Visualising Typological Relations: Using Heat Maps with WALS

Richard Littauer
University of Saarland
Computational Linguistics Department
Saarbrücken, Germany
richard.littauer@gmail.com

Rory Turnbull
Ohio State University
Department of Linguistics
Columbus, Ohio
turnbull@ling.osu.edu

Alexis Palmer
University of Saarland
Computational Linguistics Department
Saarbrücken, Germany
apalmer@coli.uni-sb.de

Abstract

This paper presents a novel way of visualising linguistic typological data. Computational methods have only recently been applied in the formation and use of large typological databases. Many studies have since focused on discovering relations between languages using typology, often using sophisticated statistical techniques. However, few papers have provided new or newly applied ways of visually presenting the resulting data. Here, we show that one can use the 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. We hope that the images will bring a new perspective to the data, resulting in interesting findings and illuminating areas of research.

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, enironmental 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.

Towards this end, the World Atlas of Language Structures (WALS) was formed. WALS is "a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors (many of them the leading authorities on the subject)."(http://www.wals.info) At the time of submission, there were 81828 datapoints for 2678 languages (an average of 28 per language). At least one feature, the most populated, had data for 1519 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.

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 statistics.

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 Material and Methods

2.1 Linguistic material

- Information about WALS data
  - Description of typological data
  - Longitude and latitude, and how it is measured
  - language families
  - WALS-language Sparseness
- Information about Multitree(for Language Information and List), 2009)
- Information about Ethnologue(Lewis, 2009) scrape

2.2 Methods

- Cleaning WALS data
- Collapsing WALS data
- Scraping Ethnologue, Multitree, formati
- Measuring Phylogenetic distance
- Measuring Geographical distance 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.
- Combined map
- Compiling python scripts, converting into R
- Sorting R output

3 Results

3.1 Heat maps

- Phylogenetic distance heatmaps
- Geographical distance heatmaps
- Combined maps

4 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.