Random Steinhaus Distances for Robust Syntax-Based Classification of Partially Inconsistent Linguistic Data

Laura Franzoi[1,4(✉)], Andrea Sgarro[2,4], Anca Dinu[3,4], and Liviu P. Dinu[1,4]

[1] Faculty of Mathematics and Computer Science, University of Bucharest (Ro), Bucharest, Romania
laura.franzoi@gmail.com
[2] DMG University of Trieste (I), Trieste, Italy
[3] Department of Modern Languages, University of Bucharest (Ro), Bucharest, Romania
[4] Human Language Technologies Research Center, University of Bucharest (Ro), Bucharest, Romania

Abstract. We use the Steinhaus transform of metric distances to deal with inconsistency in linguistic classification. We focus on data due to G. Longobardi's school: languages are represented through yes-no strings of length 53, each string position corresponding to a syntactic feature which can be present or absent. However, due to a complex network of logical implications which constrain features, some positions might be undefined (logically inconsistent). To take into account linguistic inconsistency, the distances we use are Steinhaus metric distances generalizing the normalized Hamming distance. To validate the robustness of classifications based on Longobardi's data we resort to randomized transforms. Experimental results are provided and commented upon.

Keywords: Steinhaus distance · Linguistic classification · Łukasiewicz logic · Fuzzy logic

1 Introduction

The linguist G. Longobardi and his school have an ambitious project on language classification which is based on syntactic rules rather than lexical or phonetic data [1,11–13]: the idea is that syntactic rules have a definitely slower time-drift and so, by being able to reach back deeper into the past, one might obtain precious information on linguistic macrofamilies which have been proposed, but whose adequacy is still a moot point. In a way, one should like to mimic what evolutionary bioinformatics has been able to achieve in the genetic domain of quaternary DNA strings; it is no surprise that tools of bioinformatics, e.g. those used in character-based and distance-based classifications, have been exported into linguistics, cf. e.g. [1,12], a fact we shall comment upon below. Longobardi's approach is defended in [1,11–13], to which the linguistic-minded reader is referred. In linguistics, binary strings are obtained by specifying n linguistic features which can be present $= 1$

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 17–26, 2020.
https://doi.org/10.1007/978-3-030-50153-2_2

or absent = 0 in a given language L. Since the length n is the same for all strings describing the languages L, $\Lambda, \ldots$ one intends to classify, a distance like Hamming distance, which counts the number of distinct features, appears to be adequate to gauge the dissimilarity between two languages L and Λ (between the corresponding strings), were it not that Longobardi's features are constrained by a complex network of logical implications. This has lead Longobardi's school to the use of a string distance which modifies the Hamming definition so as to get rid of positions corresponding to undefined (and undefinable) positions; the drawback is that the generalized distance they resort to is *not* metric, as instead often required by clustering techniques used to obtain the corresponding classification trees. Due to reasons discussed in next section, also a non-metric generalization the *Jaccard distance* has been used by Longobardi's school: the Jaccard distance "ignores" positions corresponding to features which are absent in both languages L and Λ, and so are linguistically "irrelevant", cf. next section. Now, the original 0–1 Jaccard distance (no inconsistency) can be obtained from the standard Hamming distance by means of a powerful mathematical tool called the *Steinhaus transform*: this transform needs the specification of a *pivot*, which in the Jaccard case is precisely the all-zero string; we stress that the Steinhaus transform of a metric distance is itself metric. In [8] one has already used Steinhaus transforms to deal with old *fuzzy* linguistic data due to Ž. Muljačić, where logical values can be intermediate between 0 and 1: this gave us the idea to represent logical inconsistency by the "ambiguous" value $\frac{1}{2}$ which is equidistant from both *crisp* logical values 0 and 1, cf. Sects. 3 and 4, and to use as pivot the "totally ambiguous" string, i.e. the all-$\frac{1}{2}$ string. The results which we obtain with this Steinhaus transform are surprisingly good, as commented upon in Sects. 3 and 4. The fact that moving from the original Longobardi's non-metric distance to our Steinhaus metric distance leaves the classification tree largely unchanged may be interpreted as a proof of the *robustness* of Longobardi's data: this is in puzzling contrast with results obtained by bootstrapping techniques described and used in [1,11,12] and suggested by bioinformatics, which seemed to show that Longobardi's original classification is not that robust. Below, in Sects. 3 and 4, we argue that this seeming non-robustness might be due to the inadequacy of tools exported from bioinformatics to linguistics. Rather than by bootstrapping, we prefer to validate the robustness of data by *randomly perturbing* our Steinhaus distance, or rather by randomly perturbing the pivot which is used: results are shown and commented upon in Sects. 3 and 4.

2 A Detour: From Muljačić to Steinhaus

In the past, the authors have been working on old and new linguistic data [6–10]; the starting point is the same: languages L, Λ, ... are described by n linguistic features f_i, $1 \leq i \leq n$, which in each language can be either present (1 = true) or absent (0 = false). The usual (crisp) Hamming distance, which counts the number of positions i where the corresponding bits are different, would be to the point, but both in the old Muljačić data and in the new ones due to Longobardi there is a stumbling block, since "ambiguous" situations are possible. Even if in both cases the symbols we will be using[1] are 0, $\frac{1}{2}$, 1, the symbol $\frac{1}{2}$, i.e. neither true nor false, neither present nor absent, has a distinct meaning.

In the case of Muljačić [14], $\frac{1}{2}$ can be interpreted as a logical value intermediate between 0 and 1 in a multivalued logic as is *fuzzy logic*, for which cf. e.g. [5]. In ampler generality one may consider strings $\underline{x} = x_1 x_2 \dots x_n$ where each component x_i may belong to the *whole* interval [0,1] allowing for all possible "shadings" of logical values. To define an adequate distance between fuzzy strings, suitably generalizing the usual Hamming distance between crisp strings, the relevant question to be posed is: if x and y are the logical values of feature f in the two languages L and Λ represented by the two strings $\underline{x}$ and $\underline{y}$, is f {present in L *and* absent in Λ} *or* {absent in L *and* present in λ}? Let $\perp$ and $\top$ be the disjunction *or* and the conjunction *and* in the multi-valued logic we choose to use; as for the negation, denoted by an overline, we will always use the 1-complement: $\overline{x} = 1 - x$ (the symbols $\top$ and $\perp$ which we are using for abstract conjunctions and disjunctions remind one of $\wedge$ and $\vee$, and are common when dealing with T-*norms*, cf. e.g. [5]). Assuming additivity w.r. to the n features, one gets for the distance $d(\underline{x}, \underline{y})$ between two strings $\underline{x}$ and $\underline{y}$, and so for the corresponding distance $d(\mathrm{L}, \Lambda)$ between languages L and Λ:

$$d(\underline{x}, \underline{y}) = \sum_{1 \le i \le n} \left(x_i \top \overline{y_i} \right) \perp \left(\overline{x_i} \top y_i \right) \tag{1}$$

In the case of standard fuzzy logic, conjunction *and* and disjunction *or* are computed through *minima* $\wedge$ and *maxima* $\vee$, respectively, $x \top y = x \wedge y = \min[x, y]$, $x \perp y = x \vee y = \max[x, y]$. The distance one obtains from (1) is a fuzzy generalization of the usual crisp Hamming distance; rather than fuzzy Hamming distance as in [15], or even Sgarro distance as in [3], we found it proper to call it *Muljačić distance*. We stress that use of the latter distance has proved to be quite successful in the case of Muljačić data, which, unlike Longobardi's, are genuinely fuzzy. The curious reader is referred to [6,7].

Already in [7] we tried several other logical operators of multi-valued logics, for example Łukasiewicz operators $x \perp y = (x + y) \wedge 1 = \min[x + y, 1]$, $x \top y = (x + y - 1) \vee 0 = \max[x + y - 1, 0]$. The results were in general uninteresting, since the distances one obtains were metrically unacceptable, cf. [7]; instead, Łukasiewicz case was surprising: as a straightforward computation shows one re-obtains the very well-known *Manhattan distance* or *taxicab distance* or *Minkowski distance*

$$d_T(\underline{x}, \underline{y}) = \sum_{1 \le i \le n} |x_i - y_i|$$

which in this context might even be called *Łukasiewicz distance.* It is precisely this distance that we shall use below, rather than Muljačić; for a more extensive discussion cf. [7,9].

[1] Longobardi instead of 0 $\frac{1}{2}$ 1 uses $-$ 0 $+$.

Before moving to Longobardi's data, we tackle *Steinhaus transforms*. The starting point was Longobardi's observation that positions i where both languages have a zero are linguistically *irrelevant*, and so should be ignored: mathematically, one has to move from Hamming distances to Jaccard distances[2]. What if the strings are not crisp? How should one go from Hamming distances or Muljačić distances to their Jaccard-like counterparts? The answer is precisely the Steinhaus transform, cf. e.g. [3]:

$$\delta_{St}(x,y) = \frac{2\delta(x,y)}{\delta(x,y)+\delta(x,z)+\delta(y,z)} \tag{3}$$

where $\delta(x,y)$ is any metric distance between objects which are not necessarily strings, and where z is a chosen fixed object called the *pivot* of the transformation. As it can be proved, the Steinhaus transform is itself a metric distance; it is normalized to 1, and is equal to 1 when x, z, y form an aligned triple $\delta(x,z)+\delta(z,y)=\delta(x,y)$ for the original distance to be transformed. Now, going back to our strings $\underline{x}, \underline{y}, \dots$, the Jaccard case corresponds to taking an all-zero pivot string $\underline{z} = 00\dots0$, in which case the distance from the pivot is nothing else but the *fuzzy weight* $w(\underline{x}) = d(\underline{x}, \underline{z}) = \sum_i x_i$ both with Muljačić and the taxicab distance.

The reason why we mentioned here irrelevance is simply that it paves the way to the use of Steinhaus transforms, even if with a different pivot, as we are going to do in the next section.

3 Dealing with Inconsistency

We move to Longobardi's ternary strings[3], where a complex network of logical implications involves features, of the type: if f_2 is false and f_4 is true, then f_6 does not make sense, it is logically inconsistent. In the case of inconsistency we use once more the symbol $\frac{1}{2}$: in the example just given $f_2 = 0$ and $f_4 = 1$ implies $f_6 = \frac{1}{2}$.

The distance used by Longobardi's school is simply a normalized Hamming distance, where the positions where one or both languages have a $\frac{1}{2}$ are *ignored*: in practice, one deals with *shorter* strings, possibly *much* shorter. Since Longobardi's distance is *not* metric, we took a bold step to preserve metricity. In Muljačić case, the numeric value $\frac{1}{2}$ represents suitably total logical ambiguity, but certainly not logical inconsistency, as however we shall now do. In the case of irrelevance an all-0 string did the job and got us rid of positions which are

[2] Actually, in as yet unpublished Longobardi's research this point of view has been relinquished and only inconsistency is taken care of, as we are doing below.

[3] Data we shall work on refer to 38 world languages described by means of 53 syntactic features, cf. [13] and Sect. 4, but Longobardi's group are constantly updating, improving and extending their database.

irrelevant. Forgetting about irrelevance, but of course not about inconsistency, here we shall take a totally ambiguous or rather totally inconsistent pivot string, which is the all-$\frac{1}{2}$ string $\underline{z} = \frac{1}{2}\frac{1}{2}...\frac{1}{2}$; this gets us rid of positions where there is inconsistency in both languages L and Λ, but not, as instead Longobardi's own distance does, of positions where only one of the two is inconsistent: actually, this turns out to be a possible source of weakness in Longobardi's choice, since few positions might survive if far-off languages are compared.

Now, rather than weights, cf. Sect. 2, one has *consistencies*, i.e. distances $d(\underline{x}, \frac{1}{2}\frac{1}{2}...\frac{1}{2})$ from the new pivot[4]: Łukasiewicz consistency turns out to be, as is proper, $\sum_i |x_i - \frac{1}{2}| = \sum_i \left(\frac{1}{2} - f(x_i)\right)$, where $f(x_i) = x \wedge (1-x)$ is often seen as the *fuzziness* of the logical value x, since it is the Euclidean distance from the totally ambiguous fuzzy value $\frac{1}{2}$.

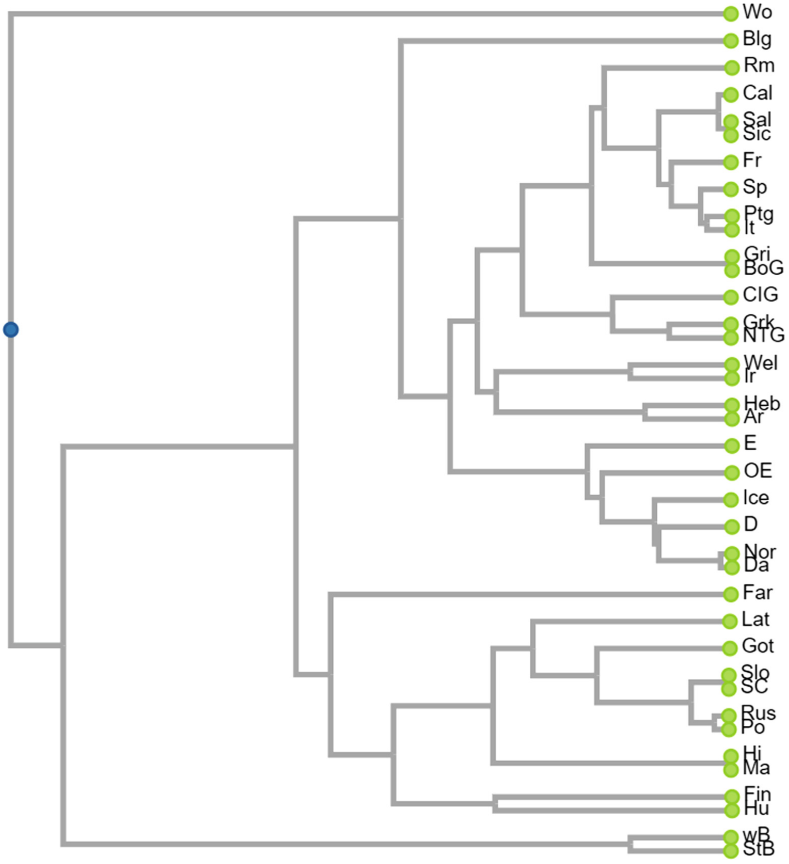
Let us move to Longobardi's data, so as to illustrate our methodology. The tree we obtain, fig. (b), is definitely and surprisingly[5] good, as it is virtually undistinguishable from the original Longobardi's tree (a) [13] based on a non-metric distance, and is linguistically equally sound. The fact that two distinct distances perform so similarly appears to be an indication that data are quite robust. Instead, use of statistical bootstrap techniques as done by Longobardi's school seemed to show that data are not that robust, cf. [1,11–13]. Actually, bootstrapping works quite well with the strings of bioinformatics, whose length is by magnitudes larger than ours, hundreds of thousands vs 53 (also in the case of DNA strings the assumptions of independence between positions, particularly if nearby, is untenable, but this weak point is smoothed out by the huge length of the strings involved). In our case strings are comparatively short and the structure of dependences is pervasive and strong, so the poor performance of bootstrapping might be simply an indication that the network of logical rather than statistical dependences makes the use of bootstrapping inadequate.

Instead, we propose an alternative to check robustness: let us *perturb* the distance, and see what happens. We shall take at random the pivot string $\underline{Z}$ (totally at random with *uniform* distribution on $[0,1]^n$), and check which sort of trees we obtain, taking also into account the taxicab distance between the *observed* random pivot and the "correct" all-$\frac{1}{2}$ pivot (capital letters denote random variables or random n-tuples).

Since the n terms in the random distance $d_T(\underline{x}, \underline{Z}) = \sum_{1 \le i \le n} |x_i - Z_i|$ are independent, not only the expectation, but also the variance is additive, and so

[4] Muljačić consistency, based on *maxima* and *minima*, is unusable, being always $\frac{n}{2}$ independent of $\underline{x}$; cf. [7,9].

[5] Farsi (modern Persian) appears to be poorly classified, which is true also with Longobardi's original tree. Also Bulgarian is poorly classified, but Longobardi uses only features relative to the syntax of nouns, and the Bulgarian noun, due to *substratum* influences, is well-known to be an outsider among Slavic languages. Be as it may, the aim of this paper is simply to check mathematical tools and robustness of data, rather than outperforming current classifications; cf. instead [4].



(a) Longobardi's tree

it will be enough to assume $n = 1$. Straightforward computations show that, for given $x \in [0,1]$:

$$\mathrm{E}\big[d_T(x,Z)\big] = x(x-1) + \frac{1}{2}\ , \quad \mathrm{var}\big[d_T(x,Z)\big] = -x^2(x-1)^2 + \frac{1}{12}$$

For XZ uniform on $[0,1]^2$ one has $\mathrm{E}\big[d_T(X,Z)\big] = \frac{1}{3}$, while for x crisp, i.e. $x = 0$ or $x = 1$, the taxicab expectation is $\frac{1}{2}$ and the taxicab variance is $\frac{1}{12}$.

Rather, we are interested in the case when x has the "correct" pivot value $\frac{1}{2}$: then expectation and variance are equal to $\frac{1}{4}$ and $\frac{1}{48}$, respectively, and so, for $n \geq 1$ the standard deviation $\sigma = \sigma\big[d_T(x,Z)\big]$ is approximately $0.144\sqrt{n} \approx 1.05$ with $n = 53$. Since the random distance $d_T(\frac{1}{2} \dots \frac{1}{2}, \underline{Z})$ is the sum of $n = 57$ i.i.d. terms, the central limit theorem allows one to resort to a normal approximation, and so the three intervals of semi-width $i\sigma$, $i = 1, 2, 3$ centered in the expected distance have probability ≈ 0.68, 0.95, 0.997, respectively. Correspondingly, the trees will be called of type α (observed distance inside the first and most probable interval), β (outside the first interval but inside the second), γ (outside the second interval but inside the third), else δ. So, trees of type α, β and γ have approximately probability $0.68, 0.27$ and 0.04, respectively.

4 Experimental Results

We reproduce ten trees for Longobardi's data, the first (b) with the correct pivot all-$\frac{1}{2}$, the others, (c) to (k), with a random pivot string; α and β-trees are virtually identical with the unperturbed tree, and in particular preserve the Indoeuropean standard groups; the γ-tree is weaker, e.g. it creates a single large family for Semitic and Celtic languages. For more random trees obtained in successive trials cf. [16].

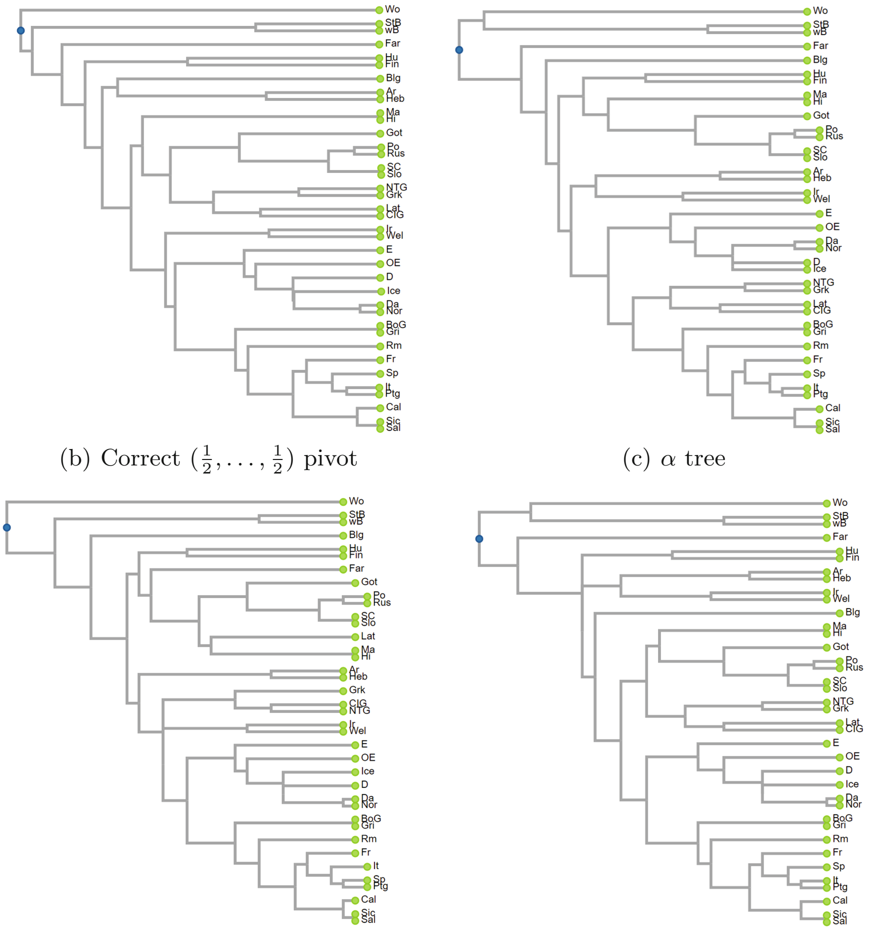
The languages are 38, namely Sic = Sicilian, Cal = Calabrese as spoken in South Italy, It = Italian, Sal = Salentine as spoken in Salento, South Italy, Sp = Spanish, Fr = French, Ptg = Portuguese, Rm = Romanian, Lat = Latin, ClG = Classical Attic Greek, NTG = New Testament Greek, BoG = Bova Greek as spoken in the village of Bova, Italy, Gri = Grico, a variant of Greek spoken in South Italy, Grk = Greek, Got = Gothic, OE = Old English, E = English, D = German, Da = Danish, Ice = Icelandic, Nor = Norwegian, Blg = Bulgarian, SC = Serbo(Croatian), Slo = Slovenian, Po = Polish, Rus = Russian, Ir = Gaelic Irish, Wel = Welsh, Far = Farsi, Ma = Marathi, Hi = Hindi, Ar = Arabic, Heb = Hebrew or 'ivrit, Hu = Hungarian, Finn = Finnish, StB = standard Basque, wB = Western Basque, Wo = Wolof as spoken mainly in Senegal.

Cf. the supplementary material [16] for more information, inclusive of the 38×53 ternary matrix with the strings of length $n = 53$ associated to the 38 languages.

Final remarks. Unsurprisingly, the trees exhibited perform all very well when compared to ours and the original Longobardi's tree, cf. [1,11–13]; cf. also footnote 5. A finer statistical analysis to gauge tree similarity might require suitable distances between trees like tree edit distances, as we are currently doing in [4]; here we have been more easy-going, since observed similarities are quite obvious to the eye of the linguist. Note that phylogenetic tree distances as we would need, cf. [3], are known to raise nasty computational problems. Unsurprisingly, thinking of Gray's classification tree [2], largely recognized by the linguistic community as a sort of reference benchmark, use of tree distances shows that Longobardi's tree and our own tree have virtually the same distance from Gray's tree, even if the distances used are quite distinct, one of the two not even metric, cf. [4]; once more, this appears to be an indication that Longobardi's data are quite robust. The statistical technique of random perturbation might be readily extended to the generalized Steinhaus distance used in [9], where one copes jointly with irrelevance and inconsistency; cf. however footnote 2.

The idea we are trying to defend in this paper is that, rather than mimicking bioinformatics, evolutionary linguistic should try to create its own new tools. This need has become more and more evident in Longobardi's research, where strings are dramatically shorter than those of bioinformatics and where the distances used, including our own metric distances, are quite different from those used in bioinformatics for distance-based classifications.

Our current work takes into account the new larger data tables provided by Longobardi's school; these data include many non-Indoeuropean languages, so as to get rid of an unwanted prominence of "usual" languages. This much enhances the linguistic significance of the results obtained.

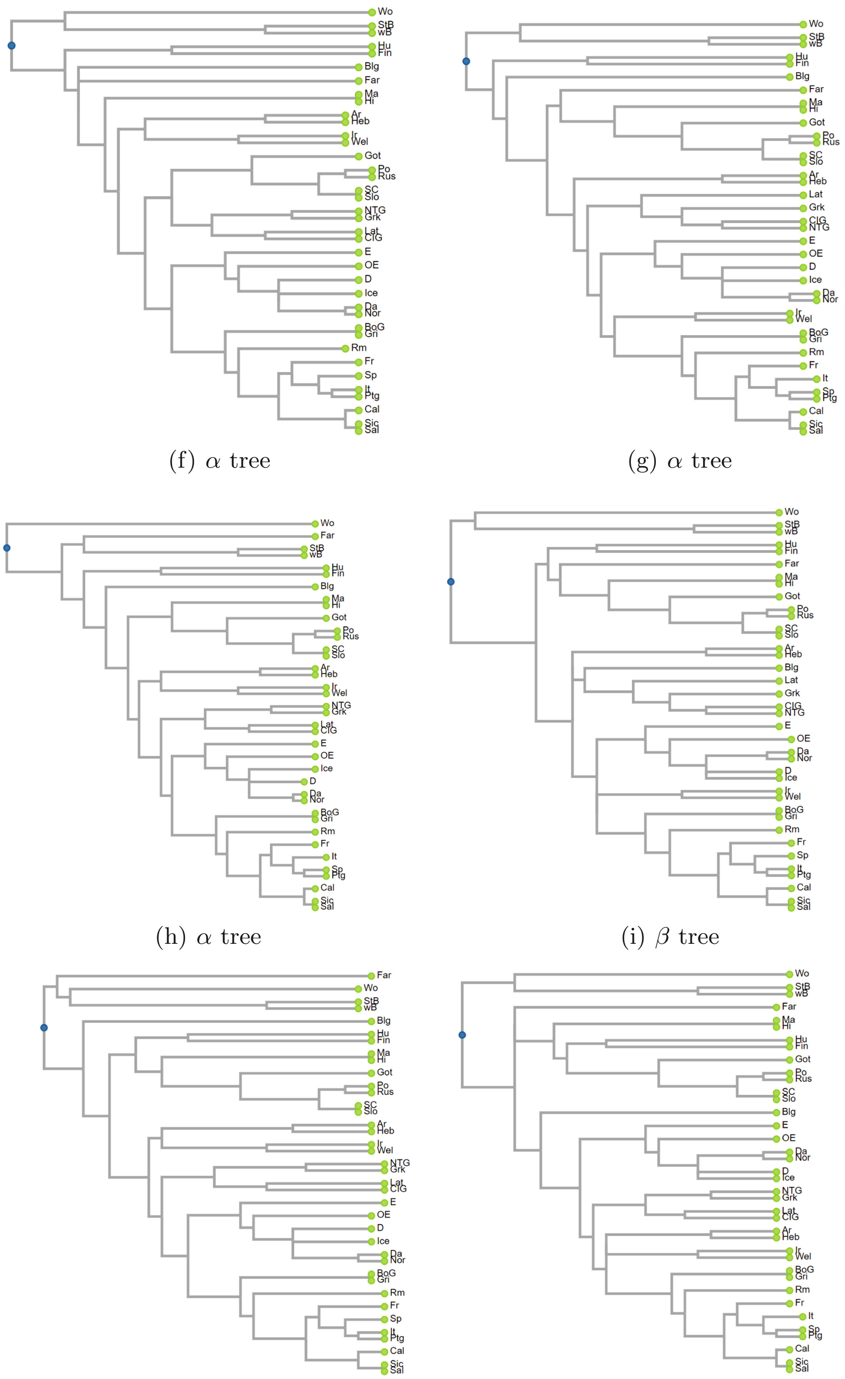


(b) Correct $(\frac{1}{2}, \dots, \frac{1}{2})$ pivot

(c) α tree

(d) β tree

(e) γ tree



(f) α tree

(g) α tree

(h) α tree

(i) β tree

(j) α tree

(k) β tree

Acknowledgment. Two of the authors, L. Franzoi and A. Sgarro, are members of the INdAM research group GNCS.

References

1. Bortolussi, L., Sgarro, A., Longobardi, G., Guardiano, C.: How Many Possible Languages are There? Biology, Computation and Linguistics, pp. 168–179. IOS Press, Amsterdam NLD (2011). https://doi.org/10.3233/978-1-60750-762-8-168
2. Bouckaert, R.: Mapping the origins and expansion of the indo-european languge family. Science **337**(6097), 957–960 (2012). https://doi.org/10.1126/science.1219669
3. Deza, M.M., Deza, E.: Dictionary of Distances. Elsevier B.V., Amsterdam (2006)
4. Dinu, A., Dinu, L.P., Franzoi, L., Sgarro, A.: Linguistic families: Steinhaus vs. Longobardi trees, ongoing work (2020)
5. Dubois, D., Prade, H.: Fundamentals of Fuzzy Sets. Kluwer Academic Publishers, New York (2000)
6. Franzoi, L., Sgarro, A.: Fuzzy hamming distinguishability. In: IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, pp. 1–6 (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015434
7. Franzoi, L., Sgarro, A.: Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities. KES Procedia Comput. Sci. **112**, 1168–1177 (2017). https://doi.org/10.1016/j.procs.2017.08.163
8. Franzoi, L.: Jaccard-like fuzzy distances for computational linguistics. Proc. SYNASC **1**, 196–202 (2017). https://doi.org/10.1109/SYNASC.2017.00040
9. Franzoi, L., Sgarro, A., Dinu, A., Dinu, L.P.: Linguistic classification: dealing jointly with irrelevance and inconsistency. In: Proceedings of Recent Advances in Natural Language Processing RANLP, pp. 345–352 (2019). https://doi.org/10.26615/978-954-452-056-4_040
10. Franzoi, L.: Fuzzy information theory with applications to computational linguistics. Ph.D. thesis, Bucharest University (2019)
11. Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., Ceolin, A.: Toward a syntactic phylogeny of modern Indo-European languages. J. Hist. Linguist. **3**(11), 122–152 (2013). https://doi.org/10.1075/jhl.3.1.07lon
12. Longobardi, G., et al.: Across language families: genome diversity mirrors language variation within Europe. Am. J. Phys. Anthropol. **157**, 630–640 (2015). https://doi.org/10.1002/ajpa.22758
13. Longobardi, G., et al.: Mathematical modeling of grammatical diversity supports the historical reality of formal syntax, University of Tübingen, Tübingen DEU, pp. 1–4. In: Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics (2016). https://doi.org/10.15496/publikation-10122
14. Muljačić, Z.: Die Klassifikation der romanischen Sprachen. Rom. J. Buch **XVIII**, 23–37 (1967)
15. Sgarro, A.: A fuzzy Hamming distance. Bull. Math. de la Soc. Sci. Math. de la R. S. de Romanie **69**(1–2), 137–144 (1977)
16. Support material for Random Steinhaus distance for robust syntax-based classification of partially inconsistent linguistic data. https://goo.gl/DMd72v