Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung

1. Einleitung

Wenn Forschungsdaten in mehreren geisteswissenschaftlichen Disziplinen Verbreitung und Anwendung finden sollen, dann kann dies über einen gemeinsamen Zugriff realisiert werden.[1] So entstehen Synergien, z.B. in Hinblick auf die nicht doppelt zu leistende Digitalisierung oder Annotation von Quellen. Repositorien müssen damit in der Lage sein, den Zugriff auf eine heterogene Menge an Forschungsdaten zu ermöglichen. Die vorliegende Arbeit zeigt, wie ein Forschungsdatenmodell für Metadaten, das in der Korpuslinguistik entwickelt wurde, auch in anderen Geisteswissenschaften für textuelle Daten einen solchen Zugriff in Verbindung mit der technischen Umsetzung realisiert. Dabei soll während des Zugriffs der Entstehungsprozess der Daten überblickt und verstanden werden, um so die Daten korrekt referenzieren oder wiederverwenden zu können.[2]

2. Forschungsdatum Korpus

In der Korpuslinguistik wird ein Korpus als ein Digitalisat unterschiedlichster sprachlicher Primärquellen, die strukturiert mit weiteren Informationen – Annotationen – angereichert sind, verstanden (vgl. Lemnitzer & Zinsmeister 2006, 7). Der Erkenntnisgewinn erfolgt bei korpuslinguistischen Studien über eine durch Annotationen gestützte, qualitative oder quantitative Analyse natürlichen und authentischen Sprachmaterials.[3] Es ist dabei nicht klar, was jeweils unter sprachlichen Primärquellen verstanden wird (vgl. z.B. Claridge 2008, Himmelmann 2012). Die Ausweisung, was in einem Korpus ein Primärdatum ist, wird über die Forschungsfrage und Theorie zur Forschung motiviert.[4] Auch die Bedeutung der Annotationen und deren Zuweisung und Auswertung kann nicht von der jeweiligen Forschungsfrage der Korpusersteller getrennt werden.[5] Beide Konzepte werden technisch in den Korpora, konkreter in diversen Annotationstypen und deren Formaten umgesetzt (vgl. dazu Zipser 2014).
Die Korpuslinguistik steht demnach in einem Spannungsfeld zwischen dem technischen Verständnis der Korpora und der theoretisch-wissenschaftlichen Motivation der Korpuserstellung.

> ‘On an abstract technical level, there are no categorical differences between a large corpus for a well-researched language with many resources and a standardized orthography and a corpus of an endangered language or small variety without codified standards: In both cases one needs to represent a source text and annotations to it.’(Lüdeling 2012, 32)

Genau diese technisch-abstrakte Perspektive ist die Grundlage in dieser Arbeit. Wenn beispielsweise die Korpuseinheit „Token“ theoretisch mit einem Konzept von „Wort“ motiviert wird, dann kann sie als eine konzeptionelle Größe eines Korpus betrachtet werden. Das „Token“ kann auch als kleinste technische und zu annotierende Einheit in einem Korpus verstanden werden und besitzt so keinen theoretisch motivierten Wert (vgl. Krause et al. 2012). Gleiches gilt für Annotationen: Die Bedeutungen der linguistischen Annotation wie Wortartenannotationen sind insofern relative Größen, als dass ihre Bedeutungen immer im Bezug zur Forschungsfrage stehen.[6] Damit gelten solche Definitionen in der Regel für nicht mehr als ein Korpus. Darüber ist ableitbar, dass Korpora allgemein nicht über ein festes Set von Annotationen und eine eigenständig identifizierbare Primärtextebene definiert werden können.

---

[1] Im Gegensatz zu einem nicht einheitlichen Zugriff auf einzelne Forschungsdaten wie Projektwebseiten oder Anwendungen für einen Typ Korpus (z.B. TIGERSearch Suchwerkzeug Lezius 2002).
[2] Unter Wiederverwendung von Korpora wird eine erneute Analyse der vorhandenen Korpora oder die Anreichung dieser mit weiteren Annotationen verstanden.
[3] Dabei kann es sich um mündliche oder schriftliche, synchrone oder diachrone, moderne oder historische Sprache handeln. Je nach Fragestellung entsteht ein Korpus zusammengestellt bspw. nach Register, Datum, Ort etc.
[4] Diese Vielfalt zeigt sich beispielsweise anhand der Korpora, die über das Such- und Visualisierungstool ANNIS (Krause & Zeldes erscheint, Zeldes et al. 2009 http://www.sfb632.uni-potsdam.de/annis/) zur Verfügung stehen.
[5] Es gibt wenige häufig genutzte Quasistandards für Annotationen wie das STTS (Schiller et al. 1999). Was genau in den jeweiligen Studien unter Wortarten, Satzgliedern, etc. und deren Annotationen verstanden wird, ist bereits Teil der Forschung und damit immer Interpretation (vgl. Lüdeling 2011).
[6] Natürlich können „Wortarten“ oder „Satzglieder“ als kategoriale Größen definiert und so annotiert werden.

Ein Forschungsdatenmodell für Korpora muss demnach theorieneutral diese Konzepte erfassen und abbilden können. Zu berücksichtigen sind dennoch die theoretischen Fragen nach dem Primärdatum sowie nach Werten und Bedeutungen von Annotationen, ohne jedoch eine feste Aussage darüber treffen zu müssen.

3. Forschungsdatenmodell

Wenn nicht die Bedeutung der Primärquelle oder die der Annotationen Grundlage für ein Forschungsdatenmodell sein können, dann sind es die technischen Eigenschaften. Dies wird in dem hier vorgestellt Modell aufgenommen. Dabei gilt folgendes (Odebrecht 2014): Ein *Korpus* ist die Summe seiner *Dokumente*. Ein *Dokument* ist die Summe seiner *Annotationen* (Abbildung 1).

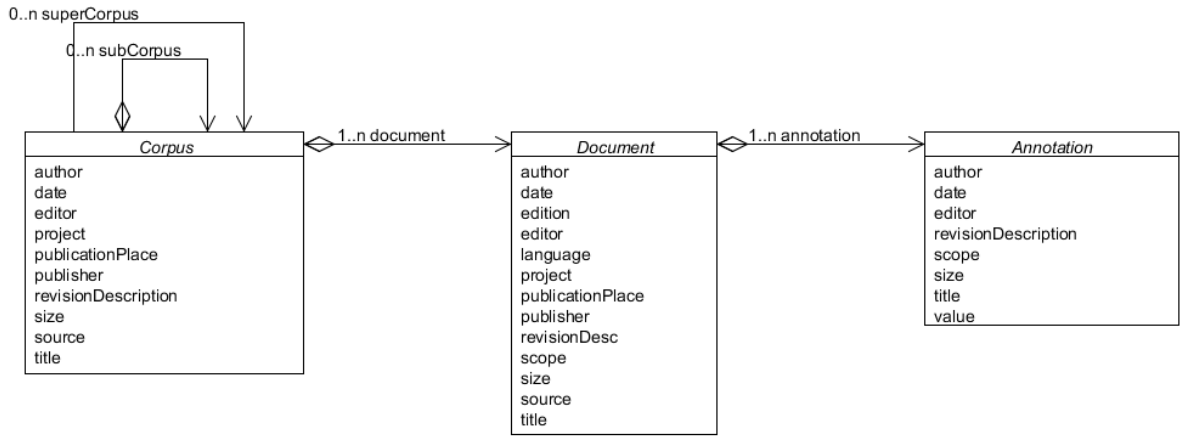

Abbildung 1 Forschungsdatenmodell für Metadaten

Die *Dokumente* definieren sich nicht zwingend nur über ihre konzeptionellen bibliographischen Eigenschaften, aber immer über ihre technische Aufbereitung. Sie bestehen wiederum aus der Summe ihrer *Annotationen*, die von *Dokument* zu *Dokument* unterschiedlich sein kann. So können mehrere *Dokumente* dieselbe Textvorlage wie bspw. ein historisches Buch des 18. Jahrhundert besitzen, aber dennoch als getrennte *Dokument* verstanden werden, da die *Annotationen* getrennt voneinander vorhanden sind. Die Einheit *Annotation* definiert sich aus dem Annotationsschlüssel und dessen Werten, die typischerweise in einem konkreten Format abgebildet sind und sich so wiederum durch ihre technischen Eigenschaften ausweisen lassen.[7] Weiterhin können Subkorpora abgebildet werden, die dann in ihrer Summe ein Superkorpus bilden. Durch die Summenmodellierung kann die Versionsgeschichte eines *Korpus*, mit steigender oder fallender Anzahl von Subkorpora, *Dokumenten* und *Annotationen* modelliert werden.

Für alle Instanzen, die mit diesem Modell abgebildet werden sollen, muss so nicht einheitlich abgebildet werden, welche theoretischen Konzepte hinter den einzelnen Summen stehen oder wie diese aus der Fachwissenschaft heraus motiviert werden. Zum Beispiel: Das diachrone *Korpus* RIDGES Herbology Corpus[8] beinhaltet neben einer Transkription auch mehrere Normalisierungen und linguistische sowie Markup-Annotationen. In diesem Fall werden die Transkription und die Normalisierungsebenen als *Annotation* neben anderen *Annotationen* verstanden und fallen unter dieselbe Summenregel. Damit stellt sich nicht die Frage nach einer Ausweisung des Primärtextes und die Bedeutungen der einzelnen Annotationsebenen sind nicht im Modell verankert. Dies wäre gänzlich irreführend, wie folgendes Minimalbeispiel zeigt: Die Annotation „lb“ wird häufig mit der Bedeutung Zeilenumbruch (‚line break‘) aus den TEI Guidelines (Burnard & Bauman 2008) verbunden – so auch in RIDGES. Sie kann jedoch auch für ein eigenständiges Konzept stehen, nämlich für eine textlinguistische Kategorie der Sachverhaltsdarstellung wie bspw. im Kasseler Junktionskorpus[9]. Das

[7] Alle Annotationen (Token-, Spannen, Baumannotationen) besitzen jeweils mindestens einen Annotationsschlüssel und -wert und können über diese Gemeinsamkeit im Modell zusammengefasst werden.
[8] http://hdl.handle.net/11022/0000-0000-2D32-6
[9] http://hdl.handle.net/11022/0000-0000-2102-8

Modell abstrahiert über alle *Annotationen*, ist damit eben nicht an Konzepte[10] gebunden, sondern überlässt letztere der jeweiligen Forschung.[11]
Zusammen mit einer Arbeitsgruppe aus der Musiksoziologie wird das Modell nun erstmals für Forschungsdaten der Editionswissenschaft getestet. So entsteht eine digitale Edition in TEI-XML für den Nachlass des „Vereins für musikalische Privataufführungen“[12]. Das Korpus wird damit Konzepte dieses Fachbereichs tragen, die sich in allen Klassen des Modells wiederfinden. Neben der nicht linguistisch motivierten Transkription werden Kategorien vergeben, die für die Erforschung der Vereinsstruktur relevant sind. Gerade die Arbeit mit Personen- oder Publikationsreferenzlisten in der Annotation ist hierfür essenziell.[13] Wir werden zeigen, wie diese Daten ebenfalls im Modell abgebildet werden können und somit für eine technisch-abstrakte Modellierung von geisteswissenschaftlichen Korpora argumentieren.

4. Forschungsdatenmodell und Metadaten

Dieses Modell regelt den Zugriff auf Korpora für das Open-Access-Forschungsdatenrepositorium LAUDATIO (Krause et al 2014) mittels Metadaten: Die Klassen *Korpus*, *Dokument* und *Annotation* werden jeweils mit Metadaten beschrieben, um ein Korpus in einer Menge von Korpora zu finden oder zu dokumentieren. Der Fokus liegt weniger auf der korpuslinguistischen Forschung sondern mehr auf der Entstehung dieser Daten und ihren Lebenszyklen (vgl. Rümpel 2011).
Zur Beschreibung wird ein Metadatenschema genutzt, das aus einer Anpassung der TEIp5 via ODD generiert wird (Burnard & Rahtz 2004). [14] Das Metadatenschema gibt vor, welche Metadatenattribute zu welcher Klasse beschrieben werden müssen. Die Werte der Metadatenattribute sind dann spezifisch für jedes Korpus. So kann dokumentiert werden, dass mehrere *Dokumente* eine gemeinsame textuelle Vorlage besitzen, wenn sie dieselben bibliographischen Metadaten teilen und somit konzeptionell als Abschnitte eines Buches interpretiert werden können. Für die Ausweisung des Primärtextes für ein einzelnes Korpus gibt es für alle *Annotationen* ein Metadatum, das besagt, ob sie jeweils technisch gesehen eigenständige Segmentierungen besitzen. Wenn dies der Fall ist, dann kann das auf eine primäre Textebene hinweisen (Krause et al 2012). Wenn es mehrere *Annotationen* gibt, die so ausgewiesen sind, kann es sich um eine Parallelkorpus handeln. Weiterhin werden die Annotationen mithilfe der Metadaten in Kategorien eingeteilt, die wiederum im Repositorium die Menge strukturieren und durchsuchbar machen. Diese Kategorien werden post hoc und für den Anwendungsfall LAUDATIO gebildet. Wenn andere Disziplinen neue Kategorien benötigen, dann können diese ohne Änderung des Modells hinzugefügt werden. So können linguistische Annotationen gemäß dem TIGER Schema[15] gleichwertig zu Markupannotationen, welche in den Editionswissenschaften benutzt werden, beschrieben werden. Die Spezifizierung erfolgt zweckgebunden immer an der Oberfläche und ist für ein breites Spektrum anderer Fachwissenschaften offen.

Referenzen

Broeder, D., Kemps-Snijders, M., et al. (2010). A data category registry- and component-based metadata framework. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC ’10) , Valletta, Malta. ELRA.
Burnard, L., Bauman, S. (Ed.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.*
Burnard, L., Rahtz, S. (2004) RelaxNG with Son of ODD. *Extreme Markup Languages Proceedings 2004*. Montréal, Québec.

[10] Einige Formate wie das EXMRALDA-Format identifizieren Primärebenen, hier verschiedene Sprechebenen (Schmid & Wörner 2009).
[11] Einen ähnlichen Ansatz besitzt das Projekt FREEBANK, das französische Korpora frei zur Verfügung stellt (Salmon-Alt 2006).
[12] Seminar „Der Nachlass des Vereins für musikalische Privataufführungen - digitale Edition in der Musikwissenschaft“ Humboldt-Universität zu Berlin, geleitet von Katrin Bicher.
[13] Solche Referenzlisten sind auch in der Linguistik gängig. So wird bspw. ISOCAT für die Referenzierung von Annotationsbedeutungen verwendet (vgl. Wright, Kemps-Snijders & Windhouwer 2007, http://www.isocat.org/).
[14] Frei verfügbar unter https://github.com/korpling/LAUDATIO-Metadata. Die Text Encoding Initiative ist in vielen Geisteswissenschaften bereits etabliert und findet viel Abdeckung. Deswegen wurde sie aus anderen Frameworks zur Beschreibung von Metadaten wie bspw. CMDI (Broeder et al. 2010) gewählt.
[15] https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-annot.pdf

Claridge, C. (2008) Historical Corpora. In Lüdeling, A., Kytö, M. (Hg.) *Corpus Linguistics. An International Handbook*. Vol 1. De Gruyter, Berlin. 242–259.
Himmelmann, N. P. (2012) Linguistic Data Types and the Interface between Language Documentation and Description. In *Language Documentation & Conservation* 6. 187-207.
Krause, Th., Zeldes, A**.** (2014) *ANNIS3: A new architecture for generic corpus query and visualization.* In Literary and Linguistic Computing. http://llc.oxfordjournals.org/content/early/2014/10/24/llc.fqu057.abstract
Krause, Th., Lüdeling, A., Odebrecht, C., Romary, L., Schirmbacher, P., Zielke, D. (2014) LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository. *Digital Humanities 2014 Conference. Poster Session. 8.7.-12.7.2014, Lausanne.* http://www.laudatio-repository.org/
Lemnitzer, L., Zinsmeister H. (2006) *Korpuslinguistik. Eine Einführung.* Gunter Narr Verlag, Tübingen.
Lezius, W. (2002) *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
Lüdeling, A. (2012) A corpus-linguistics perspective on language documentation, data, and the challenge of small corpora. In Seifart, F., Haig, G., Himmelmann, N. P., Jung, D.,; Margetts, A. & Trilsbeek, P. (Hg.) *Potentials of Language Documentation: Methods, Analyses, and Utilization.* Language Documentation & Conservation Special Publication No. 3 at the University of Hawai'i Press. 32-38.
Lüdeling, A. (2011) Corpora in Linguistics: Sampling and Annotation. In Grandin, K. (Hg.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* [Nobel Symposium 147]. Science History Publications/USA, New York. 220-243.
Rümpel, St. (2011)Der Lebenszyklus von Forschungsdaten. In Büttner, St., Hobohm, H. & Müller, L. (Hg.) Handbuch Forschungsdatenmanagement. Bock und Herchen Verlag. Bad Honnef. 25-31.
Salmon-Alt, S., Romary, L., Pierrel, J. (2006) Un modèle générique d'organisation de corpus en ligne : application à la FReeBank. Traitement Automatique des Langues, ATALA, 2006, 45, 145-169. <hal-00110970>
Schiller, A., Teufel, S., Stöckert, Ch., Thielen, Ch. (1999) Guidelines für das Tagging deutscher Textkorpora mit STTS. Technischer Report. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung & Universität Tübingen, Seminar für Sprachwissenschaft.
Schmidt, Th., Wörner, K. (2009) EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics 19/4.* 565-582.
Wright, S.E., M. Kemps-Snijders, M., Windhouwer, M. (2007) ISO-Cats: The Revised and Future TC 37 Data Category Registry. Presentation at the *Pragmatic Applications for TC 37 Standards* (TC37 2007), Provo, UT USA, August 13, 2007.
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. http://www.sfb632.uni-potsdam.de/annis/
Zipser, F. (2014) SaltNPepper und das Formatpluriversum. LAUDATIO-Workshop 07.10.2014. Berlin.