58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 15. bis 17. März 2022
(als Online-Konferenz)

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

METHODENMESSE

Mittwoch, 16. März 2022, 15:45 Uhr bis 17:45 Uhr

Das deutsch-dänische XPEROHS-Korpus: Hassrede in sozialen Medien

Eckhard Bick / Klaus Geyer (Süddänische Universität)

Das hier vorgestellte Korpus wurde mit dem Ziel erstellt, Sprachgebrauch in sozialen Medien qualitativ und quantitativ auf Hassrede gegen ethnische und religiöse Minderheiten untersuchen zu können. Das zweisprachige, deutsch-dänische Korpus bildet die datenmäßige Grundlage des XPEROHS-Projekts (*Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech*, Baumgarten et al. 2019). Es handelt sich um ein Monitor-Korpus von Twitter- und Facebook-Einträgen über den Zeitraum Ende 2017 bis 2020, wobei die Abfrageschnittstelle für Twitter einen fast 100-prozentigen Deckungsgrad ermöglichte, während Facebook eine Vorauswahl bestimmter Webseiten und damit einen thematischen Bias erzwang. Insgesamt enthalten die Twitter-Korpora ca. 2 Milliarden Wörter für Deutsch und 270 Millionen Wörter für Dänisch, die kleineren Facebook-Korpora ca. 200 Millionen Wörter für Deutsch und 60 Millionen Wörter für Dänisch.

Um gezielte linguistische Statistiken und die qualitative Identifizierung linguistischer Muster von Hassrede zu erlauben, wurde das Korpus mehrfach grammatisch und lexikalisch-semantisch mit Hilfe von Constraint-Grammar-Parsern annotiert (Bick 2020a), wobei Grammatiken und Lexika laufend an spezielle sprachliche Merkmale des Genres (z. B. unvollständige Sätze, orthographische Variationen, kreative Wortbildungen und nicht-textliche Kürzel) sowie an den inhaltlichen Distinktionsbedarf des Projektes angepasst wurden. Neben Lemmatisierung und standardmäßiger morphosyntaktischer Annotation wurde auch auf semantische Aspekte und eine genaue morphologische Analyse wert gelegt. So werden die Bestandteile von Komposita erkannt und Inhaltswörter semantisch klassifiziert und disambiguiert, z. B. Substantive als 'Berufsbezeichnung', 'Tier', 'Werkzeug', 'Fahrzeug' oder 'Gebäude'. Verben (und prädizierende Substantive) erhalten eine Framenet-Kategorie und einen Satz von semantischen Rollen für ihre Argumente (Bick 2011 und Bick 2017). Besonderes Augenmerk lag auch auf der Unterstützung von Sentiment Analysis, z.B. über die Polarität von Adjektiven und indem Text-Emoticons und bildliche Emojis morphosyntaktisch in die Satzanalyse eingebunden und als zehn verschiedene Emotionen "lemmatisiert" wurden (Bick 2020b).

Als grafische Benutzerschnittstelle wurde eine erweiterte Version von Corpuseye (corp.hum.sdu.dk) verwendet, die intern auf einer CQP-Datenstruktur (Hardie 2012) basiert. Die Projektversion enthält neben Feld- und Tag-basierten Suchoptionen auch Neuerungen wie gestaffelte Suchen, N-gram-Auswertung, die Unterstützung von Emojis und den Zugang zu Dependenzrelationen.

In unserer Präsentation besprechen wir eine Reihe von korpusspezifischen Herausforderungen datenrechtlicher, technischer und korpuslinguistischer Art und zeigen entsprechende Lösungsansätze auf. Schließlich erläutern wir an Beispielen, wie das Korpus in der Hassredeforschung eingesetzt werden kann, z.B. bei der Identifikation und Exemplifizierung von indirekter Hassrede durch konstruktionsgrammatische Muster (Bick / Geyer / Kleene 2021 und Geyer / Bick / Kleene i. Dr.), bei der Erkennung bestimmter metaphorischer Prädikationen oder bei der Unterscheidung von Erwähnung vs. Verwendung (*mention* vs. *use*) von Ethnopaulismen (Geyer 2021).

Bibliographie:

Baumgarten, Nicole / Bick, Eckhard / Geyer, Klaus / Iversen, Ditte Aakær / Kleene, Andrea / Lindø, Anna Vibeke / Neitsch, Jana / Niebuhr, Oliver / Nielsen, Rasmus / Petersen, Nedenskov, Esben (2019). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). In: *RASK – International Journal of Language and Communication* 50, 87-108.

Bick, Eckhard (2011). A FrameNet for Danish. In: *Proceedings of NODALIDA 2011*, May 11-13, Riga, Latvia. NEALT Proceedings Series, Vol 11, 34-41. Tartu: Tartu University Library.

Bick, Eckhard (2017). Propbank Annotation of Danish Noun Frames. In: *Proceedings of IWCS2017 – 12th International Conference on Computational Semantics* (Montpellier, September 2017). ACL Anthology W17-69. URL: http://aclweb.org/anthology/W17-69.

Bick, Eckhard (2020a). An Annotated Social Media Corpus for German. In: Calzolari, Nicoletta et al. (eds.), *Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC2020* (Marseille, May 2020), 6129-6137. ACL / ELRA.

Bick, Eckhard (2020b). Annotating Emoticons and Emojis in a German-Danish Social Media Corpus for Hate Speech Research. In: *RASK – International Journal of Language and Communication* 52, 1-20.

Bick, Eckhard / Geyer, Klaus / Kleene, Andrea. (2020). „Die ách so friedlichen Muslime“: eine korpusbasierte Untersuchung von Formulierungsmustern fremdenfeindlicher Aussagen in Sozialen Medien. In: Wachs Sebastian / Koch-Priewe Barbara / Zick, Andreas (Hrsg.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen*. Heidelberg: Springer VS. 81-103

Geyer, Klaus (2021). Verwendung (*use*) vs. Erwähnung (*mention*) von Ethnophaulismen: Eine Untersuchung zu Hassrede in Facebook-Beiträgen. In: Lüger, Heinz-Helmut & Giessen, Hans Werner (Hrsg.), *Text-, Diskurs- und Kommunikationsforschung: Festschrift für Hartmut Lenk*. Landau: Verlag Empirische Pädagogik, 165-184.

Geyer, Klaus / Bick, Eckhard / Kleene, Andrea (i. Dr.). “I am not a racist, but …”. A Corpus-Based Analysis of Xenophobic Hate Speech Constructions in Danish and German

Social Media Discourse. In: Natalie Knoblock (ed.), *Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse. Cambridge: C*ambridge University Press.

Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17 (3), 380-409.