Medium und Register als Einflussfaktoren intraindividueller Variation

Hannah Seemann, Universität Tübingen

Autorschaftserkennung macht sich für die Identifizierung einzelner Autor:innen individuelle Variation im Sprachgebrauch zu Nutze. Gleichzeitig wurde Variation durch extratextuelle Faktoren als Problem für die Autorschaftserkennung beschrieben (Overdorf/Greenstadt 2016). Erschwerend kommt hinzu, dass die Texte von Autor:innen sich nicht nur zwischen verschiedenen genutzten Medien unterscheiden können, sondern dass auch das gewählte Register, der situative Kontext von Kommunikation (Biber/Conrad 2019), den Sprachgebrauch beeinflusst (Scheffler et al. 2022). Andererseits wurde gezeigt, dass Unterschiede im Register genutzt werden können, um verlässlich zwischen verschiedenen Autor:innen unterscheiden zu können (Grieve 2023). Um diese verschiedenen Ergebnisse zusammenzuführen, vergleicht meine Analyse den Einfluss von Medium und Register auf den Sprachgebrauch einzelner Individuen.

Zu diesem Zweck verwende ich das TwiBloCop (Scheffler et al. 2023), ein Korpus, das deutsche Texte von 44 Autor:innen aus zwei Medien (Blogs und Tweets) umfasst. Die einzelnen Blogposts sowie die gesammelten Tweets der enthaltenen Autor:innen sind hinsichtlich des verwendeten Registers annotiert: INFORMATIV, ERZÄHLEND oder ÜBERZEUGEND. Entsprechend liegen von einer Person mindestens zwei bis maximal vier verschiedene Dokumente vor: Blogposts in bis zu drei verschiedenen Registerdimensionen und zusätzlich die Sammlung der Tweets. Basierend auf den Bi- und Trigrammen dieser Dokumente führe ich eine Principal Component Analyse (PCA) durch. Der Output dieser Analyse ist in zweierlei Hinsicht aufschlussreich: Zum einen veranschaulicht die visuelle Darstellung der Cluster, welche Texte von Autor:innen sich ähneln. So können die Texte aus einem Register oder aus einem Medium näher beieinander liegen. Zum anderen zeigt die PCA auf, welche der gewählten Input-Dimensionen die meiste Variabilität erklären. Diese Analyse führe ich sowohl auf der Ebene einzelner Autor:innen als auch gepoolt für alle Dokumente im Korpus durch. Die Ergebnisse erlauben somit Rückschlüsse darauf, wie die extratextuellen Faktoren Medium und Register individuellen Sprachgebrauch *einzelner* Autor:innen beeinflussen. Darüber hinaus zeigt die Analyse, ob die intraindividuelle Variation *verschiedener* Autor:innen in vergleichbarer Weise von den beiden Faktoren Medium und Register beeinflusst wird.

Ausgewählte Literatur

Biber, Douglas/Conrad, Susan (2019): Register, Genre, and Style (= Cambridge Textbooks in Linguistics). 2. Aufl. Cambridge: Cambridge University Press.

Grieve, Jack (2023): Register variation explains stylometric authorship analysis. In: Corpus Linguistics and Linguistic Theory 19 (1), S. 47–77. https://doi.org/10.1515/cllt-2022-0040.

Overdorf, Rebekah/Greenstadt, Rachel (2016): Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. In: Proceedings on Privacy Enhancing Technologies 2016 (3), S. 155–171. https://doi.org/10.1515/popets-2016-0021.

Scheffler, Tatjana/Kern, Lesley-Ann/Seemann, Hannah (2022): The medium is not the message: Individual level register variation in blogs vs. tweets. In: Register Studies 4 (2), S. 171–201. https://doi.org/10.1075/rs.22009.sch.

Scheffler, Tatjana/Kern, Lesley-Ann/Seemann, Hannah (2023): Individuelle linguistische Variabilität in sozialen Medien. Ein multimediales Korpus. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022 (= CLIP). Tübingen: Narr. S. 89–99. https://doi.org/10.24053/9783823396024.