Computational approaches to language variation

Yves Scherrer, Olli Kuparinen, Aleksandra Miletić, Erofili Psaltaki, Dana Roemling & Janine Siewert

The goal of this workshop is to bring together researchers working on different aspects of linguistic variation in Natural Language Processing (NLP) on the one hand, and in dialectology, sociolinguistics and related fields on the other. We aim to discuss the state of the art in processing and modeling language variation in its multiplicity, and to showcase its potential for linguistics research.

In NLP, important advances have been made in recent years thanks to neural algorithms and large language models. Nevertheless, engaging with linguistic variation remains one of the crucial research gaps. For a long time, variation was treated as noise and often dealt with at the preprocessing stage by normalising to the standard variety (van der Goot et al. 2021), but recently the NLP community has developed more interest in varied language material (Zampieri & Nakov (eds.) 2021, and more generally the VarDial workshop series). Producing models able to handle multiple axes of variation at the same time remains a challenge.

Large-scale computational processing and analysis have been successfully applied in the field of dialectometry for several decades (Wieling & Nerbonne 2015), but the wider field of sociolinguistics has only recently started to make more systematic use of computational methods (Nguyen et al. 2016). In this context, computational approaches provide relatively objective, high-level analyses. Nevertheless, tracing them back to linguistic features remains challenging (Prokic et al. 2012; Rumpf et al. 2009; Kuparinen & Scherrer 2024).

We welcome contributions on the computational and statistical analysis of linguistic variation and change, as well as on the development of NLP models and tools that handle variation. The domains of language variability discussed in this workshop include: regional (dialects), orthographic (non-standardized writing practices), stylistic (individual repertoires) and diachronic (change over time).

The workshop is organized by the CorCoDial (Corpus-based Computational Dialectology) project, funded by the Research Council of Finland from 2021 to 2025.

References

Van der Goot, R., Ramponi, A., Zubiaga, A., Plank, B., Muller, B., San Vicente Roncal, I., Ljubešić, N., Çetinoğlu, Ö., Mahendra, R., Çolakoğlu, T., Baldwin, T., Caselli, T., & Sidorenko, W. (2021). MultiLexNorm: A Shared Task on Multilingual Lexical Normalization. In

Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), 493-509.

Kuparinen, O., & Scherrer, Y. (2024). Corpus-based dialectometry with topic models. *Journal of Linguistic Geography*, 1-12.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & De Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537-593.

Prokić, J., Çöltekin, Ç., & Nerbonne, J. (2012, April). Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 72-80).

Rumpf, J., Pickl, S., Elspaß, S., König, W., & Schmidt, V. (2009). Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, (H. 3), 280-308.

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annu. Rev. Linguist.*, *1*(1), 243-264.

Zampieri, M., & Nakov, P. (eds.) (2021). *Similar Languages, Varieties, and Dialects: A Computational Perspective*. Cambridge: Cambridge University Press.