Extracting dialect features from German social media data using local spatial autocorrelation

Dana Roemling[1,2] & Jack Grieve[1,3]

[1]University of Birmingham
[2]University of Helsinki
[3]The Alan Turing Institute

Dialectology and dialectometry rely in large parts on data collected through surveys or interviews. Research working with large corpora of unstructured data, such as lexical analyses based on Twitter data (e.g., Huang et al., 2016), are still few and often rely on pre-selected features for analysis. A recent study by Louf et al. (2023) used corpus methods to find cultural regions in the US through mapping lexical variation based on geo-located social media data. They used frequency distributions to find regional hotspots in word usage and identified areas with similar patterns using a principal component analysis.

This submission extends Louf et al.'s methodology to German data. We work with a corpus of 21 million geolocated social media posts from the platform Jodel (Hovy & Purschke, 2018), a social media app structurally similar to Twitter. Employing the same methodology, we work with the 2000 most frequent tokens in the corpus. We first calculate Getis-Ord's z-scores (Ord & Getis, 1995) for all locations and words, which reduces noise and smooths the regional patterns, so that the underlying regional signal can be derived (Louf et al., 2023). We then enter the z-scores into a principal component analysis and extract the features contributing most and least to each dimension. We find that extracted features range from classical dialectal features, e.g. *nüt* (G. 'nothing') or *schau* vs. *guck* (G. 'watch/see') to region-specific usage of abbreviated, elliptical or contracted forms e.g. *ner* (standard *einer*, G. 'a'), *gmacht* (standard *gemacht*, G. 'done'), *bim* (standard *bei dem*, G. 'at the') or *weils* (standard *weil es*, G. 'because it'). Consequently, this study demonstrates the usefulness of local spatial autocorrelation for extracting dialect features from German social media data.

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255.

Louf, T., Gonçalves, B., Ramasco, J. J., Sánchez, D., & Grieve, J. (2023). American cultural regions mapped through the lexical analysis of social media. *Humanities and Social Sciences Communications*, *10*(1), 133.

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, *27*(4), 286–306.