55. Jahrestagung des Instituts für Deutsche Sprache vom 12. bis 14. März 2019

# DEUTSCH IN SOZIALEN MEDIEN

Interaktiv, multimodal, vielfältig

Mittwoch, 13. März 2019, 11:30 Uhr

## IBK- UND SOCIAL-MEDIA-KORPORA AM IDS: STAND UND PERSPEKTIVEN

Harald Lüngen und Marc Kupietz (IDS)

Der Vortrag untersucht vorhandene Lösungen und neue Möglichkeiten des Korpusausbaus aus Social-Media und internetbasierter Kommunikation (IBK) für das Deutsche Referenzkorpus (DeReKo), die Sammlung gegenwartssprachlicher Schriftkorpora am IDS, die der sprachwissenschaftlichen Öffentlichkeit über die Korpusschnittstellen COSMAS II und KorAP angeboten wird.

Anhand von Merkmalen und Beispielen gehen wir zunächst auf die Extensionen und Überlappungen der Konzepte *Social-Media-Korpora*, *Korpora internetbasierter Kommunikation* und *Webkorpora* ein. Wir betrachten die rechtlichen Voraussetzungen für einen Korpusausbau aus sozialen Medien, die sich aus dem kürzlich in relevanten Punkten reformierten deutschen Urheberrecht, aus Leistungsschutzrechten und Persönlichkeitsrechten wie der EU-DSGVO ergeben und stellen Konsequenzen und mögliche und tatsächliche Umsetzungen dar. Der Aufbau von Social-Media-Korpora in großen Textmengen stellt außerdem korpustechnologische Herausforderungen, die für traditionelle Schriftkorpora als gelöst galten oder gar nicht erst bestanden. Wir berichten, wie Fragen der Datenaufbereitung, des Corpus Encoding, der Anonymisierung oder der linguistischen Annotation von Social-Media-Korpora für DeReko angegangen wurden.

Wir betrachten die Korpuslandschaft verfügbarer deutschsprachiger IBK- und Social-Media-Korpora und geben einen Überblick über den Bestand an IBK- und Social-Media-Korpora und ihre Charakteristika (Chat-, Wiki Talk- und Forenkorpora) in DeReKo sowie von laufenden Projekten in diesem Bereich. Anhand zweier Studien zu Zeitverläufen und diskursrelevanten Kollokationen in IBK- versus Zeitungskorpora demonstrieren wir die Relevanz von IBK- und Social-Media-Korpora für die Untersuchung der deutschen Gegenwartssprache.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hrsg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA), 2018. S. 4353-4360

Lüngen, Harald/Kupietz, Marc (2017): CMC Corpora in DeReKo. In: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (Hrsg.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. Mannheim: Institut für Deutsche Sprache, 2017. S. 20-24

Margaretha, Eliza/Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Beißwenger, Michael/Oostdijk, Nelleke/Storrer, Angelika/van den Heuvel, Henk (Hrsg.): Building and Annotating Corpora of Computer-mediated Communication: Issues and Challenges at the Interface between Computational and Corpus Linguistics. Journal for Language Technology and Computational Linguistics (JLCL) 29 (2). Regensburg: GSCL, 2014. S. 59-82

Fankhauser, Peter/Kupietz, Marc (eing.):Analyzing domain specific word embeddings for a large corpus of contemporary German. Manuskript eingereicht für die Corpus Linguistics Conference 2019