Ein Wizard für die Erschließung strukturierter Textdaten

Fritz Kliche[1], **Nicolas Schmidt**[2], **Ulrich Heid**[1]

[1]Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
[2]Institut für Betriebswirtschaft und Wirtschaftsinformatik, Universität Hildesheim
{kliche,schmi032,heid}@uni-hildesheim.de

Wir stellen einen *Wizard* vor, mit dem strukturierte Textdaten in einer Browser-Anwendung erschlossen werden können, um die textlichen Inhalte und Metadaten für textwissenschaftliche Analysen nutzen zu können. Der *Wizard* ist ein interaktives Werkzeug, das es dem Benutzer erlaubt, von Beispielfällen auf größere Mengen von Daten zu generalisieren, ohne dass er dazu zu programmieren braucht.

Die Voraussetzung sind Textdaten, deren Textstruktur nicht für jeden Text unterschiedlich ist, sondern wo sich ähnliche Textstrukturen über größere Mengen von Einzeltexten hinweg beobachten lassen. Beispiele sind Sammlungen von Zeitungsartikeln, Sammlungen von Blogs und user-generated content oder Protokolle von Parlamentsdebatten. Solche Texte sind einerseits nicht standardisiert, andererseits doch relativ homogen repräsentiert, mindestens innerhalb jeder Kollektion, jedes Zeitungs- oder Blog-Archivs. Die extrahierten Inhalte werden als *Textobjekte* in einer Datenbank abgelegt und mit Labels versehen, die den Zugriff ermöglichen. Über diesen Zugriff können die Textobjekte in eine neue Datenstruktur überführt werden. Der *Wizard* entsteht innerhalb des DH-Projekts *e-Identity* (Blessing et al., 2013), in dem ein umfangreiches Sample von Zeitungsartikeln (>800.000 Artikel) aus 5 digitalen Medienportalen erschlossen und ein Korpus erstellt wurde, in dem die textlichen Inhalte der Artikel und die begleitenden Metadaten kategorisiert vorliegen.

Der *Wizard* führt den Anwender durch die Funktionen, die über eine Browser-GUI gesteuert werden. Zunächst können Textdaten in unterschiedlichen Formaten (RTF, DOCX, ODT, TXT, HTML) und Zeichencodierungen (UTF-8, ISO-8859-1) importiert werden. Die anschließende Erschließung erfolgt in zwei Schritten: (1) Die Textdaten werden zunächst in strukturelle Einheiten (z. B. in *e-Identity*: Zeitungsartikel) segmentiert; (2) in diesen Einheiten werden anschließend textliche Inhalte und Metadaten erkannt und klassifiziert, indem ihre Anfangs- und Endpunkte im fortlaufenden Textmaterial identifiziert werden. Dafür werden in einem Vorschau-Fenster Ausschnitte der importierten Textdaten angezeigt. Der Anwender erstellt anhand solcher Beispiele *Extraktionsregeln*, d. h. Muster, nach denen Textobjekte identifiziert werden. Mit der Erstellung mehrerer Extraktionsregeln entsteht ein Regelset als eine Schablone, mit der in den importierten Daten Inhalte erkannt und extrahiert werden. Für die Extraktionsregeln wurden Elemente einer Regelsprache implementiert, über die der *Wizard* computerlinguistische Konzepte (reguläre Ausdrücke, Text Mining, computerlinguistische Prozessierung) textwissenschaftlichen Anwendern möglichst intuitiv zugänglich macht. Die im Folgenden dargestellten Konzepte wurden umgesetzt:

Integrierte computerlinguistische Werkzeuge
Der *Wizard* integriert computerlinguistische Verarbeitungsschritte zur Tokenisierung, Lemmatisierung, Wortartenerkennung und zur Erkennung von Eigennamen. Weiter werden Tokens verschiedenen „Tokentypes“ zugeordnet (Versalien, groß- oder kleingeschriebene Wörter, Zahlwörter usw.). Die entsprechende computerlinguistische Verarbeitung soll einerseits im Hintergrund stattfinden; andererseits soll sie von den Anwendern gesteuert werden können. Wir trennen dazu die Erstellung der Extraktionsregeln von ihrer Anwendung. Nach der Erstellung einer Schablone validiert der *Wizard* deren Regeln und prüft, welche computerlinguistischen Vorverarbeitungsschritte sie verlangen. Der Anwender muss also nicht entscheiden, welches computerlinguistische Werkzeug an welcher Stelle zum Einsatz kommen soll, sondern welches Prozessierungsergebnis angestrebt wird. Wo nötig, wird dazu ein computerlinguistischer Verarbeitungsschritt vom Werkzeug vorgeschlagen und eingeschoben.

Unterschiedliche Textobjekte
In den Daten können unterschiedliche Textobjekte definiert werden. Als Textobjekte sind Tokens, Segmente (d. h. einzelne Zeilen) und mehrzeilige Objekte möglich.

Merkmale zur Identifikation
Zur Erstellung der Extraktionsregeln können unterschiedliche textliche Merkmale berücksichtigt werden. Als Indikator eines Textobjekts können (1) ein Ankerwort oder (2) ein regulärer Ausdruck definiert werden; (3) die maximale und die minimale Länge eines Segments können festgelegt werden; (4) um das Segment im Kontext zu definieren, können Ankerwörter und reguläre Ausdrücke zum Vorgänger- oder Nachfolgersegment des zu bestimmenden Segments definiert werden. (5) Schließlich kann die Abfolge unterschiedlicher Typen von Tokens definiert werden, die über die Wortart, einen Abgleich mit Terminologielisten (z. B. eine Liste von Monatsnamen), „Tokentypes“ oder über eine feste Zeichenkette charakterisiert werden.

Funktionen der Extraktionsregeln
Die Extraktionsregeln dienen zunächst der Identifikation von Textobjekten; sie können auch als Blocker fungieren, die die Identifizierung von Objekten durch andere Regeln verhindern.

Der Aufbau einer eigenen Datenstruktur
Die erkannten Objekte werden in einer Datenbank mit ihrem Label abgelegt. Durch das Label kann auf die Daten zugegriffen werden. Damit sind die Daten für den Aufbau einer neuen Datenstruktur zugänglich. Die Daten können in ein XML-Format konvertiert werden. Wir planen die Möglichkeit zur Konvertierung in gängige Formate: TEI, CMDI (Broeder et al., 2012), etc.

Anwendung für Textwissenschaftler in DH-Projekten
Das Poster richtet sich besonders an textwissenschaftliche Anwender. Screenshots stellen die Arbeitsschritte zur Erschließung strukturierter Textdaten dar. Aus computerlinguistischer Sicht zeigt das Poster ein Beispiel, wie linguistische Annotationen in ein Softwareprojekt für die Digital Humanities eingebunden werden. In einer Demonstration können die Benutzerschnittstellen des Systems vorgeführt werden.

Literatur

Blessing, André; Sonntag, Jonathan; Kliche, Fritz; Heid, Ulrich; Kuhn, Jonas; Stede, Manfred (2013). Towards a tool for interactive concept building for large scale analysis in the humanities. In: *Proceedings des 7. Workshops „Language Technology for Cultural Heritage, Social Sciences, and Humanities“*. Association for Computational Linguistics, Sofia, Bulgarien.

Broeder, Daan; Windhouwer, Menzo; van Uytvanck, Dieter; Goosen, Twan; Trippel, Thorsten (2012). CMDI: a component metadata infrastructure. In: *Proceedings des Workshops „Describing Language Resources with Metadata“*. LREC 2012, Istanbul, Türkei.