How digital communication mirrors socio-cultural orientation

A computational sociolinguistic study of regional variation on “Jodel”

Christoph Purschke, University of Luxembourg

christoph.purschke@uni.lu | @questoph | https://purschke.info

Regional variation is one of the basic characteristics of language use. It plays a central role in the way people interact with, perceive, and evaluate other people. Variation contributes to the structuring of social practice, for example, concerning the negotiation of group membership based on linguistic (dis)similarity. In the German-speaking area, these distinctions were traditionally bound to local communities, creating small-scale local dialects along a continuum. Over the last 100 years, these fine-grained distinctions have gradually diminished, leading to regionally-bound intermediate varieties that combine features of the old dialects and structures of the standard variety. Rather than disappearing, distinctions persist at a regional level, allowing people to still use them as a linguistic resource to stylize their language and, therefore, for social positioning in interaction.

Social media writing differs considerably from other domains of language use. In online social spaces like Twitter, Facebook, and WhatsApp, people employ a variety of linguistic resources and develop complex and often hybrid writing styles to communicate effectively and playfully. At the same time, online writing preserves all the socio-pragmatic functions of language in practice, that is, identity building, social positioning, the negotiation of relationships, and discourse organization. For German, online writing has been analyzed mostly for the use of linguistic resources typical of digital culture (like emojis, abbreviations, or non-standard spelling). Regional variation, on the other hand, has so far been considered only to a limited extent, for example, in Switzerland or local IRC groups. However, due to the hybrid nature of online writing, social media is likely to offer a range of novel insights for the study of regional variation.

Our study

Starting from this premise, we use an integrated approach that combines computational methodology with in-depth sociolinguistic analysis. Our study analyzes a corpus of more than 3 million anonymous discussions (150 million word forms) collected from the social media platform “Jodel” in the entire German-speaking area (Hovy/Purschke 2018, Purschke/Hovy 2019). We use methods from computational linguistics, mainly neural networks, representation learning (Doc2Vec), and geographic retrofitting, to model regional variation in digital writing focusing on lexical variation. In doing so, we analyze the data without assuming a specific linguistic structure for regional variation (e.g., motivated by regional differences in dialect) or the pragmatic organization of discussions (e.g., driven by the technical characteristics of the network).

Nonetheless, our analysis reveals clear-cut regional clusters of language use which largely correspond to the traditional division of the German dialects as studied in dialectology. These correspondences can be interpreted against the backdrop of other socio-cultural spatial structures, all of which show region-based patterns in the German-speaking area (e.g., socioeconomic mobility, socio-cultural orientation, attitudes).

The revealed spatial structures show that (anonymous) social media communications of young adults in the German-speaking area can indeed be characterized by forms of “digital regionality” which are

- *regionally distinct*, that is, they are bound to and typical of specific regions or countries (e.g., the characteristic use of regional dialects in Switzerland as opposed to the rest of the German-speaking area);
- structured by the use of *specific linguistic resources*, that is, the region-specific use of certain types of vocabulary (e.g., the large number of loan words and ethnolectal forms in the Frankfurt Metropolitan area or the use of newly coined words characteristic of certain regional user groups);
- closely *linked to extra-linguistic factors* mirroring the regional organization of social practice (e.g., the overall structure of the German dialects, as well as socioeconomic factors including student mobility and work commuting);
- to be seen as *digital equivalents of social practices* which are deeply rooted in traditional forms of socio-cultural orientation, that is, regions.

These findings can be substantiated by an analysis of region-specific profiles in discussion topics we find in the dataset, thus combining traditional sociolinguistic analysis of regional variation with a content-based perspective on regionality. User communities on Jodel are constituted not only by *how* they write, but also by *what* they talk about, that is, their entire socio-cultural orientation (as mirrored in regionally-bound discussion topics). Taken together, region-specific writing styles and topic profiles shed light on different aspects of social dynamics in digital communication, for example, regarding the spread and establishment of new regional or medium-specific variants in the Jodel community.

Additionally, we compare manual annotations of the 1000 most prototypical words for each cluster with a learned prediction model (based on these annotations) to evaluate the importance of region-specific vocabulary for the overall linguistic structure of each cluster. The analysis shows that, while the clusters differ considerably in terms of their most prototypical words (dialect forms, loan words, local topic words), they also share many linguistic features if we predict the distribution of linguistic resources for all words in a cluster (e.g., standard German forms; characteristics of online writing).

The study shows that the regional distribution of linguistic and thematic resources in Jodel communications can be predicted accurately with the help of a combination of computational and sociolinguistic analysis methods. It sheds light on the importance of (linguistic as well as socio-cultural) regionality for the organization of social practice. Besides, the study shows how a combination of computational and sociolinguistic methods that takes into account the specific strengths and epistemic potential of the respective disciplines opens up new possibilities for the study of linguistic variation and cultural dynamics as a whole.

Bibliography

Hovy, Dirk / Purschke, Christoph (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. Proceedings of EMNLP 2018. Brussels, 4383–4394. https://www.aclweb.org/anthology/D18–1469/

Purschke, Christoph / Hovy, Dirk (2019). Lörres, Möppes, and the Swiss. (Re)Discovering Regional Patterns in Anonymous Social Media Data. In: Journal of Linguistic Geography 7/2, 113–134. doi: 10.1017/jlg.2019.10