Short Term Mobility Research Program - 2015

Dr. Martijn Wieling

Faculty of Arts, University of Groningen, the Netherlands

Title of the work program

Estensione della risorsa dialettale online ALT-Web con funzionalità di analisi dialettometrica dei dati dialettali

Extension of the ALT-Web online resource with dialectometric analysis functionalities

1. Introduction

The goal of the visit of Martijn Wieling at ILC in Pisa was the integration of an online dialect atlas together with an online application for dialectometric analyses. By integrating these two systems, the user is not only able to conduct a qualitative fine-grained analysis of the data via the online dialectal resource, but is also able to effortlessly obtain a quantitative aggregate view of (subsets of) the data.

The dialect atlas is the *Atlante Lessicale Toscano* (ALT; Giacomelli et al., 2000), a specially designed linguistic atlas focusing on dialectal variation within Tuscany and in which lexical data have both a diatopic and diastratic characterization. The online resource, ALT-Web[1] (Cucurullo et al., 2006), provides facilities to inspect the data at the level of the individual question items and/or individual locations. However, facilities to carry out aggregate analyses of the data are not available, and visualization options are limited (see e.g., Figure 2 of Cucurullo et al., 2006).

The online application *Gabmap*[2] (Nerbonne et al., 2011; Leinonen et al., forthcoming) has been developed by the University of Groningen in the framework of CLARIN-NL as an accessible open source web application to analyze language variation data. Gabmap allows both aggregate dialectometric analyses and data inspection at the level of the individual items.

In the following, we illustrate how we have integrated the two systems and highlight the functionality of the new system, dubbed *ALT Explored*. This can be seen as an example of integration of resources and services within the pan-European CLARIN ERIC infrastructure: whereas *Gabmap* is already part of it, ALT-Web will be integrated soon, thanks to the recent participation of Italy to CLARIN. In this way, linguists and dialectologists will be able to explore, quantitatively and qualitatively analyse and visualize ALT data in order to support their research.

[1] `http://serverdbt.ilc.cnr.it/altweb/`

[2] `https://www.gabmap.nl/`

2. Background

2.1 Atlante Lessicale Toscano

The interviews of the *Atlante Lessicale Toscano* were conducted between 1974 and 1986 in 224 localities in Tuscany, with 2,193 informants selected with respect to various parameters, including age, socio-economic status, education and culture. The interviews were carried out on the basis of a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary, semantics and phonetics. A dialectal corpus with these features lends itself to investigations concerning geographic or horizontal (diatopic) variation as well as social or vertical (diastratic) variation: *ALT Explored* focuses on both dimensions of linguistic variation.

Tuscany is a region which has a special status in the complex puzzle of linguistic variation in Italy where also non-Tuscan dialects are spoken: the latter is the case of Gallo-Italian dialects spoken in Lunigiana and in small areas of the Apennines. The locations where a Tuscan dialect is spoken are 213 out of the 224 investigated locations (with a 2060 informants).

Each informant responded to a total of 460 onomasiological (eliciting a lexical item for a concept) and 285 semasiological (eliciting a meaning for a word form) questions. In what follows, we will focus on the subcorpus of answers to onomasiological questions.

2.2 ALT-Web

ALT-Web contains the digitized responses to all 745 onomasiological and semasiological questions. In ALT-Web, all dialectal responses are assigned different levels of representation, with a first level rendering the original phonetic transcription and other levels containing normalized representations of the original form encoded in standard Italian orthography. In this multi-level representation model, dialectal data are encoded in layers of progressively decreasing detail going from phonetic transcription to different levels of normalized representations abstracting away from details of speakers' pronunciation. For the phonetic transcription, the *Carta dei Dialetti Italiani* (CDI) transcription system (Grassi et al., 1997) was used.

In what follows, we will focus on the phonetic and normalized representation levels. For what concerns the former, the CDI representation was automatically converted to the International Phonetic Alphabet (IPA) on the basis of 158 ordered conversion rules encoded as regular expressions. Concerning the latter, the most abstract normalized representation was selected, i.e. the representation level abstracting away from productive phonetic variation within Tuscany and mainly reflecting lexical and (limited) morphological variation.

ALT-Web provides flexible and dynamic query facilities which permit the user to interactively define the access key to the corpus of dialectal data and to navigate through it on the basis of his/her research interests. Information can be accessed and retrieved on the basis of a wide range of parameters which can be variously combined; for example, lexical data can be searched on the basis of the question of which they represent the answer, or the location in which they were witnessed and/or the socio-economic features of the informant(s), or of their relevance with respect to a given semantic field or linguistic register. ALT-Web also contains a basic visualization facility providing, for each answer, presence/absence maps: frequency information is not taken into account in this visualization. More details about ALT-Web can be found in Cucurullo et al. (2006).

2.3 Gabmap

Gabmap is an online web application for conducting dialectometric analyses. Besides enabling a researcher to conduct aggregate analyses (such as cluster analysis, or visualizing the data using multidimensional scaling) of dialect data, Gabmap also offers facilities for visualizing and inspecting the original data. All visualizations can be downloaded as vector or bitmap graphics. Gabmap has been positively evaluated by Snoek (2014: 206): "Gabmap is excellent software that permits the mapping and comparison of linguistic data in a fast and generally painless manner".

To use Gabmap, a free account needs to be created. After logging in, language variation data can be uploaded and the results of the dialectometric analysis can be inspected. The datafile should contain the transcribed pronunciations (or lexical variants) of all items in each location. In addition, a mapping file (i.e. a kml file) with the geographical coordinates of each location is necessary to adequately visualize the results on a map.

3. Design and functionality of *ALT Explored*

As the source of Gabmap is available as open source software,[3] we have customized the original source code to allow for the integration with ALT-Web. The changes can be classified in three categories, which are illustrated in the following.

3.1 Custom Gabmap projects

The first change involved the creation of four custom Gabmap projects on the basis of the dataset gathered with all onomasiological ALT questions. We created four projects to represent both the phonetic data (converted to IPA) and the normalized data, both for all locations in the ALT dataset (224) and for the subset of 213 locations where a Tuscan dialect is spoken. For this purpose the ALT-Web data was converted to the tabular form required for Gabmap and the appropriate mapping files with the locations were created. In each location, the pronunciation (or normalized form) of each individual speaker was retained. Note that a relatively similar customization of Gabmap has been made for another dataset containing English lexical variation (i.e. *BBC Voices explored;*[4] Wieling, 2013). The four individual projects can be accessed directly via the links at http://www.gabmap.nl/~alt (see Figure 1).

3.2 Integration of Gabmap distribution maps

The second change involved the extension of the Gabmap application by allowing direct (url) access to the distribution maps. Consequently, when the user supplies both the item and the variant as parameters in the url, the corresponding distribution map is shown visualizing the relative frequency of use of the variant throughout Tuscany. This functionality is in the process of being

[3] `https://github.com/coltekin/Gabmap`
[4] `http://www.gabmap.nl/voices`

integrated in the ALT-Web application and will replace the original ALT-Web visualization, which does not allow for the visualization of relative frequencies. As the ALT-Web application uses a different transcription system than Gabmap, the CDI to IPA conversion script (see above) is used here to convert the CDI transcription of individual items (transmitted via the url) to the IPA transcription used in Gabmap. For example, the CDI transcription [suSi8na] (a dialectal variant of *susina*, 'plum') would be converted to [suz'ina] in IPA. Figure 2 visualizes the distribution map of this variant, which can be viewed online by using the following url:

http://www.gabmap.nl/~alt/bin/accountALT?username=altdemo&action=ALTDISTMAP&project=3&var=suSi8na&convert=1&item=101-susina#fig

3.3 Inspection of custom-generated data

The final change involved creating a second direct url-based interface to Gabmap. The purpose of this interface was to allow a user to inspect language variation data in (a temporary account in) Gabmap, without the need for creating an account or manually uploading data (but instead by simply clicking a link). Clearly, this interface facilitates the integration of Gabmap in online repositories of dialect atlas data, such as ALT-Web. For example, ALT-Web is currently in the process of being adapted to allow users to analyze custom subsets of ALT data in Gabmap.

Each subset is created on the basis of setting specific filters. The current filters allow the user to specify: the age range,[5] the gender, the job type and the education level of the speaker; the province(s) in which the locations should be located; and the semantic field(s) of the questions included. After setting one or more filters, the ALT-Web application creates the appropriate Gabmap datafile (transcriptions are converted to IPA), stores it on a server and generates a direct link to Gabmap. When the user clicks on this link, the data is uploaded to Gabmap, and the resulting analyses can be directly inspected in a temporary Gabmap account. If the username starts with "_tmp" and the description, the url to the data file and the url to the kml file are supplied (and the files are valid Gabmap input files), the project is created successfully. In the final version of the paper, this functionality – still under completion - will be illustrated in detail.

4. Conclusion

The main contribution of this this work can be summarized as follows. From the perspective of on-line digital resources (dialect atlas) such as ALT-Web, the integration with Gabmap in *ALT Explored* provides quantitative analyses which were not possible before. From the perspective of Gabmap, the flexible extensions represent a unique opportunity to become the visualization and exploration tool for an entire linguistic atlas. While the two online applications are both useful in their own right, their integration becomes more than the simple sum of the two parts.

Within *ALT Explored*, the dynamic combination of macro- and micro-analysis, or to put it in Moretti's words of "close and distant reading" functionalities, allows the dialectologist to alternate between views of the data, zooming in and out, searching for large-scale patterns and then focusing in on fine-grained analysis. The combination of close and distant reading of textual data is

[5] See Wieling et al. (2014) for a discussion of the importance of age in Tuscan lexical variation.

becoming more and more an open challenge in the digital humanities: to our knowledge, our contribution represents the first attempt in this direction in the sub-area of digital dialectology.

To conclude, the STM visit of Martijn Wieling was very fruitful, given that the obtained results are interesting and shed new light on the study of dialectal variation by combining "close and distant reading" functionalities. The results achieved during the stay have been illustrated in a paper submitted at the 10th International Conference on Language Resources and Evaluation (LREC 2016) which will be held next May in Portorož (Slovenia).

References

Cucurullo, N., Montemagni, S., Paoli, M., Picchi, E., Sassolini, E. (2006). Dialectal resources on-line: The ALT-Web experience. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, 1846–51.

Giacomelli, G., Agostiniani, L., Bellucci, P., Giannelli, L., Montemagni, S., Nesi A., Paoli M., Picchi E., Poggi Salani T. (2000). *Atlante lessicale Toscano*. Roma: Lexis Progretti Editoriali.

Grassi, C., Sobrero, A., Telmon, T. (1997). *Fondamenti di Dialettologia Italiana*, Roma-Bari, Laterza.

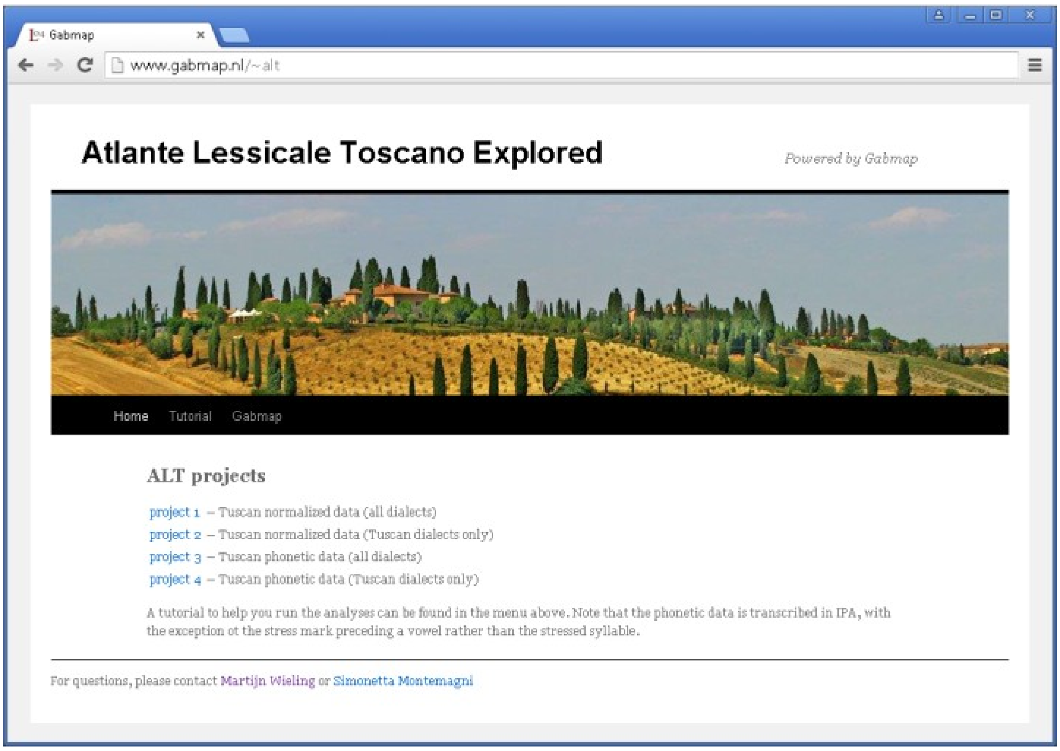
Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., Leinonen, T. (2011). Gabmap – a web application for dialectology. *Dialectologia: revista electrònica*, SI II, 65-89.

Leinonen, T., Çöltekin, Ç., & Nerbonne, J. (forthcoming). Using Gabmap. *Lingua*.

Snoek, C. 2014. Review of Gabmap: Doing Dialect Analysis on the Web. *Language Documentation & Conservation*, 8, 192-208.

Wieling, M. (2013). Voices dialectometry at the University of Groningen. In: Clive Upton and Bethan Davies (eds.) *Analysing 21st-century British English: Conceptual and methodological aspects of the BBC 'Voices' project*. London: Routledge, pp. 208-218.

Wieling, M., Montemagni, S., Nerbonne, J., Baayen, R.H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, 90(3), 669-692.



Atlante Lessicale Toscano Explored

Powered by Gabmap

ALT projects

project 1 – Tuscan normalized data (all dialects)
project 2 – Tuscan normalized data (Tuscan dialects only)
project 3 – Tuscan phonetic data (all dialects)
project 4 – Tuscan phonetic data (Tuscan dialects only)

A tutorial to help you run the analyses can be found in the menu above. Note that the phonetic data is transcribed in IPA, with the exception of the stress mark preceding a vowel rather than the stressed syllable.

For questions, please contact Martijn Wieling or Simonetta Montemagni

Figure 1. Four custom Gabmap projects in ALT Explored.

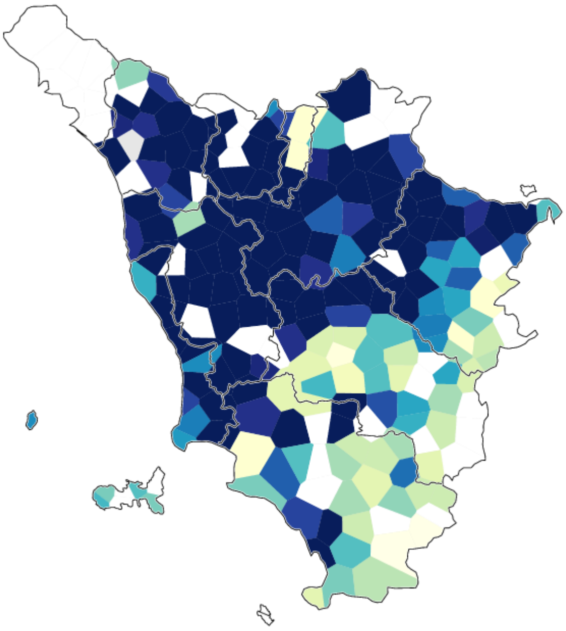


Figure 2. Distribution map of [suzˈina] for the concept susina, ‘plum’. Darker colours indicate a greater relative frequency of use.