**Linguistic Variability on Social Media - Constructing and Analysing Corpora**

Tatjana Scheffler, Universität Potsdam

Social media such as Twitter, Facebook, blogs, forums, etc. have become huge collections of user-generated data online. Since many more people participate on these channels than in traditional media, this provides unique opportunities for studying linguistic variation. At the same time, the large amounts of data and the informal types of content pose specific challenges for researchers trying to collect and analyze social media texts. In this talk, we present computational linguistic methods for collecting specialized social media corpora and for analyzing the resulting data sets linguistically.

In the presentation, we show how we constructed a corpus of German social media data from the same users in Twitter and blogs, allowing us to study linguistic variability within the same individual and across media channels. Based on comparisons of social media corpora with traditional texts and spoken language, we have identified several phenomena in which social media conversations differ from texts: For example, they contain questions, particles, fill words, informal language and alternative spellings. These non-standard features of social media text are challenging for automatic processing, and we discuss approaches for either normalizing the data or adapting the tools to the specific properties of the data.