Names | A Journal of Onomastics

Navigating Linguistic Similarities Among Countries Using Fuzzy Sets of Proper Names

Davor Lauc
University of Zagreb, CROATIA

ans-names.pitt.edu

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 1, Spring 2024 DOI 10.5195/names.2023.2569

Abstract

This paper examines the commonalities among several countries and languages through the lens of proper names, especially forenames. It posits that the investigation of these names offers a fresh perspective on language similarity due to their distinct influence from cross-cultural interactions and language contact compared to regular vocabulary. The study introduces a novel measure that generalizes the similarity between sets by considering the distances between elements. This metric is employed to assess phonetic commonalities in forenames. The results of this analysis show a notable correlation between the commonality of proper names across languages and the overarching commonality of the languages themselves. In addition, the forename commonalities also provided more insights. As this investigation shows, proper names can also serve as a potentially potent metric for language similarity and may be used to unveil additional cultural commonalities and disparities among nations. The paper concludes by addressing the constraints of this research and discussing prospects for subsequent studies.

Keywords: first name, proper name, anthroponomastics, language similarity, language distance, phonetic similarity, socioonomastics

Introduction

From historical linguistics to language learning, language distance and similarity have been measured for different purposes, using a variety of methods. This paper focuses on lexical similarity, which measures the degree to which the word sets of two given languages are similar. Lexical similarity has been calculated using different word lists. There are curated limited word sets such as the Swadesh list (Swadesh 1955) and entire lexicons (Goldhahn & Quasthoff 2014). To estimate similarities between word pairs, different measures between words have also been used. Two examples of measurement criteria are string distances and phonological similarity (Dryer & Haspelmath 2013; Müller et al. 2010; Wichmann et al. 2010).[1]

In the current investigation, the analytical focus is the similarity among proper names. To the author's knowledge, proper names are a part of the language lexicon that has not been previously used to systematically measure similarity between languages on a large-scale. This paper posits that this approach offers a fresh perspective on the issue of language distance, as proper names are influenced by linguistic and cultural exchange differently than other elements of shared vocabulary.

The Forgotten Half of the Language

With a few exceptions, mainstream linguistics does not focus on studying proper names. There are many reasons for this oversight, including the lack of regularity and predictability in onomastic meaning and the fact that proper names are extraordinarily culturally and contextually bound. However, even a simple count in available labeled corpora demonstrates that, despite the fact that proper names typically constitute only 15–21% of the text, they dominate the vocabulary: measured as distinct text tokens, they comprise 49–81% of it.[2]

Most proper names never appear in public corpora but are hidden in the deep web and databases (Liang 2008). Therefore, the method of comparing labeled corpora does not accurately compare respective vocabulary sizes. The NomoGraph database (Mondonomo 2023) of names estimates that there are more than 20 million English forenames and surnames in the countries where English is spoken.[3] When we compare this to the estimated size of the English vocabulary—excluding the names of people, organizations, locations, and other entities—, it is safe to assume that ordinary, non-proper-name words represent about one to two percent of the total number of words in the entire English lexicon. So, given the relative sizes of proper names and regular language lexicons, it is reasonable to hypothesize that measurements of language similarities based on comparing proper names can provide a new complementary perspective to the investigation of language similarity.

Some General Issues about Similarity Relation

It should be noted that the relation of similarity between languages is not a well-defined, clear-cut construct on which most experts agree. Similarity may therefore be best understood from the perspective of fuzziness and context dependency, first recognized conceptually by the philosopher Nelson Goodman (1983). To paraphrase his now famous airport luggage analogy, to a historical linguist, two languages are similar if they are of similar origin; to a phonologist, if they sound similar; to a typologist, if they share grammatical features; to a second language acquisition researcher, if they are mutually understandable, and so on.

Should we accept Goodman's argument, we must also accept that there is no one single objective way to measure similarity. The meaning of similarity is contingent on the aim of the comparison. Theoretical physicist Satoshi Watanabe later provided a similar, albeit more formal, argument through his "ugly duckling theorem" (Watanabe 1986, 1969). Watanabe's theorem is akin to Goodman's more influential "grue" argument and Wolpert's formalization "no free lunch theorem" (Lauc 2018; Wolpert & Macready 1997). According to these theories, in the absence of preference for the aspects of languages being compared, every classification is equally valid.

Empirical research provides additional arguments for the peculiarity of similarity relationships. For example, research conducted by psychologist Amos Tversky (1977) implies that human perception of similarity is, contrary to popular belief, asymmetric; is highly context-dependent; and does not satisfy the "triangle inequality." Where the latter feature is concerned, an example may be useful. Some people might consider North Korea to be more similar to China than China is to North Korea. Moreover, when the USA is taken into consideration, the perceived similarity between North Korea and China might be greater than the direct similarity between North Korea and the USA. These psychological findings from the 1970s have been corroborated by later investigations of language similarity that show asymmetry in mutual intelligibility (Gooskens et al. 2018; Schüppert 2011).

Further strong support for Goodman's and Tversky's research is the existence of a plethora of interdisciplinary publications demonstrating similarity and distance measures (Carlier et al. 2023; Almasoud et al. 2020; Bero et al. 2017; Vijaymeena & Kavitha 2016; Cha 2007). This work does not mean similarity is an unscientific relationship that should be avoided. On the contrary, the working hypothesis is that similarity relations are both epistemologically and ontologically primary and that equivalence relations—as used, for example, in language/dialect identification—are only special asymptotic cases. In this spirit, this research attempts to reveal one of many aspects of similarity among languages.

Data and Methods

Proper Name Database

This research uses the Nomograph knowledge graph developed by Mondonomo (2023) as a proxy for proper name lexica. It is the most extensive database of proper names, encompassing various semantic relationships between names and their attributes. The entire dataset is available to the public (Mondonomo 2023).

Nomograph has been perpetually constructed through an iterative process since 2020. Each iteration begins with data gathering and proceeds along various language processing steps, such as: data cleaning, Named Entity Recognition (NER), language and entity classification, data labeling, and training various name understanding models.[2] The initial dataset comprised more than 41 TB of data from 618 different sources compiled from multiple publicly available datasets, both unstructured (e.g., the C4 Multilingual dataset or scanned phone directories) and structured (e.g., Wikidata [2023] and Virtual International Authority File (VIAF) [VIAF 2023]). In compliance with privacy protection, all personal data are maximally anonymized such that only statistical data are retained. By the end of the fifth iteration in 2023, Nomograph contained nearly 200 million different name forms in 6,000 name/script/country combinations and over 3 billion data points (name features and relationships among names). The initial estimation is that the knowledge graph covers approximately 98% of human names in most living languages spoken by more than one million people.

In the context of this study, to make phonetic and similarity algorithms viable in terms of computation, exceedingly infrequent names have been removed from the dataset. However, given this study's weighted approach to similarity, this reduction should not significantly affect the output similarity matrix. The threshold is established at the projected relative frequency of a single name per million individuals in the population of each respective country. Names in each country are categorized according to all the official languages or de facto national languages of that nation. Only names in the official script of the country's language are considered to ensure the efficacy of phonetic algorithms, thereby excluding any transliterated or Romanized variants. The final list of name-country pairs consists of 5.33 million tokens, encompassing an estimated 6.32 billion namesakes in the global population.

Phonetic Transliteration of the Names

Names are transcribed into International Phonetic Alphabet (IPA) notation to facilitate the estimation of phonetic similarity. Despite recent significant advancements in grapheme-to-phoneme (G2P) algorithms, a high-quality, multilingual G2P system still does not exist. For the purposes of this research, a combination of available phonetic dictionaries and G2P algorithms has been employed.

The utilized dictionaries (Open-Dictionary-Data 2023; Taubert 2022; Zhu et al. 2022; Lee et al. 2020) include only a fraction (696,278: 15%) of forenames in the list. However, they are instrumental in evaluating and selecting different G2P algorithms. The employed G2P algorithms are from the following resources: Zhu et al. 2022; Li et al. 2020; Phatthiyaphaibun 2020; Park 2019; Mortensen et al. 2018; Llarena 2017; ESPEAK 2015. Some of these algorithms are rule-based and demonstrate a high level of accuracy (98%+) for "orthographically shallow languages." As defined by Katz & Frost (1992) such languages have a high level of predictability for print-to-speech correspondences and the derivation of word pronunciations based on their orthography. Examples include most Slavic languages, Finnish, and Italian. Conversely, for other languages like English, French, and Arabic, this research relies on neural models.

It is worth noting that even the word error rate (WER) of recent state-of-the-art systems is relatively high and averages about 20–25% (Sun et al. 2019). This rate exhibits significant variability across different algorithms and languages, however. Consequently, all applicable G2P algorithms are scored by the inverse WER, as measured by the collected IPA dictionaries for each language. All transliteration algorithms are applied to the names list, and only the transliteration with the highest score is selected.[3]

IPA transliterations are normalized by removing accents and tone markers. This normalization is performed because only a minority of systems generate accents and tones despite their critical role in assessing similarity in certain languages, particularly tonal languages. The working hypothesis is that this transliteration process is sufficient for statistically assessing language similarity, as no language-specific bias in error analysis has been identified as yet. Nonetheless, further research is necessary to corroborate this finding.

Similarity Assessment

The application of steps described above allow each language-country combination to be represented as a fuzzy set. The set members are given names in their phonetic form, and the membership function represents the country-wise normalized propensity of the name. The task of estimating similarity between national languages can be defined as the appropriate similarity measure between two such sets.

Phonetic Similarity

The similarity between elements should be defined to facilitate the comparison between the elements of such a constructed set. One possible approach entails matching only identical International Phonetic Alphabet (IPA) tokens. However, this method could overlook names that sound very similar but are represented by different IPA symbols. For example, names could be similar if they containedan unrounded vowel /i/ and close front rounded vowel /y/ or an alveolar tap /ɾ/ and alveolar trill /r/. This discrepancy can be attributed to variations among languages, such as closer or more open vowels and unintended variations among grapheme-to-phoneme (G2P) algorithms. This research employs the feature edit distance measure as implemented by Mortensen and colleagues (2016) to minimize such unintended differences.

While this similarity measure is far from ideal, it seems to be the most robust measure available. For example, the German name *Johann* /joːhan/ is 100% similar to the Dutch name *Yohan*, the Thai name โยฮัน /johan/, and the Korean 요한 /joːhan/. However, it is 85% similar to the Norwegian *Johan* /juːhɑn/ and the Hungarian *Johann* /johɒn/. Names like the Italian *Nicola*, the French *Nicolas*, the Croatian *Nikola*, the Serbian *Никола*, and others are pronounced the same and have a maximal similarity. They are, however, 75% similar to the Arabic نيكولا/ /nikula/ and the Hungarian *Nikola* /nikolɒ/.

Given the definition of language similarity used in this study, the above method of estimating similarity among names should suffice for statistical purposes.[5] The assumption is that errors stemming from phonetization and distance measuring procedures are not biased toward any specific language or features. However, further research is necessary to develop better language-agnostic and more empirically adequate approximations of phonological similarity (Li et al. 2022; Kessler 2005). It is essential to note that for the purposes of this investigation, language variants like British and US American English were considered the same language for the analyses. This procedure is an admitted limitation of this work. In addition, only official or de facto national languages were considered in this analysis.

Similarity Measure

In this study, similarity is conceptualized as the probability that a monolingual native speaker of one language will identify a name from another language as resembling a name in their own language when heard without context. This criterion is analogous to mutual intelligibility (Gooskens & Heuven 2021), but, for obvious reasons, it utilizes the recognition of resembling names instead of understanding. This definition is strongly based on the assumption of judgement made by an idealized speaker of one language who has the following characteristics: 1.) has not been exposed to the names of another language; 2.) can assess phonological familiarity or sound-alike on a scale; and has been exposed to a representative sample of the names in their native language. In this sense, the similarity is the bi-variate probability distribution of phonological similarity and the propensity of names. The more similar-sounding a name in a target language is to a more common name in a source language, the more probable it is that a source language speaker finds the name familiar. Assuming (naively) the independence of these variables, we can approximate it using the following generalization of the Tversky similarity measure (Tversky 1977; Jaccard 1908), which is the most commonly used method in similarity assessment:

$$sim(L_1, L_2) = \left(\alpha \times \frac{\sum_{\substack{(w_1,w_2)\in \\ ms(L_1 \times L_2)}} \frac{m_s(w_1,w_2)}{min\left(f(w_1), f(w_2)\right)^{-1}}}{\sum_{\substack{(w_1,w_2)\in \\ ms(L_1 \times L_2)}} \frac{m_s(w_1,w_2)}{max\left(f(w_1), f(w_2)\right)^{-1}}} \right) + \left((1-\alpha) \times \frac{\sum_{\substack{(w_1,w_2)\notin \\ ms(L_1 \times L_2)}} \frac{m_s(w_1,w_2)}{min\left(f(w_1), f(w_2)\right)^{-1}}}{\sum_{\substack{(w_1,w_2)\notin \\ ms(L_1 \times L_2)}} \frac{m_s(w_1,w_2)}{max\left(f(w_1), f(w_2)\right)^{-1}}} \right)$$

In the formula above, $m_s s(L_1 \times L_2)$ represents a subset of the Cartesian product of two lexica, where the phonetic similarity of a pair exceeds a threshold discussed earlier. m_s is the normalized phonetic similarity between words w_1 and w_2, ranging from 0 to 1. Meanwhile, f denotes the normalized log frequency[6] of the word (or name) within the corpus (or country). The α coefficient is used to represent the non-symmetrical nature of language similarity, as discussed earlier. In a hypothetical case where one language's lexicon is a proper subset of another, the similarity of the former to the latter will be maximal. However, the reverse similarity will be lesser, depending on the magnitude of α, as many word pairs will not belong to the set of approximate homophone pairs.

It should be noted that this measure resembles the Ružička similarity used in life sciences and various similarity measures used among fuzzy sets (Wang 1997). The primary difference, apart from the relatively standard weighting, is that there is no straightforward one-to-one correspondence among elements of the sets. Instead, there is a slightly more complex many-to-many correspondence between (phonetically) similar elements.

Generating Similarity Pairs Using Siamese IPA2Vec Model

Given the potentially quadratic number (6 million squared) of comparisons among the name list, a neural model was designed to vectorize names for efficient large-scale similarity evaluation. This step allows cosine similarity between vectors to represent their phonetic closeness, determined by the inverse feature edit distance,[7] making the processing feasible within days rather than years.

The training dataset consisted of word pairs from the large-scale cognates' lexical database (Batsuren et al. 2019) for the languages occurring in the name list. Only words that were contained in the previously constructed phonetic dictionary were used. Additionally, one random pair, as a negative example, was generated for each positive pair. The resulting dataset contained 2,364,433 word pairs to which feature edit distance was assigned using the method developed by (Mortensen et al. 2016). The training, development, and test splits were performed in a 98:1:1 ratio.

To create vector embeddings that would facilitate fast retrieval of phonetically similar names, a neural network model that uses standard Siamese network architecture was developed. For the shared encoder part, the ByT5 model was used (Xue et al. 2022). PyTorch cosine similarity was used as the objective function, using the mean pooling of token embeddings. The model was trained for 30 epochs, resulting in training and evaluation losses of 0.00075 and 0.00072, respectively.

The actual list of similarity pairs was then constructed using a two-stage process. First, for each name in the list, the FAISS library (Johnson et al. 2019) was used to perform an approximate nearest neighbor search which retrieved the 10,000 closest names. The results were then refined by calculating the actual feature edit distance using the method developed by Mortensen and colleagues (2016). The total number of candidate pairs was 13.87 billion. The cutoff was applied to distances greater than five to facilitate the efficient application of the similarity measure. The data preparation resulted in 2.3 billion pairs of phonetically similar names. The data was stored in sparse matrices, ensuring efficient implementation of the similarity calculation.

Results

The table below presents a ranking of the top 30 countries based on mutual similarity. Each pair of countries was assigned a similarity score, scaled from 0 to 100, along with information detailing whether the countries share the same language “L.”, either entirely (S) or partially (P). An overlap was classified as partial (P) if at least one common official language exists. Table 1 also indicates whether the countries neighbor each other (“N.”). The comprehensive list of similarities between any two country pairs is available at the following link: https://echoes.mondonomo.ai/countries.

Table 1: The 30 Countries Exhibiting the Highest Similarity Based on Forename Comparisons

Rank	Country 1	Country 2	Sim.	L.	N.
1.	United Kingdom	Australia	84.6	S	N
2.	Australia	Canada	80.5	S	N
3.	United Kingdom	Canada	78.2	S	N
4.	Mexico	Peru	77.4	S	N
5.	Chile	Ecuador	77.1	S	Y
6.	Mexico	Venezuela	75.2	S	N
7.	Austria	Switzerland	73.4	S	Y
8.	Bolivia	Guatemala	73.1	S	N
9.	Venezuela	Peru	72.3	S	N
10.	Mexico	Colombia	72.0	S	N
11.	Algeria	Morocco	71.9	S	Y
12.	Venezuela	Colombia	70.5	S	Y
13.	Peru	Colombia	69.7	S	Y
14.	Bolivia	Dominican Republic	68.3	S	N
15.	Angola	Portugal	68.2	S	Y
16.	Bolivia	Argentina	67.3	S	Y
17.	United Kingdom	South Africa	66.6	P	N
18.	Nicaragua	Honduras	66.2	S	Y
19.	Benin	Togo	66.0	S	Y
20.	Singapore	Canada	65.8	P	N
21.	Bolivia	Colombia	65.8	S	N
22.	Dominican Republic	Colombia	65.7	S	N
23.	Kazakhstan	Russia	65.4	P	Y
24.	Singapore	New Zealand	65.3	P	N
25.	Mozambique	Portugal	65.0	S	N
26.	Chile	Uruguay	64.4	S	N
27.	Costa Rica	Ecuador	64.2	S	N
28.	Palestinian Territory	Jordan	64.1	S	Y
29.	Senegal	Ivory Coast	64.0	S	Y
30.	Panama	Chile	64.0	S	N

Many pairs of countries in our similarity list can be anticipated based on a shared language or language family. For example, the three most similar countries—the United Kingdom, Canada, and Australia—share the English language, while the official languages of Algeria and Morocco belong to the Semitic branch. Cultural and political ties also play an important role, as seen between the United Kingdom and South Africa, which is further enhanced by geographic proximity. The same may be said of Russia and Kazakhstan, fostering with reference to language contact.

When we focus further on the country pairs that do not share a common language, the effect of common language families, cultural and political ties, and geographic proximity becomes more visible. The results of that analysis are displayed in Table 2.

Table 2: The 20 Countries without a Common Language that Exhibited the Highest Similarity based on Forename Comparisons

R.	Country 1	Country 2	Sim.	L.	N.
44.	Algeria	France	62.8	D	N
93.	Belgium	Sweden	56.1	D	N
103.	Germany	Netherlands	55.2	D	Y
149.	Mexico	Italy	52.2	D	N
171.	Germany	Sweden	51.0	D	N
174.	Belgium	Romania	50.6	D	N
176.	Serbia	Slovenia	50.4	D	Y
181.	Italy	Peru	50.2	D	N
215.	Croatia	Czechia	48.2	D	N
216.	Italy	Tanzania	48.1	D	N
219.	Austria	Sweden	48.1	D	N
225.	Slovakia	Slovenia	47.9	D	Y
227.	Afghanistan	Kenya	47.8	D	N
228.	Malaysia	Belgium	47.8	D	N
229.	Romania	Colombia	47.7	D	N
230.	Romania	Argentina	47.6	D	N
236.	Bolivia	Romania	47.3	D	N
241.	Czechia	Finland	47.1	D	N
242.	North Macedonia	Slovenia	47.1	D	N
247.	Romania	Guatemala	46.9	D	N
251.	Romania	Sweden	46.8	D	N

The potential influence of migrations, such as those between Algeria and France or Italy and Tanzania, may also be visible in this list, although more research is needed to test this hypothesis.

Forename Similarity-Based Clustering of Countries

Forename similarity-based clustering of countries is an approach to grouping nations based on the commonality of first names among their populations. Rather than utilizing traditional language-family similarity, this method focuses on the frequency of forenames, offering a fresh, human-centric lens to examine global patterns. Figure 1 presents a hierarchical clustering of European countries, each with a population exceeding one million. This clustering was conducted using the Voorhees algorithm (1986). Table 3 provides the country codes used in this research.

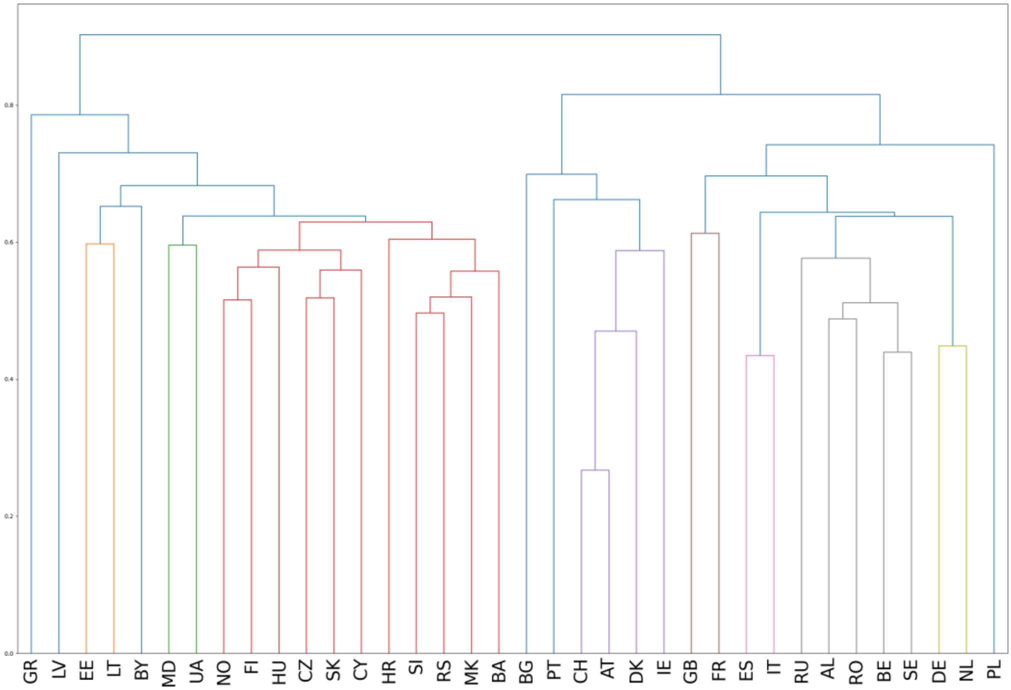

Figure 1: Hierarchical Clustering of European Countries

Table 3: The numerical codes for the countries examined

	Code	Country		Code	Country
1	AL	Albania	22	LT	Lithuania
2	AT	Austria	23	LV	Latvia
3	BA	Bosnia and Herzegovina	24	MD	Moldova
4	BE	Belgium	25	MK	North Macedonia
5	BG	Bulgaria	26	NL	Netherlands
6	BY	Belarus	27	NO	Norway
7	CH	Switzerland	28	PL	Poland
8	CY	Cyprus	29	PT	Portugal
9	CZ	Czechia	30	RO	Romania
10	DE	Germany	31	RS	Serbia
11	DK	Denmark	32	RU	Russia
12	EE	Estonia	33	SE	Sweden
13	ES	Spain	34	SI	Slovenia
14	FI	Finland	35	SK	Slovakia
15	FR	France	36	UA	Ukraine
16	GB	United Kingdom			
17	GR	Greece			
18	HR	Croatia			
19	HU	Hungary			
20	IE	Ireland			
21	IT	Italy			

These similarity pairs shown in Figure 1 might reveal groupings that may initially be unexpected. However, similarities in the forenames of countries may be due to a complex mix of factors, including religion, language, historical events, and pop culture trends that may have escaped immediate notice. For instance, it might be expected that nations with common linguistic roots, like Hungary and Finland, will display similar forename preferences. However, the analysis above also reveals that Hungary and Slovakia, despite being in different language families, also share onomastic similarity, perhaps due to their geographical proximity and language contact. The influence of globalization and migration patterns might also be traced through the evolution of forename similarity.

An SNA Analysis of Similarity Graph

Utilizing Social Network Analysis (SNA) techniques, it is possible to graphically represent countries and their linguistic similarities. The countries themselves are depicted as nodes, and the vertices between them indicate relationships. Though centrality measures are uncommon in linguistics, this method may provide insights into the roles countries and cultures play in facilitating language contacts. Table 3 showcases countries ranked by centrality measures, specifically using PageRank (P.), Closeness (C.), and Eigenvector (E.) metrics. Explained briefly, a "PageRank" is an algorithm that assigns a numerical weighting to each country, with the purpose of measuring its relative importance within the set. "Closeness" is a measure of centrality in a network, indicating the average length of the shortest path from a country to all other countries. An "Eigenvector" is a principal component that indicates the relative importance of a country within the network.

Table 4: The 20 Countries Exhibiting the Highest Centrality Based as Measured by PageRank

R.	Country	P.	C.	E.	R	Country	P.	C.	E.
1.	Austria	1.09	53.62	10.51	13.	Bosnia and Herzegovina	0.87	55.08	10.92
2.	Uruguay	1.09	52.14	9.82	14.	Namibia	0.86	49.61	9.42
3.	Burkina Faso	1.01	54.71	10.81	15.	Cameroon	0.86	52.82	10.12
4.	Togo	1.0	54.22	10.65	16.	Bahamas	0.86	53.14	10.62
5.	Portugal	0.99	52.65	10.04	17.	Belgium	0.85	47.85	8.92
6.	El Salvador	0.99	50.36	9.53	18.	Papua New Guinea	0.84	51.36	9.71
7.	Switzerland	0.98	49.94	9.83	19.	Suriname	0.84	57.17	11.21
8.	Ukraine	0.98	56.23	10.94	20	Finland	0.83	47.07	9.11
9.	Norway	0.98	50.12	9.13					
10.	Czechia	0.97	49.41	9.62					
11.	Bulgaria	0.95	55.63	11.02					
12.	Uganda	0.94	51.16	9.62					

Centrality metrics indicate that countries like Austria and Uruguay have the highest scores. These outcomes suggest that they may play a pivotal role in fostering linguistic similarities across a range of nations, attributable to their geographic locations as well as their cultural and political impacts. A deeper investigation is required to substantiate the causal relationships underlying these observations. Nonetheless, by concentrating on Austria, the country with the most pronounced centrality, it is possible to illustrate some of the reasons behind these centrality metrics. It is useful to employ a forced layout data visualization technique which arranges countries based on their overall centrality. The resultant graph is displayed in Figure 2.

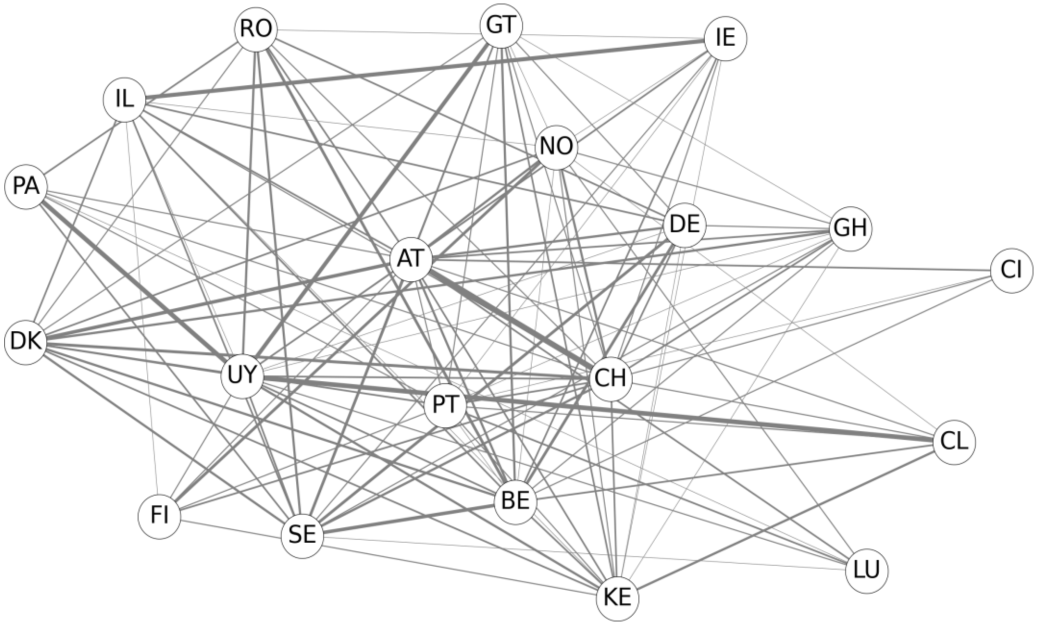

Figure 2: An Example of the Centrality Position of Austria in Force-Layout Graph

The centrality analysis of Austria, as depicted by the Force-Layout graph above, highlights its pivotal role, as measured by the similarity of first names. The degree of similarity between two countries is indicated by the thickness of the line; and the overall centrality is depicted through their positioning on the graph. Predictably, Austria, Germany, and Switzerland, where German is the prevalent language and naming traditions are commonly shared, occupy a central position with Austria in the middle. At the periphery of the graph, we find countries where Germanic languages are also spoken, such as Denmark, Belgium, and Sweden. However, African countries like Kenya and Ghana also appear in this mapping which may be a reflection of historical migration patterns. It is important to recognize that this visualization is symmetric and should not be interpreted as indicating any causal relationships.

Discussion

It is challenging to directly compare the results of this research with previous studies given the intricacies of the similarity relation and definition used in this research as well as the distinct corpora, methodology, and limitations of this study. However, when juxtaposing our results with recent lexicostatistical analyses of similarities among languages (Bella et al. 2021), which are based on the CogNet database (Batsuren et al. 2019), the correlation is statistically significant, albeit only moderately positive, with a value of 0.30. This correlation was calculated using the "high robustness" section of the database, utilizing the Spearman correlation. The p-value is 2.94e − 7. The language-to-language dataset from this study was constructed by assigning the most spoken language from each country and taking the average similarity value.

Calculating similarities among languages from a small, controlled list of carefully selected cognates—as in the work of Müller and colleagues (2010)—and comparing it to forename similarity yields similar results. When comparing the phonetic similarities of cognates using the same metrics applied in this study (Mortensen et al. 2016), the rank correlation emerges as significant yet moderately positive. The Spearman correlation factor is 0.29. It is also interesting to compare these results with those obtained from more extensive lists (Bella et al. 2021), where the Spearman correlation coefficient is 0.57.

However, additional research is needed to thoroughly analyze the relationship between this aspect of language similarity and more conventional ones to better understand the role of proper names in grasping the nuances of language similarities, language contacts, and changes. There are many aspects of this work that require further investigation. Firstly, there is a need to delve deeper into hypotheses concerning the relative size of the proper names vocabulary and to quantify it more precisely across different corpora and languages. Improving the methodology employed in this study could also be beneficial. For example, enhancing multilingual grapheme-to-phoneme (G2P) models and refining measures of phonetic similarity and distances might be advisable.

Such improvements would facilitate more accurate measurement of proper names and allow for applying the same methodology to larger, regular language corpora and comparing the results with the existing research. Moreover, it would be worthwhile to broaden the research scope to include other types of proper names, such as surnames, location names, and even organization names. The inclusion of these other name types may reveal the differential influence of linguistic and cultural factors. As this current onomastic investigation has shown, the method showcases here holds much promise for providing many new insights into language similarities and differences.

Notes

[1] Developed by linguist Morris Swadesh, the Swadesh list is a set of basic vocabulary words used to study the historical relatedness of languages. These core words are less likely to be borrowed between languages. Consequently, by comparing them, linguists can determine how closely related languages are and estimate when they diverged from a common ancestor. Methods to assess equivalence or similarity among words vary from exact string equivalence and multiple string distance measures to subjective assessment of the researcher.

[2] The statistics are calculated from the following corpora: CoNLL-2003 (Tjong Kim Sang & De Meulder 2003), WEXEA (Strobl et al. 2020), Few-NERD (Ding et al. 2021). For counts, words are tokenized and normalized by lowercasing and stemming using NLTK PorterStemmer; and all tags except “other” are counted as proper names. As the corpora above are limited to news and encyclopedia text, and the accuracy of labels is not perfect, these ratios are expected to vary in different corpora. However, it is reasonable to assume that the number of different tokens in proper names outnumbers the lemmas of regular language. Further research is needed to quantify this more precisely across different corpora and languages.

[3] For this preliminary estimation, names and surnames are deemed English if adopted by individuals residing in English-speaking countries. This doesn’t suggest they are “native” to these speakers; however, there are valid reasons to consider them as part of the vocabulary. Additional theoretical and empirical research is required for a more precise estimation of these values.

[4] The result of the process is available at *echoes.mondonomo.ai*

[5] A quick manual verification showed that minor variations in vowel types and similar consonants were well managed (see: *echoes.mondonomo.ai*). However, variations did occur. For example, differences among the rhotic sounds in English |ɹ|, Spanish |r|, and German |ʀ| seem to have generated more significant discrepancies than they should have, but more theoretical and empirical research is needed to test this informal observation.

[6] As expected, the probability distribution of name frequency was strongly skewed, so it was transformed into a natural logarithmic scale.

[7] To obtain a similarity measure on a 0-1 scale, the inverse feature edit distance is computed by subtracting the actual feature edit distance from the maximal distance and then dividing it by the maximal distance.

References

Almasoud, Ameera, Hend S. Al-Khalifa, and Abdulmalik S. Al-Salman. 2019. "Handling Big Data Scalability in Biological Domain Using Parallel and Distributed Processing: A Case of Three Biological Semantic Similarity Measures." *BioMed Research International* 2019. https://doi.org/10.1155/2019/6750296

Batsuren, Khuyagbaatar, Gábor Bella, and Fausto Giunchiglia. 2019. "Cognet: A Large-scale Cognate Database." In *ACL 2019: The 57th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, 3136—3145. Boston, Massachusetts: Association for Computational Linguistics.

Bero, S. A., A. K. Muda, Y. H. Choo, N. A. Muda, and S. F. Pratama. 2017. "Similarity Measure for Molecular Structure: A Brief Review." In *Journal of Physics: Conference Series* 892, no. 1: 012015. Bristol, UK: IOP Publishing.

Carlier, Chiara, Julian Karch, Peter Kuppens, and Eva Ceulemans. 2023. "A Comprehensive Comparison of Measures for Assessing Profile Similarity." PsyArXiv. May 10. doi:10.31234/osf.io/zbrd7

Cha, Sung-Hyuk. 2007. "Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions." *City* 1, no. 2: 1.

Ding, Ning, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. "Few-NERD: A Few-Shot Named Entity Recognition Dataset." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* 3198-3213. Boston: Association for Computational Linguistics.

Dryer, Matthew S. and Martin Haspelmath, eds. 2013. "The World Atlas of Language Structures Online." *Zenodo*. Accessed August 15, 2023. https://doi.org/10.5281/zenodo.7385533

ESPEAK. 2015. "Pronunciation Dictionary." Accessed August 15, 2023. https://espeak.sourceforge.net/index.html

Goldhahn, Dirk and Uwe Quasthoff. 2014. "Vocabulary-Based Language Similarity Using Web Corpora." In *Proceedings of the Ninth International Conference on Language Resource and Evaluation,* 26–31. Reykjavik: European Language Resources Association.

Goodman, Nelson. 1983. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Gooskens, Charlotte and Vincent J van Heuven. 2021. "Mutual Intelligibility." In *Similar Languages, Varieties, and Dialect: A Computational Perspective*, 50-95. Cambridge: Cambridge University Press.

Gooskens, Charlotte, Vincent J van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. "Mutual Intelligibility Between Closely Related Languages in Europe." *International Journal of Multilingualism* 15, no. 2: 169–193.

Jaccard, P. 1908. "Nouvelles Recherches Sur La Distribution Florale." [New Research on Floral Distribution] *Bulletin de La Société Vaudoise Des Sciences Naturelles* 44, no. 1: 223–270.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. "Billion-Scale Similarity Search with GPUs." *IEEE Transactions on Big Data* 7, no. 3: 535–547.

Katz, Leonard, and Ram Frost. 1992. "The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis." In *Advances in Psychology* 94, 67-84. North-Holland.

Kessler, Brett. 2005. "Phonetic Comparison Algorithms[1]." *Transactions of the Philological Society* 103, no. 2: 243–260.

Lauc, Davor. 2018. "How Gruesome Are the No-Free-Lunch Theorems for Machine Learning?" *Croatian Journal of Philosophy* 18, no. 54: 479–486.

Lee, Jackson L., Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. "Massively Multilingual Pronunciation Modeling with WikiPron." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4223–4228. Marseille: European Language Resources Association.

Li, Xinjian, Siddharth Dalmia, David Mortensen, Juncheng Li, Alan Black, and Florian Metze. 2020. "Towards Zero-Shot Learning for Automatic Phonemic Transcription." *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 1: 8261–8268.

Li, Xinjian, Florian Metze, David R. Mortensen, Alan W Black, and Shinji Watanabe. 2022. "Phone Inventories and Recognition for Every Language." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1061-1067. Marseille: European Language Resources Association.

Liang, Jie. 2008 "Estimation Methods for the Size of Deep Web Textural Data Source: A Survey." Accessed November 1, 2023. https://richard.myweb.cs.uwindsor.ca/cs510/survey_jie_liang.pdf

Llarena, Jose. 2017. "Britfone." *GitHub Repository*. Accessed August 15, 2023. https://github.com/JoseLlarena/Britfone

Mondonomo. 2023. "Mondonomo Knowledge Graph." Accessed August 15, 2023. https://mondonomo.ai

Mortensen, David R., Siddharth Dalmia, and Patrick Littell. 2018. "Epitran: Precision G2P for Many Languages." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 23–31. Paris: European Language Resources Association.

Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics,* 3475–3484. Boston: Association for Computational Linguistics.

Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H Brown, Pamela Brown, Sebastian Sauppe, Eric W Holman, et al. 2010. *Asjp World Language Tree of Lexical Similarity: Version 3*. Accessed August 15, 2023. https://asjp.clld.org/download

Open-Dictionary-Data. 2023. "IPA-Dict: Monolingual Wordlists with Pronunciation Information in IPA." *GitHub Repository*.Accessed August 15, 2023. https://github.com/open-dict-data/ipa-dict

Park, Jongseok, Kyubyong & Kim. 2019. "g2p: English Grapheme to Phoneme Conversion." *GitHub Repository*. Accessed August 15, 2023. https://github.com/Kyubyong/g2p

Phatthiyaphaibun, Wannaphong. 2020. "Thai-g2p." *GitHub Repository*. Accessed August 15, 2023. https://github.com/sigmorphon/2020/tree/master/task1/

Schüppert, Anja. 2011. *Origin of Asymmetry: Mutual Intelligibility of Spoken Danish and Swedish*. Accessed August 15, 2023. https://research.rug.nl/en/publications/origin-of-asymmetry-mutual-intelligibility-of-spoken-danish-and-s

Strobl, Michael, Amine Trabelsi, and Osmar Zaiane. 2020. "WEXEA: Wikipedia EXhaustive Entity Annotation." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1951-1958. Marseille: European Language Resources Association.

Sun, Hao, Xu Tan, Jun-Wei Gan, Hongzhi Liu, Sheng Zhao, Tao Qin, and Tie-Yan Liu. 2019. "Token-Level Ensemble Distillation for Grapheme-to-Phoneme Conversion." *arXiv preprint arXiv:1904.03446*. Accessed August 15, 2023. https://arxiv.org/abs/1904.03446

Swadesh, Morris. 1955. "Towards Greater Accuracy in Lexicostatistic Dating." *International Journal of American Linguistics* 21, no. 2: 121–137.

Taubert, Stefan. 2022. "Pronunciation Dictionary." *Zenodo*. Accessed August 15, 2023. https://doi.org/10.5281/zenodo.7386813

Tjong Kim Sang, Erik F., and Fien De Meulder. 2003. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. Boston: Association for Computational Linguistics.

Tversky, Amos. 1977. "Features of Similarity." *Psychological Review* 84, no. 4: 327–352.

VIAF. 2023. *Virtual International Authority File (VIAF)*. Accessed August 15, 2023. https://viaf.org/

Vijaymeena, M. K. and K. Kavitha. 2016. "A Survey on Similarity Measures in Text Mining." *Machine Learning and Applications: An International Journal*, 3, no 2: 19–28.

Voorhees, Ellen M. 1986. "Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval." *Information Processing & Management* 22, no. 6: 465–476.

Wang, Wen-June. 1997. "New Similarity Measures on Fuzzy Sets and on Elements." *Fuzzy Sets and Systems* 85, no. 3: 305–309.

Watanabe, Satoshi. 1969. "Modified Concepts of Logic, Probability, and Information Based on Generalized Continuous Characteristic Function." *Information and Control* 15, no. 1: 1–21.

Watanabe, Satoshi. 1986. "Epistemological Relativity Logico-Linguistic Source of Relativity." *Annals of the Japan Association for Philosophy of Science* 7, no. 1: 1–14.

Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. "Evaluating Linguistic Distance Measures." *Physica A: Statistical Mechanics and Its Applications* 389, no. 17: 3632–3639.

Wikidata. 2023. *Wikidata*. 2023. Accessed August 15, 2023. https://www.wikidata.org/dumps/

Wolpert, David H. and William G Macready. 1997. “No Free Lunch Theorems for Optimization.” *IEEE Transactions on Evolutionary Computation* 1, no. 1: 67–82.

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. “Byt5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models.” *Transactions of the Association for Computational Linguistics* 10, no 1: 291–306.

Zhu, Jian, Cong Zhang, and David Jurgens. 2022 “ByT5 Model for Massively Multilingual Grapheme-to-Phoneme Conversion.” *arXiv preprint arXiv:2204*.03067. Accessed August 15, 2023. https://arxiv.org/abs/2204.03067

Notes on the Contributor:

Davor Lauc is a professor of logic and AI at the University of Zagreb, Croatia. His interest in proper names, particularly human names, stems both from logical and philosophical issues related to the meaning of names, as well as from his role as a data scientist in the projects on Croatian names (actacroatica.com) and global names (mondonomo.ai). He has published a book on Croatian surnames and several articles related to onomastics. His specialization involves using logical data science and artificial intelligence methods in the humanities.

Correspondence to: Prof. Dr. Sc. Davor Lauc, Faculty of Social Sciences and Humanities, University of Zagreb, I. Lucica 3, Zagreb, Croatia. Email: dlauc@ffzg.unizg.hr