CORRELATED EVOLUTION OR NOT? PHYLOGENETIC LINGUISTICS WITH SYNTACTIC, COGNACY AND PHONETIC DATA

GIUSEPPE LONGOBARDI,[*][†]ARMIN BUCH,[‡] ANDREA CEOLIN,[§] AARON ECAY,[*] CRISTINA GUARDIANO,[¶] MONICA IRIMIA,[*] DIMITRIS MICHELIOUDAKIS,[*] NINA RADKEVICH,[*] AND GERHARD JÄGER[‡]

[*]*Department of Language and Linguistic Science, University of York, York, United Kingdom*

[†]*Laboratorio di Linguistica e Antropologia cognitiva, Università di Trieste, Trieste, Italy*

[‡]*Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany*

[§]*Department of Linguistics, University of Pennsylvania, Philadelphia, USA*

[¶]*Dipartimento di Comunicazione e Economia, Università di Modena e Reggio Emilia, Modena, Italy*

In this work we compare, on the well explored domain of Indo-European languages, the phylogenetic outputs of three different sets of linguistic characters: traditional etymological judgments, a system for phonetic alignment of lists of cognates, and a set of values for generative syntactic parameters. The correlation and relative informativeness of distances and phylogenies generated by the three types of characters can thus be for the first time accurately evaluated, and the degree of success of the last two, innovative, alternatives to the classical comparative method can be so assessed.

1. Introduction

For many decades now historical linguistics has sought taxonomic procedures alternative and complementary to the classical comparative method, in order to strengthen and expand the results achieved through the latter. Two ideal adequacy standards such new methods should aim at are the possibility of applying also beyond the limits imposed by classical etymological criteria, and that of being easily subjected to quantitative analysis and automated statistical testing. For this purpose, we compare, on a domain where significant historical knowledge is already available (IE languages), the computational outputs of three sets of linguistic characters: traditional etymological judgments, a system for phonetic alignment of lists of synonyms, and a set of values for generative syntactic parameters. The correlation and relative informativeness of distances and phylogenies generated by the three types of characters can thus be for the first time accurately evaluated.

2. Data being used

For our study we used three different data collections, providing syntactic, cognacyand phonetic information:

Syntactic parameters: The values for syntactic parameters are taken from the database of (Longobardi et al., 2013)[a] integrated with data about 4 ancient languages (Latin, Classical Greek, Gothic, and Old English). In such a database, the two opposite values of the 56 binary parameters used were represented by '+' and '−'; '0' symbolizes instead a parameter value which is uninformative as fully predictable from the values of other parameters. Hence, such '0's need to be disregarded for the purposes of taxonomic computations, a standard practice in the Parameric Comparison Method (Longobardi & Guardiano, 2009).

Lexical cognacy data: The IELex database (http://ielex.mpi.nl/) contains 207-item Swadesh lists for 157 living and extinct languages. Entries are assigned to cognate classes, based on expert judgments.

Phonetic data: The ASJP database (Wichmann et al., 2013) is a collection of 40-item Swadesh lists for more than 6,000 languages and dialects. All entries are given in a uniform phonetic transcription (see http://asjp.clld.org/ for the actual data).

We identified a sample of 22 Indo-European languages (18 contemporary and 4 ancient ones) for which all three databases provide information. Syntactic parameters were organized in a binary matrix with languages as rows and parameters as columns.

From the data supplied by IELex, we constructed a binary matrix with languages as rows and cognacy classes (such as *dog-A* etc.) as columns. A cell has entry "1" if the row-language has an entry for the column-cognacy class, "0" if the row-language does not have an entry for the column-cognacy class but an entry for another cognacy class for that meaning, and "?" (undefined) otherwise. In total, IELex contains 1,566 cognate classes for the 22 languages in our sample. We excluded those cognate classes from

[a] Available on line at: https://benjamins.com/#catalog/journals/jhl.3.1.07lon/additional

consideration that are present in all 22 languages, such as class A for the concept 'I', which comprises English *I*, Hindi *me*, Latin *ego*, Spanish *yo* etc. (all deriving from the paradigm of PIE $*h_1 eg\hat{h}_2 \acute{o}m$). This leaves us with 1,553 informative binary characters.

Regarding phonetic data, we manually created a lookup table mapping IELex entries to corresponding ASJP entries for all language/concept pair where both databases contain the same word(s). There are several ASJP entries having no counterpart in IELex but clearly belonging to one of the IELex cognate classes. The Russian word `pos` ('dog', i.e. a synonym to `sobak3`), for instance, is evidently cognate to the Serbocroatian `pas` and the Polish *piEs*. We did not include this kind of information into the cognacy character matrix, but we used it to create sound alignments. (An ASJP entry without IELex counterpart was automatically added to a cognate class if its average string similarity to the members of that class exceeds a certain threshold.)

In a next step, the T-Coffee algorithm (Notredame et al., 2000) was applied to perform *multiple sequence alignment* within each cognate class (see Jäger & List, 2015 for a fuller description of how T-Coffee was adapted to phonetic strings).[b]

This is illustrated for the concept *dog* in Tab. 1.

Ideally, sounds within the same column are cognate, i.e. they derive historically from the same ancestor. A "`-`" represents a gap, i.e. a position where a sound has been deleted or added in the lineage leading to that language. A "." indicates that the language in question does not contain a word from that cognate class.

As the multiple sequence alignment has been performed automatically, an experienced historical linguist will not agree with every detail. Most alignments arguably capture genuine sound correspondences though. As there are no gold standard data of that type, it is at present not possible to quantify the quality of our alignments.

The multiple sequence alignments can be transformed into a binary character matrix in the familiar way. For a given column, all sound types occuring there define one binary character. Gaps are not treated as characters. If a language has a "." in a column, all characters from that column are undefined for that language. Again excluding non-informative characters, we end up with a binary character matrix with 1,521 columns.

[b] In (Jäger & List, 2015) a systematic comparison is conducted between Sound Class Based Phonetic Alignment (SCA, List, 2014) and T-Coffee alignment, indicating the latter to be superior.

Table 1. ASJP entries and multiple alignment for cognate classes A, H, and E/concept ‘dog’

language	*ASJP entries*	*multiple sequence alignment*
Italian	kane	k----a-ne---..........
Spanish	-	....
French	Sia	S-----ia----..........
Portuguese	kau	k----a--u---..........
Romanian	k3ne	k----3-ne---..........
Greek	-	....
English	-	....
German	hunt	h----u-n-t--..........
Danish	hun7	h----u-n-7--..........
Icelandic	hintir	h----i-n-tir..........
Norwegian	hund	h----u-n-d--..........
Bulgarian	-	....
Serbocroatian	pas	..................p-as
Polish	piEs	..................piEs
Russian	sobak3,pos	sobak3------......p-os
Irish	ku	k----u------..........
Marathi	k3tra7	............k3tra7....
Hindi	kutta	............kutta-....
Latin	kanis	k----a-nis--..........
Classical Greek	kion	k--i-o-n----..........
Gothic	hunds	h----u-n-d-s..........
Old English	hund	h----u-n-d--..........

3. Distances and correlations

From each of those three binary matrices we computed pairwise distances between languages. For syntactic parameters, the possible values “+” and “-” are symmetric. Therefore the syntactic distance was defined as the Hamming distance between their parameter vectors. As for the cognacy and phonetic matrices, a “1” represents the presence and a “0” the absence of a certain trait in that language. Therefore the *Dice distance* was used (as proposed originally in Longobardi et al., 2015 in such a context), which is defined as

$$d(A,B) \doteq 1 - \frac{|A \cap B|}{1/2(|A| + |B|)},$$

where A is the set of characters where the first language has entry “1”, and likewise for B.

The three distance measures defined above quantify for each pair of languages on how many syntactic parameters they disagree (syntactic distances), how many basic concepts are expressed by unrelated words (cognacy distance) and how many sound pairs participating in regular correspondences are non-identical (phonetic).

We observe the correlations given in Tab. 2 and depicted in Fig. 1. All

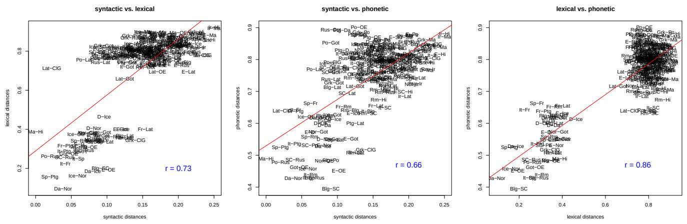

Figure 1. Direct correlations

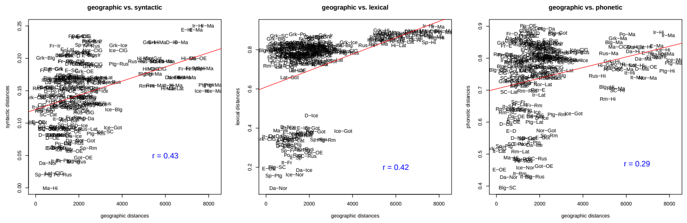

Figure 2. Correlations with geographic distance

correlations are significant ($p < 10^{-4}$ according to the Mantel test).

Table 2. Direct (left) and partial (right) correlations

	syntactic	cognacy
cognacy	0.73	–
phonetic	0.66	0.87

	syntactic	cognacy
cognacy	0.67	–
phonetic	0.62	0.85

Syntactic ($r = 0.43, p < 10^{-4}$), cognacy ($r = 0.42, p < 10^{-4}$) and phonetic ($r = 0.291, p < \times 10^{-4}$) distances are significantly correlated with geographic distance; cf. Fig. 2.

The correlations with geography may reflect both common descent (as closely related languages tend to be geographically proximate) and effects of language contact. To control for the latter, we computed the partial pairwise correlations between the three linguistic distances conditioned on geographic distance. All three correlations remain strongly significant ($p < 10^{-4}$ according to partial Mantel test) when we control for geography.

4. Phylogenetic inference

Using phylogenetic inference, we can construct the evolutionary scenario best explaining observed data. We performed such inferences for each of the three data sources separately, using two different approaches.

Distance-based inference The simplest family of methods only rely on the assumption that on average, the distance between species/languages increases after they diverge. The relation between distances and divergence times can be noisy. Distance-based inference take a pairwise distance matrix as input and find a tree (with branch lengths specification) such that the path length between two leafs is as close as possible to the input distance between these two taxa.

Here we will consider the "Kitsch" algorithm from the *Phylip* software package (Felsenstein, 1989). It uses a weighted least squares method to assess the fit between a distance matrix and a tree, and it finds a tree that minimizes this distance under the constraint that all leaves have the same distance from the root. This amounts to the assumption that evolution proceeds at the same pace across lineages (the so-called "molecular clock" assumption). It is of course only applicable if all leaves are contemporary. Therefore we will only consider the 18 recent languages in our sample in the sequel.

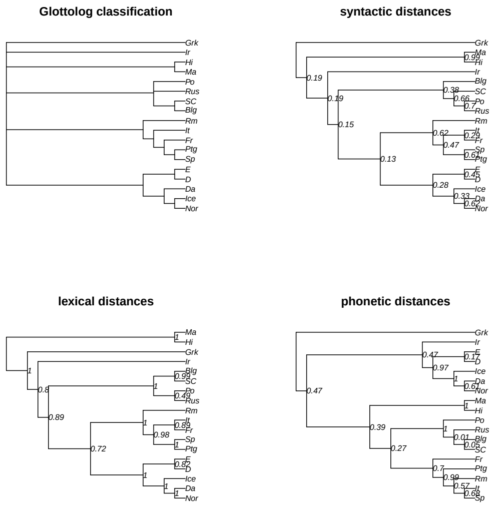

Figure 3. Glottolog expert tree and Kitsch-generated trees; interior node labels are bootstrap support values

The resulting trees are shown in Fig. 3. The numbers at the internal nodes give *bootstrap support values* for the corresponding clade (See Appendix for the technical definition). They indicate how strongly a certain clade is supported by the data.

We mostly find high support values for the established sub-groups of Indoeuropean and lower values for the other nodes. The support values in the syntactic tree are generally rather low. This does not indicate though that syntactic information is less reliable, but it reflects the relative sparseness of the syntactic data.

The topologies of the four trees display remarkable similarities; in particular, three trees (i.e. all but the phonetic one) are consistent in singling out (in slightly different clusterings) Greek, Irish and Indo-Aryan languages from the core formed by Romance, Germanic and Slavic. The phonetic tree does not follow the general pattern on this point, but is suggestive that it groups together Slavic and Indo-Aryan, the two satəm families of the sample, in agreement precisely with one of the longest known phonological differentiations within IE (the treatment of velar consonants). Both the cognacy and the syntactic tree favor a closer relation of Germanic and Romance as opposed to Slavic. At this stage we can only notice that the combination of the latter two observations (i.e. the complementarity of the various methods) may usefully capture some plausible areal effect (reflecting the intermediate geographical position of Slavic, of course).

The quality of an automatically generated tree can be measured by quantifying the degree of its disagreement with the expert classification. A suitable distance measure is the *Generalized Quartet Distance* (GQD; see (Pompei et al., 2011), where it is argued convincingly that GQD is more informative than other tree distance measures such as the Robinson-Foulds distance). Briefly put, this is a measure for the recall of an automatically generated tree being evaluated against an expert tree, i.e. it proportion of topological information in the expert tree that is not correctly recovered in the induced tree. (It is not possible to determine a corresponding precision score since the true topology is usually not known; expert trees are generally underspecified in various respects.)

The GQDs of the three Kitsch-generated trees to the Glottolog tree are 0.031 for syntactic, 0.019 for cognacy and 0.054 for phonetic distances.

Character-based inference Distance-based methods infer a tree explaining diversification patterns in a summary fashion. Character-based phylogenetic inference is a family of more advanced methods that model not just an evolutionary tree but the history of change of each feature (a.k.a. character) along the branches of this tree. We chose Maximum-Likelihood estimation as a representative of those method to infer evolutionary histories for syntactic, cognacy and phonetic characters.[c] The resulting trees are shown in Fig. 4.

The GQDs of the three Maximum-Likelihood trees to the Glottolog tree are 0.019 for syntactic and cognacy, and 0.059 for phonetic distances.

[c] The calculations were carried out using the software package *Paup** (Swofford, 2002). For all three data sets, we chose the model with molecular clock and gamma-distributed rates. Rates, base frequencies and proportion of invariant sites were estimated.

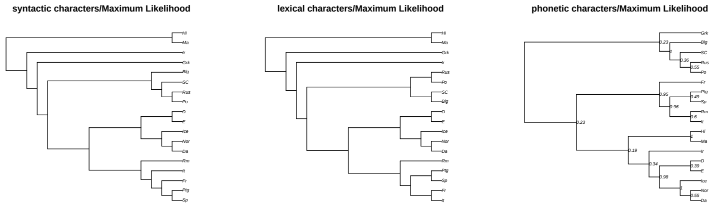

Figure 4. Maximum-likelihood trees

Generally, the syntactic and the cognacy tree are in good agreement both with each other and with the expert tree. The phonetic tree displays some unexpected groupings. To understand this effect better, we performed a Bayesian analysis with the phonetic data (using the Beast software package, `http://beast.bio.ed.ac.uk/`) and used the posterior sample to obtain confidence values for the nodes in the ML tree (displayd in Fig. 4). It turned out that almost all clades with high confidence values (≥ 0.9) correspond to established groupings (Romance, Slavic, Germanic, North-Germanic; the only exception being a group comprising all Romance languages except French). All nodes beyond the sub-families have extremely low support (< 0.5), i.e. they are artefacts of the inference algorithm without real support in the data.

5. Conclusion

By using a variety of methods (correlation studies, distance based phylogenetic inference and character based phylogenetic inference), we provided evidence (a) that different aspects of the language system (syntax, lexicon, sound system) preserve essentially the same historical signal and (b) the insights established by the comparative method are essentially supported by all three signals considered here. These results have a potential impact on future research in phylogenetic linguistics because they indicate that phylogenetic techniques might be suitable to reconstruct earlier language stages by statistical means, and that different linguistic domains inform each other in this endeavor.

Acknowledgements

This work was supported by the ERC Advanced Grant 295733 *LanGeLin* (GL, AC, AE, CG, MI, DM, NR), the ERC Advanced Grant 324246 *EVOLAEMP* (AB, GJ), as well as the DFG-Center for Advanced Studies in the Humanities 2237 *Words, Bones, Genes, Tools* (GJ), which is gratefully acknowledged.

References

Felsenstein, J. (1989). Phylip — Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164-166.

Jäger, G., & List, J.-M. (2015). Factoring lexical and phonetic phylogenetic characters from word lists. In H. Baayen, G. Jäger, M. Köllner, J. Wahle, & A. Baayen-Oudshoorn (Eds.), *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics (QITL 6).* (to appear)

List, J.-M. (2014). *Sequence comparison in historical linguistics.* Düsseldorf: Düsseldorf University Press.

Longobardi, G., Ghirotto, S., Guardiano, C., Tassi, F., Benazzo, A., Ceolin, A., & Barbujani, G. (2015). Across language families: Genome diversity mirrors linguistic variation within Europe. *American journal of physical anthropology.* (doi: 10.1002/ajpa.22758)

Longobardi, G., & Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua*, *119*(11), 1679-1706.

Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., & Ceolin, A. (2013). Toward a syntactic phylogeny of modern indo-european languages. *Journal of Historical Linguistics*, *3*(1), 122-152.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), 205-217.

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

Swofford, D. (2002). *Phylogenetic analysis using parsimony (* and other methods).* Sunderland, MA: Sinauer Associates.

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/.