Semi-automatische Differenzanalyse von komplexen Textvarianten

André Medek (né Gießler) andre.medek@informatik.uni-halle.de
Marcus Pöckelmann marcus.poeckelmann@informatik.uni-halle.de
Jörg Ritter joerg.ritter@informatik.uni-halle.de

Einer der Schwerpunkte von Projekten der Editionsphilologie ist die Untersuchung alter Texte mit Mehrfachüberlieferungen sowie die Textgenese. Dabei stellt sich für die beteiligten Wissenschaftler die Aufgabe, die Verbindungen zwischen den einzelnen Textvarianten herauszuarbeiten und dabei Gemeinsamkeiten und Unterschiede zu erkennen. Oft sind große Textmengen der verschiedenen Varianten zu sichten, einander zuzuordnen und detailliert zu vergleichen, um anschließend als Edition präsentiert werden zu können. Bisher erfolgen die bei der Edition anfallenden, teils sehr gleichartigen und zeitaufwändigen Zwischenschritte in Handarbeit, belegen damit wertvolle Arbeitszeit und setzen einen Gesamtüberblick über das Textmaterial voraus. Die Informationstechnologie bietet heute Möglichkeiten, mit denen die Durchführung vieler dieser Schritte zumindest teilautomatisiert werden kann. Den Geisteswissenschaftlern können Werkzeuge zur Verfügung gestellt werden, die ihnen die Arbeit nicht nur wesentlich erleichtern, sondern auch die Fehleranfälligkeit reduzieren und neue Formen der Auswertung eröffnen.

Das hier vorgestellte, vom BMBF geförderte Gemeinschaftsprojekt von Geisteswissenschaftlern und Informatikern mit dem Ziel, Werkzeuge und Methoden zum Textvergleich und zur Erstellung kritischer und genetischer Editionen zu entwickeln. Diese Methoden sollen generisch und damit auf viele Textformen anwendbar sein. Dazu werden zwei Repräsentanten verschiedenartiger Textformen mit ihren Überlieferungen zur Grundlage genommen, für deren Anforderungen und Eigenheiten jeweils zugeschnittene Verfahren entwickelt und evaluiert werden. Dabei werden die Prozesse von der Erkennung und Lemmatisierung der Wörter, das Auffinden sich entsprechender Textstellen, die Herausarbeitung der Unterschiede und Gemeinsamkeiten, bis hin zur Darstellung in einer genetischen Edition abgedeckt. Im späteren Verlauf des Projektes findet die Verallgemeinerung der gewonnenen Erkenntnisse für möglichst viele Textformen statt.

Ein Teil des Projektes betrachtet einen handschriftlichen Lehrbuchtext aus der Zeit des Spätmittelalters, der in Frühneuhochdeutsch verfasst wurde. Unter der Leitung von Hans-Joachim Solms wird die „Wundarznei“ des Heinrich von Pfalzpaint aus dem 15. Jahrhundert in ihren zehn verfügbaren Überlieferungen untersucht. Ziel der Altgermanisten ist hier, die Varianten zu vergleichen und in einer kritischen Edition und einer Online-Edition mit Synopse und Variantenapparat darzustellen. Ausgangspunkt sind die Handschriften, die in einem ersten Arbeitsschritt von den Geisteswissenschaftlern diplomatisch transkribiert wurden. Da diese Texte in der Sprachstufe Frühneuhochdeutsch verfasst wur-

den, gibt es keine einheitliche Graphie, wodurch dieselben Wörter in verschiedenen Überlieferungen deutlich unterschiedlich geschrieben werden und mit jeder weiteren Überlieferung neue Schreibweisen entdeckt werden. Ein Beispiel ist das Wort Pfeil, welches in den Schreibweisen „pfeil“, „pffeil“, „pfejl“, „pffejl“, „pfeyl“, „pffeyl“ auftritt. Bevor ein Textvergleich stattfinden kann, ist somit eine philologische Aufbereitung nötig, bei der die Wörter erkannt und normalisiert werden. Die Normalisierung wird mit der Lemmatisierung jedes einzelnen Wortes erreicht. Für eine möglichst präzise Abbildung einer Handschrift auf eine andere werden die Wörter zusätzlich noch mit Part-Of-Speech-Tags und morphologischen Attributen wie Kasus, Numerus und Genus versehen. Für die Aufgabe der Lemmatisierung und Annotation existieren bereits verschiedene automatisierte Ansätze, die allerdings nicht auf Handschriften aus dem Frühneuhochdeutschen anwendbar sind, da sie nur eine sehr geringe Toleranz für abweichende Schreibweise (oder Schreibfehler) von Wörtern aufweisen. Die unstetige Graphie in den Handschriften der „Wundarznei“ führt bei ihnen zu geringen Trefferquoten in Bezug auf die Korrektheit der von ihnen vorgeschlagenen Annotationen.

Im Projekt wurde das Werkzeug *LAKomp* entwickelt, das einen semiautomatischen Ansatz verfolgt. Es erlaubt die manuelle Annotierung jedes einzelnen Wortes, in dem es zu dem Wort ähnliche Wortformen ableitet, diese mit zugehörigen Annotationen in Lexika sucht und dem Benutzer so Vorschläge für das aktuelle Wort unterbreitet. Ähnlich heißt hier, dass sich die neue Wortform mittels von Altgermanisten erarbeiteten Ersetzungsregeln, die auf Äquivalenzen bestimmter Buchstabenfolgen basieren, aus dem gegebenen Wort ableiten lässt. Im Gegensatz zu automatischen Ansätzen liegt die Entscheidung für das passende Lemma und die passenden morphologischen Daten beim Anwender.

Die Benutzeroberfläche ist intuitiv verständlich gestaltet und auf die Massenverarbeitung ausgerichtet. *LAKomp* hat sich bei der Annotation der Handschriften der „Wundarznei“ durch Germanisten als sehr große Arbeitserleichterung erwiesen. Da es ein webbasiertes Werkzeug ist, können mehrere Nutzer gleichzeitig annotieren und profitieren von den gelernten Eingaben der anderen Nutzer. Bild 1 zeigt den Dialog für eine Wortform.

Nach der Lemmatisierung der Handschriften lassen diese sich nun innerhalb des Werkzeugs *LAKomp* detailliert vergleichen. Der Textvergleich auf dem Weg zu einer kritischen Edition erfolgt zweistufig: Zuerst werden in einem Vergleich auf Makroebene sich entsprechende Textstellen mit Hilfe von Signaturen identifiziert und einander gegenübergestellt. Das Ergebnis dieses Schrittes ist eine Alignierung. Im nächsten Schritt werden die innerhalb einer Alignierung einander zugeordneten Textstellen der verschiedenen Überlieferungen auf Mikroebene verglichen und durch Auflistung der Unterschiede in einem Variantenapparat präsentiert.

Das zweite Teilprojekt widmet sich einem neuphilologischen Text in Fremdsprache. Unter der Leitung von Thomas Bremer wird die im späten 18. Jahrhundert in Französisch verfasste „Histoire philosophique et politique des établissements et du commerce des Européens dans les deux indes“ von Abbé Guillaume Thomas François Raynal untersucht. Sie gilt auf Grund ihrer Thematik, die Auseinandersetzung mit der europäischen Kolonialpolitik dieser Zeit, als bedeutendes Werk der Aufklärung. Nach dem Verbot der Erstauflage von 1770 erschienen zwei weitere Auflagen sowie eine postume Textfassung. Im Rahmen

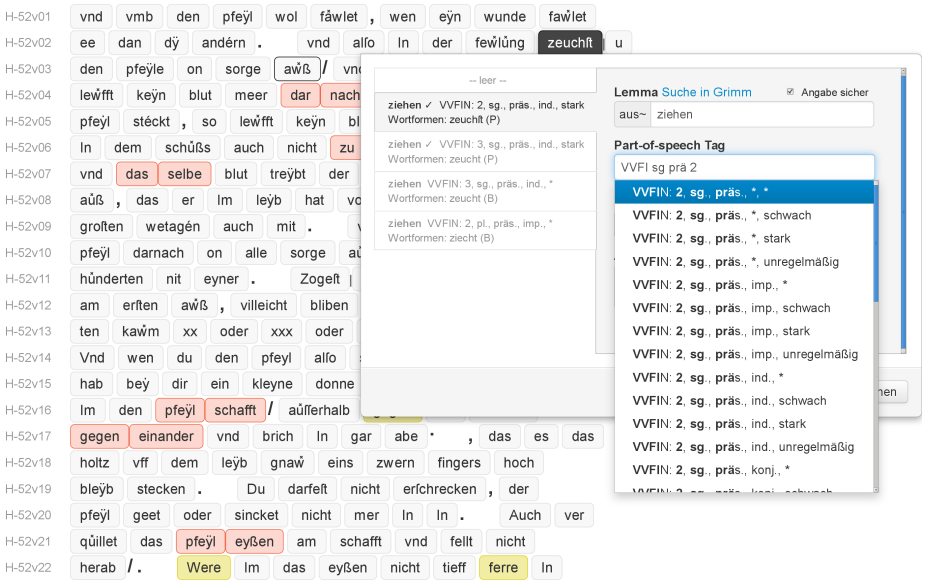

Abbildung 1: Annotationsdialog für eine Wortform in LAKomp

des Projektes soll eine genetische Edition der Lateinamerika-Bände entstehen, die insbesondere durch ihre Interaktivität die Evolution dieses Werkes nachvollziehbar macht. Dabei liegt das Hauptaugenmerk auf einer abschnittsweisen Gegenüberstellung der vier Varianten als Fließtext mit einer übersichtlichen Form des Apparats.

Ausgangspunkt hier sind digitale Faksimiles, die als Scan der Erstausgaben angefertigt wurden. Diese wurden mittels existierender Software zur Texterkennung in eine maschinenlesbare Form gebracht, anschließend von den Romanisten von fehlerhaft erkannten Stellen bereinigt sowie mit speziellen Markierungen versehen, die beispielsweise Überschriften oder Seitennummern kenntlich machen. Erleichtert wird die Suche nach Fehlern dabei durch ein softwaregestütztes Verfahren, das auffällige Kombinationen von Symbolen anzeigt. So liefert beispielsweise die Suche nach Kombinationen aus Buchstaben und Zahlen ohne trennendes Leerzeichen eine Reihe von fehlerhaft kodierten Jahreszahlen, wie „15o6“ oder „158o“. Aus den so entstehenden Textdokumenten wird automatisch eine TEI-konforme XML-Darstellung generiert, die als Grundlage für die folgenden Arbeitsschritte dient. Für einen der beiden algorithmischen Schwerpunkte aus Sicht der Informatik, die Alignierung der Absätze, werden derzeit verschiedene automatische Verfahren geprüft. Für den zweiten Schwerpunkt, die Bestimmung und Visualisierung der Unterschiede auf Absatzebene, wurde bereits ein erster Ansatz implementiert. Dieser vergleicht eine beliebige Anzahl von Textpassagen untereinander. Auf Basis der Levenshtein-Distanz werden die Differenzen zwischen den Varianten ermittelt und daraus eine synoptische Darstellung in LaTeX erzeugt (Abbildung 2). Die Philologen können für die Visualisierung der Textvarianten zwischen konfigurierbaren Darstellungsarten wählen. Die genannten Arbeitsschritte, von der Generierung der XML-Dateien bis hin zum Entwurf der elektronischen Edition, sollen perspektivisch in einer gemein-

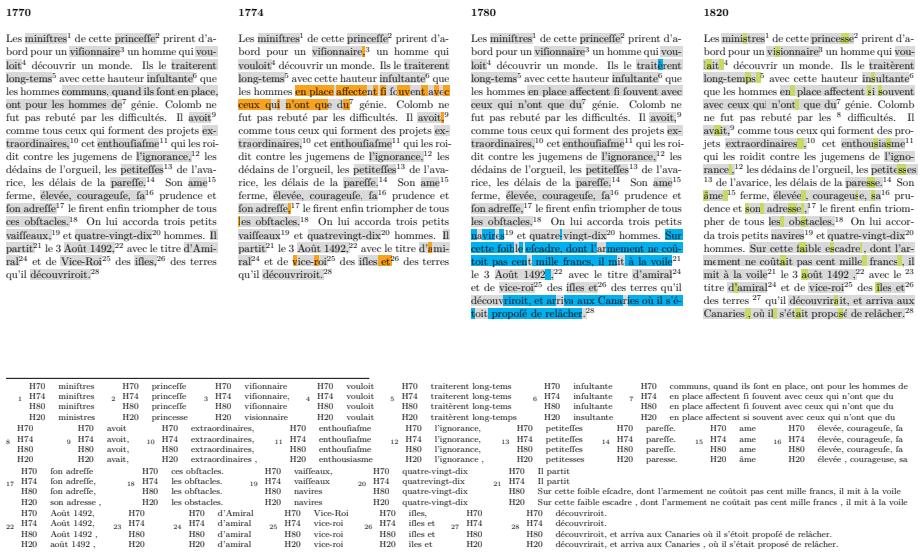

Abbildung 2: Automatisch generierte Synopse mit Variantenapparat für einen Absatz des franz. Textes

samen, webbasierten Arbeitsumgebung eingebettet werden.

Prototypische Werkzeuge, unter anderem *LAKomp*, werden in den nächsten Monaten zur Demonstration als Webanwendungen öffentlich verfügbar gemacht.

Anmerkungen

Diese Arbeit wurde durch das Bundesministerium für Bildung und Forschung (BMBF) [Projektkürzel: 01UG1247 / human-325-010 / SaDA] im Rahmen des Projekts „Semi-automatische Differenzanalyse von komplexen Textvarianten“ unter Leitung von Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter und Prof. Dr. Hans-Joachim Solms gefördert. An dieser Stelle möchten wir auch unseren Projektmitarbeiterinnen Sylwia Kösser, Dr. Aletta Leipold und Susanne Schütz danken.

Weitere Informationen zum Projekt sind auf der Website `http://www.informatik.uni-halle.de/sada/` zu finden.