Universität Zürich[UZH]

Geographisches Institut
Geographische Informationssysteme

Dr. Peter Ranacher
peter.ranacher@geo.uzh.ch

Prof. Dr. Ross Purves
ross.purves @geo.uzh.ch

Master Topic: Extracting South American Language Diversity from Text

Short description

There are two big databases of the languages of the world: Ethnologue (Eberhard et al. 2021), a paywalled publication documenting the language distribution of the world including polygon data, but no academic references; and Glottolog (Hammarström et al. 2020), which started out as academic bibliography about the languages of the world and their classification, which contains point locations for most languages, but no speaker counts or ranges.

For working with world-wide language samples that take into account geography, this situation is deeply dissatisfying. The GIS unit has initiated the "Glottography" project to collect published geographical distributions of languages and present them in a unified manner.

While linguistic publications often contain maps which we can digitize, georeference and integrate, a lot of distribution data is provided in text format, and not accessible directly for GIS applications. One such source is the "Diccionario etnolingü.stico y guía bibliográfica de los pueblos indígenas sudamericanos" by Fabre (2005), which contains textual descriptions of locations and speaker counts of nearly 200 languages. The Diccionario has some overall structure for listing the distributions, and as such is a good starting point for applying Gazetteer-based methods for extracting location data from text, similar to those taught in GEO871. Our database already contains language distribution polygons from other sources for some of the languages in the Diccionario, providing a way to validate results. While some of the necessary steps could use established techniques, which you may even have seen in GEO871, the aggregation of location data into describing a range is not well established and would be a worthwile extension, also with ecological applications (eg. Scott et al. 2021) in mind.

The goal of this master project is to demonstrate the applicability of Geographic Information Retrieval methods to provide mostly non-overlapping distribution polygons and points weighted with speaker counts from textual descriptions of language ranges. Ideally, the outcome will be the language map of South America according to Fabre, including uncertainty of the distributions and speaker counts.

Languages

The ability to work with Spanish text is a requirement, because of the sources (in particular Fabre).

Contact

If you are interested, contact

- Dr. Peter Ranacher, peter.ranacher@geo.uzh.ch
- Prof. Dr. Ross Purves, ross.purves@geo.uzh.ch

References

- Eberhard, David M. & Simons, Gary F. & Fennig, Charles D. (eds.). 2021. Ethnologue: Languages of the World. Twenty-fourth edition. Dallas, TX: SIL International. (http://www.ethnologue.com)
- Fabre, Alain. 2005. Diccionario etnolingü.stico y guía bibliográfica de los pueblos indígenas sudamericanos. Tampere, ms. (Handbook.) (http://www.ling.fi/DICCIONARIO.htm) (Accessed December 15, 2021.)
- Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2020. Glottolog 4.2.1. Jena. (doi:10.5281/zenodo.3754591) (https://glottolog.org/ accessed 2020-08-10)
- Scott, Jamie & Stock, Kristin & Morgan, Fraser & Whitehead, Brandon & Medyckyj-Scott, David. 2021. Automated Georeferencing of Antarctic Species. In Janowicz, Krzysztof & Verstegen, Judith A. (eds.), 11th International Conference on Geographic Information Science (GIScience 2021) - Part II (Leibniz International Proceedings in Informatics (LIPIcs)), vol. 208, 13:1-13:16. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. (doi:10.4230/LIPIcs.GIScience.2021.II.13) (https://drops.dagstuhl.de/opus/volltexte/2021/14772) (Accessed December 15, 2021.)