Application of Chemometrics Approaches to Dialect Data and the Tracing of Mobility of Humans in North America

R. G. Brereton[1], **B. Vaux**[2], **L. Zhu**[1]

[1]Centre for Chemometrics, School of Chemistry, University of Bristol, Bristol, U.K.
[2]King's College University of Cambridge, King's Parade, Cambridge, U.K.

Abstract

In the present study we apply to these data novel statistical methods based on chemometric principles, with the goals of: 1) identifying the most salient American dialect clusters and their geographic distributions; and 2) comparing the differences between the distribution based on the phonological data and the distribution based on both lexical and phonological data.

Introduction

In 2003 one of the present authors (Vaux) carried out an online survey of more than 47000 U.S. English speakers for 122 linguistic variables. The variables were chosen so as to cover the central components of linguistic variation: phonology, morphology, syntax, and vocabulary.

We believe that this new form of dialectometric study presents several innovations. Firstly, the analysis is based on a significantly large corpus of dialect data. Numerically-driven dialect studies typically include only a moderate number of samples (c. 300). The present study, in contrast, selected 39018 samples for the data analysis. Sampling on this scale can reduce the risk of bias resulting from paucity of data and can provide more detailed information about the distribution of the linguistic features under investigation. Secondly, the research is not based on the linguists' subjective judgement/experience. No prior assumptions are made concerning the dialect distributions that may or may not be revealed by the statistical analysis; the conclusions are based purely on the data. Data driven methods can reveal the nature of dialect distribution without any prior knowledge of spatial distributions.

Materials & Methods

The dialect survey is an expansion of an initiative begun at Harvard University, Cambridge, US. The dialect survey uses a series of questions, including rhyming word pairs and vocabulary words, to explore words and sounds in the English language. For the background of the survey, please read the [1].

The data are from one survey which has 47471 respondents from the 51 states (including Washington DC) in the USA. The sample distribution is roughly proportion to population distribution. It contains 122 questions which can be categorised into two groups: one is about phonetic differences from different places, the other one is more phrase (slang) oriented [2]. Each question has the form like below.

How do you pronounce *au*nt?

1. [ɑ] as in "ah"
2. [æ] as in "ant"
3. [ɒ] as in "caught"
4. I have the same vowel in "ah", "caught", and "aunt"
5. I pronounce it the same as "ain't"
6. f. I use [ɑ/ɒ] when referring to the general concept of an aunt, but [æ] when referring to a specific person by name.
7. I use [æ] when referring to the general concept of an aunt, but [ɑ/ɒ] when referring to a specific person by name.
8. Other

Each sample collected in the survey is represented in vector form such as [1, 1, 2, 3, 4, 2, … 2, 3, 4]. The numbers in the vector are the corresponding answers, and a similarity matrix can be formed between samples which is subject to multivariate analysis using multidimensional scaling.

Results

The results group states in the USA into various linguistic groups. Some of the conclusions can be related to population movement, for example, there are similarities between Florida, California and the North East States. There is a large movement of people from the NE that retires to Florida, and also there is significant movement of people between jobs in California and the NE of the US, especially high tech well educated people. More detailed analyses can also be performed.

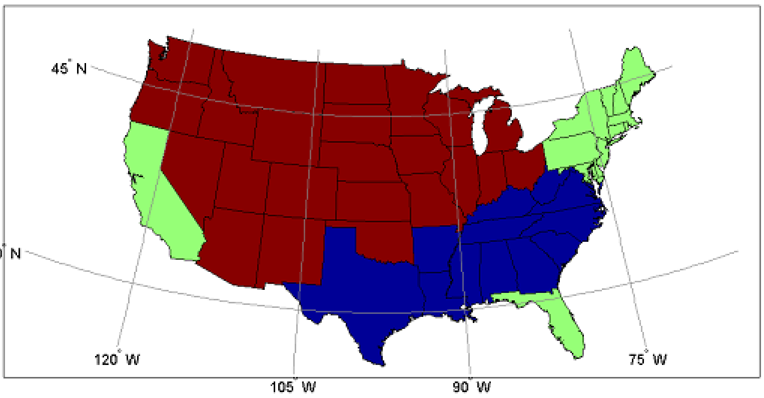

Fig. 1: Identification of three main linguistic groups in the US.

Conclusions

Multivariate methods such as MDS (or principal co-ordinates analysis) can be applied to linguistic data, as a powerful probe of similarities and is a new tool derived from chemometrics that can be employed in linguistics and as such look at population movement.

Reference

1) Harvard University, Dialect Survey at http://www.hcs.harvard.edu/~golder/dialect/ accessed March/08
2) T.F. Cox, M.A.A. Cox, Multidimensional Scaling, Chapman & Hall/CRC, (2000), ISBN: 1-584-8809-45