Visualising Typological Relationships: Plotting WALS with Heat Maps

Richard Littauer
University of Saarland
Computational Linguistics
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
Saarbrücken, Germany
apalmer@coli.uni-sb.de

Abstract

This paper presents a novel way of visualising relationships between languages. The key feature 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 the typological data by geographical proximity in order to ascertain areal typological effects. The data are displayed in heat maps which reflect the strength of similarity between languages for different linguistic features. Finally, the heat maps are annotated for language family membership. The images so produced allow a multi-faceted perspective on the data which we hope will facilitate the interpretation of results and perhaps illuminate new areas of research in linguistic typology.

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 part of a new attempt 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 (Cysouw, 2011; Daumé III and Campbell, 2007; Daumé III, 2009, among others), and another line of work uses typological data in natural language processing (Georgi et al., 2010; Lewis and Xia, 2008, for example). The task of visually presenting the resulting data in this way has been only infrequently addressed. We are aware of some similar work (Mayer et al., 2010; Rohrdantz et al., 2010) in visualising differences in linguistic typology, phylogeny (Multitree, 2009), and geographical variation (Wieling et al., 2011). Here, we present our method for addressing the visualisation gap, bringing together phylogeny, typology, and geography 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, identified a number of cross-linguistic typological properties and implications and aimed 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).

Taking a different perspective, Dunn et al. (2011) argued that a language's typology relies upon the previous generations' language more than on any biological, environmental or cognitive constraints, and that there are pathways which

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 30–34,
Avignon, France, April 23 - 24 2012. ©2012 Association for Computational Linguistics

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.

The most comprehensive computational resource for linguistic typology currently available is the World Atlas of Language Structures (WALS).[1] 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.

However, of the 2,678 languages and 192 features in WALS, only 16% of the possible data points are actually specified—the data are *sparse*, and the sparsity of the data naturally makes it difficult to perform reliable statistical analysis. One way to work around this limitation is to seek meaningful visualisations of the data in WALS, instead of simply relying on raw numbers. This is our approach.

In this paper, we first discuss in more detail the source data and the types of information extracted, followed by a discussion of some difficulties presented by the available data and our approaches for addressing those difficulties. Finally, we present a sample of the resulting visualisations.

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 would explore alternate sources such as Ethnologue (Lewis, 2009) or MultiTree (2009) for alternate phylogenetic hierarchies.

2.1 Linguistic features

At the time of writing, WALS contains information for 2,678 languages. The linguistic features covered in WALS range from phonetic and phonological features, over some lexical and morphological features, to syntactic structures, word order tendencies, and other structural phenomena. A total of 192 features are represented, grouped in 144 different chapters, with each chapter addressing a set of related features. Ignoring the fact that a language having certain features will cancel out the possibility (or diminish the probability) of others, only 15.8% of WALS is described fully. In other words, if we consider WALS to be a 2,678x192 grid, fewer than 16% of the grid's squares contain feature values.

The coverage of features/chapters varies dramatically across languages, with an average of 28 feature values per language. The most populated feature has data for 1,519 languages. Because of the extreme sparsity of the data, we restricted our treatment to only languages with values for 30% or more of the available features—372 languages, with a total of 36k feature values.

2.2 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 in WALS, as a proxy for contact.

2.3 Geographic distance

Geographic distance is an important aspect of typological study because neighbouring languages often come to share linguistic features, even in the absence of genetic relationship between the languages. Each language in WALS is associated with a geographical coordinate representing a central point for the main population of speakers of that language. We use these data to determine geographic distance between any two languages, using the haversine formula for orthodromic distance.[2] A crucial aspect of our visualisations is that we produce them only for sets of languages within a reasonable geographic proximity

[1]As of 2008, WALS is browsable online (`http://www.wals.info`).

[2]This measure is inexact, especially over long distances, due to the imperfect topography and non-spherical shape of the earth, but it is computationally simple and is accurate enough for our present purposes.

and with sufficient feature coverage in WALS.

For this study, we used two approaches to clustering languages according to geographic distance. First, we chose an arbitrary radius in order to create a decision boundary for clustering neighbouring languages. For each language, that language's location is fixed as the centroid of the cluster and every language within the given radius is examined. We found that a radius of 500 kilometres provides a sufficient number of examples even after cleaning low-coverage languages from the WALS data.

The second approach selected an arbitrary lower bound for the number of languages in the geographic area under consideration. If a sufficient percentage (enough to graph) of the total number of languages in the area remained after cleaning the WALS data, we took this as a useful area and did mapping for that area. This number is clearly under-representative of the amount of contact languages, as only half of the world's languages are present in WALS with any degree of coverage. This proxy was not as good as the radius method at choosing specific, useful examples for the n-nearest neighbours, as the languages chosen were often quite distant from one another.

3 Heat Map Visualisations

We focused on producing visualisations only for features that are salient for the maximal number of selected languages. We choose two heat maps for display here, from the least sparse data available, to demonstrate the output of the visualisation method. The remaining visualisations, along with all code used to produce the visualisations, are available in a public repository.[3]

All data was downloaded freely from WALS, all coding was done in either Python or R. The code was not computationally expensive to run, and the programming languages and methods are quite accessible.

In a two-dimensional heat map, each cell of a matrix is filled with a colour representing that cell's value. In our case, the colour of the cell represents the normalised value of a linguistic feature according to WALS. Languages with the same colour in a given row have the same value for that typological feature.[4] Below we discuss two types of heat maps, focusing first on geographic and then on phylogenetic features.

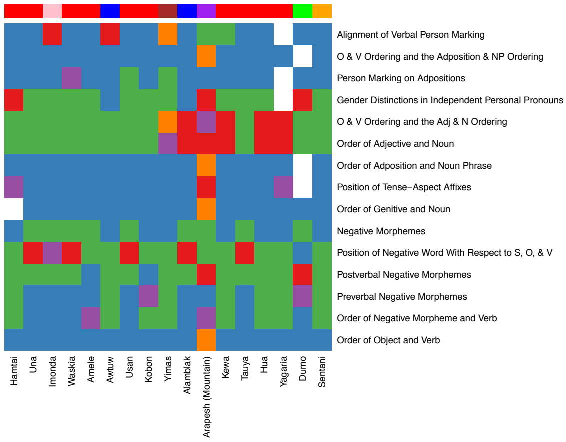

Figure 1: Geographically-focused heat map; see text for details. The bar at the top of the image represents the language family of the language in that column: Pink = Border; Red = Trans-New Guinea; Blue = Sepik; Brown = Lower Sepik-Ramu; Purple = Torricelli; Green = Skou; and Orange = Sentani.

3.1 Geographically-focused heat maps

For the geographic distance maps, for each language present in the cleaned data, we identified all possible languages that lay within 500km, and sorted these languages until only the 16 closest neighbours were selected. Once the set of languages was determined, we selected for graphing only the most commonly-occurring features across that set of languages.

To present the visualisation, we first centred the source language in the map. This decision was made in order to reduce the effect of one of the primary issues with using distance on a two dimensional graph; distance between two non-source languages is not shown, meaning that one could be to the north and another to the south. This means that the languages on the extremes of the map may be far apart from each other, and should be viewed with caution.

Figure 1 shows a geographically-focused heat map with values for various morphological and word order features. The map is centred on Yimas, a language spoken in New Guinea. The features presented represent a particularly non-

[3] https://github.com/RichardLitt/visualizing-language

[4] Due to this reliance on colour, we strongly suggest viewing the heat maps presented here in colour.

sparse section of WALS for this language area. A number of insights can be gleaned here. Most prominently, these languages are quite homogenous with respect to the selected features. Given that most of the languages do indeed belong to the same language family (cf. top bar of the graph), this is unlikely to be a chance effect. In the 5th row ('O&V Ordering and the Adj&N Ordering'), we see via the cluster of red cells a partial grouping of languages close to Yimas, with less similarity at a greater distance. The nearly alternating pattern we see for 'Position of Negative Word With Respect to S,O,&V' may suggest areal groups that have been split by the data-centring function. Also, the checkerboard pattern for this feature and the one below ('Postverbal Negative Morphemes') suggests a possible negative correlation between these two linguistic features.

3.2 Phylogenetically-focused heat maps

To produce phylogenetically-focused visualisations, for each language we identified other languages coming from the same family, subfamily, or genus. Figure 2 shows a phylogenetically-focused heat map for Niger-Congo languages, arranged from west to east. A number of the western languages show red cells for features related to relative clauses; these can be compared to mostly blue cells in the eastern languages. We also see some apparent groupings for variable word order in negative clauses (red cells in western languages) and for NegSVO Order (purple cells in western languages). For some pairs of adjacent languages (most notably Bambara and Supyire), we see clusters of shared features. Especially give the importance of Bambara for syntactic argumentation (Culy, 1985), this graph is an excellent example of visualisation pointing out an intriguing area for closer analysis.

4 Conclusion

In this paper we present a new approach to visualising relationships between languages, one which allows for the simultaneous viewing of linguistic features together with phylogenetic relationships and geographical location and proximity. These visualisations allow us to view language relationships in a multi-faceted way, seeking to work around the sparseness of available data and facilitate new insights into linguistic typology.

In this work we placed strong restrictions on

Figure 2: Phylogenetic heat-map of Niger-Congo languages, arranged from west to east.

both feature coverage and selection of salient features for representation, reducing the number of graphs produced to 6 with geographic focus and 8 with phylogenetic focus. One topic for future work is to explore other ways of working with and expanding the available data in order to access even more useful visualisations. In addition, it would be very interesting to apply this visualisation method to data from other sources, for example, data from multiple related dialects. In such cases, coverage is likely to be better, and the languages in question will have been selected already for their relatedness, thus avoiding some of the data-filtering issues that arise. Finally, we would like to investigate more principled approaches to selection, presentation, and ordering of linguistic features in the heat maps.

Acknowledgments

We are grateful to the three anonymous reviewers for helpful comments on the current paper as well as interesting suggestions for future work.

References

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

Christopher Culy. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8:345–351. 10.1007/BF00630918.

Michael Cysouw. 2011. Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of northwestern european languages. In Horst Simon and Heike Wiese, editors, *Expecting the Unexpected*, pages 411–431. De Gruyter Mouton, Berlin, DE.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Hal Daumé III. 2009. Non-parametric Bayesian model areal linguistics. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Boulder, CO.

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

Michael Dunn, Simon Greenhill, Stephen Levinson, and Russell Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Ryan Georgi, Fei Xia, and Will Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of COLING 2010*.

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

William Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of IJCNLP 2008*.

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

Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel Keim. 2010. Consonant co-occurrence in stems across languages: automatic analysis and visualization of a phonotactic constraint. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, NLPLING '10, pages 70–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Christian Rohrdantz, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel Keim. 2010. Comparative visual analysis of cross-linguistic features. In *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)*, pages 27–32. Poster paper; peer-reviewed (abstract).

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613, 09.