DOI: 10.13137/2039-8646/35609

Distanze, lingue e parole[*]

ANDREA SGARRO
Dipartimento di Matematica, Informatica e Geoscienze
Università di Trieste
sgarro@units.it

LAURA FRANZOI
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
Università di Trieste
laura.franzoi@deams.units.it

ABSTRACT

Even if surprising for many mathematicians, quite a large number of the distances described in the Encyclopedia of distances, are not metric distances, i.e., they do not comply with some or other of the metric axioms which appear to be so natural or even unexpendable to those who tackle this multifaceted geometric and topological notion. Using examples taken from linguistics and word strings, we argue that the notion of distance is so rich and fruitful that the metric axioms in some cases risk to be an unreasonably narrow cage.

PAROLE CHIAVE

DISTANZA/ DISTANCE; DISTANZA METRICA / METRIC DISTANCE; DISTANZA DI EDIT / EDIT DISTANCE; LINGUISTICA COMPUTAZIONALE / COMPUTATIONAL LINGUISTIC.

1. CHE COS'È UNA DISTANZA?

Alla domanda appena formulata molti matematici risponderebbero senza esitazione enunciando la definizione di una *distanza metrica* o più concisamente di una *metrica*, definizione che ci sarà preziosa. Si parte da un insieme o se volete uno *spazio* S, i cui oggetti (i cui elementi), in numero finito o infinito, denoteremo con lettere come x, y o z (per evitare banalità converrà pensare che gli oggetti distinti siano almeno tre, anche se i matematici più diligenti si accontenterebbero di imporre che *S* non sia vuoto).

[*] *Title: Distances, languages and words.*

Ad ogni coppia (x,y) di oggetti distinti o anche coincidenti (può darsi che x = y) viene associato un numero non negativo d(x,y) che è appunto la distanza fra x e y. Inutile dire che per poter usare un termine impegnativo come "distanza" conviene imporre qualche condizione di regolarità in modo da escludere "comportamenti inammissibili". La distanza d si dice metrica se, qualunque siano gli oggetti x, y e z, distinti o coincidenti, valgono i seguenti tre assiomi:

i) d(x,y) = 0 se e solo se x = y (solo le auto-distanze sono nulle);

ii) d(x,y) = d(y,x) (simmetria della distanza);

iii) d(x,z) + d(z,y) ≥ d(x,y) (disuguaglianza triangolare, cfr. Figura 1).

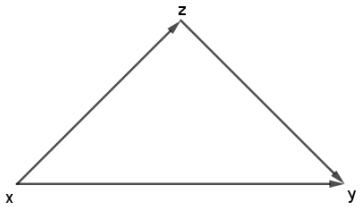

Figura 1. La disuguaglianza triangolare.

(i simboli ≥ e ≤ significano "maggiore o uguale a" e rispettivamente "minore o uguale a"); inutile osservare che la normalissima distanza euclidea nel piano o nello spazio è appunto metrica. Non cambia nulla di essenziale, ma precisiamo subito che talvolta vengono ammesse anche distanze infinite d(x,y) = +∞.

Potremmo iniziare una discussione filosofica sull'opportunità o sulla naturalezza dei tre assiomi, ma il discorso si incaglia appena partiti. In quello che è il "libro sacro" delle distanze, vale a dire nell'*Encyclopedia of distances* di Deza e Deza[1], buona parte delle distanze elencate non sono affatto metriche, e ce ne sono a centinaia provenienti dai più diversi campi teorici e applicati. A giudizio degli autori, va da sé sindacabile, la richiesta minima al di sotto della quale non si dovrebbe scendere se si vuole evitare che l'uso del termine "distanza" diventi fuorviante è l'unico debolissimo assioma per ogni coppia di oggetti distinti x e y:

[1] Cfr. DEZA, DEZA 2016.

iv) $0 \leq d(x,x) \leq \min[d(x,y), d(y,x)]$

dove min[,] indica il minimo dei due argomenti nelle parentesi quadre. Per far funzionare meglio la nostra intuizione è auspicabile che valga la disuguaglianza triangolare ordinata *iii)* dove stavolta l'ordine in cui vengono specificati gli oggetti è essenziale, non essendo più garantita la simmetria. A questo punto la discussione filosofica è inevitabile, ma prima di impegnarcisi nelle sezioni seguenti, anzi parallelamente alla discussione, vediamo di introdurre tre distanze estremamente notevoli che abbiamo scelto perché mettono in crisi gli assiomi metrici, pur essendo quelle dei §§ 2 e 3 delle distanze registrate a pieno titolo nell'*Encyclopedia of Distances*.

2. La distanza di Muljačić

Nell'*Encyclopedia* è chiamata forse troppo generosamente *Sgarro distance*; altro nome usato in letteratura è quello di *fuzzy Hamming distance* visto che ha a che fare appunto con la logica *fuzzy* o sfocata o sfumata[2]. Per non appesantire il discorso per il momento ci accontenteremo della distanza di Muljačić "monodimensionale", limitata a singoli valori logici x e y che sono numeri reali dell'intervallo [0,1]: $0 \leq x \leq 1$, $0 \leq y \leq 1$.

Nella *logica binaria* (o se volete nella logica matematica di tipo aristotelico, crisippino) esistono solo i valori logici 0=falso e 1=vero, mentre nella *logica fuzzy* sono ammessi tutti i valori intermedi: oltre al bianco e al nero ci sono tutte le sfumature di grigio. Potremmo ad esempio decidere, sulla base delle nostre opinioni, del nostro stato di conoscenze, che il valore logico della proposizione P={Alice parla correntemente il francese} sia 0,9, mentre il valore logico della proposizione Q={Bruno è un abile chitarrista} sia solo 0,3 (stiamo dando dei "voti" come a scuola, solo sulla scala da 0 a 1, ma si veda Sgarro 2019[3] per maggiori dettagli).

Se x e y sono due valori logici, x il valore logico della proposizione P e y quello della proposizione Q, come possiamo definire la loro distanza d(x,y) nello spirito della logica fuzzy? Indichiamo con AND, OR e NOT i tre *operatori logici* fondamentali di

[2] Circa la logica *fuzzy* o sfocata o sfumata si veda Sgarro 2019, pp. 37-45.
[3] Cfr. Sgarro 2019.

congiunzione (P e anche Q, sia P sia Q), *disgiunzione* (P oppure indifferentemente Q, almeno uno dei due) e *negazione*. Una risposta, argomentata in SGARRO 2019[4], potrebbe essere, usando i valori logici che sono numeri fra 0 e 1:

d(x,y) = (P è vera AND Q è falsa) OR (P è falsa AND Q è vera)

Nella logica fuzzy[5] si definisce, argomentando opportunamente le definizioni:

x AND y = min[x,y], x OR y = max[x,y], NOTx = 1–x, per cui è adeguato porre, come appunto si fa con la distanza di Muljačić:

$$d(x,y) = \max\big[\min[x, 1 - y], [\min[1 - x, y]\big]$$

Con un po' di pazienza si potrebbe dimostrare che valgono la *ii)* e la *iii)*[6], ma ovviamente non vale la *i)*, perché le auto-distanze possono essere strettamente positive. Un problema? Al contrario, un grande vantaggio nell'ambito della logica sfocata, visto che l'*autodistanza* d(x,x) diventa una preziosa misura della "sfocatezza" del valore logico x, nulla solo se x è nitido o *crisp* (verità totale 1 o falsità totale 0) e massima in corrispondenza al valore totalmente "ambiguo" $x = \frac{1}{2}$, a metà strada esatta fra il vero e il falso.

Qualche chiarimento sul nome di Žarko Muljačić, un linguista spalatino che è stato uno dei massimi esperti delle lingue romanze o neo-latine. Scelse 40 parametri lessicali o sintattici da associare a 12 lingue neolatine[7]; i parametri potevano essere presenti o assenti nella lingua rispettiva, ma in alcuni casi la loro presenza/assenza era poco chiara per cui era saggio ricorrere a valori logici sfumati, di auto-distanza strettamente positiva. Ad esempio, al parametro 15 (comparativo formato con *plus* come nell'italiano *più buono* o nel francese *plus bon*, e non con *magis* come nel catalano *més bo* o nel rumeno *mai bun*) viene dato il valore di verità $\frac{1}{2}$ nel caso del provenzale, che oscilla fra i due.

Lo scopo di Muljačić era di dimostrare che il dalmatico, ormai estinto, è una "lingua-ponte" fra la latinità occidentale (italiano, francese, spagnolo, friulano, ecc.) e quella

[4] Cfr. SGARRO 2019.
[5] Cfr. SGARRO 2019.
[6] Si veda ad esempio FRANZOI, SGARRO 2017.
[7] Cfr. MULJAČIĆ 1967.

orientale, rappresentata essenzialmente dal rumeno. In dalmatico sono scritti ad esempio i *Praecepta Rectoris* della Repubblica di Ragusa-Dubrovnik (siamo nel 1280); l'ultimo parlante di dalmatico fu Antonio Udina (Tuane Udaina), morto a Veglia-Krk nel 1898, dopo essere stato, per fortuna della linguistica, ampiamente intervistato dal linguista albonese Matteo Bartoli.

Un'osservazione tecnica: la distanza fra stringhe di valori logici, nel nostro caso di 40 valori logici, è additiva e si trova sommando le distanze fra i corrispettivi valori (nel nostro caso la somma ha 40 addendi); le proprietà metriche si conservano e una lingua ha auto-distanza strettamente positiva quando almeno uno dei 40 parametri che la descrivono è mal definito, ossia né 0 né 1.

3. La distanza di Levenštejn

D'accordo, il nome yiddish di Vladimir Levenštejn è un po' impronunciabile, ma potete chiamarla anche *distanza di edit* oppure, se siete degli incorreggibili puristi, *distanza redazionale*; la sua importanza per gli utenti del web è fin troppo ovvia: se scrivete "distamzza" il sistema vi chiederà «Volevi dire "distanza"?» o magari vi correggerà "distamzza" in "distanza" senza neppure chiedervi niente. Ciò implica che "lui" si sia accorto che la stringa non registrata "distamzza" è vicina alla stringa pienamente legittima "distanza", la loro distanza è piccola (e scusateci il bisticcio di parole).

Occupiamoci proprio di parole o se preferite di stringhe costruite, per dire, sull'alfabeto delle 21 lettere italiane maiuscole, come TRIESTE o MONFALCONE di lunghezza rispettiva 7 e 10 (numero delle lettere che compongono la stringa).

L'idea delle distanze di edit è di servirsi di trasformazioni ognuna delle quali ha il proprio costo, e di usare la sequenza di trasformazioni a costo minimo che consente di passare dalla prima alla seconda stringa (il costo di più trasformazioni usate una dopo l'altra è la somma dei singoli costi): è per l'appunto questo costo additivo minimo che è la distanza di edit fra le due stringhe. Qualche esempio chiarirà le cose.

Supponiamo di avere a disposizione due trasformazioni, il *twiddle* (scambio di due lettere

adiacenti) e la sostituzione di una lettera con un'altra, la prima trasformazione a costo unitario, la seconda a costo 2.

Trasformare CERTA in CETRA costa 1, mentre passare da CARO a CASO costa 2 - non ci sono possibilità più economiche. Passare da CARRO a CARO non si può, la distanza, se si vuole, è infinita. Se ci fossero anche gli inserimenti e le cancellazioni, diciamo a costo 1, la distanza fra CARRO e CARO sarebbe solo 1 (una singola cancellazione), come quella fra CARO e CARRO (un inserimento), ma se gli inserimenti costassero 2 e le cancellazioni soltanto 1 la simmetria si perderebbe. Se poi i *twiddle* fossero gratuiti la distanza fra le due stringhe, pur distinte, CERTA e CETRA scenderebbe a zero.

Abbiamo smontato tutti gli assiomi delle distanze metriche, tranne uno, la triangolarità ovviamente *ordinata* visto che non è più garantita la simmetria. Possiamo passare a costo minimo da x a z per poi arrivare sempre a costo minimo da z a y; questo è un modo legittimo di passare da x a y, ma non è affatto detto che sia il più economico, visto che stiamo ignorando i percorsi legittimi che non rispettano il vincolo di dover "toccare" per forza z. In altre parole: le trasformazioni e i loro costi possono essere capricciosi come vi pare, ma la disuguaglianza triangolare ordinata è sempre verificata dalle distanze di edit.

Ci corre l'obbligo di precisare che la distanza di edit di gran lunga più popolare consente tre trasformazioni, *sostituzione*, *cancellazione* e *inserimento*, tutte e tre a costo unitario, ed è perfettamente metrica. A differenza di quanto succedeva con i nostri esempietti, scovare il percorso minimo non è sempre facile "a occhio" ma per fortuna esiste uno splendido algoritmo basato sui principi della programmazione dinamica che risolve il problema a bassa complessità di calcolo: è per questo che "lui" è così bravo a calcolare le distanze in tempo reale, come vi siete ben accorti usando la rete o il vostro telefonino.

4. Le due distanze di Longobardi

Finora non siamo riusciti a smentire la triangolarità, ma neanche questa resiste come mostrano numerosi casi trattati nell'*Encyclopedia* e come mostra la situazione che

segue, situazione che nell'*Encyclopedia* non appare ancora.

La scuola di linguisti tradizionali e computazionali diretta da Giuseppe Longobardi (York, UK) ha portato a risultati straordinari nella filogenesi delle lingue[8]. I metodi sono basati sulla *sintassi* piuttosto che sul *lessico*, come si fa di norma, in base all'osservazione che il "segnale" sintattico evolve nel tempo più lentamente del segnale lessicale: è questo che spiega l'uso del termine "preistoria" nel titolo del lavoro di CEOLIN *et al.*[9].

Nel caso in questione[10] le lingue trattate, tutte del Vecchio mondo – Europa, Asia e Africa – sono 58, classificate ognuna mediante 94 parametri sintattici. A differenza di quanto accadeva con Muljačić, qui il problema non è più la *fuzziness*, ma piuttosto il fatto che certi parametri sono totalmente non informativi, è come se non esistessero.

Per intenderci, anche se siamo costretti a semplificare e a chiedere scusa ai linguisti: se al parametro «In questa lingua esiste la declinazione dei sostantivi» abbiamo risposto no è inutile chiederci, come previsto da un parametro successivo, se genitivo e dativo possano avere desinenze distinte, come in tedesco, cui spetterebbe l'1, mentre in rumeno, cui spetterebbe lo 0, sono sempre coincidenti (i sostantivi tedeschi e rumeni si declinano)[11]. Le stringhe associate a ciascuna lingua sono dunque ternarie, 0 per i parametri no, 1 per i parametri sì e una stella * per i parametri non informativi che di fatto in quella lingua sono inesistenti; nel caso appena accennato, del genitivo e del dativo, l'italiano, che non ha la declinazione dei sostantivi, avrebbe una stella.

Per giungere alle classificazioni delle lingue e cercare di capire se il giapponese e il coreano, per dire, abbiano qualche remoto antenato che i metodi tradizionali, lessicali, non riescono a intravedere, abbiamo bisogno di distanze fra stringhe ternarie.

Se non ci fossero le stelle e le stringhe fossero solo binarie potremmo usare la *distanza di Hamming* o la *distanza di Jaccard*. Nel caso di Hamming si contano le posizioni in cui le stringhe differiscono e si normalizza dividendo per la lunghezza comune delle stringhe

[8] Si veda ad esempio CEOLIN *et al.* 2021.
[9] Cfr. CEOLIN *et al.* 2021.
[10] Cfr. CEOLIN *et al.* 2021.
[11] Ad esempio in romeno *fata*, *fetei* (la ragazza, della/alla ragazza), *fetele*, *fetelor* (le ragazze, delle/alle ragazze).

(la normalizzazione non si fa sempre, come invece faremo noi), nel caso di Jaccard si accorciano le stringhe eliminando le posizioni in cui figura lo zero sia in una stringa sia nell'altra e si calcola la *distanza normalizzata di Hamming* fra le stringhe così accorciate.

Ad esempio $d_H(001, 011) = \frac{1}{3}$, $d_J(001, 011) = \frac{1}{2}$: nel caso di Jaccard la prima posizione è come se non esistesse. Va subito detto che queste due distanze sono rigorosamente metriche, anche se nel caso di Jaccard non è così semplice dimostrarlo.

Purtroppo le nostre stringhe sono ternarie, per cui dobbiamo saper gestire anche le stelle modificando quanto abbiamo detto in modo da ottenere distanze di tipo ternario; chiameremo le due nuove distanze *pseudo-Hamming* e *pseudo-Jaccard*.

Quello che si fa, date le due stringhe, è eliminare le posizioni in cui compaiono le stelle, comprese le posizioni in cui una sola delle due lingue ha una stella, accorciare di conseguenza la lunghezza e finalmente calcolare la distanza normalizzata di Hamming o di Jaccard fra le stringhe così accorciate (che nel caso di Jaccard possono essere soggette a un ulteriore accorciamento).

Ad esempio con x =1*0*01 e y =100*10 si ha $d_{PH}(x, y) = \frac{2}{4} < \frac{2}{3} = d_{PJ}(x, y)$.

Purtroppo entrambe le pseudo-distanze possono violare la disuguaglianza triangolare; si pensi alle tre stringhe "astratte" di lunghezza 21:

x = 100000000001111111111,

y = 111111111110000000000,

z = 1**********.

Le stringhe x e z sono diverse, ma d(x,z) = 0; come se non bastasse

$d(x,z) + d(z,y) = 0 + 0 < d(x,y) = \frac{20}{21}$,

e dunque quasi 1, che è il valore più grande che una distanza normalizzata possa assumere. Non abbiamo specificato il tipo di distanza perché in questo caso, in mancanza di posizioni in cui entrambe le stringhe abbiano uno 0, pseudo-Hamming e pseudo-Jaccard coincidono.

Il nostro antipatico esempio riguarda stringhe "astratte": che cosa succede con le 58 lingue "concrete" e con le loro stringhe di lunghezza 94?

Ci sono sì lingue a distanza 0, come l'irlandese e il gallese, ma questo solo perché le loro stringhe ternarie sono identiche (si rammenti che stiamo usando distanze fra stringhe che solo indirettamente sono distanze fra lingue). In quanto alla disuguaglianza triangolare le triple di lingue x, y e z che vanno controllate sono 92568, come mostra il calcolo combinatorio (basta contare il numero delle coppie non ordinate di elementi x,y distinti e moltiplicare per il numero degli elementi z che rimangono).

Con la distanza pseudo-Hamming le triple difettose sono 993, vale a dire lo 0.0107 ≈ 1% del numero totale. Le triple "peggiori", quelle con la differenza $d_{pH}(x,y) - d_{pH}(x,z) - d_{pH}(z,y)$ massima e pari a 0.1, sono (SerboCroato, Hindi, Mandarino), (SerboCroato, Hindi, Cantonese), (Sloveno, Hindi, Mandarino), (Sloveno, Hindi, Cantonese), (Polacco, Hindi, Mandarino), (Polacco, Hindi, Cantonese), (Russo, Hindi, Mandarino), (Russo, Hindi, Cantonese).

Con la distanza pseudo-Jaccard la situazione è migliore. Ci sono 143 triple difettose pari allo 0.015... = 0.15% del numero totale. Lo scarto massimo di 0.089 corrisponde a (Hindi, Tamil, Mandarino), (Hindi, Tamil, Cantonese), (Hindi, Telugu, Mandarino), (Hindi, Telugu, Cantonese)[12].

Ai linguisti poco interessa la mancata metricità: ecco piuttosto, nella Figura 2, l'*albero filogenetico* cui sono giunti. Fra le parentele che suggerisce c'è quella fra il giapponese e il coreano, con un progenitore comune che affonda nel tempo e che è "invisibile" ai tradizionali strumenti della filogenesi lessicale.

5. Conclusioni

Dati gli interessi e le preferenze dei due autori, non sorprende che gli esempi trattati rivestano tutti carattere linguistico, ma le distanze dell'*Encyclopedia*, metriche o non metriche che siano, riguardano situazioni prese dalla geometria, dall'algebra, dalla teoria dei numeri, dall'analisi funzionale, dalla statistica e dal calcolo delle probabilità,

[12] Per maggiori dettagli si veda Dinu *et. al.* 2023.

dalla teoria dei grafi e dalla combinatorica, dall'ingegneria matematica e dall'informatica (immagini, audio, internet), dalla biologia e dalla fisica, dalla chimica e dalla medicina, dalla geografia, dalla geofisica e dall'astronomia fino alla cosmologia, alla teoria della relatività e perfino alla criminologia: se c'è un concetto pervasivo, dunque, è quello di *distanza*, e neppure la gabbia della metricità, pur ampia e generosa, riesce a contenerlo.

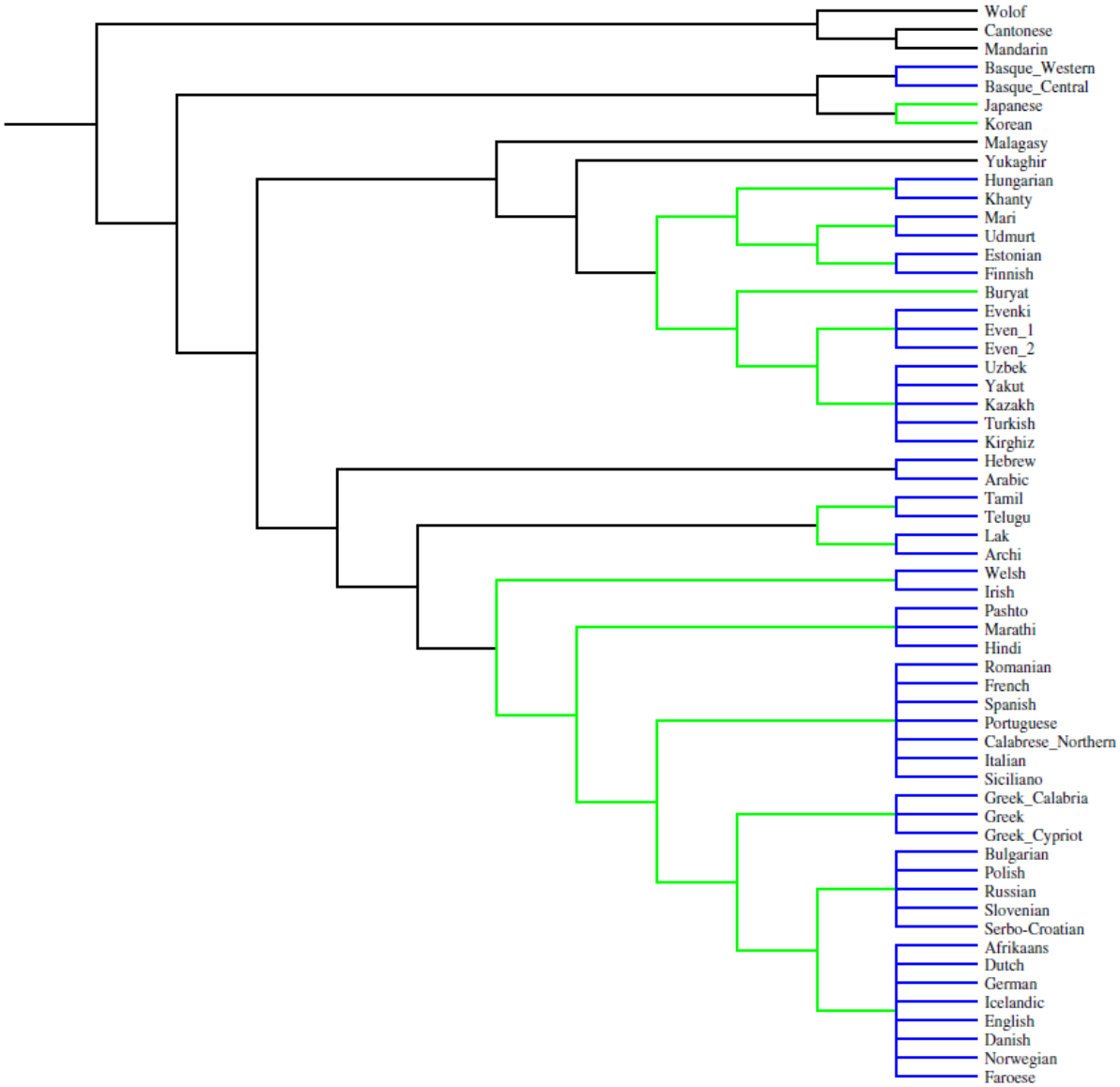

Figura 2. L'albero filogenetico basato sulla sintassi (Fonte: CEOLIN *et al.* 2021).

BIBLIOGRAFIA

CEOLIN A., GUARDIANO C., LONGOBARDI G., IRIMIA M. A., BORTOLUSSI L., SGARRO A.
2021, «At the boundaries of syntactic prehistory», *Philosophical Transactions of the Royal Society B*, 376, pp. 1-10.

DEZA M. M., DEZA E.
2016, *Encyclopedia of distances*, Springer.

DINU A., DINU L., FRANZOI L., SGARRO A.
2023, *Distances in syntax-based linguistic phylogeny*, to be submitted to a computational linguistic forum.

FRANZOI L., SGARRO A.
2017, *Fuzzy Hamming distinguishability*, Napoli, FUZZ-IEEE, pp. 1-6.

MULJAČIĆ Ž.
1967, «Die Klassifikation der romanischen Sprachen», *Romanistisches Jahrbuch*, 18, pp. 23-37.

SGARRO A.
2019, «Come contraddirsi rimanendo coerenti: il caso della logica fuzzy», *QuaderniCIRD*, 19, pp. 37-45.

SGARRO A., FRANZOI L.
2021, «Sillogismi sfocati: il modus ponens», *QuaderniCIRD*, 22, pp. 50-62.