LERA - Explorative Analyse komplexer Textvarianten in Editionsphilologie und Diskursanalyse

Schütz, Susanne
susanne.schuetz@romanistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Ausgangssituation und Projektrahmen

Der Vergleich von Textfassungen ist eine zentrale Aufgabe der Editionsphilologie und Grundlage für die Erstellung von Variantenapparaten in kritischen Editionen. Zudem lässt sich über den Textvergleich die inhaltliche Evolution von Texten nachvollziehen. Je umfänglicher oder komplexer die zu edierenden Werke sind, desto schwieriger ist es für die Editoren und die Nutzer von Editionen den Überblick über die Textfassungen zu behalten. Mit dem hier vorgestellten, in die Arbeitsumgebung LERA (Bremer et al. 2015) integrierten Ansatz, wird diesem Problem mit einer Kombination von graphischen Werkzeugen begegnet, deren Zusammenwirken den Überblick über große Textmengen und ihre inhaltliche Auswertung erleichtert und zudem die Möglichkeit zum ergbnisoffenen Erkunden der Textvarianten bietet.

LERA ist eine interaktive, webbasierte Arbeitsumgebung zur Untersuchung mehrerer Fassungen eines Textes. Sie wurde im Rahmen des vom Bundesministeriums für Bildung und Forschung geförderten Projekts: *Semiautomatische Differenzanalyse von komplexen Textvarianten (SaDA)* entwickelt (Medek et al. 2015). Als Fallbeispiel für die Entwicklung von LERA wurde ein Buch aus der 'Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes' von Thomas-Guillaume Raynal, einem der gesamteuropäisch einflussreichsten Erfolgs- und Skandalbücher des 18. Jahrhunderts, gewählt. Das Werk ist eine kritische Auseinandersetzung mit der europäischen Kolonialexpansion in den ‚beiden' Indien und liegt in vier stark überarbeiteten Fassungen aus den Jahren 1770, 1774, 1780 und 1820 vor (Schütz / Pöckelmann 2014).

Für die Kollationierung verschiedener Textfassung gibt es bereits eine Reihe digitaler Werkzeuge. Drei der bekanntesten, CollateX , Juxta und TUSTEP , liefern für viele Fragestellungen gute Ergebnisse, unterscheiden sich in ihren Anwendungsszenarien aber erkennbar von LERA. Während LERA einen zweistufigen Ansatz verfolgt, bei dem vor dem detaillierten Textvergleich zunächst eine Alignierung größerer Textpassagen berechnet wird, was denEinsatz komplexer Signaturwerte für den Vergleich notwendig macht, liegt der Fokus von CollateX auf der Alignierung (normalisierter) Token, für den auf einfache Zeichenketten als Signaturwerte zurückgegriffen wird. Juxta zeigt mit seinen hilfreichen Visualisierungen die Unterschiede zu einer ausgewählten Leithandschrift auf, während LERA für den direkten Vergleich mehrerer Textfassungen konzipiert wurde und stets die Textänderungen zwischen allen Fassungen darstellt. Der wohl vielseitigste digitale Werkzeugkasten aus dem Bereich der Geisteswissenschaften, TUSTEP, bietet ebenfalls Möglichkeiten zum Vergleich verschiedener Textfassungen. Allerdings erfordert der Umgang mit dem textbasierten Interface eine längere Einarbeitungszeit, die sich zumindest für einige Anwender durch die neu entwickelte XML-basierte Variante TXSTEP verringert. LERA wurde hingegen von Beginn an durch interaktive graphische Elemente als intuitiv bedienbare Arbeitsumgebung entwickelt. Veränderte Vergleichsparameter sollen dabei durch schnelle Neuberechnung eine direkte Präsentation des Ergebnisses ermöglichen und so zum Experimentieren einladen.

Explorative Analyse mit LERA

Die Arbeitsumgebung erzeugt eine synoptische Gegenüberstellung von größeren Textsegmenten, welche die Grundlage für den Textvergleich bilden. Für die „Histoire des deux Indes“, die in vielen Drucken mit variierendem Zeilenfall vorliegt, wurden Absätze als Vergleichsebene gewählt. Unterschiede zwischen den Textfassungen werden durch Farbmarkierungen in der Synopse und in einem gemeinsamen Variantenapparat dargestellt. Das Vergleichsergebnis kann durch Filtereinstellungen beeinflusst und so an verschiedene Fragestellungen angepasst werden. Beispielweise lassen sich orthographische Varianten ausblenden, so dass dem Nutzer hauptsächlich die inhaltlichen Unterschiede präsentiert werden.

Aufbauend auf dieser Grundfunktionalität wurden die drei im Folgenden beschriebenen Komponenten in LERA integriert, die das Überblicken und Auswerten der Textunterschiede erleichtern sollen.

Integrierte Suche

Als erster wichtiger Baustein wurde eine Suchfunktion eingebunden, die das schnelle Auffinden einzelner Schlagwörter ermöglicht. Das Suchfenster zeigt an, wie häufig der – intern normalisierte – Begriff in der gesamten Synopse vorkommt. Die Treffer der Suche sind dabei in den

Texten farbig unterlegt und können über Navigationspfeile angesteuert werden.

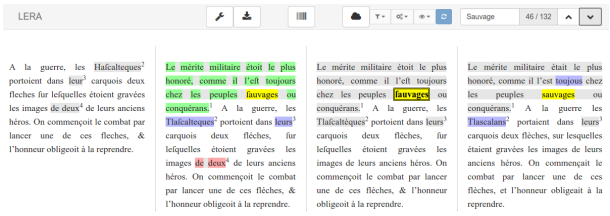

***Abb. 1**: Suchfunktion in LERA: Die gelb hervorgehobenen Suchtreffer können durch die Schaltflächen der Suchmaske (oben rechts) angesteuert werden.*

CATview: Strukturelle Übersicht

Der oben beschriebene Textvergleich bildet die Datenbasis der integrierten interaktiven Übersichts- und Navigationsleiste CATview (Pöckelmann et al. 2015). Die Synopse wird schematisch in CATview dargestellt, indem jedes Segment der Vergleichstexte als Rechteck abgebildet und entsprechend seiner Position in der Synopse angeordnet wird. Die Farbe der Rechtecke zeigt die Intensität der Überarbeitung an. Je stärker dabei ist die Veränderung des Textsegments ausfällt, desto dunkler wird das Farbfeld dargestellt. So verschafft CATview einen Überblick über die Struktur der Texte und die Verteilung der Unterschiede. Zudem erleichtern weitere Funktionen wie die Verlinkung der Rechtecke mit den entsprechenden Textsegmenten die Navigation in der Synopse.

***Abb. 2**: Die in LERA integrierte CATview stellt Struktur und Unterschiede zwischen den Textfassungen schematisch dar und erleichtert die Navigation durch Verlinkungen und einen Bildlauf-Indikator (orange).*

Wortwolken: Thematische Übersicht

Zusätzlich zum synoptischen Textvergleich und dessen Visualisierung in CATview kann die Arbeitsumgebung LERA für jede Textfassung eine interaktive Wortwolke generieren, die einen Vergleich des Auftretens und der Häufigkeit von Schlagwörtern ermöglicht und so einen Überblick über die Inhalte der jeweiligen Textfassung gibt. Welche Wörter dabei angezeigt werden, hängt von den Nutzervorgaben ab. So kann beispielsweise die Auswahl der Wörter auf einzelne Wortarten beschränkt oder eine Mindestlänge festgelegt werden. Zudem sind das Bewertungskriterium (Häufigkeit, Tf-idf-Maß) und verschiedene Darstellungsformen wählbar. Durch die Gegenüberstellung der Wortwolken aller Textfassungen wird das Erkennen einer veränderte Wortwahl und Themensetzung zwischen den Textfassungen erleichtert.

***Abb. 3**: Von LERA generierte Wortwolken für vier Textfassungen: Entsprechend der Nutzereinstellungen wurden als eine der wählbaren Darstellungsformen dreidimensionale Kugeln mit den jeweils 20 häufigsten Wörtern generiert und die Wortgröße über alle Fassungen hinweg ermittelt. So wird beispielsweise die veränderte Häufigkeit des Begriffs „espagnol“ kenntlich (67, 77, 63 und 81).*

Das Zusammenspiel der Werkzeuge

Die innovative Kombination der beschriebenen Ansätze zur Analyse und Darstellung von Varianten ermöglicht das effiziente Erkunden umfangreicher Texte und die wirkungsvolle Visualisierung von Suchergebnissen. Die Arbeitsumgebung bietet so verschiedene Zugänge zur explorativen Analyse an, von denen im Folgenden drei Kombinationsmöglichkeiten erläutert werden.

Kombination von CATview und Suchfunktion

Die Übersichtsleiste CATview zeigt die Überarbeitung der Textfassungen an und vereinfacht die Navigation in umfangreichen Synopsen. Kombiniert man dieses Werkzeug mit der Suchfunktion, werden die Treffer der Suche nicht nur farbig im Text unterlegt, sondern ebenfalls in den Rechtecken der Übersichtsleiste angezeigt. So kann die Verteilung eines Themas über den gesamten Text auf einen Blick erfasst werden. Mit dieser Funktion lassen sich zudem korrespondierende Textstellen, die nicht direkt aligniert sind, leichter auffinden.

Kombination von Suche und Wortwolken

Die Ergebnisse der Suche werden in den interaktiven Wortwolken ebenfalls farbig hervorgehoben, was das Auffinden der Schlagworte und damit den Vergleich der Relevanz bzw. Häufigkeit in den verglichenen Textfassungen erleichtert. Neben dieser intendierten Suche kann die Kombination der beiden Komponenten den Nutzern Anregungen für die weitere Textexploration geben. In den Wortwolken sind Begriffe enthalten, die auf Grund ihrer Häufigkeit eine Relevanz für den Textinhalt

vermuten lassen, ein Klick auf die angezeigten Begriffe in den Wortwolken löst die Suche aus.

CATview in Kombination mit Wortwolken und Suchfunktion

CATview bietet ein Auswahlwerkzeug an, mit dem der Textbereich für die Analyse eingeschränkt werden kann. Dieses Vorgehen bietet sich an, wenn eine Passage des Textes wegen ihrer Überarbeitung oder der Verteilung von Suchbegriffen besonders interessant erscheint. Durch die Beschränkung der Textauswahl verändert sich sofort die Zusammensetzung der Wortwolken, da auch ihre Datengrundlage eingeschränkt wird, indem nur noch die Begriffe aus der Auswahl betrachtet werden. Die Wortwolken werden für jede neue Auswahl umgehend aktualisiert. Die Auswahlbox kann durch drag-and-drop in der Übersichtsleiste bewegt werden. Durch diese Bewegung verändern sich folglich auch die Wortwolken, sodass stets die Themen der aktuell gewählten Textpassage angezeigt werden. Dadurch erzeugt die Bewegung eine Art interaktive Animation an Hand derer die inhaltliche Entwicklung der Texte veranschaulicht wird. Dieses Vorgehen kann zu neuen interessanten Fragestellungen führen.

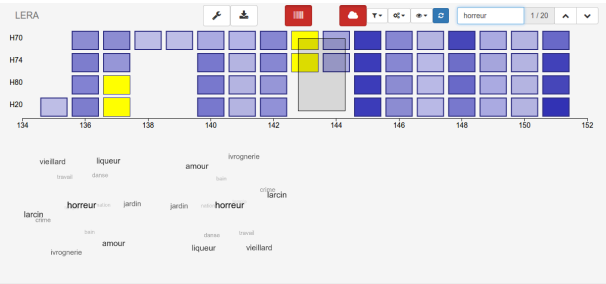

***Abb. 4**: Beispiel für das Zusammenspiel von Suche, CATview und den Wortwolken: Mit Hilfe des Auswahlwerkzeugs von CATview wurden die Absätze 143 und 144 markiert, was sogleich die Wortwolken aktualisiert und so eine grobe inhaltliche Übersicht dieser aus H80 entfernten Textpassage erzeugt. Mit einem Klick auf den Begriff „horreur“ in den Wolken wird eine Suche gestartet, die wiederum in CATview durch die Markierung der Suchtreffer einen Hinweis dafür liefert, dass Teile der Passage bei der Überarbeitung in den Absatz 137 eingeflossen sind.*

Zusammenfassung

Neben der Grundfunktionalität zum Vergleich mehrerer Fassungen eines Textes entsprechend getroffener Nutzereinstellungen wurden in die Arbeitsumgebung LERA mit der Suchfunktion, der Übersichts- und Navigationsleiste CATview sowie den interaktiven Wortwolken drei Komponenten zum Analysieren der gefundenen Unterschiede integriert, deren Kombination das traditionelle Vorgehen der systematischen Textauswertung um die ergebnisoffene Exploration erweitert. So führt das willkürliches Experimentieren mit diesen Komponenten gegebenenfalls zu unerwarteten Erkenntnissen und neuen Hypothesen. Das ist insbesondere für Nutzer von digitalen Editionen interessant, die oft andere Fragestellungen verfolgen als die ursprünglichen Editoren. So werden das Nachvollziehen von Argumentationsketten und die historische Veränderung von Diskursen im Zuge der Überarbeitung von Texten effektiv unterstützt.

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.

Weitere Informationen samt Demonstratoren zu LERA finden Sie auf der Projektseite von SaDA: http://sada.uzi.uni-halle.de

Fußnoten

1. Für einen Absatz fließen beispielsweise dessen Position und enthaltene signifikante Wörter in die Bestimmung des Signaturwerts ein. Ein Wort gilt dabei als signifikant, wenn es in mindestens zwei der zu vergleichenden Texte vorkommt und in der allgemeinen Sprachverwendung relativ selten ist.
2. Die Bestimmung der Wortarten für die Wortwolken erfolgt automatisch mit Hilfe des TreeTaggers . Siehe „TreeTagger - a language independent part-of-speech tagger“.

Bibliographie

ARP (2012): *Juxta*. Compare - Collate - Discover. http://www.juxtasoftware.org/ [letzter Zugriff 15. Oktober 2015].

Bremer, Thomas / Molitor, Paul / Pöckelmann, Marcus / Ritter, Jörg / Schütz, Susanne (2015): "Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen - Das Fallbeispiel der Histoire philosphique des deux Indes von Guillaume Thomas Raynal" in: Nutt-Kofoth, Rüdiger / Plachta, Bodo / Woesler, Winfried (eds.) *Editio*. Internationales Jahrbuch für Editionswissenschaften 29, 1: 29–51.

Bremer, Thomas / Molitor, Paul / Ritter, Jörg / Solms, Hans-Joachim(2012-2015): *Semi-automatische*

Differenzanalyse von komplexen Textvarianten http://sada.uzi.uni-halle.de [letzter Zugriff 15. Oktober 2015].

Medek (*Gießler), André / Pöckelmann, Marcus / Bremer, Thomas / Solms, Hans-Joachim / Molitor, Paul / Ritter, Jörg (2015): "Differenzanalyse komplexer Textvarianten - Diskussion und Werkzeuge", in: *Datenbank-Spektrum* 15, 1: 25-31.

Pöckelmann, Marcus / Medek (*Gießler), André / Molitor, Paul / Ritter, Jörg (2015): "CATview - Supporting The Investigation Of Text Genesis Of Large Manuscripts By An Overall Interactive Visualization Tool", in: *Digital Humanities, DH2015, Sydney, Australia, 29.06.-03.07.2015*.

Pöckelmann, Marcus / Molitor, Paul / Ritter, Jörg (2015): *CATview*. The Colored and Aligned Texts view http://catview.uzi.uni-halle.de/ [letzter Zugriff 15. Oktober 2015].

Schmid, Helmut(1994-): *TreeTagger*. A language independent part-of-speech tagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 15. Oktober 2015].

Schütz, Susanne / Pöckelmann, Marcus (2014): "IT-Werkzeuge zur Unterstützung elektronischer Edition am Beispiel eines französischen Textes aus dem 18. Jahrhundert", in: *9. Kongress des Frankoromanistenverbands, Münster, 24.-27.09.2014*.

The Interedition Development Group (2010-2013): *CollateX*. Software for Collating Textual Sources http://collatex.net/ [letzter Zugriff 15. Oktober 2015].

Zentrum für Datenverarbeitung der Universität Tübingen (1978-): *TUSTEP*. Tuebingen System of Text Processing tools http://www.tustep.uni-tuebingen.de/ [letzter Zugriff 15. Oktober 2015].

Zentrum für Datenverarbeitung der Universität Tübingen / pagina GmbH Publikationstechnologien / Hochschule der Medien Stuttgart (2010-): *TXSTEP*. Die XML-Version von TUSTEP http://www.txstep.de/ [letzter Zugriff 15. Oktober 2015].