KorDaF Korpora Deutsch als Fremdsprache

Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 4, Nummer 2 (Dezember 2024)
ISSN 2749-4411

DAS TWIBLOCOP
Ein multimediales Korpus aus Blogposts und Tweets

Tatjana Scheffler, Hannah J. Seemann
Ruhr-Universität Bochum

Abstract

Das Korpus TwiBloCoP besteht aus deutschsprachigen Blogposts sowie Tweets von 44 Autor:innen, die parallel in beiden Plattformen aktiv waren. Es wurde 2017 erhoben und enthält Beiträge zum Familienalltag sowie zur Elternschaft. Insgesamt umfasst das Korpus über 81.000 Tweets und fast 500 Blogposts. Die Posts wurden anonymisiert, jedoch bleiben Autorschaftsbezüge medienübergreifend durch zufällige IDs erhalten. Im Korpus wurden drei zu den Medien quer stehende Register identifiziert: informierend, erzählend und überzeugend. Modalpartikeln und Intensivierer wurden zusätzlich manuell annotiert. Alle Daten stehen auf Anfrage im XML-Format für weitere Forschungen zur Verfügung.

Keywords: Soziale Medien; Twitter; Blog; Familie; Register; Modalpartikeln; Intensivierer

Abstract

The TwiBloCoP corpus consists of German blog posts and tweets from 44 authors who were active on both platforms at the same time. It was collected in 2017 and contains posts centered around everyday family life and parenting. In total, the corpus comprises over 81,000 tweets and almost 500 blog posts. The posts were anonymized, but common authorship information was retained across the media using random IDs. Three cross-media registers were identified in the corpus: informative, narrative and persuasive. In addition, modal particles and intensifiers were manually annotated. All data is available on request in XML format for further research.

Keywords: social media; Twitter; blog; family; register; modal particles; intensifiers

1. Einleitung

Das TwiBloCoP (*Twitter+Blog Corpus – Parenting*) enthält deutschsprachige Tweets und Blogposts von 44 Autor:innen aus der Elternblogger-Blogosphäre. Die Autor:innen schreiben in beiden Kommunikationskanälen über kinder- und familienbezogene Themen oder berichten von ihrem (Familien-)Alltag. Eine spezifische Community von Autor:innen wurde gewählt, um medienspezifische Variation im Inhalt der Texte zu minimieren. Das Thema Elternschaft bot sich an, da zum Zeitraum der Datenerhebung viele Personen zu diesem Themenbereich auf Twitter aktiv waren, die ebenfalls einen Blog betreiben.

2. Datensammlung und Inhalt

Die Daten wurden im Februar 2017 gesammelt und umfassen Posts aus den vier vorhergegangenen Monaten. Um jeweils sowohl Blogs als auch Tweets derselben Personen zu sammeln, extrahierten wir aus der Twitter-Liste *Elternbloggerkarte* eine Menge von Autor:innen, die sowohl auf Twitter aktiv waren, als auch einen eigenen Blog betreiben. Die Rohdaten aus beiden Medien wurden mithilfe von Python-Skripten gesammelt: Blogposts wurden (wenn möglich) über einen RSS-Feed ausgelesen, die Tweets wurden durch die Twitter-API gesammelt. Die Verlinkung von Blogposts sowie

Tweets erfolgte basierend auf der Selbst-Verlinkung des Blogs in der Twitter-Biografie der Autor:innen. Pro Autor:in sind ca. 10 Blogposts und ca. 1800 Tweets vorhanden.

Da die vorliegenden Daten möglicherweise sensible Inhalte enthalten, haben wir zwei Schritte zur Sicherung des Datenschutzes unternommen. Alle Autor:innen wurden per Opt-Out-Verfahren um Zustimmung zur Nutzung ihrer Texte gebeten. Nach Ausschluss der Personen, die nicht kontaktiert werden konnten oder der Datennutzung widersprachen, befinden sich Texte beider Medien von 44 Personen im Korpus. Die verbleibenden Texte wurden teils automatisch (Usernamen, E-Mailadressen, Telefonnummern, URLs), teils manuell (Personennamen, Ortsnamen) anonymisiert und Namen der Blogs pseudonymisiert. Namen von Personen der Öffentlichkeit wie Politiker:innen oder historische Personen wurden nicht ersetzt.

Alle Texte wurden mithilfe des Python-Pakets SoMaJo[1] (vgl. Proisl / Uhrig 2016) automatisch satzsegmentiert und tokenisiert. Tabelle 1 zeigt eine Übersicht über die Größe des resultierenden Korpus.

	Twitter	Blog
Posts	81.440	468
Sätze	137.914	24.981
Token	1,2 Mio.	360.000

Tabelle 1
Aufbau und Inhalt des TwiBloCoP

3. Annotationen

Das Korpus umfasst Texte aus zwei unterschiedlichen Medien, die zusätzlich in ihrem Register variieren, wobei Register den situativen Kontext einer Kommunikation beschreibt (vgl. Biber / Conrad 2019). Autor:innen passen sich an ihre Gesprächspartner:innen und den Gesprächskontext an, ebenso variiert der Sprachgebrauch je nach kommunikativem Ziel. Basierend auf dieser Definition haben wir im Korpus drei Register identifiziert: INFORMIEREND, ERZÄHLEND und ÜBERZEUGEND. Diese unterscheiden sich im Grad der Involviertheit der Autor:innen sowie im emotionalen Gehalt der behandelten Themen. Die Registerdimensionen sind manuell für jeden Blogpost annotiert, Tweet-Sammlungen (pro Autor:in) erhalten das Register als Label, welches den Großteil der einzelnen Tweets am besten beschreibt.

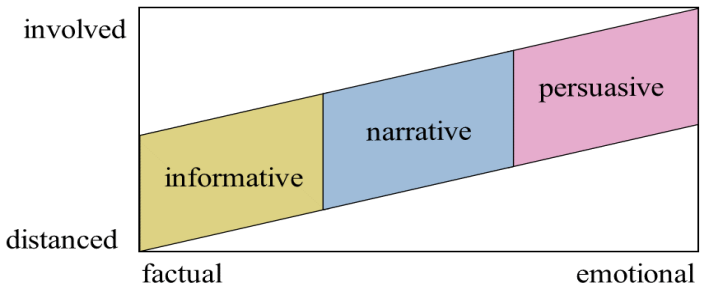

Abbildung 1
Annotierte Registerdimensionen im Korpus (aus Scheffler / Kern / Seemann 2022).

[1] https://github.com/tsproisl/SoMaJo (07.11.2024).

Darüber hinaus sind im Korpus Modalpartikeln sowie Intensivierer manuell annotiert. Modalpartikeln (1)-(2) drücken Einstellungen von Autor:innen aus oder werden genutzt, um Erwartungen und Wissen im Diskurs zu verhandeln (vgl. Zimmermann 2011). Intensivierer (3)-(4) verstärken oder schwächen die Intensität einer Aussage (vgl. Breindl 2007). Beide Phänomene sind typisch für informelle Sprache in sozialen Medien. Die Häufigkeit einzelner Modalpartikeln wie Intensivierer variiert zwischen den Registerdimensionen, teilweise auch zwischen den beiden im Korpus enthaltenen Medien (vgl. Seemann / Scheffler 2022).

(1) @[USERNAME] Hier ist's **ja** auch warm. Aber Frühling ist ab Mai 😔
[tweets-6317]

(2) […] Schaut euch **doch** nur diese Designs an!
[blogposts-9065]

(3) Einschlafgespräch K2: 'K1 und mich hast du **ganz doll** lieb.
[tweets-1611]

(4) Sie ist **so** süß!
[blogposts-4308]

Eine detailliertere Beschreibung und Auswertung der Annotationen ist in Scheffler / Kern / Seemann (2022) zu finden.

4. Formate und Verfügbarkeit

Das Korpus liegt im Textformat und in einer XML-Struktur vor, welche sich am TEI-CMC Schema (vgl. Beißwenger et al. 2012; Beißwenger / Lüngen 2020) orientiert. Als Metadaten sind der Titel des Dokuments, die den Autor:innen zugewiesene ID, das Medium des Dokuments, sowie Zeitraum der Erstellung sowie Erhebung des Dokuments verzeichnet. Weiterhin wurden in Blogposts Absätze und Sätze, bzw. in Tweets sowohl Posts als auch Sätze sequenziell nummeriert und mit einer eindeutigen ID versehen. Verlinkte Medien wie Bilder, Videos und GIFs wurden nicht gespeichert und die zugehörigen Links wurden entfernt. Ein Beispiel des Datenformats ist in Scheffler / Kern / Seemann (2023) abgebildet. Die annotierte Registerdimension ist am jeweils zugehörigen Eintrag (Blogpost oder Tweetsammlung) vermerkt. Die Annotation von Modalpartikeln und Intensivierern liegen im CoNLL-Format[2] mit Bezug auf die Dokument- und Satz-IDs vor.

Die Daten stehen zur wissenschaftlichen Forschung zur Verfügung und können per Anfrage an tatjana.scheffler@rub.de erhalten werden. Aktuelle Informationen sind auf der Webseite des Korpus zu finden[3].

5. Nutzungsbeispiel mit DaF-Bezug

Durch die Sammlung von Texten verschiedener Medien und Register ermöglich TwiBloCoP Untersuchungen von sprachlicher Variation auf verschiedenen Ebenen. So kann bspw. untersucht werden, wie einzelne Autor:innen ihren Sprachgebrauch an das Medium oder das kommunikative Ziel anpassen. In den Beispielen (5) und (6) behandelt Autor:in 6794 denselben Inhalt, die genutzte Sprache allerdings variiert zwischen den beiden Medien Tweet und Blogpost.

[2] siehe: https://universaldependencies.org/format.html (07.11.2024).
[3] https://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/ (07.11.2024).

(5) "Es ist nur eine Phase..." - wie mich dieser Satz nervt! [URL]

[tweets-6794]

(6) Natürlich geht es nach der Geburt gleich weiter, all die gut gemeinten Ratschläge und so. Mit der Zeit kam ich damit zurecht, aber einen Satz konnte und kann ich nicht leiden: „Es ist nur eine Phase!“ [...] Ich meine dieses altkluge, veteranenmäßige Es ist nur eine Phase, das kam wenn ich wirklich meine Sorgen und Nöte schilderte.

[blogposts-6794]

Weiterhin bietet die vorhandene manuelle Annotation von Modalpartikeln eine Ressource, um diesen häufig für Lerner:innen schwierigen Gegenstand an authentischen Beispielen zu üben (s. bspw. Kresić / Batinić 2014 oder Schoonjans 2021 für entsprechende Überlegungen).

Literatur und Ressourcen

Beißwenger, Michael / Ermakova, Maria / Geyken, Alexander / Lemnitzer, Lothar / Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: *Journal of the Text Encoding Initiative* 3. https://doi.org/10.4000/jtei.476.

Beißwenger, Michael / Lüngen, Harald (2020): CMC-core: a schema for the representation of CMC corpora in TEI. In: *Corpus* 20. https://doi.org/10.4000/corpus.4553.

Biber, Douglas / Conrad, Susan (2019): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Breindl, Eva (2007): Intensitätspartikeln. In: Hoffmann, Ludger (Hrsg.): *Handbuch der deutschen Wortarten*. Berlin /New York: de Gruyter, 397-422.

Kresić, Marijana / Batinić, Mia (2014): *Modalpartikeln: Deutsch im Vergleich mit dem Kroatischen und Englischen*. Zadar: Universität Zadar.

Proisl, Thomas / Uhrig, Peter (2016): SoMaJo: State-of-the-Art Tokenization for German Web and Social Media Texts. In: Cook, Paul / Evert, Stefanie / Schäfer, Roland / Stemle, Egon (eds.): P*roceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Berlin: Association for Computational Linguistics, 57-62.

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, 171-201.

Scheffler, Tatjana / Kern, Lesley-Ann / Seemann, Hannah (2023): Individuelle linguistische Variabilität in sozialen Medien. Ein multimediales Korpus. In: Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*. Tübingen: Narr, 89-99.

Schoonjans, Steven (2021): Abtönungspartikeln im Deutschunterricht für Niederländischsprachige. In: *Germanistische Mitteilungen* 47: 47, 87-119.

Seemann, Hannah / Scheffler, Tatjana (2022): Differentiating Social Media Texts via Clustering. In: Karsdorp, Folgert / Lassche, Alie / Nielbo, Kristoffer (eds.): *CHR 2022: Computational Humanities Research 2022*, 177-188. https://ceur-ws.org/Vol-3290/ (07.11.2024).

Zimmermann, Malte (2011): Discourse Particles. In: Maienborn, Claudia / Heusinger, Klaus v. / Portner, Paul. (eds.): *Semantics*. Berlin: de Gruyter, 2011-2038.

Biographische Notiz: Tatjana Scheffler studierte Computerlinguistik in Saarbrücken, Shanghai und Peking und promovierte in Linguistik an der University of Pennsylvania, USA. Nach einer Zeit in der außeruniversitären Forschung lehrte sie an den Universitäten Potsdam und Konstanz. Seit 2020 ist sie Professorin für Digitale Forensische Linguistik an der Ruhr-Universität Bochum und widmet sich der korpus- und computerlinguistischen Analyse von sprachlichen Daten aus sozialen Medien.

Kontaktanschrift:
Prof. Dr. Tatjana Scheffler
Digitale Forensische Linguistik
Germanistisches Institut
Ruhr-Universität Bochum
Universitätsstraße 150
44801 Bochum
tatjana.scheffler@rub.de

Biographische Notiz: Hannah J. Seemann studierte Germanistik an der Ruhr-Universität Bochum und ist dort seit 2022 wissenschaftliche Mitarbeiterin am Lehrstuhl von Tatjana Scheffler. In ihrer Dissertation untersucht sie den Einfluss von Modalpartikeln auf die Interpretation von Diskursrelationen. Darüber hinaus gilt ihr Forschungsinteresse dem Sprachgebrauch und der Variation in sozialen Medien.

Kontaktanschrift:
Hannah J. Seemann
Germanistisches Institut
Ruhr-Universität Bochum
Universitätsstraße 150
44801 Bochum
hannah.seemann@rub.de