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

DEUTSCH IN SOZIALEN MEDIEN
Interaktiv, multimodal, vielfältig

KALEIDOSKOP

Mittwoch, 13. März 2019, 15:00 Uhr

DIE WEBKORPORA IM DWDS – STRATEGIEN DES KORPUSAUFBAUS UND NUTZUNGSMÖGLICHKEITEN

Adrien Barbaresi und Alexander Geyken

Strategien des Korpusaufbaus

Alle betrachteten Webkorpora basieren auf einer Auswahl von Webseiten auf Deutsch (vor allem aus Deutschland, Österreich und der Schweiz). Die Seiten werden durch Webcrawlingverfahren „entdeckt“, also durch die maschinell erkundete deutschsprachige Websphäre. Dabei wird ein Gleichgewicht durch Merkmale (Sampling von Seiten für jede Homepage) und formale Kontrollen angestrebt. Erstens werden qualitativ bessere Dokumente berücksichtigt, die in Textform vorkommen. Außerdem spielen die Metadaten eine wichtige Rolle, beispielsweise müssen die Texte im Kontext der lexikographischen Forschung datierbar sein.

Die sich daraus ergebende Dokumentenbasis besteht aus mehreren hunderttausenden unterschiedlichen Webseiten, die ein Datum aufweisen, das Korpus enthält also vergleichsweise viele Blogeinträge. Die Webseiten werden sowohl professionell (z.B. Nachrichten- und Firmenseiten) als auch privat (Vereine, Gemeinschaften, Hobbys) betrieben, so dass das Korpus Sprechsituationen unterschiedlichster Art abdeckt. Diese Ressource wird fortlaufend verbessert, u.a. im Sinne einer qualitativ feineren Kalibrierung, sowohl inhaltlich als auch auf der Metadatenebene (z.B. Extraktion des Titels und Heuristiken zur Bestimmung des Veröffentlichungsdatums einer Webseite).

Integration in die DWDS-Plattform und Nutzungsmöglichkeiten

Eine Voraussetzung für die Integration von Korpustexten in das DWDS ist deren strukturelle und linguistische Annotation und die Bereitstellung von Metadaten. Die einzelnen Textwörter werden darüber hinaus mit weiteren, für die linguistische Suche relevanten Informationen versehen. Zur Zeit werden für jedes Textwort die Grundform (Lemma) und die Wortart angegeben und von der Suchmaschine indiziert. Die primär für die Zwecke der lexikographischen Arbeit der Projektgruppe erstellten Korpora haben seit der Veröffentlichung der DWDS-Webseite eine weit über diesen Kreis hinausgehende Nutzung erfahren, insbesondere bei den Nutzerinnen und Nutzern des Wörterbuchs, die die Wörterbucheinträge mit den Textquellen vergleichen wollen, aber auch bei Wissenschaftlerinnen und Wissenschaftlern, die die Korpora des DWDS als Quelle korpuslinguistischer Studien nutzen.

Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics, 7-16.

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., & Lemnitzer, L. (2017). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). Zeitschrift für germanistische Linguistik, 45(2), 327-344.