Das Tool LAKomp und seine Anwendung auf Texte nichtstandardisierter Sprachstufen

,
Barbara.Aehnlich@uni-jena.de
Friedrich-Schiller-Universität Jena, Deutschland

,
sylwia.koesser@germanistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Die Verarbeitung historischer Sprachdaten des Deutschen birgt zahlreiche Probleme: Sie weisen einen hohen Grad an Variation auf, insbesondere auf den Ebenen Phonologie und Graphematik, aber auch in den Bereichen der Morphologie, Syntax und Lexik. Die bisher entwickelten Tools, z. B. im Bereich der automatischen Wortarten-Annotation, sind auf Daten des Gegenwartsdeutschen trainiert und können deshalb nur bedingt oder gar nicht auf Daten historischer Sprachstufen angewandt werden.

Für die Lemmatisierung und Annotierung mit Part-of-Speech-Tags existieren bereits linguistische Werkzeuge, die nach einer Trainingsphase auf bereits annotierten Texten weitere Texte automatisch annotieren können. Angewendet auf frühneuhochdeutsche Texte liefern diese Werkzeuge aber hohe Fehlerraten, denn eine Voraussetzung für ihr Funktionieren ist hier schwer erfüllbar: das Erkennen von Wortformen. Hier stellt die stark variierende Graphie ein Hindernis dar.

Im Projekt SaDA (Semiautomatische Differenzanalyse von komplexen Textvarianten) (Bremer et al. 2012-2015) werden deshalb elektronische Werkzeuge entwickelt, die der Aufbereitung eines historischen Korpus dienen sollen und zur Anwendung in verschiedenen philologischen Bereichen gedacht sind. Zur Erstellung eines strukturierten Korpus ist die Anreicherung der Überlieferungszeugen mit verschiedenen Informationen Voraussetzung. Zu diesem Zweck wurde das Werkzeug LAKomp entwickelt, mit dessen Hilfe alle im Zuge der Bearbeitung dem Text hinzugefügten Informationen gespeichert und für die spätere Nutzung aufbereitet werden.

LAKomp wird unter anderem an der "Wundarznei" des Heinrich von Pfalzpaint (weiter)entwickelt. Nach der Transkription der Überlieferungszeugen nach den Konventionen und Kodierungen der Mittelhochdeutschen Grammatik, des Referenzkorpus Mittelhochdeutsch und des Referenzkorpus Frühneuhochdeutsch werden die Texte lemmatisiert und annotiert.

Die morphologische Annotation reichert das Textmaterial zunächst mit der Angabe der Wortart an, wobei Verben und Nomina weiter spezifiziert, also mit Angaben zu den verbalen und nominalen Kategorien versehen werden. Syntaktische Informationen werden teilweise durch die Unterscheidung attributiver, prädikativer oder adverbialer Verwendung bei Adjektiven und Partizipien geliefert.

Durch Lemmatisierung und Annotation werden die Wortformen der einzelnen Handschriften einem tertium comparationis gegenübergestellt. Durch diese Abstraktion, die Zuweisung einer der Einzelgraphie übergeordneten Wörterbuchform (bei parallelem Erhalt der konkreten Handschriften-Graphie), wird ein sehr konkreter maschineller Vergleich möglich.

Mit der vorgenommenen Kodierung des Quellenmaterials ist ein semi-automatischer Textzeugenvergleich möglich. Zunächst durch die Segmentierung, aber vor allem durch die Lemmatisierung und noch stärker durch die grammatische Auszeichnung können die einzelnen Handschriften konkret aufeinander abgebildet werden, sodass Abweichungen und damit Filiationsverhältnisse deutlich sichtbar werden. Für die Darstellung der Unterschiede und Gemeinsamkeiten der Textzeugen werden diese in einem sogenannten Partiturtext vertikal dargestellt, miteinander verglichen und die Unterschiede zusätzlich farbig markiert. Der Partiturtext wird von LAKomp unter Zuhilfenahme der vorher beigegebenen Informationen automatisch erzeugt.

Neben der einfachen Suchfunktion kann das zuvor im textspezifischen Wörterbuch abgelegte und mit Informationen angereicherte Wortmaterial auch mit der Analysefunktion gezielt durchsucht werden. So bietet sich dem Nutzer beispielsweise die Möglichkeit, alle Graphieformen eines Lemmas abzurufen und ihre statistische Verteilung in den Handschriften und Drucken abzufragen. Neben der prozentualen Verteilung werden ebenso die Belegzahlen und die einzelnen Graphieformen ausgegeben.

Im Rahmen eines an der MLU Halle geplanten Projekts zu medizinischen Sachtexten des Mittelalters soll LAKomp weiterentwickelt werden, um die Untersuchung der medizinischen Inhalte (Texte und Objekte) hinsichtlich verschiedener Fragestellungen (Verschlagwortung, Datenbank, Verknüpfung von Informationen) und eine optimierte nutzerbezogene Darstellung der Ergebnisse gewährleisten (Analysefunktion, Satzprogramm zum Edieren der Texte, kartographische Darstellung) zu können. Die Überlieferung der Zeit von 1350 - 1650 ist vor allem durch Kompilationen medizinischer Texte geprägt, was eine Einordnung einzelner Texte in Überlieferungswege und -zusammenhänge bedeutend erschwert. Grundvoraussetzung für die Entwicklung und Verifizierung von Werkzeugen ist ein geeignetes Korpus. Text- und Objektbasis dieser Pilotstudie ist die "Wundarznei" des Heinrich von Pfalzpaint aus dem Jahre 1460. Anhand dieses Textes sollen die Möglichkeiten zur Beantwortung verschiedenster Fragen exemplarisch erprobt und Werkzeuge zur Umsetzung und Darstellung entwickelt werden.

Ein weiteres Projekt, das sich auf das Tool LAKomp stützt, befasst sich mit Rechtstexten aus der Rezeptionszeit des römischen Rechts (Aehnlich 2016). Es beruht auf einem Korpus zweier frühneuhochdeutscher Rechtsbücher des 15. und 16. Jahrhunderts. Der Klagspiegel ist das mit Abstand älteste populärwissenschaftliche Rechtsbuch der Rezeptionszeit und bildet mit dem Laienspiegel zusammen die wichtigste Grundlage an rechtswissenschaftlichen populären Texten des 15. und 16. Jahrhunderts. Davon ausgehend ist ein Projektantrag zu einem Korpus von Strafrechtstexten der frühen Neuzeit in Arbeit, welches ebenfalls mithilfe von LAKomp strukturiert und aufbereitet werden soll. Durch semantische und linguistische Annotationen soll eine umfassende Forschungsgrundlage geschaffen werden, die für die Schließung rechts- und sprachhistorischer Forschungslücken einen zentralen Beitrag leistet.

Das Poster stellt das Werkzeug LAKomp mit seinen Einsatzmöglichkeiten und -gebieten vor. Am Beispiel des Pfalzpaint und des Laienspiegels wird gezeigt, dass das Tool einfach und intuitiv bedienbar ist.

Fußnoten

1. **L**emmatisierung, **A**nnotation, **Komp**aration.

Bibliographie

Aehnlich, Barbara (2016): *Sprachwissenschaftliche Untersuchungen zum Klagspiegel Conrad Heydens (1436) und zum Laienspiegel Ulrich Tenglers (1509)*. Universität Jena http://www.sprachwissenschaft.uni-jena.de/Lehrbereiche/Geschichte+der+deutschen +Sprache/Dr_+Barbara+Aehnlich/Projekt-p-1881.html [letzter Zugriff 28. Januar 2016].

Bremer, Thomas / Molitor, Paul / Ritter, Jörg / Solms, Hans-Joachim (eds.) (2012-2015): *SaDA*. Semi-automatische Differenzanalyse von komplexen Textvarianten. Martin-Luther-Universität Halle http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/ [letzter Zugriff 08. Januar 2016].