Instant Swedish dialect maps

Robert Östling and Mats Wirén

Department of Linguistics, Stockholm University
robert@ling.su.se, mats.wiren@ling.su.se

Abstract

The department of linguistics at Stockholm University has a web service to visualize linguistic variation within the Swedish-speaking area, based on metadata and a billion words of text from blogs. We here describe the dataset and the statistical model behind the tool.

1. Introduction

Our present goal is to briefly summarize the Swedish dialect mapping tool at the department of linguistics, Stockholm University. A more thorough description from a linguistic point of view has been published previously (Östling, 2015). Here we focus mainly on some of the technical details.

2. Data

During the period 2010–2014 we have received blog data crawled by the private company Twingly, in total about six billion words. A large portion of the blogs are served by a small number of services, of which a few collect metadata about their users and publish it in a format that is easy to extract automatically. In total, we have geographic information about around 150,000 authors, with a total production of about a billion words.

3. Smoothing

The basic function of the mapping tool is to produce a color-coded map of Sweden and Finland with municipal-level resolution, according to the relative frequencies of the search term n-grams. Since the counts for individual words are very sparse at this geographical resolution, we also use a simple smoothing mechanism to estimate the frequency $f_{k,o}$ with which authors in area k use the word (or phrase) o. According to the model:

$$\begin{aligned} P(f_{k,o} = x) \propto & x^{n_{k,o}} \cdot (1-x)^{n_k - n_{k,o}} \\ & \cdot e^{\sum_{k' \in G(k)} \sqrt{|x - f_{k'}|}} \cdot e^{-50x} \end{aligned}$$

where $n_{k,o}$ is the observed frequency of word o by all authors in area k, n_k is the total number of words from that area and $G(k)$ is the set of neighboring areas k' of k.

The first two factors correspond to the probability of frequency x generating the observation, the third one corresponds to our prior probability for geographical patterns (the intuitive understanding is that neighboring areas tend to have similar preferences), while the fourth and last factor says that words in general are low-frequent (x is small).

In response to a query for word o, the tool runs 100 iterations of Gibbs sampling to estimate $f_{k,o}$ for all areas k, and the result is used to color-code the map.

4. Access

Our tool is freely available at http://www.ling.su.se/kartverktyg

Referenser

Robert Östling. 2015. Svenska dialektkartor på sekunden. *Språkbruk*, 3:10–13.