Available online at www.sciencedirect.com

ScienceDirect

Lingua 178 (2016) 71–83

Lingua

www.elsevier.com/locate/lingua

Using Gabmap☆

Therese Leinonen [a], Çağrı Çöltekin [b], John Nerbonne [b,c,*]

[a] University of Turku, Finland
[b] University of Groningen, Netherlands
[c] University of Freiburg, Germany

Received 20 May 2014; received in revised form 2 February 2015; accepted 9 February 2015
Available online 12 March 2015

Abstract

Gabmap is a freely available, open-source web application that analyzes the data of language variation, e.g. varying words for the same concepts, varying pronunciations for the same words, or varying frequencies of syntactic constructions in transcribed conversations. Gabmap is an integrated part of CLARIN (see e.g. http://portal.clarin.nl). This article summarizes Gabmap's basic functionality, adding material on some new features and reporting on the range of uses to which Gabmap has been put. Gabmap is modestly successful, and its popularity underscores the fact that the study of language variation has crossed a watershed concerning the acceptability of automated language analysis. Automated analysis not only improves researchers' efficiency, it also improves the replicability of their analyses and allows them to focus on inferences to be drawn from analyses and other more abstract aspects of that study.

Keywords: Dialectology; Language variation; Mapping; Quantitative linguistics

1. Introduction

Gabmap is a freely available, open-source web application that analyzes the data of language variation, e.g. varying words for the same concepts, varying pronunciations for the same words, or varying frequencies of syntactic constructions in transcribed conversations.

Other possibilities exist as well, but these are by far the most frequent uses to which Gabmap has been put. Nerbonne et al. (2011) reports on Gabmap's basic functionality and its implementation, so that this article can build on that, adding material on new functionality and reporting on the range of uses to which Gabmap has been put. Gabmap is modestly successful, and its popularity underscores the fact that the study of language variation has crossed a watershed concerning the acceptability of automated language analysis. Automated analysis not only improves researchers' efficiency, it also improves the replicability of their analyses and allows them to focus on inferences to be drawn from analyses and other more abstract aspects of that study.

2. A Gabmap session

In this section, we show an example of a typical Gabmap session and the types of analyses that can be conducted. For this purpose we use data from the Goeman-Taeldeman-Van Reenen-project (GTRP; Goeman and Taeldeman, 1996).

☆ We are grateful to CLARIN-NL and for their support of the project ADEPT (http://www.clarin.nl/node/70#ADEPT), which in turn produced Gabmap. CLARIN-NL was supported by the Netherlands Organization for Scientific Research (NWO).
* Corresponding author. Tel.: +31 503635815; fax: +31 503636855.
E-mail address: j.nerbonne@rug.nl (J. Nerbonne).

http://dx.doi.org/10.1016/j.lingua.2015.02.004

The data consist of phonetic transcriptions of Dutch dialects from the Netherlands and Belgium gathered during the period 1980–1995. These data are available as demo data on the Gabmap web site, which makes it possible for users to try out the analyses described here directly in Gabmap.

2.1. Data

The dialect data can be prepared in a spreadsheet where rows represent sites and columns represent linguistic variables. In the demo data, the columns are words and each cell in the spreadsheet shows the pronunciation of a word in the International Phonetic Alphabet (IPA) at one specific site[1]:

	boter	broden	zout
Aalsmeer	botər	brojə	zaut
Baardegem	botər	bruəs	zat
Coevorden	bœtər	brodn	sɔl̪t

Gabmap accepts tab-separated Unicode text files as input data, and most spreadsheet software allow exporting data to text files with Unicode encoding.

Analysis in Gabmap is not restricted to transcribed pronunciation data; instead, any kind of binary or numeric data can be used. When uploading data into Gabmap, the type of data is specified, so that the data can be processed appropriately. For the phonetic transcriptions in the example we choose *string data* as the type of data and *string edit distance* as the type of processing (more about data processing in section 2.3).

In order to create dialect maps, the data file should be accompanied by a map file with the geographical coordinates of the data sites and optionally borders of the country or language area. The map file is a.kml or.kmz file that can be created in Google Earth or using the Google Maps service through any standard web browser. Using a map file is, however, not compulsory. Users might want to analyze language variation related to other factors than geography. The data rows might, for example, be individual speakers instead of sites. For analysis of this type of data, no map file is needed and Gabmap will create a pseudo map instead of real maps in the mapping functions. The statistical analyses, like cluster analysis and multidimensional scaling (see below), will, then, show how individual speakers group together based on their language use.

When a project is created, Gabmap offers several ways of inspecting the data. Summaries are created of the number of sites, number of words (or other linguistic variables), number of characters and number of tokens. In *Data overview* in Gabmap, we can, for example, see that the demo data has data from 613 places and that the number of different words (items) is 562. The total number of word transcriptions (instances) is 331,690, which is less than 613×562 due to some missing data in the input table.

2.2. Distribution maps

Several types of distribution maps are offered in Gabmap. Fig. 1 shows a map of one specific phonetic character in the data set. The character maps are part of the data overview function in Gabmap, where maps can be created of any character or token in the data set. Fig. 1 shows the distribution of the velarized lateral approximant [ɫ]. White color means no instances at all of the character from a site, and the darker the color the higher the relative frequency of the character in the data at the given site. A map like this only gives a rough picture of the distribution of a speech sound, since the result depends on how well each data point has been sampled.[2] Still, the map can give a rough overview of the distribution of a dialect feature and/or of the quality of the data. It is striking that the chosen phonetic symbol in Fig. 1 is almost completely lacking in the data from Belgium. When a pattern like this is found, it could either mean that the distribution of the specific feature very closely follows the national border, or, it could mean that it was not transcribed with the same phonetic symbol by transcribers of the Flemish and Netherlandic Dutch data. In fact this is one of the indications that the Dutch and Flemish fieldworker-transcribers did not use the phonetic alphabet (Wieling et al., 2007) in the same way; it turned out that the Flemish fieldworker-transcribers used many fewer symbols. See Wieling and Nerbonne (2011) for a suggestion on how to correct for the differences in phonetic alphabet using dialectometric techniques.

Distribution maps of specific words can also be created in Gabmap. By first choosing a variable (word) and then a specific variant (pronunciation) a map is created which shows where the chosen variant can be found. Regular

[1] If there are several pronunciations available of a word from one site, these can be separated by ''space slash space'' in the data file.

[2] Sites with a lot of missing data could by coincidence get too high or too low relative frequencies compared to other sites.