two steps. First the two distances matrices (PMI distances and bigram inventory distances) are aggregated into a Consensus Distance Matrix using the *super distance matrix* (SDM) method (implemented in *ape*) from (Criscuolo, Berry, Douzery, & Gascuel, 2006). The relative weight of lexical to bigram inventory distances was, somewhat arbitrarily, set to 10:1. In this way it was assured that phylogenetic inference is dominated by the information in the PMI distances, and bigram inventory distances only act as a kind of tie breaker in situations where lexical distances do not provide a detectable signal.

SDM computes an aggregated distance matrix and a variance matrix associated to that distance matrix, which in turn serve as input for MVR. MVR is a modification of the well-known *Neighbor-Joining* algorithm which uses both distances and their estimated variances to compute a tree.

5. The world tree

The full tree is made available online.[b] It is summarized in Fig. 1. All clades comprising doculects from the same family (according to the WALS classification, cf. Haspelmath, Dryer, Gil, & Comrie, 2008), with maximally one outlier, are collapsed into a triangle.

Generally, the automatically induced tree captures the established expert classification of languages into families fairly well. Of the 52 WALS families for which ASJP contains at least 10 doculects, 40 families correspond to a clade in the tree with an F-score[c] ≥ 0.95 (meaning: the binary classification of taxa induced by that clade has an F-score ≥ 0.95 when evaluated against the extension of that family according to WALS). The 12 poorly recognized families are Trans-New Guinea (maximal F-score 0.61), Sko (0.67), Macro-Ge (0.68), Marind (0.71), Penutian (0.80), Otomanguean (0.81), Torricelli (0.82), Nilo-Saharan (0.85), West Papuan (0.89), Sepik (0.92), Sino-Tibetan (0.92), and Hokan (0.93). Most of these families are controversial.

[b] It can be inspected at http://www.sfs.uni-tuebingen.de/~gjaeger/ODljNT/worldTree.svg, using any standard web browser.

[c] The *F-score* is a statistic measuring the goodness of fit of a binary classification. It is defined as the harmonic mean between *precision* and *recall*, where

$$\begin{aligned} \text{precision} &\doteq \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}} \\ \text{recall} &\doteq \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false negatives}}. \end{aligned}$$

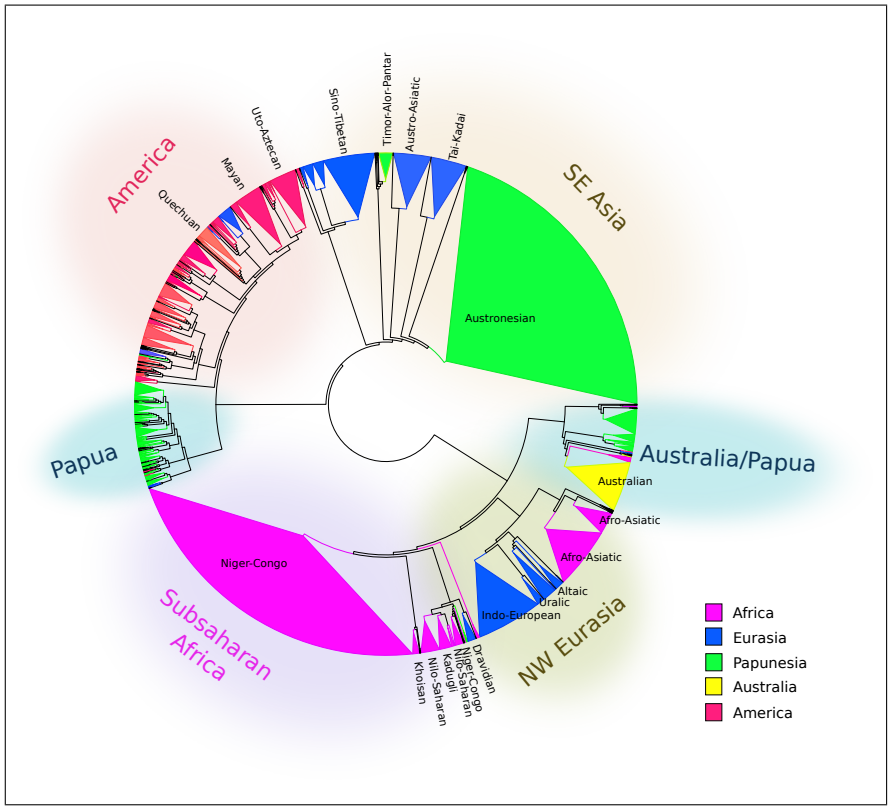

Figure 1. The world tree of languages

For the more conservative Glottolog classification (Hammarström, Forkel, Haspelmath, & Bank, 2015), only 9 of the 59 families with at least 10 members have an F-score below 0.95: Nuclear Trans-New Guinea (0.61), Sko (0.67), Otomanguean (0.81), Nuclear Torricelli (0.82), Nuclear Macro-Je (0.83), Sepik (0.85), Pama-Nyungan (0.88), Afro-Asiatic (0.93), and Sino-Tibetan (0.93).

The internal classification of language families, as assumed by experts, is also recaptured with high accuracy. The *Generalized Quartet Distance* (Pompei, Loreto, & Tria, 2011) between the automatically induced tree and the WALS classification is as low as 0.033 (0.066 for the Ethnologue Lewis, Simons, & Fennig, 2015 and 0.046 for the Glottolog classification).

6. Final remarks

Comparing extant languages in order to infer the evolution of the presently observed linguistic diversity only allows us to see the top of the iceberg in some

detail: we can reconstruct ancestral languages with relatively high precision down to around 5000 years before present. From around 5000 BP to around 10,000 BP the signal becomes increasingly more noisy and eventually gets lost. In this situation, the best we can do in order to reach further back in time is to simultaneously compare the bulk of the worlds languages reaching the kind of result shown in Fig. 1. This result indicates the existence of four distinct areas: Africa + Western and Northern Eurasia, SE Asia (including Island SE Asia occupied by Austronesians), the Americas, and Sahul. The Papuan languages of New Guinea are distributed in three separate clusters. One of these is adjacent to Australia and may represent the earliest stratum of Papuan languages; we regard the Australian languages and these Papuan languages as belonging to one Sahul cluster. The two other Papuan clusters resist meaningful interpretations in terms of the regions they occupy in the tree.

Many aspects of the topology in the upper regions of the tree — above the level of established families — evidently reflect sustained contact rather than vertical transmission. For instance, since Vajda (2010) it is widely believed that the Yeniseian languages of Central Siberia are genealogically related to the Athapaskan-Eyak-Tlingit languages of North America, but in the tree they are located within the NW Eurasian cluster. The Eskimo-Aleut languages form a clade with the Chukotko-Kamchatkan languages (which could be interpreted as indicative of common descent by proponents of the Nostratic/Eurasiatic macro-family), but this clade is part of a larger clade comprising languages from the North-American Pacific Northwest, including Wakashan and the Salish. While the geographic proximity points to possible contact, there are no good reasons to assume that the involvement of Eskimo-Aleut in this larger clade should be of a genealogical nature. This list of examples could be increased.

On the other hand, several instances of known intense contact in shallow time (up to ca. 5000 BP) are not detectable in the tree. For instance, we observe neither an affinity between the Papuan languages and Austronesian nor between Dravidian and Indo-European or between the SE Asian families and Indic or Mongolic or Tungusic languages. Based on these, admittedly preliminary, considerations we tentatively conclude that language contact is reflected in the world tree mostly if it has been sustained since deep time.

In a few cases — surprisingly few, we dare to say —, the observed patterns can only be interpreted as the result of the accumulation of chance similarities. For instance, Siouan (North America) and Alor-Pantar (Papunesia) are embedded in the SE Asia part of the tree, Ainu (East Asia) in the American part, and Dravidian (South Asia) in (or neighboring to) the Subsaharan African part. Finally, as mentioned earlier, the loci in the tree of two large clusters containing Papuan languages are not meaningful, so they should probably be regarded as random.

To sum up, with some caveats we can say that the four distinct areas identified by the tree represent ancient zones of diffusion and interaction, and a more

tentative hypothesis, subject to further testing, is that there are also deep genealogical relations among some of the languages within these four regions, reaching far beyond conventionally established families. The fact that our data and methods produce clear geographical clusters shows that the deep branchings in the world tree, at least for a large part, are not due to chance. Thus, through a single, consistent, and novel method of comparative linguistics we have obtained a framework for tracing the evolution of language back to minimally four intermediate geo-genealogical aggregates. More trivially, the method also produces fairly accurate results for more recent phylogenetic evolution. Apart from its value as a contribution to historical linguistics, the tree also represents a potentially useful framework for studying cultural evolution at both large and small scales, and the distance measure on which it is based can be employed in cross-disciplinary correlational studies of many different kinds.

Acknowledgements

This work was supported by the ERC Advanced Grant 324246 *EVOLAEMP* (GJ), the ERC Advanced Grant 295918 *MESANDLIN(G)K* and support from the Russian government to further the competitive development of Kazan Federal University (SW), as well as the DFG-Center for Advanced Studies in the Humanities 2237 *Words, Bones, Genes, Tools* (both authors), which is gratefully acknowledged.

References

Criscuolo, A., Berry, V., Douzery, E. J. P., & Gascuel, O. (2006). SDM: a fast distance-based approach for (super) tree building in phylogenomics. *Systematic Biology*, *55*(5), 740-755.

Gascuel, O. (2000). Data model and classification by trees: the minimum variance reduction (MVR) method. *Journal of Classification*, *17*(1), 67-99.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2015). *Glottolog 2.5.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (available online at `http://glottolog.org`, Accessed on 2015-09-18.)

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2008). *The World Atlas of Language Structures online.* Max Planck Digital Library, Munich. (`http://wals.info`)

Holman, E. W., Wichmann, S., Brown, C. H., & Eff, E. A. (2015). Inheritance and diffusion of language and culture: A comparative perspective. *Social Evolution & History*, *14*(1), 49-64.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, *42*(2), 331-354.

Isphording, I. E., & Otten, S. (2011). Linguistic distance and the language fluency of immigrants. *Ruhr Economic Papers*, *274*.

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, *3*(2), 245-291.

Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, *112*(41). (doi: 10.1073/pnas.1500331112)

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2015). *Ethnologue: Languages of the world* (Eighteenth ed.). Dallas, Texas: SIL International. (Online version: `http://www.ethnologue.com`)

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology and Evolution*, *20*(3), 116121.

Melitz, J., & Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, *93*, 351-363.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289-290.

Pompei, S., Loreto, V., & Tria, F. (2011). On the accuracy of language trees. *PLoS ONE*, *6*(6), e20109.

Vajda, E. (2010). A Siberian link with Na-Dene languages. *Archaeological papers of the University of Alaska*, *5*, 75-156.

Walker, R. S., Wichmann, S., Mailund, T., & Atkisson, C. J. (2014). Cultural phylogenetics of the Tupi language family in Lowland South America. *PLoS ONE*, *7*(4), e35025.

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389*(17), 3632-3639. (doi:10.1016/j.physa.2010.05.011)

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., & Valenzuela, P. (2013). *The ASJP Database (version 16).* `http://asjp.clld.org`.