Making sense of multivariate analyses of linguistic variation

Stefan Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

Multivariate analysis, which exploits correlational patterns among a large number of linguistic variables, has been established as an important and successful technique in many subfields of corpus linguistics such as register variation (Biber, 1988), dialectology (Speelman, Grondelaers, & Geeraerts, 2003), translation studies (De Sutter, Delaere, & Plevoets, 2012) and authorship attribution (Jannidis, Pielström, Schöch, & Vitt, 2015). In this approach, texts (or other linguistic samples) are represented as high-dimensional vectors of quantitative features, the distance between vectors is understood as an indicator of linguistic dissimilarity, and a small number of latent dimensions are identified to capture the main patterns of variation in the data (usually corresponding to correlations between large groups of features). Individual studies differ in their choice of quantitative features – ranging from measurements grounded in a specific linguistic theory (Diwersy, Evert, & Neumann, 2014) to plain frequency counts in a "bag of words" approach (Jannidis et al., 2015) – and in the particular mathematical algorithm used to identify latent dimensions – usually an unsupervised technique such as principal component analysis (PCA; Diwersy et al., 2014), correspondence analysis (CA; De Sutter et al., 2012) or factor analysis (FA; Biber, 1988); some authors also apply linear discriminant analysis (LDA; Baayen, van Halteren, Neijt, & Tweedie, 2002) or another supervised algorithm to exploit additional information about the texts.

A methodological key problem lies in the difficulty of making linguistic sense of the results of a multivariate analysis. Typical approaches include (i) a hermeneutic interpretation of the weighted feature combinations corresponding to each latent dimension, based on human intuition, (ii) visualizing the average coordinates of external categories (such as text types, authors or translated vs. original texts) in the latent dimensions, or (iii) comparing an unsupervised clustering of the text vectors to these external categories. Such attempts are prone to over-interpretation and researcher bias (i), fail to show whether the features contributing to a latent dimension are correlated or complementary (i), or they establish that a multivariate analysis differentiates successfully between external categories, but do not explain how these differences arise from the original features (ii, iii).

This poster explores novel approaches to the linguistic interpretation of multivariate models. First, the feature weights of a latent dimension should not be taken at face value: researchers need to consider their statisical uncertainty (determined by cross-validation or bootstrapping) as well as the distribution of feature values (especially wrt. external categories such as text types). Second, relevant features can be identified by measuring their contribution to the separation of unsupervised clusters or to groupings based on external categories (using techniques such as random forests or recursive feature elimination). Third, a secondary multivariate analysis within clusters or groups reveals how different features combine in correlated or complementary ways into a latent dimension, giving deeper insights about the interactions between individual features.

The new approaches are illustrated with examples from authorship attribution and translation studies. R code implementing the case studies will be made available at http://www.stefan-evert.de/PUB/Evert2017CL/.

References

Baayen, R. H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In *Proceedings of the 6es journées internationales d'analyse statistique des données textuelles (jadt 2002)*. Saint Malo, France.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

De Sutter, G., Delaere, I., & Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In M. P. Oakes & J. Meng (Eds.), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research* (Vol. 51, pp. 325–345). Studies in Corpus Linguistics. John Benjamins.

Diwersy, S., Evert, S., & Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech* (pp. 174–204). Linguae et Litterae: Publications of the School of Language and Literature, Freiburg Institute for Advanced Studies. Berlin, Boston: De Gruyter.

Jannidis, F., Pielström, S., Schöch, C., & Vitt, T. (2015). Improving Burrows' Delta. An empirical evaluation of text distance measures. In *Proceedings of the digital humanities conference 2015*. Sydney, Australia.

Speelman, D., Grondelaers, S., & Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, *37*, 317–337.