Visualising Typological Relationships: Plotting WALS with Heat Maps

Abstract

This paper presents a novel way of visualising relationships between languages. The key innovation of the visualisation is that it brings geographic, phylogenetic, and linguistic data together into a single image, allowing a new visual perspective on linguistic typology. The data presented here is extracted from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2011). After pruning due to low coverage of WALS, we filter data about linguistic structures and attributes by geographical proximity in order to get at areal typological effects. The data are displayed in heat maps which reflect the strength of similarity between languages for different linguistic aspects. Finally, these are annotated for language family membership. The images so produced allow a new perspective on the data which we hope may facilitate interesting findings and perhaps even illuminate new areas of research.

1 Introduction

This paper presents a novel way of visualising relationships between languages. Relationships between languages can be understood with respect to linguistic features of the languages, their geographical proximity, and their status with respect to historical development. The visualisations presented in this paper are the first to bring together these three perspectives into a single image. One line of recent work brings computational methods to bear on the formation and use of large typological databases, often using sophisticated statistical techniques to discover relations between languages (**?**; **?**; **?**, among others), and another line of work uses typological data in natural language processing (**?**; **?**, for example)but we are unaware of any previous approaches to visually presenting the resulting data. Here, we address this gap by using data from the World Atlas of Language Structures (Dryer and Haspelmath, 2011) to develop heat maps that can visually show the interconnected relationships between languages and language families.

The main envisioned application of our visualisations is in the area of linguistic typology. Typology has been used to derive implications about possible languages, and about the ordering of the human mind. Different theorists have taken different views on the relationship between typology and the universality of languages. For example, Greenberg (1963), a foundational work in the field, identifies a number of cross-linguistic typological properties and implications and aims to present them as truly universal – relevant for *all* languages. In a similar vein, typological universals have been employed as evidence in a generative story regarding language learning (Chomsky, 2000). Dunn *et al* (2011) argued that a language's 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. What these arguments have in common is a reliance on a view of linguistic typology that is potentially restricted in its scope, due to insufficient access to broad-scale empirical data, covering many features of many languages of the world.

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

The visualisations described here bring together three types of information: linguistic features, geographical distance, and phylogenetic distance. For the current study, all three types of information are extracted from the WALS database. In future work, we may explore the use of alternate sources such as Ethnologue (Lewis, 2009) or MultiTree (for Language Information and List), 2009) for alternate phylogenetic hierarchies.

2.1 Linguistic features

2.2 Geographic distance

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 under-representative 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.

2.3 Phylogenetic distance

Languages are related phylogenetically either vertically, by lineage, or horizontally, by contact. In WALS, each language is placed in a tree hierarchy that specifies phylogenetic relations. In the WALS data files, this is specified by linking at three different levels: family, such as 'Sino-Tibetan', sub-family, such as 'Tibeto-Burman', and genus, such as 'Northern Naga', .

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.

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.

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

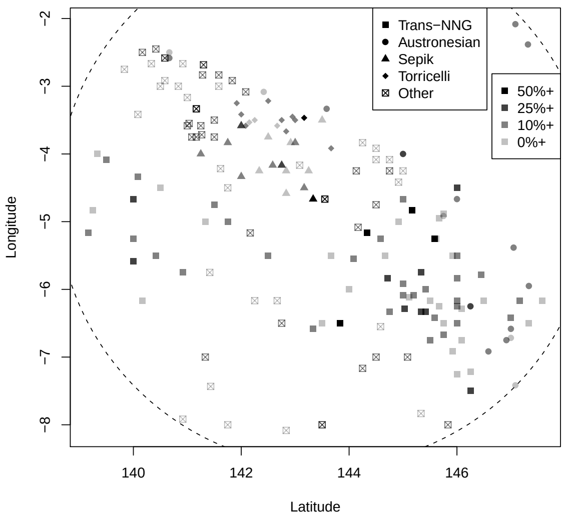

Figure 1: Sparse Data

In Fig. 2, we have plotted out the neighbouring languages for the most populous language area represented on WALS by longitude and latitude. The center language, Yimas, is a Trans New Guinean language, and is the center for the graph. The odd clustering is due to the Pacific Ocean - such geographical features limit the usefulness of distance as a measure. Another limit is the range of each language, which is not represented here, as each language is given only a single geographical coordinate that does not indicate the area over which the language resides, nor how many other languages coexist with it.

For each language within 500km, represented here by the dotted circle, the amount of features filled in WALS is shown. The majority have less than 50% filled - there are only 171 such languages in WALS. This graph does not include languages not in WALS. It ought to be clear from this that the data in WALS is too sparse to draw immediate conclusions from, which furthers the need for visualizations.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris massa elit, facilisis quis elementum quis, tincidunt et massa. Suspendisse et lacus augue, sit amet molestie est. Aliquam luctus vestibulum tellus, nec consectetur elit porttitor ut. Aenean tellus velit, placerat ac placerat nec, ultricies quis nunc. Vivamus eget neque id diam posuere tincidunt ac vel nibh. Morbi vitae eros turpis. Vestibulum vitae nulla diam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum elit odio, luctus eget elementum a, interdum id sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Sed scelerisque est eget lorem eleifend vestibulum. Ut dui dolor, tincidunt ac fringilla vitae, mollis tristique ligula. Quisque urna tortor, venenatis eu bibendum id, condimentum vel neque. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at ligula nunc, et tincidunt mi. Donec pellentesque mauris rutrum odio vulputate a egestas neque mattis. Curabitur molestie adipiscing leo, sed posuere elit luctus id. Ut arcu diam, scelerisque at hendrerit eu, venenatis eget urna. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec feugiat, tellus quis viverra porttitor, dolor dolor vehicula tortor, ac porta nisi elit non ante. Proin ligula odio, condimentum nec pellentesque a, cursus ac nulla. Cras aliquet magna ut dolor suscipit congue.

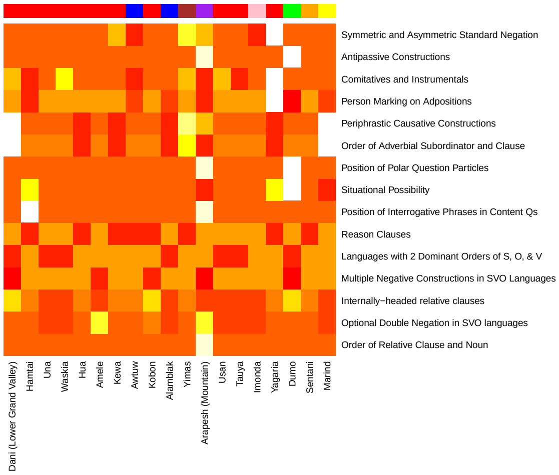

Figure 2: Example 1

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.