Mapping Language: From data to diaspora

John Pritchard[1], Eric Atwell[2], Mark Newman[3], Danny Dorling[1], Fabien Hall[2]
1 University of Sheffield, 2 University of Leeds, 3 University of Michigan

This paper presents a series of language maps created for the online *Worldmapper* map repository, together with different methods of collecting the data that underlie the maps, and the problems associated with those methods.

Worldmapper

Worldmapper (www.worldmapper.org) is a web-based project, which currently hosts over 600 world maps, each depicting how the world would look if each country was sized not according to the area taken up by its land mass, but by some other variable. The simplest example, and one which is useful as a stepping-stone to viewing all the other maps, is that of population.

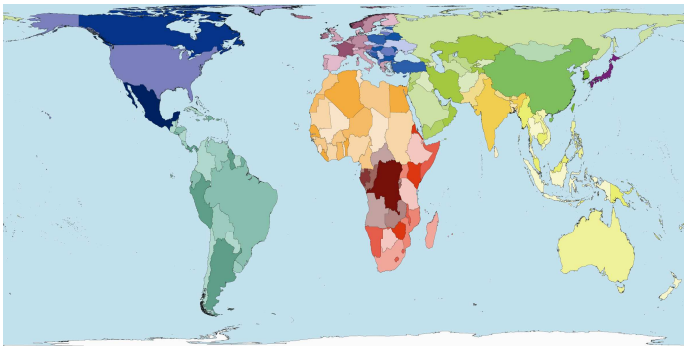

Figure 1a: Worldmapper *Land Area* map

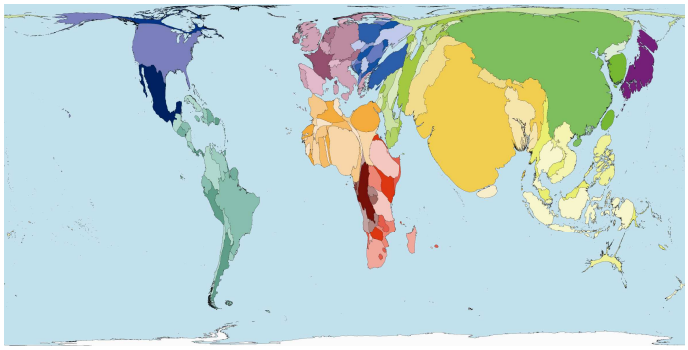

Figure 1b: Worldmapper *Population* map

The maps all use the same colour scheme. Map 1a is a cylindrical equal-area projection. Map 1b is a cartogram; the countries show population size rather than land area. The cartogram uses the algorithm devised by Gastner and Newman (2004). A way of conceptualising what is happened is to imagine everyone on earth being allocated the same space. Everyone remains in the same country, so each country has to expand or contract, depending on the density of their population; borders are pushed and pulled until and equilibrium is reached whereby each person has the same space (about one-six billionth of the land area, which remains the same overall).

Mapping language

This concept can be extended to any other variable that has been measured (usually a count of people). The latest addition to Worldmapper is a collection of 113 maps on language. To use the example of English, each territory is sized according to an estimate of the number of people who live there for whom English is their first language.

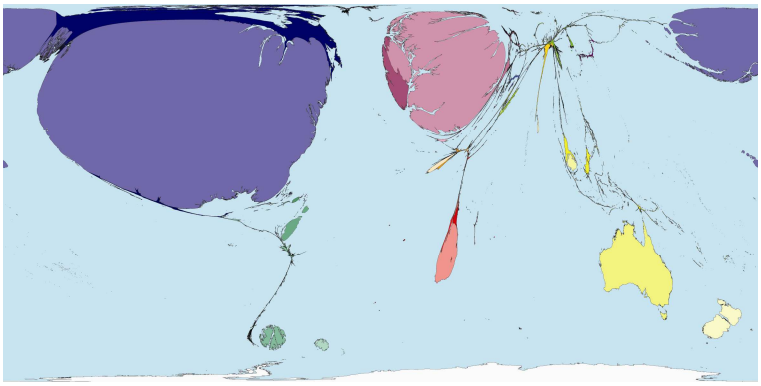

Figure 2: Worldmapper *English as a first-language* map

Which languages?

The Ethnologue reports that there are nearly 7000 languages, so there was some work required to determine which languages to map. The criteria used was that a language should be recorded as being spoken in four or more territories, and have at least half a million speakers in total. This restricted the number of languages mapped to 112, many of which are spoken in considerably more than 4 territories.

Issues of the definition of a single language have also presented a problem; the boundaries between the definitions of 'language' and 'dialect' are blurred, and there is also a political aspect to the definition of a language.

Data sources

The main data source was The Ethnologue (2005), compiled by SIL International, which is an attempt to catalogue all known languages, and provides information on the number of speakers of a language in a particular country. As reported by Paolillo and Das (2006), due to the scale of the task, and a small staff, Ethnologue is often incomplete and outdated in its reporting of speakers of a language . Hence the Ethnologue data has been supplemented with data from a variety of other sources, including national censuses, books, and personal contact with experts.

Seeing the spread

One problem with these maps is that even the languages that have become widely used tend to be dominated by just a few territories, making it hard to see on the map where the language has spread to in smaller numbers. A solution is to set those territories where a language is used by most of the population to zero, to allow the pattern of spread of the language to be viewed. The example of English is shown in Figure 3.

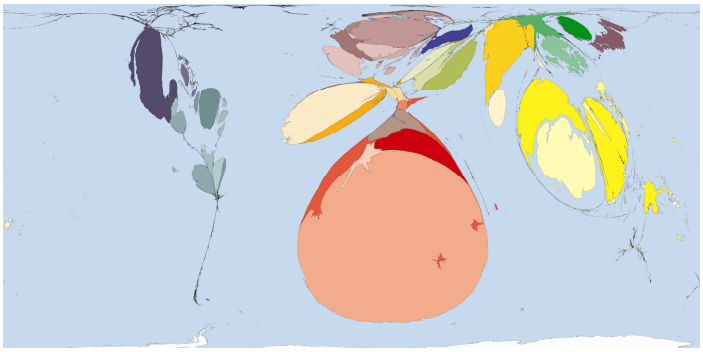

Figure 3: Worldmapper *English language spread* map

Another way of collecting data

A collaboration between Sheffield and Leeds Universities has enabled us to also produce some language maps using an alternative method of collecting data. The method utilises the ability of search engines to report on the use of certain words contained within websites, broken down by country. With careful choice of word, as a proxy for use of a certain language, data has been gathered for four different languages. The raw data tells us as much about the use of the internet within each country as it does about the use of a language. Hence different options have been explored for working with the data, including combining it with data on population and internet use; comparisons can then be made between the language maps made with the different methods of obtaining data.

References

- The Ethnologue. (2005). *Languages of the World (15th ed.).* Dallas, TX: SIL International.
- Gastner, M.T. and Newman, M. E. J. (2004) Diffusion-based method for producing density equalizing maps *Proc. Natl. Acad. Sci.* USA 101, 7499-7504.
- Paolillo, J. C. and Das, A. (2006). *Evaluating Language Statistics: The Ethnologue and Beyond*. Report prepared for the UNESCO Institute for Statistics. Retrieved 27.5.2009 from http://ella.slis.indiana.edu/~paolillo/research/u_lg_rept.pdf