Visualising Typological Relationships: Plotting WALS with Heat Maps

Abstract

This paper presents a novel way of visualising relationships between languages.

1 Introduction

.[Filler][Filler]

Typology has been used to derive implications about possible languages, and about the ordering of the human mind. Greenberg (1963) is famous for finding many cross-linguistic typological implications, such as , and claiming that they represented universals of grammar, and that they would hold true for all languages. In a similar vein, Chomsky (2000) has argued as a generativist that typological universals show that language is clearly an artefact of constraints involved with the learning process, and that some typological features are impossible due to this. Dunn *et al* (2011) argued that a languages typology relies upon the previous generations' language more than biological, environmental or cognitive constraints, and that there are pathways which are generally followed in language change based on the previous parent language.

Outside of the scope of theory, language typology is useful for other reasons.

The World Atlas of Language Structures (WALS) is a large database of details of structural properties of several thousand languages (Dryer and Haspelmath, 2011). The properties were collected from descriptive sources by the project's 55 authors. As of 2008, WALS is browsable online (`http://www.wals.info`).

However, for the several thousand languages in WALS, there is only about 16% of the possible features filled—the data are *sparse*.

Dealing with sparse data is a computational problem, as any statistical information drawn from the database will, to a large extent, be an artefact of the database. For instance, if half of the languages were marked as having uvular stops (unlikely), and then in reality if all other languages not in the database had uvular stops, then the knowledge we would glean from the database would be significantly false and misleading. Given that there are around 6,000 languages in the world, the amount of languages on WALS means that this is a serious concern. Many researchers in recent years have been developing work-arounds for sparse databases; often because languages with low resources have a similar problem.

A solution to dealing with this issue is to visualise the data in WALS, instead of relying on raw numbers.

In this paper, we will first discuss more fully our starting data, before going on to discuss the problems with analysing this data and how our methodology dealt with them. We will then present several graphs that highlight the possibilities of graphing WALS data.

2 Aspects of the Visualisations

2.1 Linguistic features

2.2 Geographical distance

2.3 Phylogenetic distance

3 Material and Methods

3.1 WALS

At the time of submission, there were 81,828 datapoints for 2,678 languages (an average of 28 per language). At least one feature, the most populated, had data for 1,519 languages. There were 144 different chapters, each containing values for different, related features. It can be seen easily from these numbers that the data on WALS is *sparse*. Ignoring the fact that a language having certain features will cancel out the possibility or probability of others, that means that the WALS possible data is only 15.8% represented in the database.

Drawing a heat map where only 16% of the features are available would be of little use. There are two options for dealing with this: to collapse the feature values in some way, or to select for languages that have a higher percentage of data filled than the average language. We opted for the second choice, and *cleaned* the file until it contained only languages that had either 30% as a lower bound, or 50% as a higher bound, of all of their entries filled. This cleaned data was then used in the other functions.

Languages are related phylogenetically either vertically, by lineage, or horizontally, by contact. On WALS, each language is placed in a tree hierarchy that specifies phylogenetic relations. In the data files, this is specified by linking to three groupings: genus, such as 'Northern Naga', family, such as 'Sino-Tibetan', and sub-family, such as 'Tibeto-Burman'. Provisional tree hierarchies were also drawn from Ethnologue (Lewis, 2009) and MultiTree (for Language Information and List), 2009), but due to the low amount of overlap between non-sparse WALS entries and the possible alternative hierarchies drawn from Ethnologue or MultiTree, only the phylogenetic relations from WALS were used.

The WALS phylogenetic hierarchies do not take into account language contact. For that, we used geographic coordinates, which are present on WALS, as a proxy for contact.

We measured geographic distance by choosing an arbitrary radius that would would create a decision boundary to cluster n-nearest neighbours. For our purposes, we went with 500km, as this proved to provide sufficient examples to draw from from cleaned WALS data. We also measured distance by selecting an arbitrary lower threshold for languages in the general area, and printed the results if the amount of cleaned languages in the area was a certain percentage over the amount of languages in the area total as specified by WALS. This number is clearly underrepresentative of the amount of contact languages, as only half of the world's languages are present in WALS. This proxy was not as good at choosing specific, useful examples, as the n-nearest nieghbours, as the languages chosen were often too far away.

Each final list was then resorted, so that the source language was centered in the map. This was due to one of the primary issues with using distance on a two dimensional graph. On the one hand, we would want close languages to be close together on the heat map. However, given the source language Egyptian Arabic, this would mean that Zulu, Saami, Mohawk, and Japanese might all roughly be placed next to each other on the map. This also means that Japanese might be placed next to Mohawk, and Mohawk again next to Korean. This is not ideal, and was the main justification for limiting the sphere of possible geographical languages to a reasonable distance, given the data.

All of the code was done either in Python, or in R. The code was not computationally intensive, and did not take more than a couple of minutes to run.

4 Results

4.1 Heat maps

- Phylogenetic distance heatmaps
- Geographical distance heatmaps
- Combined maps

5 Discussion

- What these tell us (map by map)
- What these tell us, overall - implications
- Warnings: sparse data, data not there, etc.
- Future work

References

N. Chomsky. 2000. *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge, UK.

Matthew S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011 edition.

Institute for Language Information and Technology (LINGUIST List), editors. 2009. *Multitree: A digital library of language relationships*. Eastern Michigan University, Ypsilanti, MI, 2009 edition.

J.H. Greenberg. 1963. ‘some universals of grammar with particular reference to the order of meaningful elements. In J.H. Greenberg, editor, *Universals of Language*, pages 58–90. MIT Press, Cambridge, MA.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, sixteenth edition.