New sources of large-scale evidence: Introducing a real-time global Bluesky stream

Research on language diversity increasingly relies on social media data, however, two major challenges persist. First, data collection often relies on APIs, which have recently undergone significant closures (Bruns 2019). Second, data obtained through APIs is typically a sample distributed at the discretion of commercial tech companies, often in ways that lack transparency. One widely used source of data has been geolocated Twitter (X), and numerous studies over the past decade have greatly advanced our understanding of linguistic diversity (Grieve et al. 2018; Laitinen & Fatemi 2020; Würschinger 2021). However, free access to this data has ended, forcing researchers to seek alternative sources.

This presentation introduces data collection from Bluesky, a social networking site resembling Twitter. Bluesky addresses a key disadvantage of earlier data collection, since it captures all interactions globally. This enables the construction of a massively complex and extensive corpus of social media data. Data collection began in August 2025 and produces a global, real-time stream of all public user actions and user-generated content, allowing for the observation of interactional dynamics across regions, communities, and time. Currently, the dataset includes activity from about 9 million users worldwide and grows by roughly 1.5 billion words each month, making it one of the largest real-time linguistic data sources available.

The presentation outlines how the data are collected and structured. We also describe our data enrichment process, in which we integrate network information to build a large-scale ego network dataset. This dataset captures complete personal networks, including user connections and the textual content within those networks. Together, these resources provide a new large-scale social media data source to complement existing ones, enabling analysis not only of which linguistic forms appear, but also where and with whom they circulate. This presentation will be of interest to scholars exploring how emerging digital data can be responsibly and effectively used in the study of language diversity.

References

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Grieve, J., Nini, A., & Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Laitinen, M., Fatemi, M., & Lundberg, J. (2020). Size Matters: Digital Social Networks and Language Change. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00046

Würschinger, Q. (2021). Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.648583