Annotation und Analyse eines deutschsprachigen Korpus der Alltagskommunikation mithilfe eines *Query-Enabled Annotation Framework for Text Processing*

Gaby Axer, Universität Mannheim, gabyaxer@mail.uni-mannheim.de
Reinhold Schlager, OTH Regensburg, reinhold.schlager@st.oth-regensburg.de

Qualitative Analysemethoden zur Autorenerkennung basieren auf einer Einschätzung der Aussagekraft gefundener Merkmale auf Basis einer Fehler- und Stilanalyse. Zu einer solchen Einschätzung fehlen jedoch in vielen Fällen noch systematisch erhobene Populationsdaten zur Verteilung solcher Merkmale. Um diese Lücke zu füllen, soll im Rahmen einer Promotionsarbeit ein für die Autorenanalyse aufbereitetes Korpus der Alltagskommunikation mit Hauptfokus auf Instant Messaging erstellt werden, welches möglichst viele Autoren mit möglichst vielen verschiedenen Gesprächspartnern abbildet und entsprechende demographischen Daten beinhaltet (Alter, Geschlecht, (Erst-)Sprache(n), Dialekte, Wohnort, Bildungsgrad, Beruf). Somit sollen Intra- und Inter-Autor-Variation untersucht werden können, um Grundlagenforschung zur Belastbarkeit von Autorschaftsmerkmalen für die forensische Autorenerkennung zu ermöglichen.

Hierzu soll das Korpus so erstellt werden, dass es auf mehreren Ebenen annotiert werden kann: Nicht nur die klassische Wortform-/Lemma-/POS-Annotation, sondern ebenso Annotationen zur Fehler- und Stilanalyse auf den Ebenen Orthografie, Interpunktion, Wortwahl, Grammatik und Syntax sollen umgesetzt werden, auf deren Basis später Populationsverteilungen einzelner Merkmale, Merkmalsklassen und -kombinationen im (Sub-)Korpus erhoben werden können.

Zur Unterstützung von Annotation und Analyse wird ein anfragebasiertes Open-Source-Projekt zur Textverarbeitung eingesetzt, welches im Rahmen der Abschlussarbeit "A Query-Enabled Annotation Framework for Text Processing" entwickelt wird. Diese Software ermöglicht eine flexible, zielgerichtete und komplexe Abfrage sprachlicher Merkmale von Instant-Messaging-Nachrichten. Ebenfalls können diese Nachrichten auf Basis solcher Abfragen annotiert werden, um relevante sprachliche Merkmale strukturiert zu erfassen und für weiterführende Analysen nutzbar zu machen. Dies unterstützt somit sowohl die allgemeine korpuslinguistische Erschließung als auch die Validierung von Hypothesen zur Häufigkeit stilbezogener und grammatikalischer Phänomene in relevanten Populationen.

Das Poster wird den Entwicklungstand dieses Open-Source-Projects und die daraus resultierenden Anwendungs- und Erweiterungsmöglichkeiten im Bereich der Autorenanalyse illustrieren.

Literatur

Dern, Christa. 2009. *Autorenerkennung: Theorie und Praxis der linguistischen Tatschreibenanalyse*. Stuttgart/München: Richard Boorberg Verlag.

Fobbe, Eilika. 2011. *Forensische Linguistik: Eine Einführung*. Tübingen.

Krieg-Holz, Ulrike & Udo Hahn. 2016. CodE Alltag: Ein deutsches E-Mail-Korpus für die Forensische Linguistik. In Lars Bülow, Jochen Bung, Rüdiger Harnisch & Rainer Wernsmann. *Performativität in Sprache und Recht*. 245-264. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110464856-013