Meike Meliss und Christine Möhrs

Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe: Eine Projektpräsentation

Am IDS Mannheim wurde im September 2016 die Arbeit an dem von der Leibniz Gemeinschaft geförderten Forschungsprojekt „Lexik des gesprochenen Deutsch“ (LeGeDe) aufgenommen. Zur Durchführung des Vorhabens besteht am IDS eine Kooperation der Abteilungen Lexik und Pragmatik, die die Erstellung einer empirisch fundierten lexikografische Ressource des gesprochenen Deutsch ermöglicht, die in dieser Form sowohl für das Deutsche, als auch für andere Sprachen ein Novum darstellt.
So beruht einerseits die Erstellung der korpusbasierten elektronischen LeGeDe-Ressource auf der Datengrundlage des „Forschungs- und Lehrkorpus gesprochenes Deutsch“ (FOLK, vgl. Schmidt 2014a; 2016; Kupietz/Schmidt 2015), das als erstes großes Gesprächskorpus des Deutschen am IDS aufgebaut wird. Die Datenbank gesprochenes Deutsch (DGD, vgl. Schmidt 2014b), über die FOLK recherchierbar ist, ist mit innovativen korpustechnologischen Funktionalitäten ausgestattet und beinhaltet vielfältige Optionen der Erschließung mündlicher Daten nach linguistischen und interaktionalen Merkmalen. Andererseits verfügt das IDS sowohl über die Expertise in der Konzeption und Realisierung komplexer multimedialer Internetwörterbücher (OWID) in Verbindung mit der Entwicklung entsprechender empirischer Methoden als auch über die Erfahrung in der lexikologischen und semantischen Analyse gesprochener Sprache in der Interaktion.
Unser Beitrag verfolgt neben einer allgemeinen Projektpräsentation (Projektziele, Abgrenzung des Gegenstandsbereiches, theoretische Grundlagen, empirische Untersuchungen zu Erwartungen etc.) v.a. das Ziel, sowohl unsere korpusbasierten methodologischen Grundlagen zur Datengewinnung als auch die lexikografische Strukturierung der Daten zur Diskussion zu stellen. Dabei sollen die Möglichkeiten und Grenzen der korpusbasierten Erstellung einer lexikografischen Ressource des gesprochenen Deutsch an einer Reihe von ausgewählten Beispielen präsentiert werden. Dazu werden unterschiedliche korpusbasierte Methoden zur Erfassung relevanter lexikologischer Informationen vorgestellt: Einerseits dienen diese Methoden dazu, den Gegenstandsbereich – typische lexikalische Einheiten der gesprochenen vs. geschriebenen Sprache – empirisch zu erfassen. Andererseits sollen besondere Eigenschaften der gesprochenen Lexik, die sich als divergent zur geschriebenen Sprache erwiesen haben, auf formaler, inhaltlicher und kommunikativ-funktionaler Ebene anhand der Korpusdaten beschrieben werden. Frequenzgesteuerte Daten zu Lemmata und Wortformen des gesprochenen vs. geschriebenen Deutsch in Verbindung mit Information zu ihrer Kombinatorik (Bi- und Trigramme) und zu ihren automatisch erstellten Kookkurrenzprofilen (Perkuhn et al. 2012; Serean 2010) sowie die detaillierte semiautomatische Kodierung von zufallsgenerierten Stichproben zu ausgewählten Lemmata aus der Gesamtmenge des FOLK-Korpus (Westpfahl/Schmidt 2016) in Zusammenspiel mit interaktionsspezifischen Metadaten sind einige der methodologisch relevanten korpusbasierten Verfahren, die sowohl für die Erstellung der Stichwortliste als auch für die Makro-, Mikro- und Mediostruktur der geplanten LeGeDe-Ressource eingesetzt werden.

Literatur:

Kupietz, Marc/Schmidt, Thomas (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In: Eichinger, Ludwig (Hg.): Sprachwissenschaft im Fokus. Berlin/Boston: de Gruyter, 297-322. (= Jahrbuch des Instituts für Deutsche Sprache 2015).

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= UTB 3433). Paderborn: Fink, 2012.

Schmidt, Thomas (2014a): The Research and Teaching Corpus of Spoken German - FOLK. In: Proceedings of LREC'14, Reykjavik, Iceland: ELRA.

Schmidt, Thomas (2014b): The Database for Spoken German - DGD2. In: Proceedings of LREC'14, Reykjavik, Iceland: ELRA.

Schmidt, Thomas (2016): Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. In: Compilation, transcription, markup and annotation of spoken corpora, by Kirk, John M. and Gisle Andersen (eds.), Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], 396-418.

Seretan, Violeta (2010): Syntax-Based Collocation Extraction. Text, Speech and Language Technology. Dordrecht.

Westpfahl, Swantje/Schmidt, Thomas (2016): FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 1493-1499.