Linguistic Classification: Dealing Jointly with Irrelevance and Inconsistency

Laura Franzoi
Faculty of Mathematics and Computer Science
University of Bucharest
laura.franzoi@gmail.com

Andrea Sgarro
DMG
University of Trieste
sgarro@units.it

Anca Dinu
Faculty of Foreign Languages and Literatures
University of Bucharest
ancaddinu@gmail.com

Liviu P. Dinu
Faculty of Mathematics and Computer Science
University of Bucharest
liviu.p.dinu@gmail.com

Abstract

In this paper we present new methods for language classification which put to good use both syntax and fuzzy tools, and are capable of dealing with irrelevant linguistic features (i.e. features which should not contribute to the classification) and even inconsistent features (which do not make sense for specific languages). We introduce a metric distance, based on the generalized Steinhaus transform, which allows one to deal jointly with irrelevance and inconsistency. To evaluate our methods, we test them on a syntactic data set, due to the linguist G. Longobardi and his school. We obtain phylogenetic trees which sometimes outperform the ones obtained by Atkinson and Gray (Gray and Atkinson, 2003; Bouckaert et al., 2012).

1 Introduction

According to Ethnologue (Eth, 2018), there are around 7000 living natural languages in the world, and one of the most interesting topics (not only in the academic field, but also in the general public) is their classification. While the comparative method was the main method of classifying natural languages until the 90s, the last decades brought an increasing number of computational approaches for estimating the historical evolution of languages and their relationships. Most of the computational historical linguistics approaches rely on the use of lexical items. In contrast, very few of them take into account syntactic aspects. Moreover, fuzzy tools and information theory were employed quite sparsely in language classification tasks (Ciobanu et al., 2018), in spite the inherent fuzzy nature of the natural language data.

This paper is based on previous work on fuzzy string distances and linguistic classification started in (Franzoi and Sgarro, 2017a,b; Franzoi, 2017), and inspired by the path-breaking ideas put forward back in 1967 (Muljačić, 1967) by the Croat linguist Ž., Muljačić. The technical tool which will be used in this paper is the *general Steinhaus transform*, or *biotope transform*, applied to crisp strings which are however affected by irrelevance and inconsistency, as happens with data due to the linguist G. Longobardi and his school. Fuzziness in linguistics has been seldomly treated (Franzoi and Sgarro, 2017a,b; Dinu et al., 2018), as compared to crisp approaches.

In his 1967 paper Muljačić, even if only rather implicitly, had introduced what appears to us as a natural *fuzzy* generalization of crisp Hamming distances between binary strings of fixed length n, and this only two years after Zadeh's seminal work (Zadeh, 1965): the aim was showing that Dalmatic, now an extinct language, is a bridge between the Western group of Romance languages and the Eastern group, mainly Romanian. The situation is the following: Romance languages $L, \Lambda, \dots$ are each described by means of n features, which can be present or absent, and so are encoded by string $s(L) = \underline{x} = x_i \dots x_n$, where x_i is the truth value of the proposition *feature i is present in language L*; however, presence/absence is sometimes only vaguely defined and so each $x = x_i$ is rather a truth value $x \in [0, 1]$ in a multi-valued logic as is fuzzy logic; $x = x_i$ is *crisp* only when either $x = 0$ = false = absent or $x = 1$ = true = present, else x is *strictly fuzzy*. So, the mathematical objects one deals with are *strings* $\underline{x}, \underline{y}, \dots$ of length n, each of the n components being a real number in the interval $[0, 1]$, and moreover *distances* between such objects, since

Proceedings of Recent Advances in Natural Language Processing, pages 345–352,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_040

the classifications are all distance-based. In what follows, rather than Muljačić distance, we need string distances obtained by use of the *Steinhaus transform*, cf. (Dinu et al., 2018), and the *generalized Steinhaus transform*; they are all *metric* distances, in particular they verify the triangle equality. Unlike the case of Muljačić distances, which span the interval $[0, n]$, these distances are *normalized* to the interval $[0, 1]$. Steinhaus transforms allow one to deal with *irrelevance* and *inconsistency* in linguistics, as we already argued in (Dinu et al., 2018), and not only with vagueness, or fuzziness, as in Muljačić case, cf. (Muljačić, 1967; Franzoi and Sgarro, 2017a); the reason to use the *generalized* Steinhaus transform, as we do here, is that it allows one to deal *jointly* with both irrelevance and inconsistency.

Based on arguments defended by the linguist G. Longobardi and his school, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), if a feature i has a low truth value in two languages L and Λ, then that feature is scarcely relevant: in fact, in the practice of linguistics the values 0 and 1 have a very *asymmetric* use, and the fact that languages L and Λ both have zero in a position i means that such an irrelevant feature i should *not* really contribute to the distance between the two languages. Technically, one should move from Hamming distances to (normalized) Jaccard distances. To achieve the goal, the convenient tool we have used was the *Steinhaus transform*, cf. (Dinu et al., 2018), which is known to preserve metricity and which is general enough so as to amply cover also the fuzzy situation: one starts from a distance like Muljačić distance $d_M(\underline{x}, \underline{y})$, and obtains its Steinhaus transform, in this case a *fuzzy Jaccard distance* $d_J(\underline{x}, \underline{y})$ for fuzzy strings $\underline{x}$ and $\underline{y}$; starting from the usual *crisp* Hamming distance the transform gives the usual *crisp* Jaccard distance.

In general, to apply a Steinhaus transformation one needs a *pivot string*, which in the Jaccard case is the all-0 string $\underline{z} = \underline{0} = (0, \ldots, 0)$. In the transform, actually, any other string $\underline{z}$ might be used, cf. (Dinu et al., 2018), as we do here so as to cover the case of *logical inconsistency*, as appears in the data due to G. Longobardi: his school is involved in an ambitious and innovative project on language classification based on *syntax*, cf. (Bortolussi et al., 2011; Longobardi et al., 2016); languages are represented through yes-no strings of length 53, each string position corresponding to a syntactic feature which can be present or absent. In his notation Longobardi uses + if a feature is present, - if it is absent, 0 if it is undefined; in our case, cf. Tables 1, 2, we write 1 if a feature is present, 0 if it is absent, * if it is undefined. Actually, due to a complex network of logical implications which constrain features, some positions might be undefined (logically inconsistent). For example, in Longobardi's classification, feature 34 is defined if and only if feature 8 is set to + and either feature 9 is set to + or feature 18 is not set to + (or both); otherwise it will be "neutralized" (*inconsistent*)[1]. This property does not hold true for Ptg (Portuguese), OE (Old English) and Ice (Icelandic).
All this establishes an extremely complex network of logical dependencies in Longobardi's data, and makes it necessary, if one wants to cover also this new intriguing facet, to suitably generalize crisp Hamming distances, or crisp Jaccard distances, respectively: in Longobardi's approach, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), the two distances for ternary strings one defines and uses are quite useful, but unfortunately they violate the triangle property, and so are not metric. In this paper we propose one *metric* alternative based on the generalized Steinhaus transform (or generalized biotope transform): the star $*$ will be replaced by the totally ambiguous truth value $\frac{1}{2}$, and the pivot strings in the transform will be given by the set compound by the all-$\frac{1}{2}$ string, i.e. the totally ambiguous string $\underline{z} = \left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$ (which stands for inconsistency) and all-0 string $\underline{z} = (0, \ldots, 0)$, i.e. the totally false string, which

[1]Feature 34 stands for *checking possessives*: it opposes languages like French, wherein possessives occur without any visible article (*mon livre* vs. *le mon livre*), to those like Italian, in which a visible determiner is possible and normally required instead (*il mio libro* vs. *mio libro*). This feature seems to conceptually and typologically depend on full grammaticalization of definiteness (feature 8). Also, it is relevant only in languages with strong Person in D (feature 9) or without strong article (feature 18), because otherwise the language would have GenS with determiner-like function, cf. (Longobardi et al., 2013). Feature 8 asks if a language generalizes the overt marking of definiteness to all relevant cases. Feature 9 (*Strong Person*) defines whether attraction to the D area of referential nominal material (e.g. proper names) is overt (e.g. Romance) or not (e.g. English). Feature 18 (*Strong Article*) is presence of an indefinite article, i.e. of an obligatory marker on singular indefinite count argument nominals, distinct from those used for definite and mass indefinite, cf. (Longobardi et al., 2013).

stands for irrelevance. The idea is to play down not only the contribution of 0's and $\frac{1}{2}$'s separately, as we have done in (Dinu et al., 2018), but rather the contribution of both 0's and $\frac{1}{2}$'s *jointly*. It will turn out that in this case, which is not genuinely fuzzy, rather than to Muljačić distances, the generalized Steinhaus transform had been better applied to the usual *taxicab distance* (Manhattan distance, Minkowski distance), re-found when the standard fuzzy logical operators of *min* and *max* for conjunctions and disjunctions are replaced by Łukasiewicz T-norms and T-conorms, cf. (Franzoi and Sgarro, 2017b; Dinu et al., 2018).

The paper is divided as follow: in Section 2 we shortly re-take both fuzzy Hamming distances, or Muljačić distances, and taxicab distances stressing how the latter relate to Łukasiewicz T-norms; in Section 3 we introduce Steinhaus transform and we apply it to taxi-cab or Łukasiewicz distances; in Section 4 we introduce the general Steinhaus transform to deal with irrelevance and inconsistency *jointly* and we comment on our linguistic results; in Section 5 we sum up our results.

2 Fuzzy Hamming Distances vs. Łukasiewicz or Taxicab Distances

We need some notations and definitions: we set $x \wedge y \doteq \min[x, y]$, $x \vee y \doteq \max[x, y]$ and $\overline{x} \doteq 1-x$; these are the truth values of conjunction AND, disjunction OR and negation NOT, w.r. to propositions with truth values x and y in *standard fuzzy logic*, a relevant form of multi-valued logic; $x \in [0, 1]$. Define the *fuzziness* of the truth value x to be $f(x) \doteq x \wedge (1-x)$. For the truth values x and y in $[0, 1]$ we say that x and y are *consonant* if either $x \vee y \leq \frac{1}{2}$ or $x \wedge y \geq \frac{1}{2}$, else they are *dissonant*; let $\mathcal{D}$ and $\mathcal{C}$ denote the set of dissonant and consonant positions i, respectively. We define the following distance for strings $\underline{x}, \underline{y} \in [0, 1]^n$:

$$d_M(\underline{x}, \underline{y}) \doteq \sum_{i \in \mathcal{D}} [1 - [f(x_i) \vee f(y_i)]] + \sum_{i \in \mathcal{C}} [f(x_i) \vee f(y_i)] \tag{1}$$

This expression stresses the link with *crisp* Hamming distances for binary strings $\in \{0, 1\}^n$, but its meaning is better understood due to the following fact: each of the n *additive* terms summed is the truth value of the statement:

[(feature f_i is present in L and absent in Λ) or (feature f_i is absent in L and present in Λ)]

since, as soon proved, cf. e.g. (Franzoi and Sgarro, 2017a), for two truth values x and y one has $(x \wedge \overline{y}) \vee (\overline{x} \wedge y)$ equal to $f(x_i) \vee f(y_i)$ or to $1 - [f(x_i) \vee f(y_i)]$ according whether there is consonance or dissonance. This distance, called henceforth Muljačić distance (and called *Sgarro distance* in (Deza and Deza, 2009), cf. also (Sgarro, 1977)) is simply a natural generalization of crisp Hamming distances to a fuzzy setting. As for alternative logical operators for conjunctions and disjunctions (different T-norms and T-conorms, for which cf. e.g. (Dubois et al., 2000)), they have been discussed in (Franzoi and Sgarro, 2017b). From a metric point of view, the only attractive choice, beside fuzzy Hamming distances, turned out to be Łukasiewicz T-norms for conjunctions and the corresponding T-conorms for disjunctions:

$$x \top y \doteq (x + y - 1) \vee 0, \quad x \perp y \doteq (x + y) \wedge 1$$

One soon checks that in this case, rather curiously, $(x \top \overline{y}) \perp (\overline{x} \top y)$ turns out to be simply $|x - y|$, and so the string distance one obtains is nothing else but the very well-known taxicab distance $d_T(\underline{x}, \underline{y}) = \sum_i |x_i - y_i|$, which in our context, when it is applied to fuzzy strings of length n, might be also legitimately called *Łukasiewicz distance*.

If we consider the fuzziness $f(x) \doteq d(x, x)$ of a logical value x and if we use the Muljačić distance, then we get $f_M(x) = x \wedge (1 - x)$; if we use instead the Łukasiewicz distance, then the fuzziness is always 0.

However, if we consider another equally legitimate definition of fuzziness, namely "ambiguity - crispness", which can be formalized as $\frac{1}{2} - d\left(x, \frac{1}{2}\right)$, then if we use the Muljačić distance the new fuzziness is 0, but if we use the Łukasiewicz distance it is $f_T(x) = \frac{1}{2} - d_T\left(x, \frac{1}{2}\right) = x \wedge (1-x)$: the result of the competition Muljačić distance vs. Łukasiewicz distance turns out to be a tie. In the next Section we explain why, with Longobardi's data, we decided to resort to taxicab distances.

The distance in (1) is a *fuzzy metric distance*, cf. (Sgarro, 1977; Franzoi and Sgarro, 2017a), from which a standard metric distance is soon obtained by imposing that self-distances $d_M(\underline{x}, \underline{y})$ should be 0, while, unless $\underline{x}$ is crisp (i.e. belong to $\{0, 1\}^n$, the set of the 2^n binary strings of length n), the value given by (1) would be strictly positive.

As for taxicab or Łukasiewicz distances, the self-distance $d_T(\underline{x}, \underline{y})$ is always zero even when the argument $\underline{x}$ is not crisp, a possibly unpleasant fact in a fuzzy context (but not in ours), as argued in (Franzoi and Sgarro, 2017b).

3 Steinhaus Transforms

In the general situation, one has objects $x, y, \ldots$, not necessarily strings, a *metric* distance $d(x, y)$ between objects, and a special object z called the "*pivot*-object". The Steinhaus transform, cf. (Deza and Deza, 2009), itself proven to be a metric distance, is:

$$\mathcal{S}_d(x,y) \doteq \frac{2d(x,y)}{d(x,y)+d(x,z)+d(y,z)}$$

set equal to zero when $x = y = z$.
In our case the objects are strings and *pivots* $\underline{z}$ will always be *constant* strings $\underline{z} = (z, \ldots, z)$, $z_i = z$, $\forall i, z \in [0, 1]$.
If one starts with the crisp Hamming distance, one obtains the usual crisp Jaccard distance (distances from the pivot are then Hamming *weights*); starting with the more general fuzzy Hamming distance, or Muljačić distance, one has an appropriate Jaccard-like generalization, which weighs only "little" a position where both x and y are "almost 0", and which accounts for irrelevance in itself, but not for inconsistency, as instead we need.
If the term $d_M(\underline{x}, \underline{z})$ is equal to the *fuzzy Hamming weight* $w(\underline{x}) \doteq \sum_i x_i$ for $\underline{z} = \underline{0}$, it is equal to $\frac{n}{2}$ independent of $\underline{x}$ when $\underline{z} = \underline{\frac{1}{2}}$, a constant pivot string which we shall need to deal with inconsistency. The fact that $d_M(\underline{x}, \underline{z})$ with $\underline{z} = \underline{\frac{1}{2}}$ is independent of $\underline{x}$ is a serious drawback, indeed. This is why in the case of Longobardi's data, we have applied the Steinhaus transform, rather than to the fuzzy Hamming distance or Muljačić distance, directly to the taxicab distance or Łukasiewicz distance $d_T(\underline{x}, \underline{y})$. In this case, in the denominator of the corresponding Steinhaus transform, the fuzzy Hamming weight $w(\underline{x})$ is replaced by $d_T(\underline{x}, \underline{z}) = \sum_i \left| x_i - \frac{1}{2} \right|$. In the next Section, more ambitiously, we shall deal *jointly* with both irrelevance and inconsistency.

4 Dealing with Irrelevance and Inconsistency

In (Franzoi and Sgarro, 2017a,b; Franzoi, 2017; Dinu et al., 2018) one has presented new methods for language classification, testing them on data sets due to Muljačić and Longobardi. So far we have dealt separately with irrelevance and inconsistency, but a question arises spontaneously: can we consider jointly both irrelevance and inconsistency? Does a mathematical tool which takes into account both of them exist? The answer is yes and the tool we are looking for is the *generalized Steinhaus transform* or *generalized biotope transform*, cf. (Deza and Deza, 2009).
Prompted by arguments defended by G. Longobardi and his school, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), the novelty of this section is that, since in the language classifications features can be irrelevant or inconsistent, we want to consider both aspects together.
As we said above the idea is to play down not only the contribution of 0's, as in the case of irrelevance, but also the contribution of the $\frac{1}{2}$-positions. Unlike ours, Longobardi's non-metric distance gets rid of irrelevant and inconsistent positions in quite a drastic way, possibly a serious draw-back, as we comment in our Conclusions.
The generalized Steinhaus transform, or generalized biotope transform, is:

$$\mathcal{S}_d(x,y) = \frac{2d(x,y)}{d(x,y)+\inf_{z \in M}(d(x,z)+d(y,z))} \tag{2}$$

where M is the set of pivots we are considering, cf. (Deza and Deza, 2009).
We tackle Longobardi's data (or rather to a sample of his languages, since the data he and his school are providing are steadily improving and extending), data which are not really fuzzy, even if we have decided to "simulate" logical inconsistency by *total fuzziness*. In this case the number of features is 53, and the languages are: Sic = Sicilian, Cal = Calabrese as spoken in South Italy, It = Italian, Sal = Salentin 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, 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. For comparison reasons, we have selected a part of Longobardi's data set compound by 38 languages; taking $M = \left\{\underline{0}, \underline{\frac{1}{2}}\right\}$ in (2), the UPGMA tree we obtain is given in the following figure:

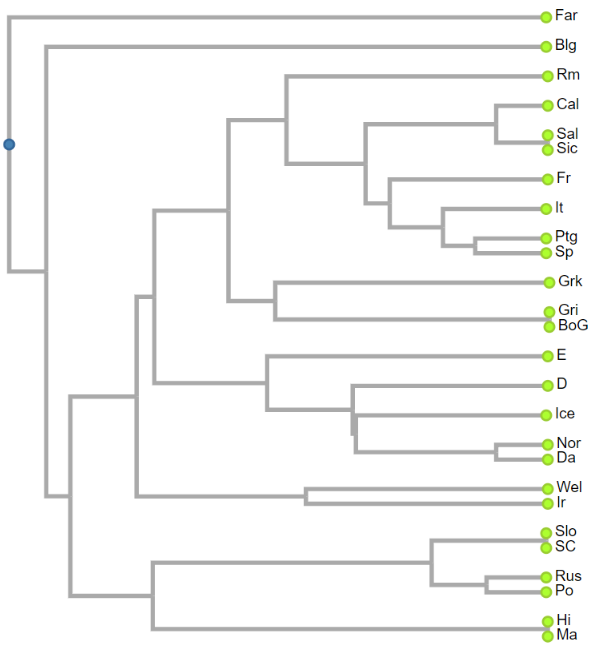

Figure 1: Generalized Steinhaus transform with taxi-cab distance and Longobardi's data

while the Longobardi's original tree is the following one:

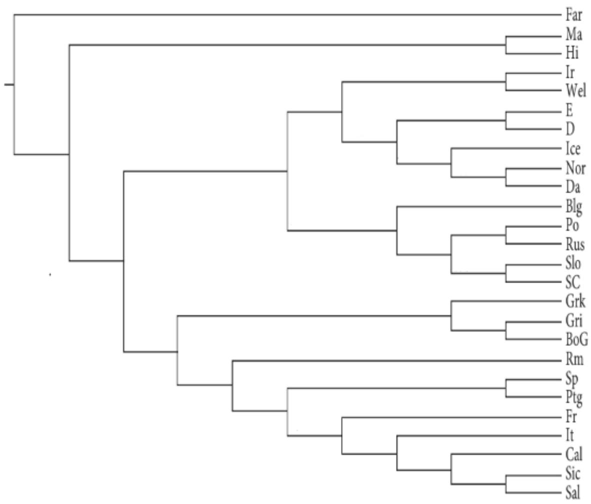

Figure 2: Longobardi's classification tree

We can observe that the Romance languages are grouped together. However there are some differences between the two trees: in our tree (Fig. 1) the big Romance languages (i.e. Italian, Spanish, Portuguese and French) are grouped together and Italian is more integrated with the Ibero-Romance languages (i.e. Portuguese ans Spanish), which are clustered together like in the standard language classifications. The three Italian dialects (i.e. Salentine, Sicilian and Calabrese) are external to this cluster in our case in Fig. 1, while in the original Longobardi's tree (Fig. 2) they are integrated with Italian and then the entire group is linked with French and after with the Ibero-Romance group. In both trees the Romanian is grouped with Romance languages, but is the most exterior with the languages from this group. In both trees the Celtic languages Gaelic (Ir) and Welsh (Wel) and Germanic languages are grouped together, but in the Longobardi's tree in Fig. 2 the Celtic group is more integrated with the Germanic group. There are two main differences between the two trees: the first one is that in Longobardi's tree in Fig. 2 Bulgarian is grouped with Slavic languages; the second one is the moving of the entire Slavic group from a closet proximity with the Germanic group (in the Longobardi's tree) to a more distance linkage with them in our case.
Our classification compares with the one obtained by Longobardi's school with these data, cf. comments in the Conclusion, where we argue why our distance is quite promising for the new and ambitious data Longobardi's school are now providing. Actually, our distance compares rather well also with the classification obtained by Q. D. Atkinson and R. D. Gray, cf. (Gray and Atkinson, 2003; Bouckaert et al., 2012).

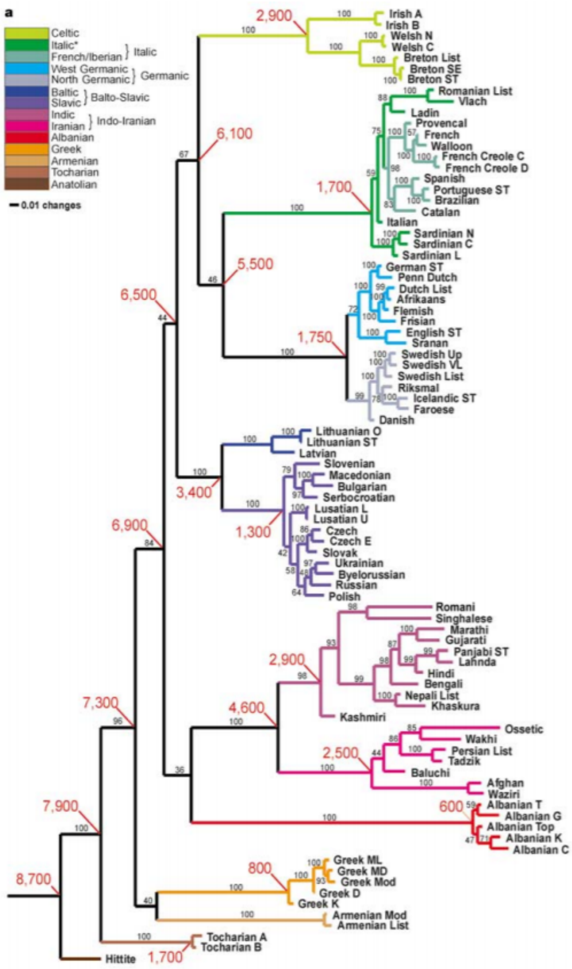

Figure 3: Q. D. Atkinson and R. D. Gray classification tree, cf. (Gray and Atkinson, 2003; Bouckaert et al., 2012)

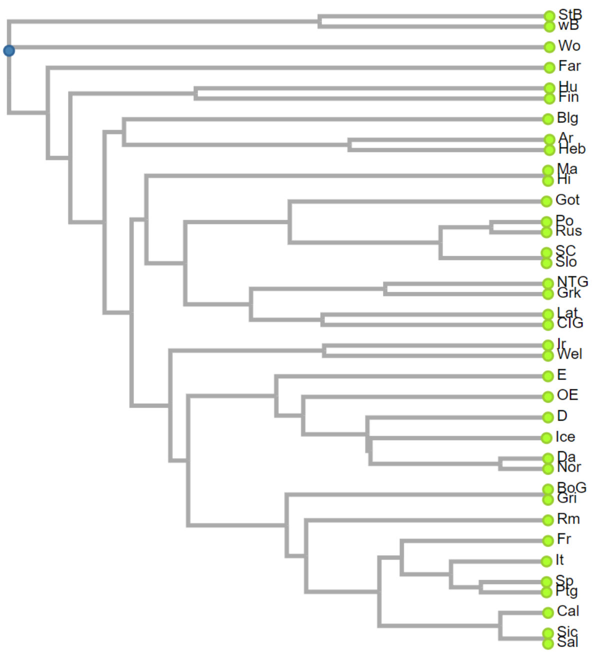

Figure 4: Classification obtained with the generalized Steinhaus transform applied to the taxi-cab distance with Longobardi's data

First of all for the classification we have used Longobardi's dataset, while Atkinson and Gray have used their own dataset. If we look to Marathi and Hindi we can notice that they are grouped together in both trees; also Polish, Russian, Serbo Croatian and Slovenian are grouped together in both trees; the same is for New Testament Greek, Greek and Classical Attic Greek. Also the Celtic languages (i.e. Gaelic and Welsh) and Germanic languages are grouped together. Our misclassification of Bulgarian is not that worring, since Longobardi covers only the syntax of the noun, and the Bulgarian noun is well-known to behave in quite a non-Slavic way, due possibly to its Balcanian substratum.

5 Conclusions

In this paper we have investigated the language classification problem by using original tools inspired by fuzzy logic. In the literature fuzzy tools and information theory have been used only quite sparsely. We have exhibited a metric distance which allows one to deal jointly with both irrelevance and inconsistency, and which is based on the generalized Steinhaus transform. Our classification compares quite well both with the one obtained by Longobardi and the one obtained by Atkinson and Gray. The merits of our metric proposal should not be underestimated, as we now comment. In more recent datasets, Longobardi and his school introduce families and macrofamilies which are quite apart. Now, think of two languages L and Λ such that the following occurs (and this does occur with "remote" languages): in most position i at least one of the two languages has a star signalling non-definition of the corresponding features. Since such positions are totally ignored by Longobardi's non-metric distance, the value obtained for the distance relies on a handful of positions only, and it is no surprise that the two languages end up being poorly classified, a sourse of worry, indeed. Now, our metric distances are not that drastic, and so might be used as a sort of companion to Longobardi's non-metric distances, useful when the latter have a low significance due to the fact that only few features "survive".

We are confident that the fuzzy ideas and methods discussed in this paper and in (Franzoi and Sgarro, 2017a; Franzoi, 2017; Dinu et al., 2018) will prove to be useful not only in linguistic classification and linguistic phylogeny, but also outside linguistic, first of all in coding theory cf. (Franzoi and Sgarro, 2017a), or even in bioinformatics. Irrelevance and inconsistency appear to be features which are dealt with quite sparsely, if ever, outside Longobardi's school; actually, these flexible features might prove to be quite useful not only in linguistic classification phylogeny, cf. (Franzoi and Sgarro, 2017a,b), but also in the investigation of the history of texts. So far, we are just providing technical tools to be used in Longobardi's research, which, in its turn, is methodically matched with the *current state of the art*, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015; Longobardi, 2017; Kazakov et al., 2017).

Table 1: Longobardi original data

ft.	Sic	Cal	It	Sal	Sp	Fr	Ptg	Rm	Lat	CIG	NtG	BoG	Gri	Grk	Got	OE	E	D	Da	Ice	Nor
1.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6.	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7.	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
8.	1	1	1	1	1	1	1	1	*	1	1	1	1	1	0	1	1	1	1	1	1
9.	1	1	1	1	1	1	1	1	*	1	1	1	1	1	*	0	0	0	0	0	0
10.	0	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11.	0	0	0	0	0	0	0	0	*	0	0	0	0	0	*	0	0	0	0	0	0
12.	0	0	0	0	0	0	0	1	*	0	0	0	0	0	*	0	0	0	1	1	1
13.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0	1	1
14.	0	0	0	0	0	0	0	0	*	0	0	0	0	0	*	0	0	0	0	0	0
15.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18.	1	1	1	1	1	*	1	1	*	0	0	1	1	1	*	0	1	1	1	0	1
19.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
20.	0	0	0	0	1	0	1	*	*	1	1	0	0	0	*	*	*	*	*	*	*
21.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
23.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
24.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	1	.*	1	1	1	1
25.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
27.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
28.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29.	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	1	1	1	1	1
30.	0	0	0	0	0	0	0	*	1	1	*	*	*	*	*	*	0	0	0	0	0
31.	1	1	1	1	1	1	1	*	1	1	*	*	*	*	*	*	1	1	1	1	1
32.	0	1	0	0	0	0	0	0	*	*	1	1	1	1	1	1	0	1	0	1	0
33.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34.	0	0	0	0	1	1	*	0	*	0	0	0	0	0	*	*	*	*	*	*	*
35.	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	*	1	1	1	1
36.	0	0	0	0	0	0	0	*	0	0	*	*	*	*	*	*	*	0	0	0	0
37.	0	0	0	0	0	0	0	0	0	0	0	*	*	*	0	0	0	0	0	0	0
38.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39.	*	*	*	*	*	*	*	*	1	1	1	*	*	*	1	1	*	*	*	1	*
40.	0	0	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0
41.	1	1	1	1	1	1	1	1	*	0	0	1	1	0	*	1	1	1	1	*	1
42.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44.	1	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
45.	*	*	1	*	1	1	1	1	0	0	0	*	*	0	0	0	0	0	0	0	0
46.	*	*	*	*	*	*	*	*	0	0	0	*	*	0	0	0	0	0	0	0	0
47.	*	*	*	*	*	*	*	*	0	0	1	*	*	1	1	1	*	1	*	1	*
48.	*	*	*	*	*	*	*	*	1	1	*	*	*	*	*	*	0	*	0	*	0
49.	*	*	*	*	*	*	*	*	1	1	1	*	*	1	1	0	0	0	0	0	0
50.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0	0	0	0	0	0
51.	0	0	0	0	0	0	0	0	*	1	1	0	0	1	1	*	0	*	*	*	*
52.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*	*	*	*
53.	*	*	1	*	1	1	1	1	0	0	0	*	*	0	*	1	1	1	1	1	1

Table 2: Longobardi original data

ft.	Blg	SC	Slo	Po	Rus	Ir	Wel	Far	Ma	Hi	Ar	Heb	Hu	Fin	StB	wB	Wo
1.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3.	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	1
4.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
5.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
6.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	*	*	*
7.	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1
8.	1	*	*	*	*	1	1	*	*	*	1	1	1	*	*	*	1
9.	1	*	*	*	*	0	0	*	*	*	1	1	1	*	*	*	*
10.	0	0	0	0	0	0	0	0	0	0	0	0	0	1	*	*	*
11.	0	*	*	*	*	0	0	*	*	*	0	0	0	*	0	1	1
12.	1	*	*	*	*	0	0	*	*	*	0	0	0	*	*	*	0
13.	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
14.	0	*	*	*	*	0	0	*	*	*	1	0	0	*	*	*	1
15.	0	0	0	0	0	0	0	0	0	0	1	0	0	0	*	*	*
16.	1	1	1	1	1	1	0	0	1	1	1	1	0	0	*	*	*
17.	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	1
18.	0	*	*	*	*	0	0	*	*	*	0	0	*	*	*	*	*
19.	*	*	*	*	*	*	*	1	0	0	*	*	0	*	*	*	0
20.	*	*	*	*	*	*	*	*	*	*	0	0	0	*	1	1	*
21.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
22.	1	1	1	1	1	1	0	0	1	1	1	1	0	1	*	*	*
23.	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
24.	*	*	*	*	*	0	*	*	*	*	*	*	*	*	*	*	*
25.	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1
26.	*	*	*	*	*	*	*	*	1	1	*	*	*	*	0	0	*
27.	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	*
28.	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	*
29.	1	0	0	0	0	1	1	1	1	1	1	1	0	0	1	1	0
30.	0	*	*	*	*	0	0	0	0	0	0	0	*	*	0	0	*
31.	1	*	*	*	*	1	1	1	*	*	1	1	*	*	*	*	*
32.	0	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1
33.	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
34.	0	*	*	*	*	1	1	*	*	*	0	0	0	*	*	*	0
35.	1	1	1	1	1	0	*	*	1	1	0	0	*	0	*	*	0
36.	0	*	*	*	*	0	*	*	1	1	*	*	*	*	*	*	*
37.	1	1	1	0	1	*	0	0	0	0	*	*	*	*	0	0	*
38.	0	0	0	0	0	0	0	0	0	0	1	1	*	*	0	0	*
39.	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	1
40.	0	0	0	0	0	*	*	0	0	0	1	*	0	0	0	*	0
41.	1	*	*	*	*	*	*	*	*	*	0	*	0	*	1	*	1
42.	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
43.	0	0	0	0	0	0	0	1	0	0	*	*	0	0	0	0	0
44.	0	0	0	0	0	1	1	*	0	0	*	*	0	0	*	*	*
45.	0	0	0	0	0	*	*	*	0	0	*	*	0	0	*	*	*
46.	0	0	0	0	0	*	*	*	0	0	*	*	0	0	*	*	*
47.	*	1	1	1	1	*	*	*	*	*	*	*	*	*	*	*	1
48.	1	*	*	*	*	*	*	*	1	1	*	*	1	1	*	*	*
49.	0	0	0	0	0	*	*	*	0	0	*	*	0	0	*	*	*
50.	0	0	0	1	1	*	*	*	0	0	*	*	0	0	*	*	*
51.	*	*	*	*	*	0	0	*	*	*	1	1	*	*	*	*	0
52.	*	*	*	0	0	0	0	1	*	*	0	0	*	*	0	0	1
53.	0	0	0	0	0	*	*	0	*	*	*	*	1	1	1	1	*

References

2018. Ethnologue. https://www.ethnologue.com/.

L. Bortolussi, A. Sgarro, G. Longobardi, and C. Guardiano. 2011. How many possible languages are there? *Biology, Computation and Linguistics* pages 168–179.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science* 337(6097):957–960.

A. M. Ciobanu, L. P. Dinu, and A.Sgarro. 2018. Towards a map of the syntactic similarity of languages. In *CICLing 2017*. volume LNCS 10761, pages 1–15.

M. M. Deza and E. Deza. 2009. *Encyclopedia of Distances*. Springer Dordrecht Heidelberg, London New York.

A. Dinu, L. P. Dinu, L. Franzoi, and A. Sgarro. 2018. Steinhaus transforms of fuzzy string distances in computational linguistics. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018*. volume Proceedings, Part I, pages 171–182.

D. Dubois, H. T. Nguyen, and H. Prade. 2000. Possibility theory, probability and fuzzy sets: Misunderstanding, bridges and gaps. In *Fundamentals of Fuzzy Sets*. Kluwer Academic Publishers, pages 343–438.

L. Franzoi. 2017. Jaccard-like fuzzy distances for computational linguistics. In *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, Timişoara, Romania, September 21-24, 2017*. pages 196–202.

L. Franzoi and A. Sgarro. 2017a. Fuzzy hamming distinguishability. In *2017 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2017, Naples, Italy, July 9-12, 2017*. pages 1–6.

L. Franzoi and A. Sgarro. 2017b. Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference KES-2017, Marseille, France, 6-8 September 2017.*. pages 1168–1177.

R. D. Gray and Q. D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature* 426:435–439.

D. Kazakov, G. Cordoni, A. Ceolin, M. A. Irimia, S. Kim, D. Michelioudakis, N. Radkevich, C. Guardiano, and G. Longobardi. 2017. Machine learning models of universal grammar parameter dependencies. *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP-17* pages 31–37.

G. Longobardi. 2017. Principles, parameters, and schemata: A radically underspecified ug. *Linguistic Analysis* 41(3–4):517–557.

G. Longobardi, A. Ceolin, L. Bortolussi, C. Guardiano, M. A. Irimia, D. Michelioudakis, N. Radkevich, and A. Sgarro. 2016. Mathematical modeling of grammatical diversity supports the historical reality of formal syntax. *University of Tübingen, online publication system Tübingen DEU* pages 1–4.

G. Longobardi, S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, and G. Barbujan. 2015. Across language families: Genome diversity mirrors language variation within europe. *American Journal of Phisical Anthropology* 157:630–640.

G. Longobardi, C. Guardiano, G. Silvestri, A. Boattini, and A. Ceolin. 2013. Toward a syntactic phylogeny of modern indo-european languages. *Journal of Historical Linguistics* 3:11:122–152.

Ž. Muljačić. 1967. Die Klassifikation der romanischen Sprachen. *Rom. Jahrbuch 18* pages 23–37.

A. Sgarro. 1977. A fuzzy hamming distance. *Bullettin Math. de la Soc. Sci. Math. de la R. S. de Roumanie* 69(1-2):137–144.

L. A. Zadeh. 1965. Fuzzy sets. *Information and Control* 8(3):338–353.