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 (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), 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).

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 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, for the several thousand languages in WALS, there is only about 16% of the possible features filled—the data are *sparse*, and the sparsity of the data of course 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 to be extracted, followed by a discussion of some difficulties presented by the available data and our approaches for addressing those difficulties. Finally, we show 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 may explore the use of alternate sources such as Ethnologue (Lewis, 2009) or MultiTree (2009) for alternate phylogenetic hierarchies.

2.1 Linguistic features

At the time of submission, WALS contained information for 2678 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 questions. A total of 192 features are represented, grouped in 144 different chapters, with each chapter addressing a set of related features.

The coverage of features/chapters varies dramatically across the languages, with an average of 28 feature values per language. The most pop-

ulated feature has data for 1519 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.

2.2 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 figures for determining geographic distance between two languages. A crucial aspect of our visualisations is that we produce them only for sets of languages within a reasonable geographic proximity (and with sufficient featural coverage within WALS).

For this study, we try two approaches to clustering languages according to geographic distance. First, we choose an arbitrary radius in order to create a decision boundary for clustering neighbouring languages. For each language, we fix that language's location as the centroid of the cluster and look at all other languages within the given radius. We found that a radius of 500 kilometers proves to provide a sufficient number of examples even after cleaning low-coverage languages from the WALS data.

The second approach selects an arbitrary lower bound for the languages in the general area. If a sufficient percentage 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. 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 Producing the heat maps

Because of the sparse coverage of the WALS data, we focus our visualisations on languages with a reasonable number of filled features. 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. The methodology discussed in this section is relevant only for languages with values for at least 30% of all possible feature categories. 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. All code is available in a public repository, links to which will be released once anonymisation is no longer required.

3.1 Geographically-organized heat maps

For the geographic distance maps, for each language present in the cleaned data, we selected all possible languages that lay within 500km, and sorted these languages until only the 15 closest neighbors were selected. For phylogenetic distances, for each language we searched for other languages coming from the same family, subfamily, or genus. In either case, the subset that fit these requirements was a mere handful. We have chosen to graph these examples, as they most fully show the use of our visualisations. We picked features to graph from among the resulting languages based on how common they were across the selected languages.

Each final list was then resorted. In the geographic case, 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. Language families were put as a bar at the top of the image, in order to show relations that might otherwise not be obvious.

3.2 Phylogenetically-organized heat maps

The phylogenetic example was sorted by geographical position, as well. In this case, we first plotted the subset of languages on a world map, and examined which groups had the most linear structure. We then arranged that group according to cardinal directions; here, we mapped from North to South. This is a better proxy in this case than centering the data, as it more clearly shows languages that would be in contact with each other.
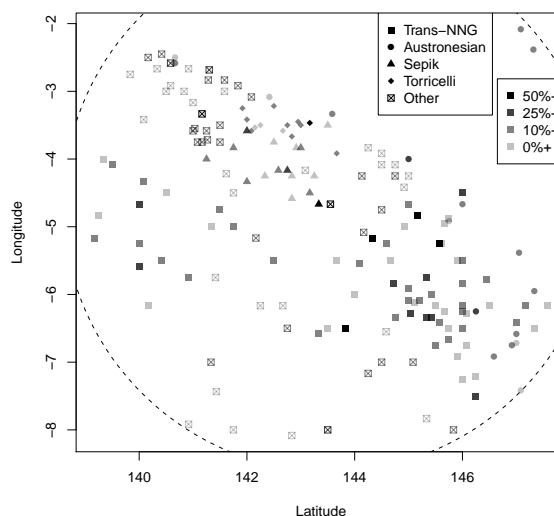
3.3 Sparse data

Figure 1: Sparse Data

In Fig. 3, 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 geograph-

ical 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.

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 at least 30% as a lower bound of all of their entries filled. This cleaned data was then used in the other functions.
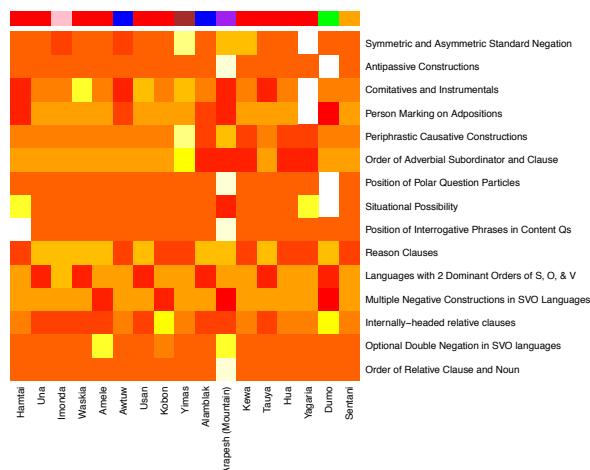
4 Results

Figure 2: Example 1

5 Conclusion

In this paper we present a new method for visualising relationships between languages, representing
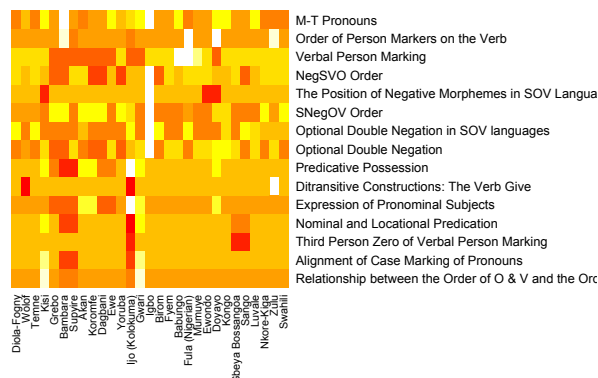
Figure 3: Example 1

References

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

M. Cysouw. 2011. Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of northwestern european languages. In H. 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 S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011 edition.

Michael Dunn, Simon J Greenhill, Stephen C Levinson, and Russell D 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*.

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.

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.

LINGUIST List, editor. 2009. *Multitree: A digital library of language relationships*. Eastern Michigan University, Ypsilanti, MI, 2009 edition.