From Feature Extraction to Measuring Dialect Typicality

Matthew Sung
Leiden University

One of the main goals in dialectometry is to automatically classify dialects into different groups in a data-driven fashion, with methods such as cluster analysis and multidimensional scaling (Heeringa 2004). Cluster analysis has been criticised for not being able to identify linguistic features found in each cluster, nor being able to identify "exclusively dominant areas, subordinate, non-dominants areas that are determined by smaller numbers of features" (Pickl 2016: 81). In other words, cluster analysis does not indicate which dialects are more 'typical' in the cluster, nor does it show the gradual change from one 'core' area to another.

This presentation will introduce a novel dialectometrical method which addresses the problems raised above with cluster analysis, using the data from the *Syntactic Atlas of Dutch Dialects* (*SAND*, Barbiers 2005, Barbiers 2008). To determine dialect typicality, one has to first find out what 'typical' dialect features are. The new methodology involves two major steps: 1) feature extraction (finding 'typical' features) and 2) typicality measurement (find out how typical a dialect is to a dialect group).

To begin with, dialect distances between all pairs of dialects in SAND are calculated using *Relative Distance Value* (Goebl 2018); six cluster have been identified using *Ward's method* (Ward 1963). Next, features associated with each cluster ('typical' features) are extracted via the application of *Normalised Pointwise Mutual Information* (nPMI, Sung and Prokic 2024). The next step involves calculating the number typical features each dialect has divided by the total number of typical features of a particular group in the feature extraction process. Finally, a Getis-Ord Gi* z-score (Ord and Getis 1995) is calculated based on the raw typicality score of each dialect in order to identify clusters of high typicality values and determine where the focal areas are for each dialect group.

The results of this study show that each dialect group has a core area (cluster of dialects with a high typicality value), and they contain dialects which show gradual decrease of membership of a cluster (typicality) as the distance from the core area increases. In addition, this methodology will allow us to address further theoretical questions in the direction of transitional dialects and grammar, such as "*what type of features are more*

likely to be lost when we move to the periphery" and "*what kind of features are more likely to be retained?*".

References:

Barbiers, S., Bennis, H., De Vogelaer, G., Devos, M. and van der Ham, M. (2005). Syntactische Atlas van de Nederlandse Dialecten Volume 1.

Barbiers, S., van der Auwera, J., Bennis, H., Boef, E., De Vogelaer, G. and van der Ham, M. (2008). Syntactische Atlas van de Nederlandse Dialecten Volume 2.

Goebl, H. (2006). Recent advances in Salzburg dialectometry. Literary and linguistic computing, 21(4), 411-435.

Heeringa, W. J. (2004). Measuring dialect pronunciation differences using Levenshtein distance.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. Geographical analysis, 27(4), 286-306.

Pickl, S. (2016). Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation. In Cote, M.-H., Knooihuizen, R. & Nerbonne, H. (eds.) The future of dialects, 75-98.

Sung, H. W. M. & Prokic, J. (2024). A Comparison between 3 Feature Extraction Methods in Dialectometry. Computational Linguistics in the Netherlands Journal, Vol. 13.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.