Marc Kupietz (Mannheim) / Thomas Schmidt (Mannheim)

Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung

Die Programmbereiche „Korpuslinguistik“ und „Mündliche Korpora“ haben am IDS die Aufgabe, Grundlagen für die empirische Erforschung der deutschen Sprache zu schaffen. Unter anderem sammeln und erstellen sie schriftliche und mündliche Korpora, bereiten sie für eine wissenschaftliche Nutzung auf und stellen sie über Web-Oberflächen (COSMAS, demnächst KorAP, DGD2) zur Verfügung. Unser Beitrag gibt zunächst einen Überblick über Entstehungsgeschichte und aktuellen Stand dieser Arbeiten. Mit einem Blick in die Zukunft widmen wir uns auch der Frage, ob und in welcher Weise das Schlagwort „Big Data“ für diese Arten linguistischer Ressourcen relevant ist.

In Bezug auf die schriftlichen Korpora wird dabei insbesondere über die diesjährige DEREKO-Erweiterung um über 17 Milliarden Wörter und die damit verbundenen Arbeiten in den Bereichen Korpusausbau, Analysemethodik und Korpustechnologie berichtet. In diesem Zusammenhang werden u.a. Überlegungen zur Dipsersion und Stratifizierbarkeit von DEREKO entlang verschiedener Dimensionen und zur Analyse seltener Phänomene skizziert und Konsequenzen für die praktische linguistische Arbeit diskutiert.

Die spezifischen Herausforderungen, die sich beim Aufbau eines großen Gesprächskorpus stellen, werden am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), das aktuell 100h Aufnahmen oder 1 Million Wörter umfasst, diskutiert. Dabei steht außer Frage, dass sich angesichts des Aufwandes, den Feldzugang sowie Transkription und Annotation der im Feld gewonnenen Audio- und Videodaten bedeuten, vergleichbare Datenmengen und Wachstumsraten wie bei Textkorpora nicht zu erreichen sind. Für die adäquate Erschließung, Bereitstellung und Analyse umfangreicher mündlicher Korpora ist daher die Entwicklung eines eigenen Methodeninstrumentariums notwendig, aus dem einige Elemente im Vortrag vorgestellt werden.