From Archives to Algorithms: Linguistic Data Extraction with AI

Christa Schneider (University of Bern)

The advent of machine learning has profoundly reshaped historical sociolinguistics, particularly in the extraction and analysis of information from premodern records. This presentation explores the intersection of historical sociolinguistics and machine learning, emphasizing the application of contemporary computational methods to vast and unstructured historical datasets. I address key challenges posed by historical texts, including linguistic variation, sparse and incomplete data, and semantic ambiguity characteristic of premodern corpora. These challenges are especially acute when employing machine learning models typically pre-trained on modern language corpora, which often fail to account for the linguistic diversity and diachronic changes inherent in historical records. Through a combination of tailored approaches and domain-specific adaptations, I illustrate how these challenges can be mitigated to enhance the utility of computational tools for historical linguistic analysis.

The talk will focus on Information Extraction (IE) methodologies, particularly Named Entity Recognition (NER) and Sentiment Analysis (SA), with an additional focus on the limitations encountered in Part-of- Speech (POS) tagging. Using the Bernese Witch Paper Corpus and the Salem Witch Trial Papers as primary sources, I demonstrate how NER can effectively trace the evolution of personal names, social roles, and gender markers in early modern texts. This analysis offers not only a diachronic linguistic perspective but also a deeper understanding of how socio-cultural shifts are reflected in language use over time. Sentiment Analysis, though less conventional in historical sociolinguistics, presents an innovative avenue for research. Applied to the Salem Witch Trials Papers, SA reveals quantifiable insights into emotional tones and subjective language, providing a nuanced perspective on the affective dimensions of these historical texts. However, this raises significant questions regarding the reliability and validity of modern sentiment lexicons in the analysis of premodern corpora.

Moreover, I discuss the potential of POS tagging as a tool for historical sociolinguistics. Despite its promise, my experiments revealed challenges of applying POS tagging to historical texts due to linguistic variability and data sparsity, highlighting the limitations of pre-trained models in handling historical language. This reflects a broader need for linguistic tools designed explicitly for historical data contexts.

Finally, I explore future directions for integrating large language models, such as GPT, with historical sociolinguistic corpora. While these models offer remarkable potential for uncovering patterns in historical texts, their application demands a critical examination of biases, ethical considerations, and their adaptability to the unique characteristics of historical data.

This presentation underscores the methodological innovation and interdisciplinary collaboration required to effectively leverage machine learning for historical sociolinguistics. By bridging computational tools and linguistic expertise, we can illuminate previously inaccessible aspects of historical language, offering more nuanced insights into the socio-cultural dynamics of past societies.