eIdentity – Werkzeuge zur Erschließung und Exploration von Textdaten

Cathleen Kantner[1], **Fritz Kliche**[2], **Jonas Kuhn**[3]
[1]Institut für Sozialwissenschaften
Universität Stuttgart
[2]Institut für Informationswissenschaft und Sprachtechnologie
Universität Hildesheim
[3]Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Im Rahmen des BMBF-Verbundprojekts *eIdentity* arbeiten wir aktuell an einem mehrsprachigen Korpus von Zeitungstexten über Kriege und humanitäre militärische Interventionen aus unterschiedlichen Medienarchiven im Umfang von ca. 700.000 Dokumenten. Dieses Korpus wird aus einer politikwissenschaftlichen Perspektive daraufhin untersucht, welche kollektiven Identitäten, beispielsweise *europäische*, *nationale* oder *religiöse Identitäten*, im Zusammenhang mit internationalen Krisen ausgedrückt, beschworen oder kritisiert werden.

Im Projekt *eIdentity* entwickeln wir dazu 1) eine *Explorationswerkbank* für die Aufbereitung und das Management von potentiell heterogen strukturierten Textdaten für die Verwendung sprachtechnologischer Werkzeuge sowie 2) unseren *Complex Concept Builder*, ein System interaktiver Tools zur inhaltlichen Arbeit mit dem Korpus.

1) Explorationswerkbank

Wissenschaftler in den Digital Humanities stehen bei der Nutzung von zuvor nicht bearbeiteten elektronischen Textdaten vor dem Problem, dass es bislang keine benutzerfreundliche Software für die komplexen Aufgaben der Korpuserstellung und -aufbereitung, der Strukturierung von Texten und zugehörigen Metadaten sowie sich anschließende Aufgaben der Samplebereinigung und des Datenmanagements gibt. Doch erst wenn all diese Aufgaben erledigt sind, kann die eigentliche Textanalyse beginnen. Mit der *Explorationswerkbank* entwickeln wir kombinierbare Werkzeuge, die Wissenschaftlern in den Digital Humanities dabei helfen, diese massive Hürde zu Beginn ihrer eigentlichen empirischen Forschung zu bewältigen.

Allerdings stellen verschiedene Wissenschaftler in den Digital Humanities ganz unterschiedliche Anforderungen an das Datenmanagement. Aus der Perspektive der jeweiligen fachlichspezifischen Forschungsfrage sind eine Vielzahl von Entscheidungen zu treffen. Unser Ansatz besteht folglich darin, wo möglich den Anwendern selbst die Steuerung der Verarbeitungsschritte zu ermöglichen, so dass sie dazu nicht auf die Mitwirkung der Werkzeugentwickler angewiesen sind. Die Nutzer können zu verschiedenen Verarbeitungsschritten eigene Zielwerte festlegen. Dieses Konzept umfasst die Import-Werkzeuge, die Werkzeuge zur Datenbereinigung und -Filterung, die Auswahl verwendeter sprachtechnologischer Werkzeuge zum Auffinden relevanter Texte und Textstellen sowie Werkzeuge anhand des gewünschten Typs von Output.

Bereits bei der Korpuserstellung finden wir auf der einen Seite digitale Daten in unterschiedlichen Formaten und Datenstrukturen vor. Auf der anderen Seite lassen sich sprachtechnologische Werkzeuge meist nicht unmittelbar auf beliebiges Datenmaterial anwenden. Dies gilt besonders für die Analyse großer Datenmengen. Die *Explorationswerkbank* bringt beide Aufgaben zusammen. DH-Wissenschaftler können damit ihr rohes Datenmaterial aus verschiedenen Quellen aufbereiten und Texte und Metadaten in einem Repositorium ablegen, das anschließend die Einbindung sprachtechnologischer Werkzeuge erlaubt.

Für den Import roher Daten stellen wir eine Wizard-Funktion bereit. Die Anwender importieren zunächst ein einzelnes Dokument in ein „Vorschau-Fenster“und erstellen anschließend Regeln, um im Dokument Metadaten und Textstrukturbausteine zu definieren. Daraufhin wenden die Import-Werkzeuge die Regeln auf die Dokumente der Datenquelle an (Generalisierung). Zur weiteren Verarbeitung entwickeln wir Werkzeuge für die Bereinigung des erstellten Samples. Die Bereinigung umfasst die Filterung um Dubletten, Semi-Dubletten sowie leere und

defekte Artikel. Der Wert, ab dem ein Artikelpaar als Semi-Dublette ausgezeichnet wird, kann ebenso wie Einstellungen zur Erfassung defekter und leerer Artikel vom Benutzer festgelegt werden. Die Anwender können sich Artikel anzeigen lassen, die die Zielwerte nicht erfüllen. Sie prüfen, ob es sich um defekte oder leere Artikel handelt, passen die Zielwerte entsprechend an und generalisieren anschließend auf das Sample. Die Artikel im Repositorium werden um Prozessmetadaten bereichert. Die Prozessmetadaten halten die Verarbeitungsschritte fest, die ein Dokument zu einem Zeitpunkt bereits durchlaufen hat.

2) Complex-Concept-Builder

Eine zentrale Rolle kommt einer Funktion zur Suche im Repository und der Darstellung von Ergebnissen zu. Wir stehen vor der Herausforderung, dass gesellschafts- und geisteswissenschaftliche Fragestellungen mit abstrakten Konzepten operieren, die sich nicht direkt in der Alltagssprache der aktuell untersuchten Zeitungstexte äußern. Kollektive Identitäten können ganz unterschiedlich ausgedrückt werden: „*wir Europäer* sind wir verpflichtet, den Völkermord im Land X zu stoppen“ ist ein sehr einfacher Fall. Typischer sind Ausdrücke wie „Deutschland sollte endlich Farbe bekennen“ oder „Washington kann in dieser Frage nicht über seinen Schatten springen“. Solche Appelle an unterschiedliche *kollektive Identitäten* verschiedener *politischer Akteure* sind zudem in den Zeitungsberichten verhältnismäßig selten. Um diese Instanzen im Meer der Worte leichter identifizieren zu können, entwickeln wir den *Complex-Concept-Builder*. Er umfasst Werkzeuge zur Topic-Analyse für die Bereinigung um Off-Topic-Artikel und für die Erstellung von Sub-Samples potentiell inhaltlich relevanter Artikel. Wir integrieren dazu über das Mallet-Tool eine Methode zur Textklassifikation, basierend auf Latent Dirichlet Allocation. Die Anwender wählen eine Anzahl n zu differenzierender Topics für ihr Sample. Die Werkzeuge klassifizieren die Texte in n thematisch definierte Gruppen, die den Anwendern präsentiert werden. Die Anwender können nun anhand dieser Beispiele bestätigen oder verneinen, ob die Artikel für ihre Fragestellung relevant sind oder nicht. Auf dieser Basis ergibt sich eine Wahrscheinlichkeit für jeden Artikel zum gesuchten thematischen Bereich zu gehören.

Schließlich haben wir ein Werkzeug zur Berechnung und Visualisierung der Medienaufmerksamkeit im Zeitverlauf („Issue Cylces“) in den Medien behandelter Themen entwickelt. Anwender können hier Themen und zu beobachtende Zeiträume definieren und in einer Zeitreihe darstellen. Mit dem Tool kann die Medienaufmerksamkeit für verschiedene Themen angezeigt werden und Perioden höherer Aufmerksamkeit für ein Thema können identifiziert werden.

Über die *Explorationswerkbank* werden sprachtechnologische Werkzeuge der CLARIN-Infrastruktur als Webservices eingebunden. Sie umfassen Wortartenerkennung, syntaktisches Parsing, Named Entity Recognition, Koreferenzanalyse und Sentimentanalyse. Diese Werkzeuge bilden Bausteine für die Erfassung komplexer Konzepte und können rechenintensiv sein. Wir wenden daher hier das Prinzip der modularen Ablaufketten an: Die Anwender wählen ein verfügbares Analysewerkzeug oder ein angestrebtes Ereignis aus, woraufhin die *Explorationswerkbank* die benötigten sprachtechnologischen Verarbeitungsschritte nennt. Die durchgeführten sprachtechnologischen Analyseschritte werden als Prozessmetadaten zusammen mit den bearbeiteten Texten im Repositorium festgehalten.

Unsere Werkzeuge eröffnen weitreichende Möglichkeiten zur Exploration digitaler Textdaten. *Explorationswerkbank* und *Complex-Concept-Builder* versetzen Wissenschaftler aus den Digital Humanities zudem in die Lage, ohne die Mitarbeit von Tool-Experten die Potentiale der sprachtechnologischen Analyse ihrer Daten auszuprobieren und anzuwenden. Damit weisen sie Wege zu einer noch stärkeren Verbreitung sprachtechnologischer Methoden in den unterschiedlichsten mit großen Textmengen arbeitenden Fächern.