# Modellierung linguistischer Forschungsdaten


| | |
|---|---|
| Name: | Carolin Odebrecht |
| Affiliation: | Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin |
| Betreuung: | Prof. Anke Lüdeling, Dr. Laurent Romary |
| Dissertationstitel (Arbeitstitel): | Modellierung linguistischer Forschungsdaten |
| Email: | carolin.odebrecht@hu-berlin.de |


Wie können linguistische Forschungsdaten für ein Repository[1] unter Berücksichtigung disziplinnaher Nutzerszenarien modelliert werden, um einheitlich und umfassend einen Zugriff, eine Dokumentation oder einen Import verschiedener Korpora zu ermöglichen? Wie müssen Datenstrukturen und deren Beschreibungen dafür semantisch und technisch erfasst und abstrahiert werden? Das in diesem Dissertationsprojekt zu entwickelnde Forschungsdatenmodell (Odebrecht 2014), das linguistische, schriftsprachliche Korpora[2] abbilden kann, soll erstens einen strukturierten Zugriff und eine umfangreiche Dokumentation dieser in einem Forschungsdatenrepositorium (Odebrecht et al. 2014) ermöglichen. Hierfür werden die Metadaten der historischen Korpora (vgl. Claridge 2008, Zinsmeister et al. 2008) in diesem Forschungsdatenmodell integriert. Zweitens muss das Modell die Daten selbst, also die textuelle Basis und deren Annotation, einheitlich und strukturiert erfassen können. Jede Datenaufbereitung wird durch die linguistische Forschungsfrage motiviert und sollte demnach auch abstrahiert werden. Dafür ist es notwendig, dass das Modell abstrahierte semantische Kategorien nutzt, die theorieunabhängig und unabhängig von linguistischen Interpretationen funktionieren, diese aber dennoch abbilden können. In diesem Beitrag werde ich den ersten Teil der Anwendung – Modellierung von Korpusmetadaten für ein Forschungsdatenrepositorium - mit ersten Ergebnissen diskutieren. Vorgestellt wird, wie Metadaten modelliert und technisch für ein Forschungsdatenrepository mit TEI ODD (Burnard & Rahtz 2004) implementiert werden können. Diese Korpusmetadaten werden den zugrundliegenden Texten, den Annotationen und den jeweiligen Korpusprojekten zugeordnet. Dabei wird ein Fokus auf die noch offenen Punkte hinsichtlich nicht linguistischer Datenaufbereitung wie beispielsweise die Abbildung verschiedenster Annotationsformen gelegt.

**Referenzen**


Burnard, Lou, Rahtz, Sebastian (2004) RelaxNG with Son of ODD. *Extreme Markup Languages Proceedings 2004.* Montréal, Québec.

Claridge, Claudia (2008) Historical Corpora. In Anke Lüdeling and Merja Kytö (Hrsg.) *Corpus Linguistics. An International Handbook* Bd. 1 (Reihe Handbücher zur Sprach- und Kommunikations-wissenschaft). Berlin: Mouton de Gruyter. 242-258.

Odebrecht, Carolin (2014) Modeling Linguistic Research Data for a Repository for Historical Corpora. *Digital Humanities 2014 Conference. 8.7.-12.7.2014, Lausanne.*

Odebrecht, Carolin, Zielke, Dennis, Krause, Thomas, Weißenfels, Benjamin, Belz, Malte, Schernikau, Tino, Voigt, Vivian (2014) Wissenschaftliche Nutzung der korpuslinguistischen Infrastruktur LAUDATIO. *DGfS-CL Poster Session 2014. 36. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS). 6.3.2014, Marburg.*

Zinsmeister, Heike, Andreas Witt, Sandra Kübler und Erhard Hinrichs (2008) Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In Anke Lüdeling and Merja Kytö (Hrsg.) *Corpus Linguistics. An International Handbook* Bd. 1 (Reihe Handbücher zur Sprach- und Kommunikations-wissenschaft). Berlin: Mouton de Gruyter. 759-776.


[1] LAUDATIO-Repository www.laudatio-repository.org

[2] Das Forschungsdatenmodell wird auf Grundlage historischer Korpora wie u.a. KAJUK (http://hdl.handle.net/11022/0000-0000-2102-8), GerManC (http://hdl.handle.net/11022/0000-0000-24E3-7), RIDGES (http://hdl.handle.net/11022/0000-0000-24EC-E) und HSJ (http://hdl.handle.net/11022/0000-0000-24F9-F) entwickelt.