Die Webkorpora im DWDS -Strategien des Korpusaufbaus und Nutzungsmöglichkeiten

Adrien Barbaresi, Alexander Geyken

Adrien Barbaresi & Alexander Geyken

Die Webkorpora im DWDS – Strategien des Korpusaufbaus und Nutzungsmöglichkeiten

Die Kernaufgabe der Projektgruppe des DWDS besteht darin, den in den Korpora enthaltenen Wortschatz lexikographisch und korpusbasiert zu beschreiben. In der modernen Lexikographie werden die Aussagen zu den sprachlichen Aspekten und Eigenschaften der beschriebenen Wörter und zu Besonderheiten ihrer Verwendung auf Korpusevidenz gestützt. Empirisch können riesige Textsammlungen Hypothesen genauer oder ausführlicher belegen. Dabei wird deutlich, wie vielfältig Sprache im Gebrauch tatsächlich realisiert wird. Zu diesem Zweck bieten wir auf der DWDS-Plattform neben den zeitlich und nach Textsorten ausgewogenen Kernkorpora und den Zeitungskorpora eine Reihe von Spezialkorpora an, die hinsichtlich ihres Gegenstandes oder ihrer sprachlichen Charakteristika von den erstgenannten Korpora abweichen. Die Webkorpora bilden einen wesentlichen Bestandteil dieser Spezialkorpora.

Strategien des Korpusaufbaus

Im Vergleich zu klassischen Textkorpora bestehen relevante Fragestellungen des Aufbaus von Webkorpora darin, Genres jenseits der Kategorien konventioneller Korpora zu berücksichtigen, die Ergebnisse zu kalibrieren (insbesondere hinsichtlich von Artefakten und unvollständigen Texten) sowie darin, die Texte mit den richtigen Metadaten zu versehen und gegebenenfalls zu klassifizieren. Alle betrachteten Webkorpora basieren auf einer Auswahl von Webseiten auf Deutsch (vor allem aus Deutschland, Österreich und der Schweiz). Auf der einen Seite gibt es allgemeingültige, universell einsetzbare Korpora im Sinne einer Einheitsgröße, die für eine Vielzahl von Nutzungsszenarien nützlich sein soll. Auf der anderen Seite gibt es spezifische Korpora aus bereits bekannten oder händisch überprüften Quellen, die möglicherweise reichere Metadaten

beinhalten und auf bestimmte Forschungsziele ausgerichtet sind, wie beispielsweise Studien zu internetbasierter Kommunikation oder Sprachvariation.

Es gibt kein umfassendes Verzeichnis von Webseiten oder Blogs, außerdem können sich Webstrukturen schnell ändern. Die Seiten werden also zunächst „entdeckt“, indem die deutschsprachige Websphäre maschinell erkundet wird (*Webcrawling*), und daraufhin bezüglich ihrer Qualität bewertet, Bei diesem Vorgehen wird ein Gleichgewicht durch Merkmale (Stichproben für jede Homepage) und formale Kontrollen angestrebt. Dabei werden qualitativ bessere Dokumente bevorzugt, die zum Beispiel Fließtext beinhalten. Außerdem spielen Metadaten eine wichtige Rolle, beispielsweise müssen die Texte im Kontext der lexikographischen Forschung datierbar sein.

Ergebnisse

Ein „kleineres“ (ca. 100 Mio. laufende Wortformen) Korpus besteht aus Beiträgen und Kommentaren, die in Blogs veröffentlicht worden sind, deren Betreiber die Wiederveröffentlichung der Texte mittels Creative-Commons-Lizenzen ausdrücklich gefördert haben. Für weitere Details zur Erhebungsmethode vgl. Barbaresi & Würzner (2014). Die Dokumentenbasis für ein größeres Webkorpus (derzeit mehr als 3 Mrd. laufende Wortformen) besteht aus mehreren hunderttausenden unterschiedlichen Webseiten, die zumindest eine Datumsangabe aufweisen müssen, um damit die zeitliche Datierung der Korpustreffer zu ermöglichen. 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 auf inhaltlicher als auch auf der Metadatenebene (z.B. Extraktion des Titels und

Heuristiken zur Bestimmung des Veröffentlichungsdatums einer Webseite[1]).

Integration in die DWDS-Plattform und Nutzungsmöglichkeiten

Eine Voraussetzung für die Integration von Korpustexten in die DWDS-Plattform 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. Derzeit 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.

Referenzen

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

Barbaresi, Adrien & Würzner, Kay-Michael (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In: Proceedings of KONVENS 2014, NLP4CMC workshop. Hildesheim, S. 2-10.

Geyken, Alexander, Barbaresi, Adrien, Didakowski, Jörg, Jurish, Bryan, Wiegand, Frank, & Lemnitzer, Lothar (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). Zeitschrift für germanistische Linguistik, 45(2), S. 327-344.

[1] Frei verfügbare Komponente: https://github.com/adbar/htmldate