DH-Toolvergleich im Hinblick auf Texte historischer Sprachstufen

Aehnlich, Barbara
barbara.aehnlich@uni-jena.de
FSU Jena, Deutschland

Seidel, Henry
hnrseidel@gmail.com
HU Berlin, Deutschland

Mittlerweile versprechen zahlreiche Tools eine mehr oder minder problemlose Lemmatisierung und Annotierung mit Part-of-Speech-Tags von Texten; viele sollen auch für historische (oder andere nicht-standardisierte) Sprachdaten nutzbar sein. [1] Dabei birgt die Verarbeitung historischer Sprachdaten des Deutschen zahlreiche Probleme aufgrund des hohen Grads an Variation, insbesondere auf den Ebenen Phonologie und Graphematik, aber auch in den Bereichen der Morphologie, Syntax und Lexik. Bei einer automatischen Verarbeitung solcher Daten stellen vor allem die Variationen in Phonologie, Graphematik und Morphologie ein besonderes Hindernis dar.

Das Poster stellt verschiedene Werkzeuge überblicksartig vor und befasst sich genauer mit zwei Tools, deren Anwendung auf Texte nicht-standardisierter Sprachstufen exemplarisch anhand zweier Textsorten aufgrund bestimmter Kriterien verglichen wurde. Zum einen handelt es sich um gedruckte deutschsprachige Rechtstexte der Frühen Neuzeit, also aus der Rezeptionszeit des römischen Rechts in Deutschland, deren Sprache in einem Projekt erforscht werden soll, zum anderen um Fürstinnen-Briefe aus einem an der Friedrich-Schiller-Universität Jena erstellten digitalen Korpus. [2] In beiden Fällen weisen die Quellen frühneuhochdeutschen Sprachstand auf. Anhand ausgewählter Beispiele aus den vorliegenden Texten sollen zwei gängige elektronische Werkzeuge miteinander verglichen werden – EXMARaLDA und LAKomp.

EXMARaLDA wurde für das computergestützte Arbeiten mit überwiegend mündlichen Korpora entwickelt, wird aber regelmäßig auch für schriftliche Sprachdaten verwendet, so auch bei den *Frühneuhochdeutschen Fürstinnenkorrespondenzen im mitteldeutschen Raum*. Das Tool besteht im Wesentlichen aus einem Transkriptions- und Annotationseditor, einem Werkzeug zum Verwalten von Korpora und einem Such- und Analysetool. [3]

Das Werkzeug LAKomp [4] wurde im Projekt SaDA (Semiautomatische Differenzanalyse von komplexen Textvarianten) [5] entwickelt und dient der Aufbereitung eines historischen Korpus. Nach der Transkription können die Texte hier lemmatisiert und annotiert werden. Aufgrund der Besonderheiten bei frühneuhochdeutschen Handschriften und Drucken wird der Lemmatisierungs- und Annotationsvorgang komplett manuell durchgeführt. Dabei ist dem Benutzer mit LAKomp ein Werkzeug an die Hand gegeben, das ihn sehr schnell und präzise große Textmengen bearbeiten lässt und damit den Mehraufwand händischer Annotation nahezu ausgleicht.

Damit wurden zwei Werkzeuge ausgewählt, bei denen manuell annotiert werden muss, die aber dennoch bestimmte Unterschiede aufweisen, die für Nutzerinnen und Nutzer, die mit nicht-standardisierten Sprachdaten arbeiten, je nach Arbeitsziel vor- oder nachteilig sein können. So sind etwa bei LAKomp die halbautomatische Annotation auf der Grundlage des DWB und die Ausgabefunktion besonders gelungen, leider kann hier aber bisher nur lemmabasiert annotiert werden; bei EXMARaLDA ermöglichen die flexiblen Annotationskriterien eine besondere Breite von möglichen Annotationen, eine automatische oder halbautomatische Annotation des frühneuhochdeutschen Textmaterials ist jedoch bislang auch mit Hilfsmitteln wie dem Treetagger [6] nicht ohne weiteres möglich.

Die beiden genannten Tools werden auf dem Poster hinsichtlich ihrer Anwendbarkeit auf Texte historischer Sprachstufen anhand folgender Kriterien verglichen: Funktionalitäten, Nutzerfreundlichkeit (Technischer Support, Qualität des Handbuchs, Verständlichkeit der Benutzeroberfläche, Verfügbarkeit eines Editors, Umgang mit Metadaten, Exportmöglichkeiten …) und Nachnutzbarkeit. Das Poster wird diesen Tool-Vergleich anhand ausgewählter Beispiele aus einem Rechtsbuch sowie einem Fürstinnenbrief aus der Mitte des 16. Jahrhundert präsentieren und stellt somit Überblick und Evaluation der Werkzeuge gleichermaßen dar.

Fußnoten

1. In Auswahl: CATMA, CorA, EXMARaLDA, GATE, LAKomp, WebAnno.
2. https://archive.thulb.uni-jena.de/hisbest/content/below/index.xml?XSL.DisplayComponentBrowse=true ; http://www.laudatio-repository.org/repository/corpus/LAUDATIO%3AFuerstinnenkorrespondenz/TEI-header_version4_Schema7_2017-03-06T08%3A38%3A26%3A247Z
3. Für genauere Informationen vgl. http://exmaralda.org/de/ueber-exmaralda/ und http://exmaralda.org/de/exmaralda-nutzer/ .
4. LAKomp steht für **L**emmatisierung, **A**nnotation, **Komp**aration.
5. http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/
6. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Bibliographie

B. Aehnlich / S. Kösser (2016): „Das Tool LAKomp und seine Anwendung auf Texte nichtstandardisierter Sprachstufen“. In: DHd 2016. Konferenzabstracts, Leipzig, S. 263–264.

V. Faßhauer (2017): „Compilation, Transcription, Multi-Level Annotation and Gender-oriented Analysis of a Historical Text Corpus:Early Modern Ducal Correspondences in Central Germany“. In: Advances in Digital Scholarly Editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp, hg. v. Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini & Dirk van Hulle. Leiden, S. 269–274.

A. Medek (*Gießler) / M. Pöckelmann / T. Bremer / H.-J. Solms / P. Molitor / J. Ritter (2015): „Differenzanalyse komplexer Textvarianten - Diskussion und Werkzeuge“. In: Informationsmanagement für Digital Humanities, hg. v. G. Heyer und A. Henrich. In: Datenbank-Spektrum 2015. http://dx.doi.org/10.1007/s13222-014-0173-y.

A. Leipold / J. Ritter / H.-J. Solms: „Neue Wege zu Textzeugenvergleich und Edition am Beispiel der Wundarznei des Heinrich von Pfalzpaint“. In: Jahrbuch für Germanistische Sprachgeschichte 2014, Band 5, Heft 1, S. 335-358.

D. Prutscher / H. Seidel (2012): „Mehrebenenannotation frühneuzeitlicher Fürstinnenkorrespondenzen“. In: G. Brandt (Hg.): Bausteine weiblichen Sprachgebrauchs. X. Texte – Zeugnisse des produktiven Sprachhandelns von Frauen in privaten, halböffentlichen und öffentlichen Diskursen vom Mittelalter bis in die Gegenwart. Stuttgart, S. 109–124.