COrLiCo – Empirische Analysen im Nähe-Distanz-Kontinuum

Sarah Broll & Roman Schneider
Leibniz-Institut für Deutsche Sprache, Mannheim
broll@ids-mannheim.de, schneider@ids-mannheim.de

Gesprochene und geschriebene Sprachäußerungen unterscheiden sich hinsichtlich Repertoire und Verwendung von syntaktischen Konstruktionen oder Vokabular, stilistischen Merkmalen und einigem mehr. Typisch lautliche Ausdrucksmöglichkeiten und Strategien auf der einen Seite kontrastieren mit einer (zumindest angenommenen) größeren Nähe zu Sprachstandards und Konventionen. Eine Überwindung des medialen Binarismus erlaubt der von Koch und Oesterreicher (1985) eingeführte Ansatz eines mehrdimensionalen Kontinuums zwischen den beiden Polen "Sprache der Nähe" und "Sprache der Distanz". Dieses Kontinuum erforschen wir empirisch anhand einer heterogen stratifizierten Datengrundlage.

Wir stellen das im Aufbau befindliche COrLiCo (*Corpus for the Oral-Literate Continuum*) vor, das sich mit einer Zielgröße von 100 Millionen Wort-Tokens zu ungefähr gleichen Teilen auf knapp 20 textsortenspezifische Subkopora aufteilt. Abgedeckt werden nicht nur die Pole des Kontinuums, sondern auch solche Sprachäußerungen, die sich nicht eindeutig einem dieser Extreme zuordnen lassen: Wissenschaftskommunikation, Interviews, Reden, Liveticker, Songtexte, Social Media, E-Mails, Online-Diskussionen, Podcasts etc. Das Korpus repräsentiert damit eine sehr breit gefächerte Datengrundlage und füllt eine Lücke als aggregierende Ressource für die vergleichende Erforschung schriftlicher und mündlicher Diskurse. Sämtliche Texte sind mit Metadaten und Annotationen angereichert, die zum einen anhand der Textoberfläche, zum anderen auf der Basis des Outputs maschineller Tagger berechnet werden.

Wir implementieren und evaluieren ein automatisiertes Klassifikationsverfahren, das unter Nutzung von Random-Forest-Entscheidungsbäumen Aggregationen einzelner Vorhersagen durchführt (Broll & Schneider, o. J.). Für die Identifizierung der Pole definieren wir einen Merkmalskatalog aus Sprachphänomenen, die als Markierer für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit diskutiert werden. Basierend auf der sehr guten Klassifikationsgüte verorten wir eine Reihe weiterer Textsorten. Die Ergebnisse werden zur besseren Interpretierbarkeit visuell aufbereitet.

References: • Broll, S., & Schneider, R. *Empirische Verortung konzeptioneller Nähe/*Mündlichkeit *inner- und außerhalb schriftsprachlicher Korpora*. (in Vorbereitung) • Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. Romanistisches Jahrbuch, 36(1), 15–43. https://doi.org/10.1515/9783110244922.15.