DGfS-CL

Carolin Odebrecht*, Dennis Zielke*, Thomas Krause*, Benjamin Weißenfels*, Malte Belz*, Tino Schernikau*, Vivian Voigt*
*Humboldt-Universität zu Berlin

Wissenschaftliche Nutzung der korpuslinguistischen Infrastruktur LAUDATIO

Empirische Grundlagen für diachrone oder historische Studien zu linguistischen Phänomenen sind intensiv und aufwändig hinsichtlich der Datenaufbereitung und -auswertung (Claridge 2008). Eine auf bereits vorhandene Korpora aufbauende Forschungstradition im Sinne einer kollaborierenden Wiedernutzung von Korpora hat sich auch aufgrund eines mangelnden Zugangs und fehlender Dokumentation der Datenaufbereitung nicht etabliert. Wiederverwendung verstehen wir als Suche und Finden von bereits bestehenden geeigneten Datengrundlagen für die eigene Forschungsfrage, die weitere Bearbeitung der Daten (wie z.B. weitere Annotationen oder Dokumente) und die anschließende Analyse. Für eine diachrone Analyse von Argumentstrukturen des Deutschen müssen beispielweise normalisierte Datengrundlagen gefunden werden, die die Annotation mit einem eigenen auf modernen Texten trainierten Parser erlauben. Auch Auswahlkriterien wie Register, Entstehungsort und -zeit müssen häufig berücksichtigt werden.

Das LAUDATIO-Repository (Krause et al. 2012) bildet in Verbindung mit ANNIS (Zeldes et al. 2009), SaltNPepper (Zipser & Romary 2010) und weiteren frei verfügbaren korpuslinguistischen Werkzeugen eine virtuelle Forschungsumgebung, die einen Open-Access-Zugang und eine (Wieder-)Nutzung im genannten Sinne für verschieden aufbereitete historische Korpora ermöglicht. Eine strukturiert einheitliche Facetten- und Freitextsuche und Anzeige für historische Korpora, Dokumente von Korpora und einzelner Annotationen werden durch ein Metamodell, das technisch durch das TEI-Metaschema ODD (Lou & Rahtz 2004) umgesetzt und im Repository implementiert ist, realisiert. Dieses Modell umfasst Metadaten zum Korpusprojekt inklusive Herausgeber und Annotatoren, zu den verwendeten Primärtexten mit bibliographischen Angaben und Registern, den Annotationen für jedes Korpus und jedes Dokument und den einzelnen Datenaufbereitungsschritten sowie Lizenzen.

Alle Korpora im LAUDATIO-Repository können unter einer freien Creative-Commons Lizenz heruntergeladen werden und auch von einem registrierten Nutzer selbstständig unter denselben Bedingungen hochgeladen werden. Durch eine Schnittstelle zum Konverterframework SaltNPepper kann ein Korpus von einem Format in ein anderes zur weiteren Datenaufbereitung konvertiert werden. Für den letzten Schritt der Datenanalyse ist eine Schnittstelle zum Such- und Visualisierungstool ANNIS im Repository eingebunden, die ebenfalls frei zugänglich für alle Korpora ist.

Referenzen

Claridge, C. (2008) Historical corpora. In Lüdeling, A. and Kytö, M. (Eds.) *Corpus Linguistics,* Berlin, Mouton de Gruyter, pp. 242–259.

Lou, B., Rahtz, S. (2004) RelaxNG with Son of ODD. Extreme Markup Languages.

Krause, T., Odebrecht, C., Zielke, D. (2013) Wie kann der Zugriff, die Wiederverwendung und langfristige Speicherung von linguistischen Korpora realisiert werden?. *35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS).* 12. März - 15. März 2013, Universität Potsdam.

Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009) ANNIS: A search tool for multi- layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.

Zipser, F., Romary, L. (2010) A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta.