Phoneme? I hardly know 'em!
Complexity in phonological structure of Romance and Germanic names

Alan Ispani
IT University of Copenhagen
alai@itu.dk

Abstract

This paper investigates the complexity of names in the two largest European language families: Germanic and Romance. It analyzes similarities and differences in name distributions and the proportion of names shared between countries. As another metric, it uses phonetic representations of countries' names and calculates the most common sequences to create a distribution of the phonemic system for each language. Finally, it calculates the phonetic distributional similarity of each language to the distribution of its language family. The analysis does not show one language family to be more complex than the other and contradicts findings of a previous study using letter distributions.

1 Introduction

Languages in a language family develop and branch from a common ancestral language. But to which extent does a language family influence its countries' names? Language plays a significant role in shaping a country's identity, analogous to how a name shapes an individual's identity. The objective of this study is to examine the usage of names in the Germanic and Romance language families and identify potential patterns in name complexity. Specifically, the research question is "To what extent do language families influence the names of people in a country? Which language family produces more complex names, Germanic or Romance?". To gain insights into these questions, the study employs statistical analysis: by comparing the distribution, frequency, and structure of names within these language families, the aim is to better understand their characteristics and identify any potential differences between them.

Statistical analysis has been applied in the research of a variety of phenomena in linguistics. Language complexity has especially received a lot of attention (Vulanović, 2007; Carpena et al., 2009). Previous work (Matushansky, 2008) has looked into syntactic complexity of proper names. Sommerlund et al. (2023, Structural complexity of Romance and Germanic names, IT University of Copenhagen) have taken a statistical approach and looked at the structures of letters within names across countries and language families. This study chooses to investigate phonological structures to determine if the name configurations of a language family are more predictable, and thereby less complex.

2 Data

This project uses the names dataset (Remy, 2021), a dataset of names of Facebook users grouped by country which was extracted from a Facebook data leak containing 533 million users. It includes 730.000 unique first names and 983.000 unique last names across 105 countries. The full dataset is available here[1].

3 Methods

3.1 Preprocessing & weighing

To simplify the analysis, the text is normalised to contain only lowercase Latin characters, and a delimiter symbol "-" as a substitute for space. This delimiter symbol is required, as tokenizing libraries will split by spaces. It is important to preserve the meaning of spaces, since first names can include more than one "word". To reduce errors, where e.g. users append a middle- or last name or made a typo, at least 3 occurrences of a name are required to include it. Any character not in a-z or a dash is removed.

Occurrences of each name for each country are counted and converted to a probability distribution. In the analysis, frequent names get assigned a higher weight when evaluating the complexity of a language for two reasons:

[1] https://pypi.org/project/names-dataset/

1. Few but popular names in themselves indicate that a country/language is less diverse, and thereby simpler. Weighing includes this dimension.

2. Weighing reduces the influence of very long-tailed distributions, which might include "names" that are actually nicknames or names which include middle- or last names. This also limits how much diversity one can capture, which subsection 3.4 quantifies.

Language adoption has less to do with its complexity and more to do with geography. Since it is important to represent the language family and not just the most populous countries in that family, countries are weighed equally by comparing the frequency percentage within each country.

3.2 Name selection (tf-idf)

The number of names included from each country are limited to reduce computation time. It is also important to select names specific to that country, since they are more likely to originate from that country's language compared to e.g. cross-cultural names or names introduced by immigration. Therefore tf-idf (Martin and Jurafsky, 2009, Chapter 6.5) is applied, which assigns a score to each name weighted by frequency, but inversely weighted by frequency in other countries. Due to the high computational cost, only the ten thousand most common names in each country are included. Then, for each country, the 1000 highest scored names are chosen, which represent names that are the most frequent in that country, while being infrequent in other countries. Tf-idf works as a form of dimensionality reduction, keeping the most important names for each country.

3.3 Language families

There are 7151 living human languages distributed across 142 language families (Eberhard, 2022). To focus the analysis and avoid too many simplifying assumptions, this paper focuses on the Germanic and Romance language families. The research takes the four most populous European countries with one primary language for each language family, which can be seen in Table 1. The analysis is limited to European countries to reduce influence of colonialism such as residents adopting names of their colonizers vs. keeping their native names. Two countries are excluded from the definition:

Language family	Country	Main language	Population (M)
Germanic	Germany	German	83.13
	Great Britain*	English	65.12
	Netherlands	Dutch	17.53
	Sweden	Swedish	10.42
Total			176.2
Romance	France	French	67.5
	Italy	Italian	59.07
	Spain	Spanish	47.33
	Portugal	Portuguese	10.3
Total			184.2

Table 1: Chosen countries for the analysis.
While Great Britain is not a country, our dataset had Great Britain names instead of United Kingdom names, so Northern Ireland is excluded.

Romania, which does not exist in the dataset; and Belgium, which contains a mix of languages from the two families. All countries are weighed equally, as mentioned previously.

3.4 Name distributions

As a simple first metric, the study looks at the distribution of names for each country and language family. The analysis checks how many unique names are required to account for 50% of the population of a country by using the proportion of the names in the dataset and adding up the largest until the 50% threshold is reached. This gives a basic comparison of how frequent the most common names are for each language family. This can also be used as a naive approach to measure the complexity of the names of a country, as the variety of unique names is a basic indicator of diversity and therefore complexity.

3.5 Shared name frequency

Another way of looking at similarities across countries is to look at the commonality of their names, that is, what percentage of people between countries share a name. For this, a pairwise calculation of the shared name frequency between each country in the same family is performed. After taking the mean within families, this metric can be compared between them.

For a pair of languages with name frequency distributions P and Q, and for the set of all names X, the shared name frequency is:

$$shared(P, Q) = \sum_{x \in X} min(P(x), Q(x)))$$

It is assumed that languages of the same language family share some structure in their names even when they are not spelled exactly the same. The shared name frequency approach, then, misses small differences in naming conventions such as the Spanish 'Alejandro'/'Máximo' vs. the Italian 'Alessandro'/'Massimo'. This was investigated by looking at structural similarities within names (Sommerlund et al., 2023, Structural complexity of Romance and Germanic names, IT University of Copenhagen), and this work will take a different approach to this explained in subsection 3.6.

3.6 Kullback–Leibler divergence of phonetic distributions

Previous work (Sommerlund et al., 2023, Structural complexity of Romance and Germanic names, IT University of Copenhagen) used Byte Pair Encodings (Sennrich et al., 2015) and KL divergence (Bishop and Nasrabadi, 2006, Chapter 1.6.1) to build an n-gram distribution for each language and determine the distance between each family of names and specific countries' names. However, it only looks at the structures of letters, which are not necessarily faithful representations of the pronunciations of names.

This project builds on their work by using epitran[2] (Mortensen et al., 2018), a library for transliterating orthographic text as IPA (International Phonetic Alphabet), before creating the Byte Pair Encodings. By first converting names to their phonetic transcription, the analysis can for example equate n-grams such as the Italian 'gi' in 'Giulia' and the English 'j' in 'Julia' to the speech sound [$\widehat{\text{dʒ}}$].

4 Results

4.1 Name distributions

The initial analysis calculates how many names countries need to account for 50% of the population. Table 2 shows that the Romance family has more frequent names when compared to the Germanic family.

[2]https://github.com/dmort27/epitran

Family	min	max	μ	σ
Romance	50	204	116	65
German	143	298	193	71

Table 2: Number of names to account for 50% of the population in a country, grouped by language family.

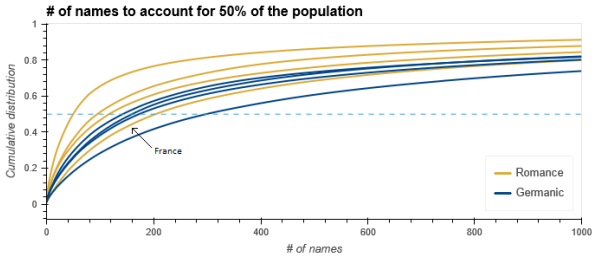

Figure 1: French as an outlier, appears closer to the Germanic languages

Looking at Figure 1 reveals the only exception, France, which requires 204 names to account for 50% of the population. For comparison, the next Romance country is Spain, which needs almost half as many names at 117. The cumulative distribution functions are discrete and long-tailed for all countries.

4.2 Shared name frequency

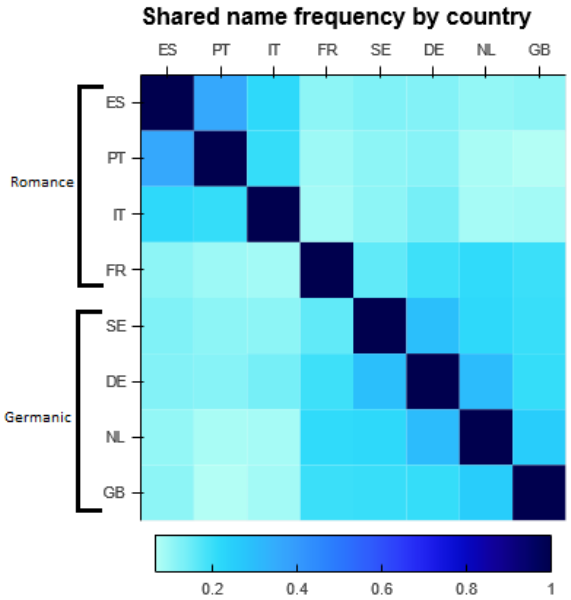

Figure 2: French names appears to be closer to the Germanic language family names

There is a clear pattern of shared name frequency for both families in Figure 2. However, France is again an exception and shares an average of 19.2%

Family	Mean shared frequency (%)
Romance	17.9
Germanic	24.7

Table 3: Mean shared frequency for each family

of names with members of the Germanic language family, but only 9.7% with members of the Romance family.

The difference in shared frequency in Table 3 can be explained by France. When France is removed, the mean shared frequency for Romance jumps to 26%.

4.3 KL-divergence

Family	Country	KL	Mean
Germanic	Great Britain	1.17	0.79
	Germany	0.76	
	Sweden	0.64	
	Netherlands	0.59	
Romance	Portugal	0.95	0.75
	France	0.76	
	Spain	0.71	
	Italy	0.57	

Table 4: Lower is better/more similar

When analyzing phonetic transcriptions instead of written names, France ceases to be an outlier and has an average contribution to KL divergence. The results in Table 4 show that the Germanic and Romance names are about as similar, with United Kingdom names contributing the most dissimilarity. They also show a marked difference from previous work (Sommerlund et al., 2023, Structural complexity of Romance and Germanic names, IT University of Copenhagen) which focused on letters instead of phonemes. As opposed to their analysis, France's phonetic KL contribution is average for the Romance family, since the rare suffixes 'ne' and 'ie' get reduced to the common 'n' and 'i'.

Previous work also states that the Romance family vocabulary matches the Germanic countries, but not the other way around. The phonetic analysis in this paper shows no statistically significant difference (1.53 KL divergence from Romance to Germanic vs 1.56 vice versa).

5 Limitations and Future Work

Due to the Unicode implementation of the IPA, speech sounds such as [d͡ʒ] and [t͡ʃ] are considered 3 individual characters while representing only one sound. This makes it so that the BPE vocabulary needs to use 3 slots for one speech sound (the strings 'd', 'ʤ' and 'd͡ʒ' will all be saved in the BPE vocabulary, and any other n-grams containing the sound will suffer from this as well). This may punish some languages more than others, and may be why Great Britain has a higher KL divergence, since names which contain [d͡ʒ], such as John and James, account for 14.5% of all British names. In contrast, this speech sound is not present in any of the other Germanic names. Future work could try to find a better representation for these sounds that does not suffer from this issue.

Facebook as a data source is more unreliable than e.g. national name registries, since people are free to choose any string as their name. This was partially mitigated by the use of tf-idf, but there still may be some artifacts of the collection medium. Future work could perform the same analysis on governmental name statistics to reduce data irregularities.

6 Conclusion

This paper looked at name distributions and showed that the Romance language family has overall less diverse names. One only needs ~100 names to cover half of the Romance population, whereas one would need almost twice that to cover the Germanic population. When strictly looking at whole names, France appears as an outlier in both name distributions and shared name frequency.

On the other hand, a more thorough measure, namely KL-divergence of the n-gram distributions of phonemes, shows that Germanic and Romance countries are near to equally complex in the structure of their names. This further analysis does not reveal France as an outlier and shows no conclusive results for any one language family being more complex than the other.

References

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Pedro Carpena, Pedro Bernaola-Galvan, Michael Hackenberg, Ana Victoria Coronado Jiménez, and Jose

Oliver. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79:035102.

David M. Eberhard. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

James H Martin and Daniel Jurafsky. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

Ora Matushansky. 2008. On the linguistic complexity of proper names. *Linguistics and philosophy*, 31(5):573–627.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Philippe Remy. 2021. Name dataset. https://github.com/philipperemy/name-dataset.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Christoffer Sommerlund, Alan Ispani, and Johan Laursen. 2023. Structural complexity of romance and germanic names. *Unpublished Manuscript*.

Relja Vulanović. 2007. On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics*, 20:399–427.