At the boundaries of syntactic prehistory: metric and non-metric distances

Andrea Sgarro
sgarro@units.it

Dept. of Mathematics and Geosciences, University of Trieste (I)
Human Language Technologies Research Center, Bucharest (Ro)

RADH 2021
Bucharest, October 2021

Andrea Ceolin, Cristina Guardiano, Monica Alexandrina Irimia, **Giuseppe Longobardi**, Luca Bortolussi, Andrea Sgarro

At the boundaries of syntactic prehistory

Philosophical Transactions B, Royal Society (2021)

Andrea Ceolin, Cristina Guardiano, Monica Alexandrina Irimia, **Giuseppe Longobardi**, Luca Bortolussi, Andrea Sgarro

At the boundaries of syntactic prehistory

Philosophical Transactions B, Royal Society (2021)

Laura Franzoi, Andrea Sgarro, Anca Dinu, Liviu P. Dinu

Random Steinhaus distances for robust syntax-based classification of partially inconsistent linguistic data

IPMU 2020, Lisbon (Pt)

Parametric Comparison Method PCM

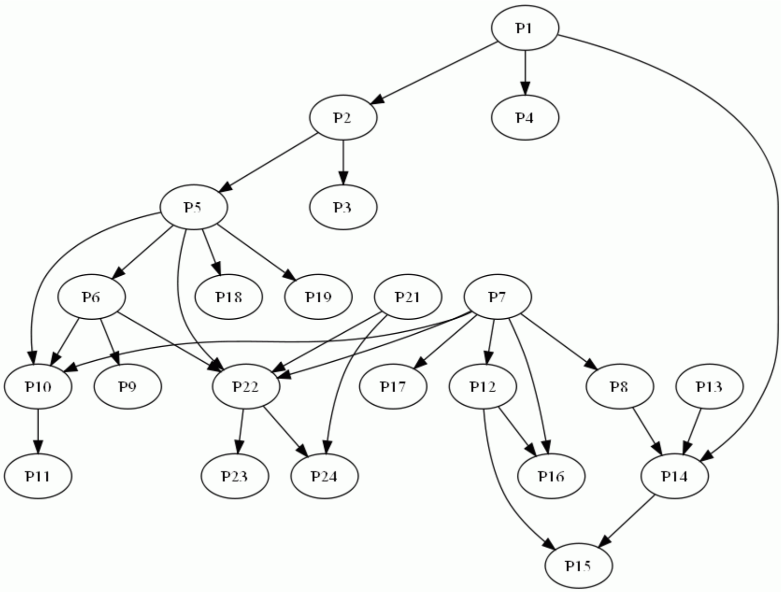
94 syntactic parameters, 58 languages from the Old World

possible languages

94 parameters as before, 5000 possible languages

results

controversial clusters such as Altaic (Japanese, Korean, Mongolian, ...) or Uralo-Altaic were signifcantly supported, while other possible macro-groupings as Indo-Uralic or Basque-Caucasian were not


P1
P2
P4
P5
P3
P6
P18
P19
P21
P7
P10
P9
P22
P17
P12
P8
P13
P11
P23
P24
P16
P14
P15

Longobardi distances, Hamming-like and Jaccard-like

L = 0 | 1 | * | 1 | 0 | 1
Λ = 0 | 0 | * | * | 1 | 1

$$\text{dist}_H(\Lambda, L) = \frac{\#\ \text{bit differences}}{\text{"sound" bit length}} = \frac{2}{4} = \frac{1}{2}$$

Longobardi distances, Hamming-like and Jaccard-like

L = 0 | 1 | * | 1 | 0 | 1
Λ = 0 | 0 | * | * | 1 | 1

$$\mathrm{dist}_H(\Lambda, L) = \frac{\#\text{ bit differences}}{\text{"sound" bit length}} = \frac{2}{4} = \frac{1}{2}$$

$$\mathrm{dist}_J(\Lambda, L) = \frac{\#\text{ bit differences}}{\text{sound length} - \#\text{ "irrelevant" positions}} = \frac{2}{4-1} = \frac{2}{3}$$

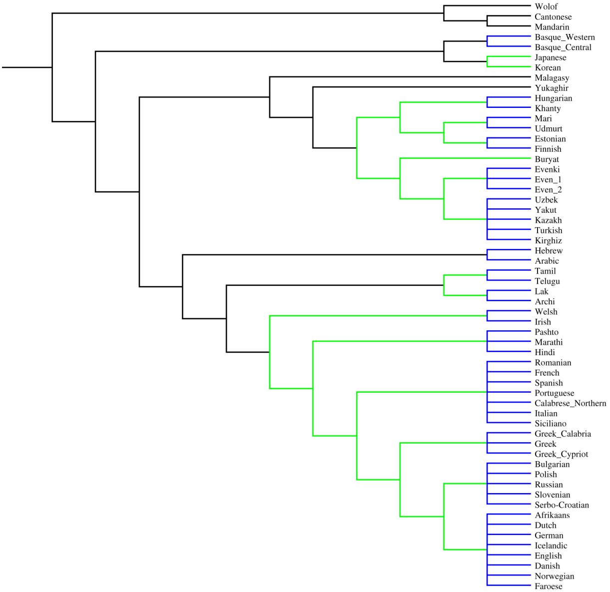
Longobardi distances, Hamming-like and Jaccard-like

$L = 0 \mid 1 \mid * \mid 1 \mid 0 \mid 1$
$\Lambda = 0 \mid 0 \mid * \mid * \mid 1 \mid 1$

$$\mathrm{dist}_H(\Lambda, L) = \frac{\#\text{ bit differences}}{\text{"sound" bit length}} = \frac{2}{4} = \frac{1}{2}$$

$$\mathrm{dist}_J(\Lambda, L) = \frac{\#\text{ bit differences}}{\text{sound length} - \#\text{ "irrelevant" positions}} = \frac{2}{4-1} = \frac{2}{3}$$

both might violate the triangle inequality


Wolof
Cantonese
Mandarin
Basque_Western
Basque_Central
Japanese
Korean
Malagasy
Yukaghir
Hungarian
Khanty
Mari
Udmurt
Estonian
Finnish
Buryat
Evenki
Even_1
Even_2
Uzbek
Yakut
Kazakh
Turkish
Kirghiz
Hebrew
Arabic
Tamil
Telugu
Lak
Archi
Welsh
Irish
Pashto
Marathi
Hindi
Romanian
French
Spanish
Portuguese
Calabrese_Northern
Italian
Siciliano
Greek_Calabria
Greek
Greek_Cypriot
Bulgarian
Polish
Russian
Slovenian
Serbo-Croatian
Afrikaans
Dutch
German
Icelandic
English
Danish
Norwegian
Faroese

what should a distance be?

at least...

- $d(x, y) \geq 0$
- $d(x, x) \leq \min\left[d(x, y), d(y, x)\right]$

(ordered) triangle inequality

$$d(x, y) \leq d(x, z) + d(z, y)$$

Steinhaus transform or biotope transform of the distance d:

$$S_d(x,y) \doteq \frac{2d(x,y)}{d(x,y)+d(x,z)+d(y,z)}$$

where:

- $x, y, \ldots$ are objects (possibly strings)
- $d(x,y)$ is their distance
- z is a fixed object called the pivot z

We'll have to generalize to several pivots
$S_d(x,y)$ preserves metricity

Steinhaus transform or biotope transform of the distance d:

$$S_d(x,y) \doteq \frac{2d(x,y)}{d(x,y)+d(x,z)+d(y,z)}$$

where:

- $x, y, \ldots$ are objects (possibly strings)
- $d(x,y)$ is their distance
- z is a fixed object called the pivot z

We'll have to generalize to several pivots
$S_d(x,y)$ preserves metricity

From (normalized) Hamming to Jaccard:
the objects are n-lenght strings,
the pivot $z = \underline{z}$ is the all-0 string

x, y strings of n logical values

$$d(x,y) = \sum_i \left[x_i \text{ AND } \neg y_i\right] \text{ OR } \left[\neg x_i \text{ AND } y_i\right]$$

x, y strings of n logical values

$$d(x,y) = \sum_i [x_i \text{ AND } \neg y_i] \text{ OR } [\neg x_i \text{ AND } y_i]$$

standard fuzzy logical operators, $\text{OR} = \max$, $\text{AND} = \min$

Solomon Marcus (1925-2016)

x, y strings of n logical values

$$d(x,y) = \sum_i [x_i \ \text{AND} \ \neg y_i] \ \text{OR} \ [\neg x_i \ \text{AND} \ y_i]$$

standard fuzzy logical operators, $\text{OR} = \max$, $\text{AND} = \min$

Solomon Marcus (1925-2016)

why do not start from the fuzzy Hamming distance?

x, y strings of n logical values

$$d(x,y) = \sum_i [x_i \text{ AND } \neg y_i] \text{ OR } [\neg x_i \text{ AND } y_i]$$

standard fuzzy logical operators, $\text{OR} = \max$, $\text{AND} = \min$

Solomon Marcus (1925-2016)

why do not start from the fuzzy Hamming distance?

Łukasiewicz: $\text{OR} = \min\left[(x+y), 1\right]$, $\text{AND} = \max\left[(x+y-1), 0\right]$

taxicab or Minkowski or Łukasiewicz distance:

$$d(x, y) = \sum_i |x_i - y_i|$$

taxicab or Minkowski or Łukasiewicz distance:

$d(x, y) \; = \; \sum_i |x_i - y_i|$

$$* \implies \frac{1}{2}$$

$$d(bit, *) = d(*, bit) = \frac{1}{2}, \; d(*, *) = 0$$

pivot of the Steinhaus transform: the "totally unsound" all-$*$ sequence

consistency $\chi(x)$ of the string x: its taxicab distance from the all-$*$ string

$$S_d(x, y) \doteq \frac{2d(x, y)}{d(x, y) + \chi(x) + \chi(y)}$$

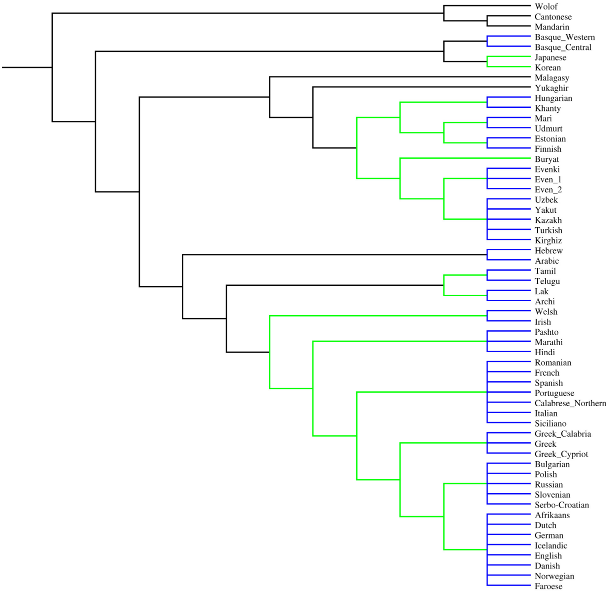
pivot of the Steinhaus transform: the "totally unsound" all-$*$ sequence

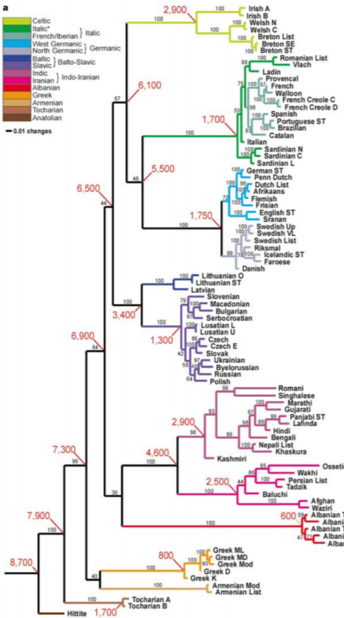
consistency $\chi(x)$ of the string x: its taxicab distance from the all-$*$ string

$$S_d(x,y) \doteq \frac{2d(x,y)}{d(x,y)+\chi(x)+\chi(y)}$$

weight $w(x)$ of the string x: its taxicab distance from the all-0 string

$$S_d(x,y) \doteq \frac{2d(x,y)}{d(x,y)+\min\left[\chi(x)+\chi(y), w(x)+w(y)\right]}$$


Wolof
Cantonese
Mandarin
Basque_Western
Basque_Central
Japanese
Korean
Malagasy
Yukaghir
Hungarian
Khanty
Mari
Udmurt
Estonian
Finnish
Buryat
Evenki
Even_1
Even_2
Uzbek
Yakut
Kazakh
Turkish
Kirghiz
Hebrew
Arabic
Tamil
Telugu
Lak
Archi
Welsh
Irish
Pashto
Marathi
Hindi
Romanian
French
Spanish
Portuguese
Calabrese_Northern
Italian
Siciliano
Greek_Calabria
Greek
Greek_Cypriot
Bulgarian
Polish
Russian
Slovenian
Serbo-Croatian
Afrikaans
Dutch
German
Icelandic
English
Danish
Norwegian
Faroese


a
Celtic
Italic*
French/Iberian
Italic
West Germanic
North Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish
Lithuanian O
Lithuanian ST
Latvian
Slovenian
Macedonian
Bulgarian
Serbocroatian
Lusatian L
Lusatian U
Czech
Czech E
Slovak
Ukrainian
Byelorussian
Russian
Polish
Romani
Singhalese
Marathi
Gujarati
Panjabi ST
Lahnda
Hindi
Bengali
Nepali List
Khaskura
Kashmiri
Ossetic
Wakhi
Persian List
Tadzik
Baluchi
Afghan
Waziri
Albanian T
Albanian G
Albanian Top
Albanian K
Albanian C
Greek ML
Greek MD
Greek Mod
Greek D
Greek K
Armenian Mod
Armenian List
Tocharian A
Tocharian B
Hittite
2,900
6,100
1,700
5,500
6,500
1,750
3,400
1,300
6,900
2,900
4,600
7,300
2,500
7,900
600
8,700
800
1,700

thanks, mulțumesc, grazie