At the Boundaries of Syntactic Prehistory

Supplementary Information

Andrea Ceolin[A], Cristina Guardiano[A], Giuseppe Longobardi[B*], Monica Alexandrina Irimia[A], Luca Bortolussi[C], Andrea Sgarro[C*]

[A] Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia, Viale Allegri 9, 42121 Reggio Emilia
[B] Department of Language & Linguistic Science, University of York, Vanbrugh College, Heslington, York YO10 5DD
[C] Dipartimento di Matematica e Geoscienze, Università di Trieste, Via Weiss 2, 34128 Trieste

Online Repository

The source code to replicate all the figures and the experiments presented in the paper and in the Supplementary Material is found in the following online repository (along with other relevant data and information): https://github.com/AndreaCeolin/Boundaries.

Table S1. List of the 58 languages

The 58 languages of the dataset, along with their associated Glottolog (https://glottolog.org/glottolog/language) and ISO 639-3 codes, the family and subfamily they traditionally belong to, their location and geographic coordinates, are listed in **TableS1**.

The database partly differs from the one employed in Ceolin et al. (2020). Since our focus here is on macro-comparison and not on micro-variation, on the one hand we removed some varieties from the Romance, the Greek and the Finno-Ugric families which were minimally different from the other related languages of the dataset, and on the other hand we expanded the typological coverage by including two Afroasiatic (Semitic) languages (Arabic, Hebrew) and a Niger-Congo (West Atlantic) one (Wolof).

Language	Label	Glottocode	Iso 639-3 Code	Top-level family	Family	Location	Latitude	Longitude
Afrikaans	Afk	afri1274	afr	Indo-European	Germanic	Cape Town	-33.91	18.42
Arabic	Ar	stan1318	arb	Semitic	West Semitic	Riyad	24.71	46.72
Archi	Arc	arch1244	aqc	NE-Caucasian	/	Machačkala	42.01	47.26
Basque_Central	cB	guip1235	eus	Basque	Guipuzcoan	Vitoria-Gasteiz	42.85	-2.68
Basque_Western	wB	bisc1236	eus	Basque	Biskayan	Bilbao	43.26	-2.93
Bulgarian	Blg	bulg1262	bul	Indo-European	Slavic	Sofia	42.7	23.32
Buryat	Bur	buri1258	bua	Mongolic	Eastern Mongolic	Ulan-Ude	51.82	107.61
Calabrese_Northern	NCA	sout3126	nap	Indo-European	Romance	Verbicaro	39.75	15.19
Cantonese	Can	cant1236	yue	Sino-Tibetan	Sinitic	Hong Kong	22.4	114.11
Danish	Da	dani1285	dan	Indo-European	Germanic	Copenhagen	55.68	12.57
Dutch	Du	dutc1256	nld	Indo-European	Germanic	Amsterdam	52.37	4.89
English	E	stan1293	eng	Indo-European	Germanic	London	51.51	-0.13
Estonian	Est	esto1258	ekk	Uralic	Balto-Finnic	Tallinn	59.44	24.75
Even_1	Ev1	even1260	eve	Tungusic	Northern Tungusic	Kustur	67.79	130.4
Even_2	Ev2	even1260	eve	Tungusic	Northern Tungusic	Sebyan-Kyuyol	65.29	130.01
Evenki	Ek	even1259	evn	Tungusic	Northwestern Tungusic	Bomnak	54.71	128.86
Faroese	Fa	faro1244	fao	Indo-European	Germanic	Tórshavn	62.01	-6.77
Finnish	Fin	finn1318	fin	Uralic	Balto-Finnic	Helsinki	60.17	24.94
French	Fr	stan1290	fra	Indo-European	Romance	Paris	48.86	2.35
German	D	stan1295	deu	Indo-European	Germanic	Berlin	52.52	13.4
Greek	Grk	mode1248	ell	Indo-European	Greek	Athens	37.98	23.73
Greek_Calabria	CG	aspr1238	ell	Indo-European	Greek	Bova Marina	37.93	15.55
Greek_Cypriot	CyG	cypr1249	ell	Indo-European	Greek	Larnaca	34.09	33.62
Hebrew	Heb	hebr1245	heb	Afroasiatic	Semitic	Tel Aviv	32.11	34.85
Hindi	Hi	hind1269	hin	Indo-European	Indo-Aryan	New Delhi	28.61	77.21
Hungarian	Hu	hung1274	hun	Uralic	Ugric	Budapest	47.5	19.04
Icelandic	Ice	icel1247	isl	Indo-European	Germanic	Reykjavik	64.14	-21.94
Irish	Ir	iris1253	gle	Indo-European	Celtic	Dublin	53.35	-6.26
Italian	It	ital1282	ita	Indo-European	Romance	Rome	41.9	12.5
Japanese	Jap	nucl1643	jpn	Japonic	/	Tokyo	35.69	139.69
Kazakh	Kaz	kaza1248	kaz	Turkic	Kipchak	Almaty	43.22	76.85
Khanty	Kh	khan1279	kca	Uralic	Ugric	Kazym	63.7	67.24
Korean	Kor	kore1280	kor	Koreanic	/	Seoul	37.57	126.98
Kirghiz	Kyr	kirg1245	kir	Turkic	Kipchak	Bishkek	42.87	74.57
Lak	Lak	lakk1252	lbe	NE-Caucasian	/	Kumukh	42.54	47.89
Malagasy	Mal	plat1254	plt	Austronesian	Malayo-Polynesian	Antananarivo	18.88	47.51
Mandarin	Man	mand1415	cmn	Sino-Tibetan	Sinitic	Beijing	39.9	116.41
Marathi	Ma	mara1378	mar	Indo-European	Indo-Aryan	Mumbai	19.08	72.88
Mari	mM	mari1278	chm	Uralic	Volgaic	Shap	56.44	47.96
Norwegian	Nor	norw1258	nor	Indo-European	Germanic	Oslo	59.91	10.75
Pashto	Pas	pash1269	pus	Indo-European	Iranian	Khyber Pass	34.09	71.16
Polish	Po	poli1260	pol	Indo-European	Slavic	Warsaw	52.23	21.01
Portuguese	Ptg	port1283	por	Indo-European	Romance	Lisbon	38.72	-9.1
Romanian	Rm	roma1327	ron	Indo-European	Romance	Bucharest	44.43	26.1
Russian	Rus	russ1263	rus	Indo-European	Slavic	Moscow	55.76	37.62
Serbo-Croatian	SC	sout1528	hbs	Indo-European	Slavic	Zagreb	45.82	15.98
Siciliano	Sic	cent1963	scn	Indo-European	Romance	Mussomeli	37.57	13.75
Slovenian	Slo	slov1268	slv	Indo-European	Slavic	Ljubljana	46.06	14.51
Spanish	Sp	stan1288	spa	Indo-European	Romance	Madrid	40.42	-3.7
Tamil	Ta	tami1289	tam	Dravidian	/	Madras	13.08	80.27
Telugu	Te	telu1262	tel	Dravidian	/	Hyderabad	17.39	78.49
Turkish	Tur	nucl1301	tur	Turkic	Oghuz	Ankara	39.93	32.86
Udmurt	Ud	udmu1245	udm	Uralic	Permic	Chur	57.07	53.03
Uzbek	Uz	uzbe1247	uzb	Turkic	Turkestan Turkic	Tashkent	41.3	69.24
Welsh	Wel	wels1247	cym	Indo-European	Celtic	Cardiff	51.48	-3.18
Wolof	Wo	nucl1347	wol	Niger-Congo	West Atlantic	Dakar	14.69	-17.44
Yakut	Ya	yaku1245	sah	Turkic	North Siberian Turkic	Yakutsk	62.04	129.68
Yukaghir	Yu	yuka1259	yux	Yukaghir	Kolmic (Southern Yukaghir)	Kolyma	65.5	151.09

Table S2. The dataset (attached, also available at: https://github.com/AndreaCeolin/Boundaries/blob/main/TableS2.pdf**).**

TableS2 contains the 94 binary nominal parameters used for the experiments presented in the paper, set in the 58 languages of **TableS1**.

The table should be read as follows:
1st column: progressive number of the parameters (p1, p2, p3, …)
2nd column: acronym of the parameter
3rd column: name of the parameter
4th column: implicational constraints specifying the conditions for setting the parameter. They are expressed in a Boolean form, either as simple values of other parameters, or as conjunctions (written ‘,’), disjunctions (‘*or*’), or negation (‘¬’) thereof.
All critical data used to set the parameters have been collected or checked with the help of trained native speakers, except for Irish, which has been parameterized based on specialized literature.
The list of questions used to determine the state of the parameters and instructions is available in Crisma et al (2020).
The order of the parameters is not motivated except for the ease of expression of cross-parametric dependencies (see directly below), which are organized from top-down. The alternative parameter states are encoded as ‘+’ and ‘-’.
The neutralizing effect of implicational dependencies across parameters is encoded as ‘0’: the content of each parameter in such cases is entirely predictable or altogether irrelevant (the total amount of null states is 2534 out of 94x58=5452).
The parametric database is a refined version of that employed in Ceolin et al. (2020), with some of the parameters, their implications, and their relative order reformulated in a descriptively more accurate way.

Table S3. Table of Jaccard distances from Table S2 (attached, also available at: https://github.com/AndreaCeolin/Boundaries/blob/main/TableS3).
The matrix was derived using a Jaccard-type distance, based on the Jaccard formula described below. The comparison between the distance matrix used in this study and that used in Ceolin et al. (2020), for the overlapping languages, yields a Mantel correlation of 0.975 (see Mantel 1967). Therefore, it is expected that the exploratory analyses and the phylogenetic modeling of the distance matrix obtained from the two datasets will largely overlap. In fact, the results show just minor differences. **Figures S1-S3** illustrate the major taxonomic results obtained from the distances in **TableS3**.

The distance measures most commonly used for two perfectly aligned binary strings of the same length are the Hamming distance (counting the number of positions where the two strings differ) and the normalized Hamming distance (obtained dividing the Hamming distance by the string length, so that all the distances are within the range [0,1]). If we use ‘+’ and ‘-’ as the binary symbols in the strings, as we do in this study (rather than the more usual 0 and 1), the formula for the latter distance is:

$$\Delta\text{Hamming}(A,B) = [\, N_{-+} + N_{+-} \,] \,/\, [\, N_{-+} + N_{+-} + N_{++} + N_{--}]$$

where N_{XY} indicates the number of positions where the string A has value X and B has value Y. When the binary strings are interpreted as indicative of the presence ('+') or absence ('-') of traits (one per position in the string), the Jaccard (or Jaccard/Tanimoto) distance encodes an additional refinement: the loci where both strings lack the trait (have a value '-') are considered irrelevant and are ignored. The formula thus removes N_{--} from the denominator:

$$\Delta\text{Jaccard}(A,B) = [\ N_{-+} + N_{+-}\ ] \ / \ [\ N_{-+} + N_{+-} + N_{++}\ ]$$

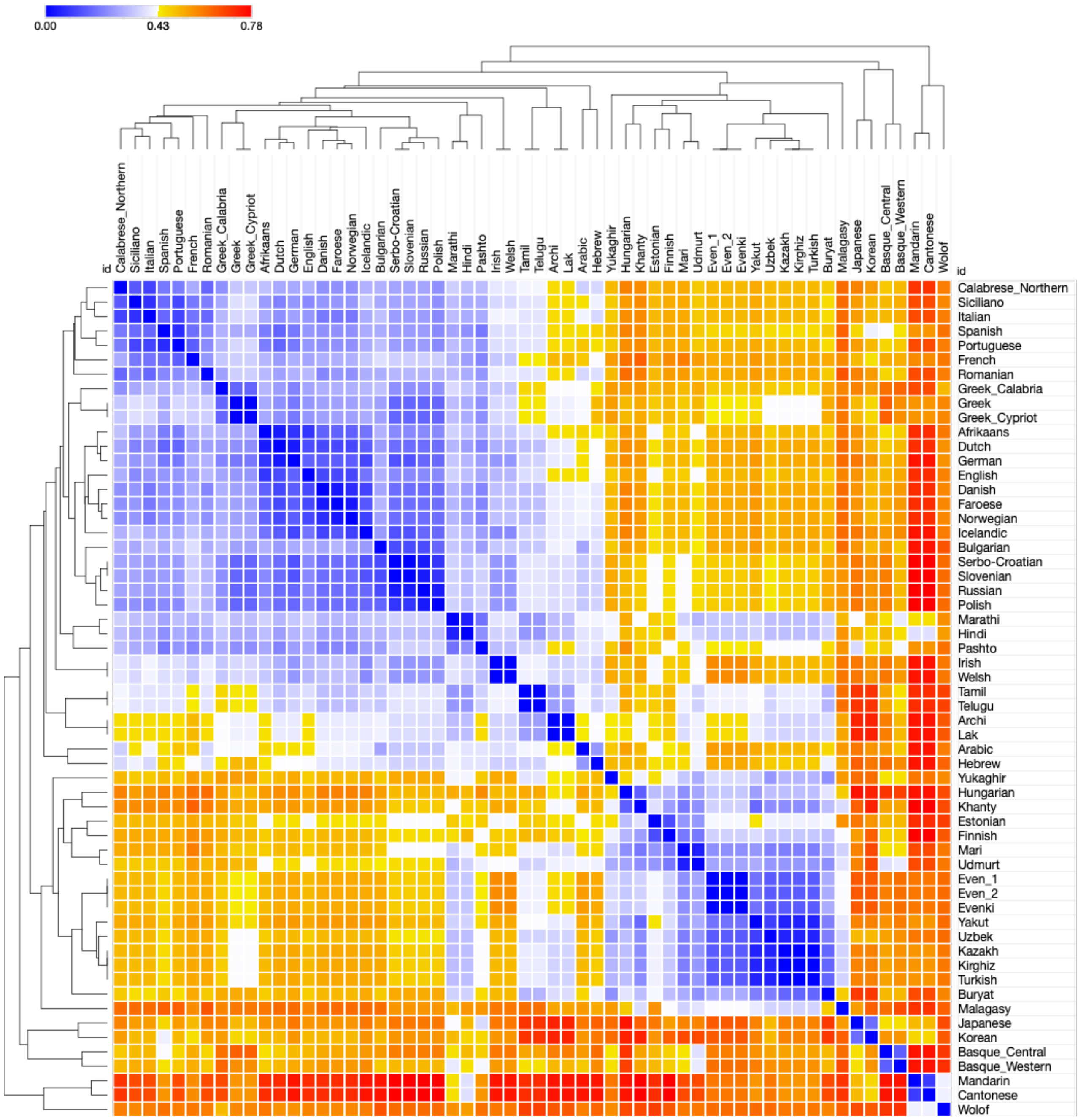
Note that in addition to '+/-', syntactic characters display a third state, '0', which indicates that the parameter is redundant or irrelevant in a language. Normalised Hamming or Jaccard distances could be used to compute linguistic distances, by removing every pair involving a '0' from the computation and, crucially, considering '+/-' as the relevant binary values. In a way, when computing distances, binary strings are shortened by getting rid of positions where '0s' are present.

In Longobardi and Guardiano (2009) and Longobardi et al. (2013) a normalized Hamming distance has been used to compute linguistic distances. This choice was appropriate given that in the framework of Principles & Parameters '+/-' were treated as being of equal markedness status. However, recent developments have re-addressed this type of assumption. Specifically, it is agreed that certain parameter values are marked, while others can be considered as 'default' settings. In our system, one of the two opposite values of all parameters (namely '-') represents a default setting, which can be interpreted as the absence of a trait, while the other '+' always requires some specified empirical evidence to be set (Crisma et al. 2020). Given this asymmetry, we find a Jaccard-type metric like the one defined above, rather than a Hamming-type metric, to be more appropriate to encode syntactic distances. Therefore, adopting a Jaccard distance corresponds to making the idealization that if two languages both change a '+' value into a '-' value in the same parameter, this does not constitute evidence of a shared innovation; for, it only represents a resetting to the unmarked state of mental grammars. On the contrary, convergence in the opposite change (from '-' to '+') is taken as stronger evidence of shared innovation.

Consider also that Franzoi et al. (2020) have developed metric distances alternative to ours in order to capture structural dependencies among characters. Their work interestingly shows that variation in the choice of distance formulae produces limited perturbations of the robustness of the signal when applied to syntactic data.

Figure S1. Heatmap

The distance matrix in Table S3 is visualized through the heatmap in **FigureS1**. The languages have been juxtaposed following the output of a hierarchical clustering algorithm, so that groups of languages sharing low distances (in blue) form squares along the diagonal.



Instructions to visualize the heatmap in the text.

1. Go to the page https://software.broadinstitute.org/morpheus/
2. Upload to the page the file *jaccard_distances.txt* (Table S3) and click the “OK” button to visualize the heatmap
3. In the “Tools” menu, select the option “Hierarchical clustering”, and then the following options:
   a. Metric > Matrix values (from a precomputed distance matrix)
   b. Linkage method > average

c. Cluster > Rows and columns

Click the "OK" button.

4. To visualize the same color distribution as Figure 1, follow the instructions below:
   a. In the "View" menu, select "Options"
   b. In the "Color Scheme" window:
      i. Uncheck the "Relative color scheme" choice
      ii. "Maximum" > 0.778
      iii. "Add color stop"
      iv. "Selected color" > yellow
      v. "Selected value" > 0.426 (the mean of the distance matrix)

Figure S2. UPGMA tree.

The tree in **Figure S2** has been produced using PHYLIP (https://evolution.genetics.washington.edu/phylip.html, Felsenstein 1989), and visualised using the Mesquite software (https://www.mesquiteproject.org, Maddison and Maddison 2018).

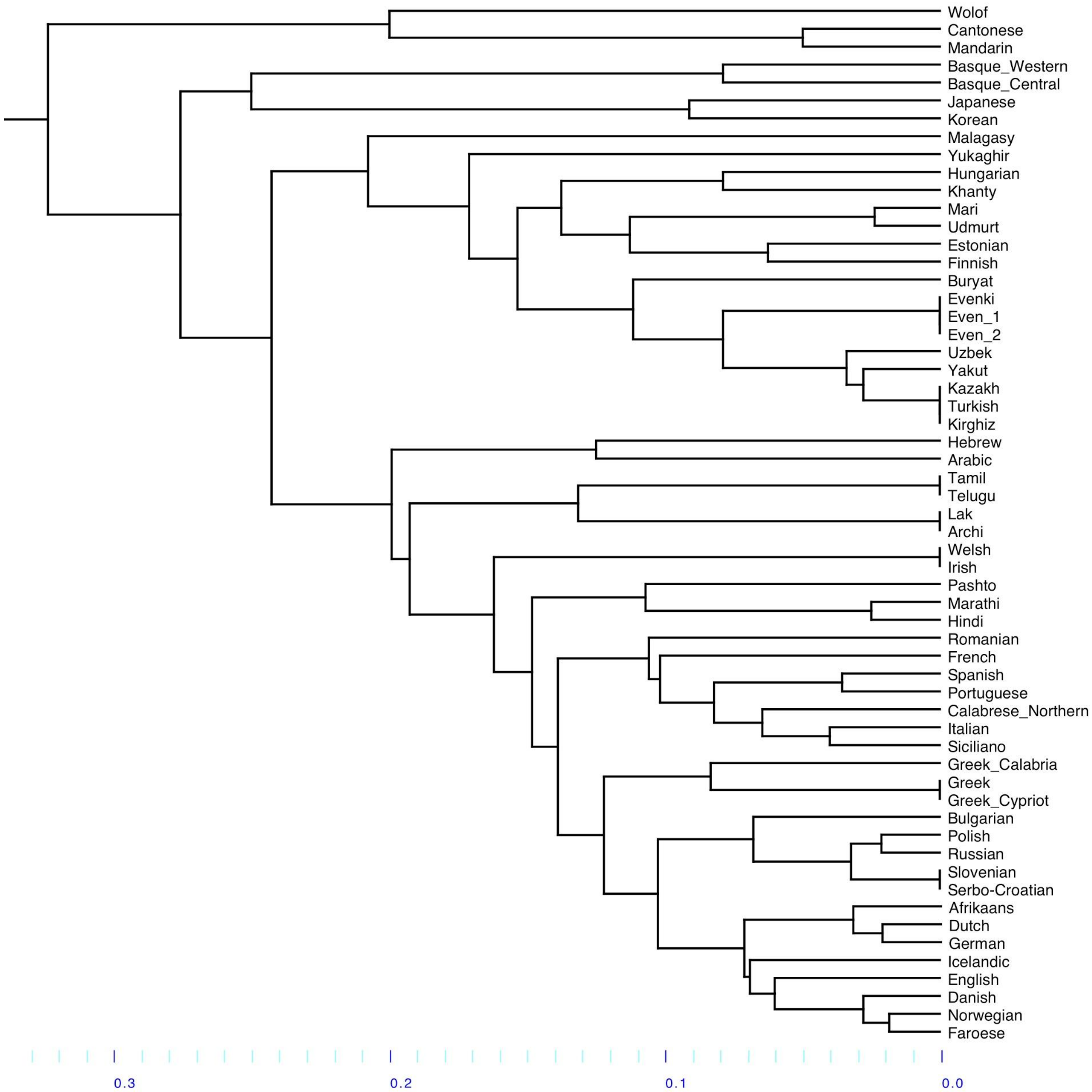
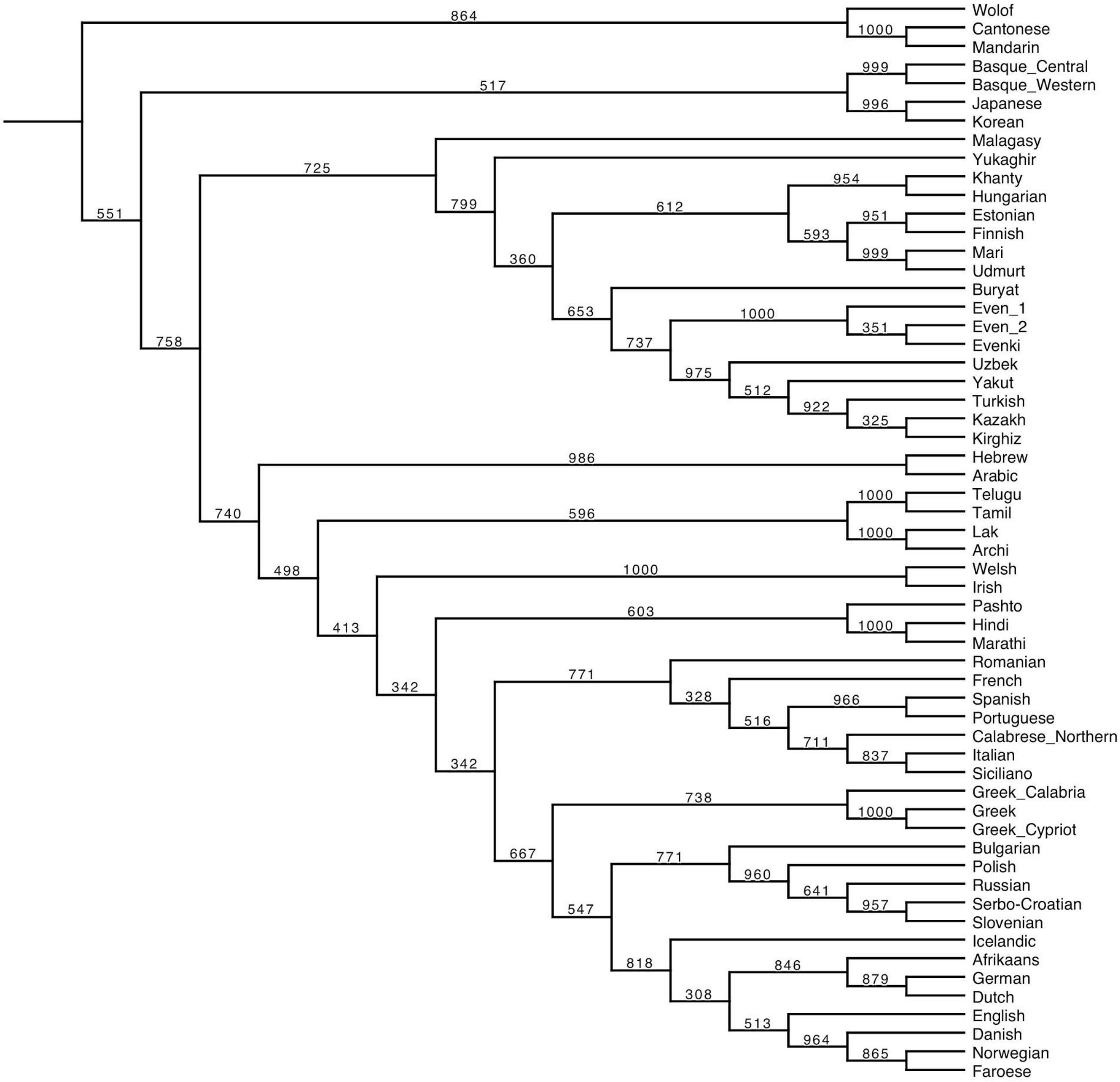


Figure S3. UPGMA (bootstrapped) tree.



The UPGMA tree in **Figure S3** has been generated using a modified bootstrapping procedure. Bootstrapping is used to establish the robustness of the nodes, and to determine whether the internal topology of the tree is robust to resampling.

The bootstrapping technique resamples the whole dataset by selecting each character with equal probability and recreating a matrix of the same length. The content of the new matrix is different from the original matrix, because some characters might be absent and some others might be present multiple times as a consequence of the sampling procedure. This allows one to estimate the robustness of the dataset by repeating the same analysis on different samples of the dataset.

Since the Jaccard distance between two languages excludes all parameters that are set to '0' in either one of them, a standard bootstrapping procedure runs the risk of making a pair of languages not comparable, because in some replicas the number of identities plus differences can reduce to zero, and then yield a zero denominator for the Jaccard formula. For this reason, we decided to adopt a

moderated bootstrap procedure, by creating 1000 datasets in which only six parameters are resampled. Since the minimum number of comparable parameters between any two languages in the dataset is seven, a resampling of six parameters will assure that any two strings are technically comparable by means of the Jaccard distance.
The UPGMA tree presented in the text is a consensus tree resulting from applying UPGMA to the 1000 replicas of the dataset.
The bootstrapping technique is insufficient as a device to assess the robustness of clusters with our data, and this is one reason to develop the statistical testing strategy presented in the article. The divergence between the outcomes of the two procedures is evidenced in **Table S2**.
In **Table S4**, we singled out the nodes which have been tested in the article (1st column), along with the result of the statistical test (2nd column) next to their bootstrapping score (3rd column).The rows of **TableS4** are arranged in decreasing order of the value of the test statistic.

Table S4. The groups of languages tested in the paper and their bootstrapping scores.
In **blue**: clusters which test positive to our statistical procedure but receive bootstrap scores <500; in **red**: clusters which test negative to our statistical procedure but receive bootstrap scores >500.

Volgaic/Permic	d=0.048	999
Tungusic/Turkic	d=0.158	737
Korean/Japanese	d=0.182	996
Germanic/Slavic	d=0.205	547
Tungusic-Turkic/Buryat	d=0.223	653
Volgaic-Permic/Balto-Finnic	d=0.225	593
Germanic-Slavic/Greek	d=0.244	667
NE Caucasian/Dravidian	d=0.263	596
Volgaic-Permic-Balto-Finnic/Ugric	d=0.275	612
Greek-Slavic-Germanic/Romance	**d=0.277**	**342**
Greek-Slavic-Germanic-Romance/Indo-Iranian	**d=0.296**	**342**
Balto-Finnic+Volgaic-Permic-Ugric/Tungusic-Turkic-Buryat	**d=0.307**	**360**
Greek-Slavic-Germanic-Romance-Indo-Iranian/Celtic	**d=0.324**	**413**
Uralo-Altaic/Yukaghir	**d=0.342**	**799**
Wolof/Cantonese-Mandarin	**d=0.4**	**864**
Basque/Japanese-Korean	**d=0.5**	**517**

It is immediately obvious that the two outcomes only partially correlate. In particular they are quite complementary in the following cases:

1) all the three nodes that include Romance display a bootstrap score below 500, though their mean distances are below the statistical threshold. This suggests that, although the significance testing algorithm clearly recognizes these groups as families because they are similar enough to each other, they also exhibit some accidental similarities with languages outside of their groups.
2) the case of Uralo-Altaic best exemplifies this case: its bootstrap value is 360, but goes up to 799 if we include Yukaghir; however the statistical algorithm suggests that only the former group can be safely established. This depends on the fact that Yukaghir exhibits some similarities with Uralic and Altaic languages, but not outside of the group, which means that although occasionally UPGMA will place Yukaghir within either group, it would rarely place it farther than the Uralo-Altaic node. But at the same time, Altaic and Uralic are sufficiently similar to pass the test, though different enough from Yukaghir for the whole set not to test positive to it (also cf. the similar case of bootstrap values for the three Tungusic languages, among the lower nodes in **Fig. S3**).
3) The opposite case is exemplified by two other nodes which are remarkably above 500 but far from passing the statistical test: Basque/Japanese-Korean (517) and especially Wolof/Cantonese-Mandarin (864). The only explanation for the high bootstrap score of these groups is long-branch attraction (Bergsten 2005), because the languages exhibit internal distances higher than the overall mean of the sample (0.444 and 0.556, respectively, thus insignificant from the viewpoint of the statistical test), but also much lower than with the rest of the dataset.

In conclusion, with this type and amount of characters, a statistical testing procedure such as we present in the text resists the effects of accidental similarities and random sampling of taxonomic units better than bootstraping techniques.

Table S5. Great Circle geographical distances of the languages of the sample (attached, also available at: https://github.com/AndreaCeolin/Boundaries/blob/main/TableS5).

This table contains a matrix of Great Circle Distances (in nautical miles) calculated using the coordinates in **Table S1**. Afrikaans was not included.

Section S1: Generating possible languages

Since the characters we used are not independent, the probability of occurrence of each pair using the binomial coefficient cannot be calculated. The binomial formula is based on independent trials, and therefore does not account for the fact that a specific result of an event might determine the outcome of a subsequent event. Therefore, we devised a method to statistically test the probability of relatedness for larger language groups using a posterior distribution generated by a population of randomly generated strings, thereby broadly following a Bayesian framework.

Bortolussi et al. (2011) was the first attempt to elaborate a way to randomly generate admissible ternary strings of type {+,-,0} compatible with the implicational constraints. The naive idea to generate a string at random and discard it when it was not admissible did not work because the probability of hitting an admissible string was too low. However, some of the implication rules were simple enough to be directly built into the random generator, which thus yielded 'quasi-admissible' strings totally at random: the result was that the probability of hitting a quasi-admissible string that was also admissible became manageable.

A key property of the algorithm is that it assumes a uniform distribution of admissible languages in string selection. The hypothesis of a uniform distribution among possible languages is not unproblematic. In Table S6, for instance, languages with a '-' at the first parameter are selected by the algorithm with a probability of 0.25, disregarding any information arising from the sample (for instance, the fact that they can be as frequent in the world as languages with '+' at P1). This is because, owing to the implicational rules, out of the eight combinatorial possibilities only four different languages exist which in reality represent the entire space of variation, and therefore each one is chosen with a probability of 1/4.

Therefore, a uniform distribution over languages tends to include languages bearing exceptional similarity to each other as the parameter values that activate many other parameters tend to be overrepresented with respect to those that neutralize them. This results in the production of too low a mean distance between the random language pairs.

We decided to modify the algorithm to account for all the implicational constraints in the random generator: we first set the independent parameters and created the strings incrementally, and then explored only those parameters that were compatible with the implicational structure, while automatically assigning a '0' value to the other parameters.

This strategy requires that a probability be associated to each value for each parameter. Therefore, we estimated the probabilities using the empirical distribution of the parameter values in our sample. This empirical estimate should also help us better control for biases towards certain parameter settings produced by general and external factors, to the extent they are detectable from the real-language sample. However, the 58 languages of our sample fall into 15 well-established families, therefore one can safely assume that these languages have ultimately evolved from 15 ancestors, and the probability of parametric values must be calculated considering this fact. Since such families are instantiated by an unbalanced number of languages, we took into account the cardinality of each language family in our sample, so as to enable the probabilistic information arising from each of them to be equally weighted in the generation of possible languages. We defined a 'family-ratio' as the ratio of '+' values for a certain parameter in the languages of a family over the total number of non-zero values. Every hypothetical language is generated with each parameter value having a '+' with probability equal to the arithmetic average of the family-ratios for '+' in that parameter within our sample. This means that, for the purposes of our algorithm, each language family is represented as an independent observation. All the implied values are automatically assigned a '0' by the algorithm. Thus, we ensure that in the case of a sample like the one shown in Table S6, the languages are '+' or '-' with p=0.5, using the distributional information of the sample as an approximation of the space of variation (see Table S7). Note that the actual variation in our real sample is almost always different from what would be expected from the equiprobability assumption, and that each parameter might exhibit different average ratios. Our algorithm takes both facts into consideration while generating the strings.

Table S6 – The sampling algorithm of Bortolussi et al. (2011). Each language is sampled with the same probability, implying that the space of the distribution is biased towards those parameters which have a lot of dependencies. In this case, +P1 languages cover 75% of the space of variation, while -P1 languages cover 25% of it.

	L1	L2	L3	L4	L5	L6	L7	L8
P1	+	+	+	+	**-**	**-**	**-**	**-**
P2 (only if +P1)	+	+	-	-	0	0	0	0
P3 (only if +P2)	+	-	0	0	0	0	0	0
Probability of L_	**0.25**	**0.25**	**0.25**		**0.25**			

Table S7 - The new sampling algorithm. Languages are created with '+' values assigned following the average ratio. Therefore, the languages which are set on parameter values that activate several others parameters are not overweighted, and the distribution is determined by the average of the empirical values of the languages of the real sample.

	L1	L2	L3	L4	L5	L6	L7	L8	Average '+' ratio
P1	+	+	+	+	**-**	**-**	**-**	**-**	**0.50**
P2 (only if +P1)	+	+	-	-	0	0	0	0	**0.50**
P3 (only if +P2)	+	-	0	0	0	0	0	0	**0.50**
Probability of L_	**0.125**	**0.125**	**0.25**		**0.50**				

SUPPLEMENTARY REFERENCES

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163-193.

Bortolussi, L., Longobardi, G., Guardiano, C., Sgarro, A. (2011), How many possible languages are there? *Biology, computation and linguistics,* eds Bel-Enguix, G., Dahl, V., Jiménez-López, M.D. (IOS Press, Amsterdam), pp. 168-179.

Ceolin, A., Guardiano, C., Irimia, M.A., Longobardi, G. (2020). Formal syntax and deep history. *Frontiers in psychology* 11:1-21.

Crisma, P., Guardiano, C., Longobardi, G. (2020). Syntactic parameters and language learnability. *Studi e Saggi Linguistici* 58: 99–130. doi: 10.4454/ssl.v58i2.265

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

Franzoi, L., Sgarro, A., Dinu, A., Dinu, L.P. (2020). Random Steinhaus Distances for Robust Syntax-Based Classification of Partially Inconsistent Linguistic Data. Information Processing and Management of Uncertainty in Knowledge-Based Systems. 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15-19, 2020, Proceedings, Part III, Communications in Computer and Information Science Series 1239, eds. Lesot, M. J, Vieira, S. M., Reformat, M. Z. Reformat, Carvalho, J.P., Wilbik, A., Bouchon-Meunier, B., Yager, R. R. (Springer): 17-26.

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.

Maddison, W. P., Maddison D. R. (2018). Mesquite: a modular system for evolutionary analysis. Version 3.40. http://mesquiteproject.org

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209-220.