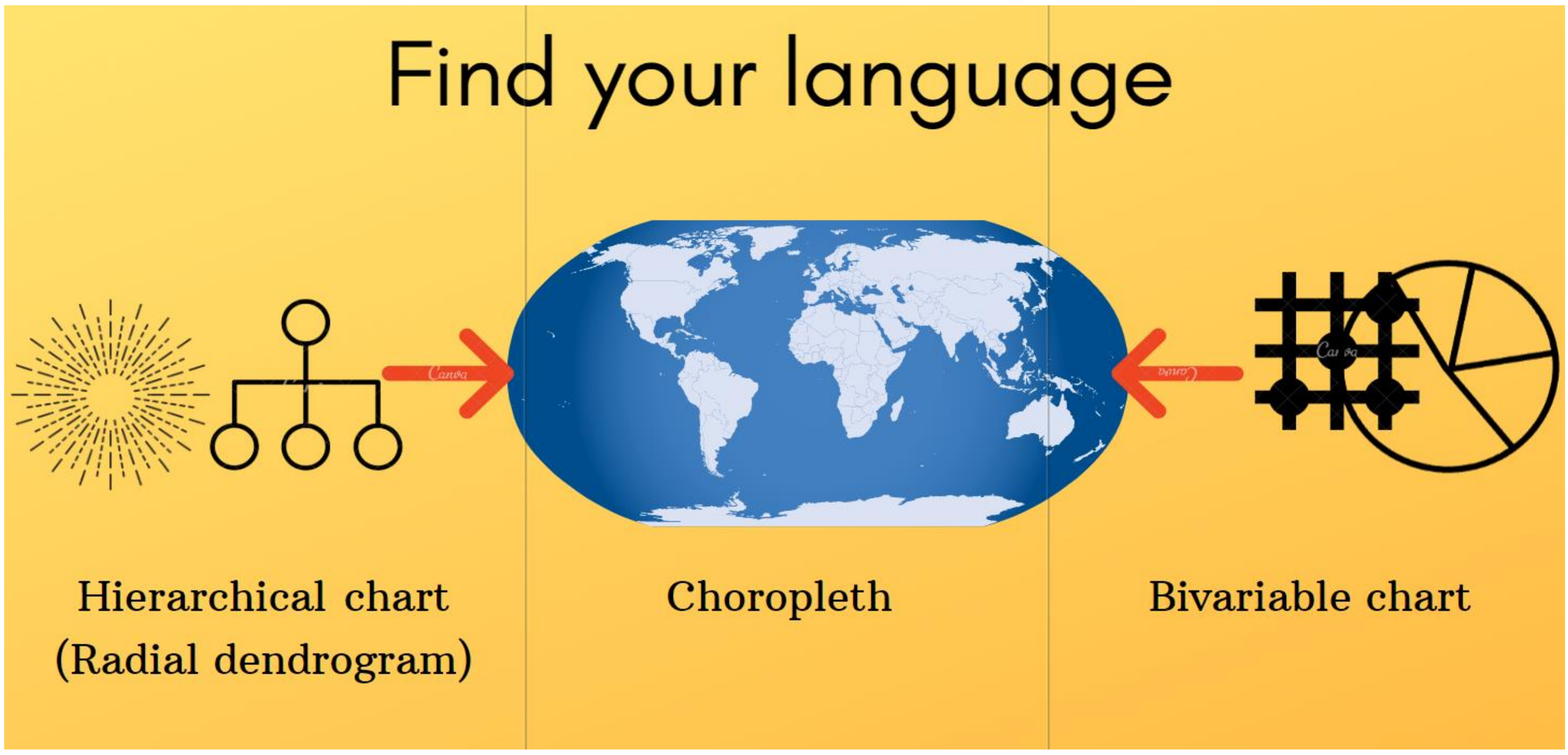
Find your Language: Geographic mapping of languages

Vasudha Varadarajan



Introduction

There are around 7200 languages spoken in the world today, but only 25 of them are spoken by more than 50% of the world's population. It is a sad fact in today's world that a number of languages have less than 1000 speakers in the world, which means that it is likely to be extinct very soon, vanishing unique sounds, expressions and ideas unique to the language. It becomes difficult to coalesce all the information for easy visualization. Linguistic information could be used for anthropological and historical explorations, but more importantly, recognizing lesser known languages across the world will help dissipate knowledge among those who do not speak any of the major languages that carry useful knowledge that could lead to significant technological progress in these less accessible communities. Furthermore, recognizing closely related languages and clustering them could help in identifying languages that could be most easily picked by a certain speaker- and could reduce the effort needed to translate information to all languages. Visualizing the language's geographical spread, its origins and other information would help simplify the problem of understanding how to place the language in the context of all of world's languages and recognizing the state of endangerment. Here we analyze the geographical aspect of data from Glottolog. This could also be a useful pedagogical tool to introduce different language families or to delve into analyzing relationships between languages

Hypothesis

More number of children languages there are, more likely it is for the language family to be dominant language. This means that more and more people switch from having one language over the generations. As this happens, the process blends sounds and eventually leads to new languages of same family.

Results

Contrary to the expected result, I find that the most languages actually come from Atlantic-Congo family. It is surprising considering that a lot of popular languages that has dominated the world over the last 2 centuries come from Indo-European family.

Discussion

Even as a lot of languages from the Indo-European family or the Sino-Tibetan family (English/ Spanish/ Mandarin/ Japanese) seem to become increasingly popular as a native speak, they do not take the lead. But the hypothesis should not be completely dismissed as the most spoken languages at least fall into the top 10 among the ~450 language families currently discovered so far. Close geographical association is also not necessarily conducive to same-family classification- the New Guinea region has the highest concentration of language families.

Other than that, the language finding tool could also help in understanding the hierarchies, in a more visual fashion. The data size is large and to visualize efficiently we would need to perform lossless data reduction.

Conclusions

We can conclude that the visual tool could be helpful in understanding hierarchical and geographic relationships between the dataThe tool provides a way to get an overview of a certain language family and understand its origins.