Gabmap: A web application for dialectology

Martin Wieling and John Nerbonne, University of Groningen, reporting on joint work with Charlotte Gooskens, Peter Kleiweg and Therese Leinonen

http://www.gabmap.nl/

We frequently ask in linguistics, especially in dialectology and comparative linguistics, how similar linguistic varieties are to one another, effectively asking how similar linguistic culture is from one site to another. We operationalize the question more specifically by asking e.g. how similar the vocabulary of one variety is to another, or more interestingly how similar the pronunciations of a set of varieties are, sampled via the pronunciations of the same set of at least 30 words at a range of sites. Since there may be thousands of words and hundreds of sites, the questions must be addressed computationally. The techniques embodied in the web application have been used in dozens of scholarly papers on dialectology (see references).

At the University of Groningen the Gabmap application has been developed that is capable of measuring differences in linguistic samples, including in particular sets of phonetic (or phonemic) transcriptions, and of projecting the results graphically onto maps. Gabmap has a graphical user interface that implements not only the comparison of vocabulary or other categorical data (essentially as percentage overlap or percentage difference) but also that of pronunciations via edit distance. Because the software is implemented as a web application users are not required to download it nor to keep it up to date by following releases. It is fairly user friendly and easily accessible and therefore enables experimentation with different techniques popular among linguists from various fields, especially dialectology and variationist linguistics.

During the workshop we will give some theoretical background about dialectometry followed by a tutorial where the theory is put into practice with exercises showing how to use the web-application. The workshop will be include hands-on work.

The workshop will be structured as follows:

- Introduction to dialectometry
- Data entry: uploading dialect data, creating and uploading maps
- Data inspection: data distribution and error detection
- Measuring linguistic distances
- Graphical presentations of linguistic distances: dialect maps
- Statistical analyses: multidimensional scaling and clustering
- Data mining, identifying influential individual variables (words, pronunciation variants)

We can accommodate up to 20 participants.

We add a note to potential participants from non-linguistic fields. In theory one might ask the same questions of non-linguistic culture that we ask of linguistic culture, and we would welcome the chance to do this in a data-intensive way. If such studies are carried out, we suspect that at least the mapping facilities we demonstrate in this tutorial will be useful.

Where and when?

Göteborg University, Humanisten
Wednesday, October 28, 2011, 09:15 – 12:00
Data Lab 4 in the G corridor, 2nd floor.

http://spraakbanken.gu.se/nclav
Recommended reading:


Other References:


Kessler, B. 1995 iComputational dialectology in Irish Gaelic,i In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, 60–67 Dublin EACL.

Leinonen, Therese 2008 iFactor Analysis of Vowel Pronunciation in Swedish Dialects,i International Journal of Humanities and Arts Computing, 2(1–2), 189–204.

Nerbonne, J. 2009 iData-driven dialectology,i Language and Linguistics Compass, 3(1), 175–198.


Prokic, J., et al 2009 iTThe Computational Analysis of Bulgarian Dialect Pronunciation,i Serdica Journal of Computing, 3(3) 269–298

Spruit, M. 2006 iMeasuring syntactic variation in Dutch dialects,i Literary and Linguistic Computing, 21(4) 493–506

Yang, C. & Castro, A. 2008 iRepresenting Tone in Levenshtein Distance,i International Journal of Humanities and Arts Computing, 2(1–2) 205–219

http://spraakbanken.gu.se/nclav