Webkorpora der zweiten Generation: Erstellung, Evaluation und Ergebnisse

Felix Bildhauer (IDS Mannheim)
Roland Schäfer (FU Berlin)

Webkorpora – also linguistisch aufbereitete Korpora aus WWW-Daten – gibt es seit über zehn Jahren. In diesem Vortrag fassen wir zunächst zusammen, was die wichtigen Entwicklungen in diesen Jahren gewesen sind. Die Zusammenfassung betrifft konzeptuelle und technologische Aspekte, Aspekte der Evaluation von Webkorpora, die Arbeit mit Webdaten in der Linguistik sowie rechtliche Aspekte. Im Mittelpunkt stehen dabei unsere eigenen COW- und COCO-Korpora (Deutsch, Englisch, Niederländisch, Spanisch, Schwedisch) im Vergleich zu anderen Webkorpora (z.B. Derik, Glowbe, SketchEngine, WaCky). Einige erfolgreiche Studien zur Variation in deutscher Morphosyntax und Graphematik, die anhand von DECOW durchgeführt wurden, werden vorgestellt. Im Weiteren geben wir einen Ausblick auf zukünftige Entwicklungen. In der Korpuserstellung betrifft dies zum Beispiel: neue Datenquellen (z.B. CommonCrawl), größere Korpora (ENCOCO1507 wird über 100 Mrd. Tokens enthalten), gezieltes Durchsuchen des WWW nach spezifischen linguistischen Daten, automatische Textklassifikation. Auf der konzeptuellen/linguistischen Seite stehen Fragen nach der Korpuszusammensetzung, der Validität von Ergebnissen aus Webdaten und effiziente Möglichkeiten der Analyse und statischen Auswertung im Vordergrund. Zum Abschluss demonstrieren wir praktisch die verschiedenen Möglichkeiten, mit den COW- und den zukünftigen COCO-Webkorpora zu arbeiten. Dabei wird differenziert zwischen den Zugriffsmöglichkeiten für FU-MitarbeiterInnen, kooperierende Institutionen und die breitere linguistische Öffentlichkeit.