Carolina Zucchi

**Extracting morphological features from published grammars of African Arabic dialects: methodological considerations**

In Africa, Arabic dialects are spoken in a vast geographical area spanning roughly from Morocco and Mauritania in the west to Egypt and Sudan in the east, including the Lake Chad area. Quantitative comparisons of the morphological systems of these dialects through linguistic distance measurements could provide valuable insights into linguistic variation in the region and the historical relationships among varieties, while providing a test for traditional linguistic classifications. Typically, such dialectometric studies are based on data drawn from dialect atlases (e.g. Spruit et al. 2009, Jeszenszky et al. 2017, Derungs et al. 2019), where the selection of categorical variables is based on the taxonomies created by the authors of the atlases themselves. Here, I take a different approach and turn to published grammatical descriptions as a data source for the comparison of morphological systems pertaining to Arabic dialects spoken throughout Africa. This choice is motivated partly by the lack of more suitable published sources — the only dialect atlas available for this region being for Egypt (Behnstedt and Woidich 1985) — and by practical difficulties in conducting detailed dialect surveys in such a vast area. I composed a dataset based on 46 published dialect descriptions produced in the course of the 20th and early 21st centuries and extracted linguistic data for comparison across 50 dialects. This presentation will discuss the challenges involved in compiling a dataset from published sources produced by different scholars and discuss how such feature extraction process, despite not being automated per se, facilitates subsequent automated analyses, such as the identification of dialectal clusters and extraction of distinctive features using Pointwise Mutual Information (PMI). Drawing examples from this dataset, I will discuss issues related to creating a taxonomy suitable for dialectometry based on different sources, the choice of a suitable type of encoding and the potential problems collinearity and contingencies between variables can cause for the computation of dialect distances. The discussion aims to highlight the advantages and constraints of this approach, seeking to enrich the dialogue on refining automatic methods in linguistic feature extraction.

## References:


- Derungs, Curdin, Christian Sieber, Elvira Glaser, Robert Weibel. 2020. “Dialect borders—political regions are better predictors than economy or religion”. *Digital Scholarship in the Humanities*, vol. 35(2): 276–295.
- Behnstedt, Peter and Manfred Woidich. 1985. *Die ägyptisch-arabischen Dialekte Band. 2, Dialektatlas von Ägypten.* Wiesbaden: Reichert.
- Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser and Robert Weibel. 2017. “Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German”. *Journal of Linguistic Geography*, 5(02): 86-108.
- Spruit, Marco René, Wilbert Heeringa and John Nerbonne. 2009. “Associations among linguistic levels”. *Lingua*, 119: 1624-1642.