Eine Basis-Architektur für den Zugriff auf multimodale Korpora gesprochener Sprache

Batinic, Josip
josip.batinic@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Frick, Elena
frick@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Gasch, Joachim
gasch@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Das Projekt ZuMult – „Zugänge zu multimodalen Korpora gesprochener Sprache – Vernetzung und zielgruppenspezifische Ausdifferenzierung“ (zumult.org) – hat sich zum Ziel gesetzt, eine Architektur zu entwickeln, die einen einheitlichen Zugriff auf verschiedene Korpora gesprochener Sprache (Audio- und Videoaufzeichnungen mündlicher Interaktion mit zugehörigen Metadaten, Transkripten, Annotationen) an verschiedenen Standorten ermöglicht, und auf deren Basis Zugangswege gestaltet werden können, die für die Bedarfe spezifischer Nutzergruppen (z.B. Sprachlehrforschung, Variationslinguistik) optimiert sind. Mit unserem Poster stellen wir das technische Konzept und eine prototypische Implementierung einer solchen Basisarchitektur vor.

Ausgehend von einer vergleichenden Analyse vorhandener Plattformen (u.a. Datenbank für Gesprochenes Deutsch, Schmidt 2016; GeWiss-Korpus-Interface, Fandrych, Meißner & Wallner 2017; Repositorium des Hamburger Zentrums für Sprachkorpora, Hedeland et al. 2014; sowie mehrere Lösungen, die außerhalb des deutschsprachigen Raums entwickelt wurden, z.B. Eshkol-Taravella et al. 2012, Komrsková et al. 2018) und einer Bestandsaufnahme existierender Standards im Bereich multimedialer Daten (vgl. dazu auch Schmidt 2014 und Schmidt et al. 2010) haben wir eine Dreiebenen-Lösung entwickelt, die so weit wie möglich auf etablierte (De Facto-)Standards aufbaut und anschlussfähig an existierende Lösungen ist. Damit wird eine transferfähige Basis für einen flexiblen Zugriff auf multimodale Korpora geschaffen.

Kern der Architektur ist zum einen eine objektorientierte Modellierung der Korpus-Bestandteile (Aufnahmen, Metadaten zu Sprechereignissen und Sprechern, Transkripte, Annotationen und Zusatzmaterialien) und ihrer Beziehungen zueinander. Für deren digitale Repräsentation (Serialisierung) werden Standards verwendet, soweit sie existieren. Für Medienobjekte können wir auf industrielle Standards insbesondere aus dem Kontext der Moving 133_final-* Expert Group (MPEG) zurückgreifen. Die Repräsentation von Transkripten und Annotation folgt dem in ISO (2016) definierten und auf den Richtlinien der Text Encoding Initiative (TEI) basierenden Format für „Transcriptions of Spoken Language“. Metadaten werden grundsätzlich in XML repräsentiert; in Ermangelung eines echten Standards, der in der Lage wäre, der Bandbreite und Komplexität von Metadaten im Bereich multimodaler Korpora vollständig gerecht zu werden, orientieren wir uns in diesem Bereich an CMDI-Profilen, die im CLARIN-Kontext für solche Korpora entwickelt wurden (z.B. Hedeland & Wörner 2012).

Zum anderen beinhaltet die Architektur ein vereinheitlichtes Konzept zur Query auf Transkriptions- und Annotationsdaten. Dieses baut auf Überlegungen zu einer „Corpus Query Lingua Franca“ (Banski et al. 2016, ISO 2018) auf und berücksichtigt somit in der Korpuslinguistik verbreitete Suchsprachen wie CQP, ANNIS-QL, Poliqarp und weitere, die allerdings für die Besonderheiten angepasst werden müssen, die spontansprachliche Daten gegenüber schriftsprachlichen Korpora aufweisen.

Die Basisarchitektur besteht somit aus zwei gleichberechtigten Komponenten: Aus der Modellierung der Korpus-Bestandteile ergeben sich Zugriffs- und Navigationsmöglichkeiten für ganze Objekte bzw. Objekthierarchien, die auf Nutzerseite vor allem für ein explorativ Browsing auf den Daten eingesetzt werden. Die Query-Komponente ermöglicht hingegen eine gezielte Auswahl von (Teilen) von Objekten und damit systematische Recherchen im Sinne einer korpuslinguistischen Methodik. Beide Komponenten werden technisch als „Locators“ bzw. „Filters“ in einer REST API umgesetzt. Diese wird in der weiteren Projektarbeit die Basis darstellen, um zielgruppenspezifisch optimierte Zugänge zu den Daten zu entwickeln.

Neben einem Überblick über diese Basis-Architektur wird unser Poster auch auf die konkrete Implementierung eingehen, die am Institut für Deutsche Sprache für den Zugriff auf die Daten aus dem Archiv für Gesprochenes Deutsch entwickelt wurde. Diese setzt auf ein vorhandenes Backend auf, das die Grundlage für die Datenbank für Gesprochenes Deutsch bildet und XML-basierte Daten in einer objektrelationalen Oracle-Datenbank hält. Für die Arbeiten in ZuMult wird dieses Backend für die im Projekt definierten Bedarfe angepasst und erweitert. Prototypische Applikationen, die den Einsatz der REST

API illustrieren, werden als Software-Demonstrationen die Posterpräsentation ergänzen.

Bibliographie

Banski, Piotr / Frick, Elena / Witt, Andreas (2016): *"Corpus Query Lingua Franca (CQLF)"*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia 2804-2809. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-50405

Eshkol-Taravella, I. / Baude, O. / Maurel, D. / Hriba, L. / Dugua, C. / Tellier, I., (2012): *"Un grand corpus oral ‚disponible' : le corpus d'Orléans 1968-2012. "* In: Ressources linguistiques libres, TAL. 52,3/2011, 17-46.

Fandrych, Christian / Meißner, Cordula / Wallner, Franziska (eds.) (2017): *"Gesprochene Wissenschaftssprache – digital Verfahren zur Annotation und Analyse müdlicher Korpora."* Deutsch als Fremd- und Zweitsprache. Tübingen: Stauffenburg.

Hedeland, Hanna / Wörner, Kai (2012): *"Experiences and Problems creating a CMDI profile from an existing Metadata Schema"*. Proceedings of LREC-Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, Istanbul, European Language Resources Association (ELRA) 37-40. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): *"Multilingual Corpora at the Hamburg Centre for Language Corpora". In: Ruhi, #ükriye/Haugh, Michael/Schmidt, Thomas/Wörner, Kai (Hrsg.):* Best Practices for Spoken Corpora in Linguistic Research. Newcastle: Cambridge Scholars Publishing, 2014. S. 208-224.

ISO (ed.) (2016): *ISO 24624:2016 Language resource management – Transcription of spoken language.* https://www.iso.org/standard/37338.html

ISO (ed.) (2018): *ISO 24623-1:2018 Language resource management – Corpus query lingua franca (CQLF) -- Part 1: Metamodel.* https://www.iso.org/standard/37337.html

Komrsková, Zuzana / Kop#ivová, Marie / Lukeš, David / Poukarová, Petra / Golá#ová, Hana (2018): *"New Spoken Corpora of Czech: ORTOFON and DIALEKT."* Journal of Linguistics 68:2, 219-228.

Schmidt, Thomas (2014): *"(More) Common Ground for Processing Spoken Language Corpora?"* In: **Ruhi, #ükriye/Haugh, Michael/Schmidt, Thomas/Wörner, Kai (eds.)**: *Best Practices for Spoken Corpora in Linguistic Research.* Newcastle: Cambridge Scholars Publishing, 2014 249-265. http://pub.ids-mannheim.de/autoren/divers/3119.html

Schmidt, Thomas (2017): *"DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim."* In: Zeitschrift für Germanistische Linguistik 45(3), S. 451-463.

Schmidt, Thomas / Elenius, Kjell / Trilsbeek, Paul (2010): *"Multimedia encoding and annotation"*. In: **Hinrichs, Erhard (ed.)**: *Interoperability and standards*. Utrecht: Utrecht University, 2010 121-124. http://www.exmaralda.org/files/CLARIN_Standards.pdf