Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Sonderband 1 der ZfdG: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum und Thomas Stäcker. 2015. DOI: 10.17175/sb01

Titel:
Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien

Autor/in:
André Blessing

Kontakt: andre.blessing@ims.uni-stuttgart.de
Institution: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
GND: 1058601865 ORCID:

Autor/in:
Fritz Kliche

Kontakt: kliche@uni-hildesheim.de
Institution: Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
GND: 1084131943 ORCID:

Autor/in:
Ulrich Heid

Kontakt: uli@ims.uni-stuttgart.de
Institution: Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
GND: 111873967 ORCID:

Autor/in:
Cathleen Kantner

Kontakt: cathleen.kantner@sowi.uni-stuttgart.de
Institution: Institut für Sozialwissenschaften, Universität Stuttgart
GND: 129464872 ORCID:

Autor/in:
Jonas Kuhn

Kontakt: jonas.kuhn@ims.uni-stuttgart.de
Institution: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
GND: 172996090 ORCID:

DOI des Artikels:
10.17175/sb001_013

Nachweis im OPAC der Herzog August Bibliothek:

Erstveröffentlichung:
30.06.2015

Medienlizenzen:
Medienrechte liegen bei den Autoren

Letzte Überprüfung aller Verweise:
24.05.2016

GND-Verschlagwortung:
Computerlinguistik | Textanalyse |

Zitierweise:
André Blessing, Fritz Kliche, Ulrich Heid, Cathleen Kantner, Jonas Kuhn: Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb001_013.

André Blessing, Fritz Kliche, Ulrich Heid, Cathleen Kantner, Jonas Kuhn

Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien

Abstracts

Die Digital Humanities stoßen auf neuartige Probleme, wenn sie Fragen der theoriegeleiteten sozialwissenschaftlichen Forschung bearbeiten. Hier kann die Textanalyse nicht unmittelbar auf linguistischen Merkmalen aufsetzen, sondern sucht den Zugang zu komplexen Begriffen und deren Bedeutungen. Unser Beitrag zur Methodologie der Digital Humanities argumentiert, dass hermeneutisch sensible Korpusanalyseverfahren entwickelt werden können, wenn die eingesetzten Werkzeugkomponenten für die Zieldisziplin transparent bleiben und sich für eine interaktive Exploration des Datenmaterials eignen. Wir stellen interaktive Werkzeuge zur Texterschließung und -analyse vor, die sich flexibel auf fachwissenschaftliche Theorien und Forschungsfragen abstimmen lassen, jedoch gleichzeitig in ihrer Architektur so generisch sind, dass sie breit einsetzbar sind.

Theory-driven research in the social sciences is confronting the Digital Humanities with new challenges. Here, text analysis can not be built directly on linguistic features, but rather the research questions require access to more complex concepts and their meaning. Our methodological contribution to DH argues that it is possible to develop hermeneutically sensitive methods for corpus analysis when the tool components remain transparent for the target discipline and allow for interactive exploration of the underlying data. We provide an overview of interactive tools for text processing and analysis that can be flexibly adjusted to specific disciplinary theories and research questions in the target disciplines, with an underlying architecture that is generic and therefore broadly applicable.

1. Einleitung

Die vielfältige Verfügbarkeit von größeren digitalen Textsammlungen eröffnet grundsätzlich den Zugang zu Fragestellungen für die geistes- und sozialwissenschaftliche Forschung, die mit traditionellen Verfahren nicht oder nur mit unrealistisch großem Zeitaufwand untersucht werden könnten. So kann beispielsweise anhand eines datierten Textkorpus die zeitliche Entwicklung von Textinhalten im größeren Maßstab betrachtet werden – unter zusätzlicher Berücksichtigung von Parametern (wie beispielsweise dem Kontext der untersuchten Texte), zu denen Metadaten wie Land, Autoren o.ä. vorliegen. Das entsprechend datierte und kategorisierte textuelle Datenmaterial scheint über digitale Archive zunächst leicht zugänglich für quantitative Analysen zu sein. Doch dieser erste Eindruck täuscht.

Tatsächlich erweist sich eine systematische, durch geistes- und sozialwissenschaftliche Theoriebildung fundierte Analyse von Textsammlungen als sehr anspruchsvoll. Dies gilt besonders dann, wenn die Textanalyse nicht wie in den sprachwissenschaftlichen Disziplinen unmittelbar auf die Konzepte der linguistischen Theoriebildung aufsetzen kann (beispielsweise durch eine quantitative Analyse der Verwendung von Passivkonstruktionen bei einer Klasse

von Verben), sondern lediglich vermittelnde Funktion beim Zugang zu den Textinhalten bzw. zu nicht-linguistischen Texteigenschaften hat.

Der vorliegende Beitrag stellt die Prozeduren und Tools zur Texterschließung und -analyse im Projekt e-Identity[1] vor, in dem ein großes multinationales und somit mehrsprachiges Zeitungskorpus im zeitlichen Verlauf von über 20 Jahren aus politikwissenschaftlicher Sicht analysiert wird: Wie entwickelten sich die Debatten über Kriege, bewaffnete Konflikte und humanitäre militärische Interventionen in unterschiedlichen Ländern? Welche Rolle spielten kollektive Identitäten (ethnische Zugehörigkeiten, ›wir Europäer‹, nationale Identitäten, Religionsgemeinschaften etc.) in der Diskussion über internationale Krisensituationen?

Obgleich eine (computer-)linguistische Analyse der Zeitungstexte keinerlei direkte Bedeutung für die untersuchten Fragen hat, kommt ihr als Zwischenschritt eine zentrale Rolle bei der Systematisierung des Vorgehens und der validen Verankerung der quantitativen Untersuchungen in der sozialwissenschaftlichen Theoriebildung zu. Das Ergebnis ist eine sorgfältig abgestimmte interaktive Umgebung für die fachwissenschaftliche Textanalyse, die sich sehr stark auf theoriespezifische Teilfragestellungen zuschneiden lässt, gleichzeitig jedoch in ihrer Werkzeugarchitektur so generisch bleibt, dass sie sich ebenso systematisch auf ganz andere inhaltsanalytische Fragestellungen in großen Textsammlungen anpassen lässt.

Aufgrund dieser über das konkrete Projekt hinausreichenden methodologischen Herausforderungen versteht sich dieser Artikel über die spezifisch technische Vorstellung des entwickelten Werkzeuginventars[2] hinaus vor allem als Beitrag zur Methodendiskussion in den Digital Humanities und argumentiert für die These, dass die Untersuchung von großen digitalen Sammlungen dann neuartige, hermeneutisch sensible Verfahren hervorbringen kann, wenn die eingesetzten Analyseverfahren und Werkzeuge für die Zieldisziplin transparent und kritisch hinterfragbar bleiben und sich für eine interaktive Exploration des Datenmaterials unter fachwissenschaftlichen Auspizien eignen. Hierfür ist eine intensive Abstimmung der technischen Modellierung mit den fachwissenschaftlichen Fragestellungen und Hypothesen von entscheidender Bedeutung.

Dieser Artikel skizziert in Abschnitt 2 den fachwissenschaftlichen Hintergrund der Fragestellungen, die im e-Identity-Projekt verfolgt werden, und legt die Grenzen einer textuell-oberflächennahen quantitativen Analyse auf den Artikelsammlungen dar. Abschnitt 3 stellt den auf die theoretischen Bedürfnisse zugeschnittenen technischen Werkzeugeinsatz im

[1] Diese Studie entstand im Rahmen des von Prof. Dr. Cathleen Kantner, Prof. Dr. Jonas Kuhn, Prof. Dr. Manfred Stede und Prof. Dr. Ulrich Heid durchgeführten interdisziplinären Verbundprojekts »Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften (e-Identity)«. Wir danken dem Bundesministerium für Bildung und Forschung für die Förderung in den Jahren 2012 bis 2015 im Rahmen der eHumanities-Initiative (Förderkennzeichen: 01UG1234A).

[2] Der Skopus des vorliegenden Beitrags ist nicht das vollständige Inventar der möglichen computerlinguistischen Werkzeugunterstützung, das im e-Identity-Projekt verwendet und entwickelt wird. Die (halb-)automatische Unterstützung beim Auffinden von Textpassagen, in denen relevante komplexe Konzepte der politikwissenschaftlichen Theoriebildung artikuliert werden, wird hier beispielsweise nicht thematisiert (wenngleich die interaktive Unterstützung der manuellen Textauszeichnung eine wichtige Teilkomponente hierfür darstellt). Weitere Projektbeiträge, u.a. auch von den Projektpartner an der Universität Potsdam, finden sich auf der Projektwebsite

Projekt konkret vor: die computerlinguistisch fundierte systematische Textsammlung und -aufbereitung, die in der sogenannten *Explorationswerkbank* bereitgestellt wird (Abschnitt 3.1), die technische Unterstützung einer anspruchsvollen Annotation bzw. Kodierung von Passagen aus der Textsammlung durch Politikwissenschaftler (Abschnitt 3.2) und die Bereitstellung von computerlinguistischen Teilanalysen für weitergehende automatische Analyseschritte (Abschnitt 3.3). Abschnitt 4 schließt mit einem knappen Resümee zur disziplinübergreifenden Zusammenarbeit im Projekt.

2. Inhaltsanalytische Studien auf großen Textsammlungen und das e-Identity-Projekt

Die zunehmende, vielfältige und bequeme Verfügbarkeit großer Mengen digitalisierter Texte über umfangreiche elektronische Textarchive bietet den Sozialwissenschaften im weitesten Sinne viele neue Chancen für die Forschung. Die netzbasierte Verfügbarkeit solchen Textmaterials ermöglicht zudem einen ländervergleichenden Ansatz, der den Bedürfnissen moderner gesellschaftswissenschaftlicher Fragestellungen in einer sich globalisierenden Welt entgegenkommt. Kontinuierlich gesammelte Jahrgänge von Zeitungsinhalten, Parlamentsdebatten, offizielle Dokumente, Rechtstexte, literarische Gesamtausgaben, historische Archive – um nur einige zu nennen – stehen in immer mehr Sprachen umfassend und gut aufgearbeitet bereit. Hinzu kommt eine reiche Auswahl zunehmend anwenderfreundlicherer Textanalysesoftware,[3] deren kommerzielle, eher für die qualitativ-interpretierende Inhaltsanalyse geeignete Angebote noch relativ teuer, aber doch meist finanzierbar sind, sowie viele kleinere Anwendungen und umfassende Plattformen (z.B. CLARIN, DARIAH), die computer- oder korpuslinguistische Verfahren der Textanalyse einem breiten Spektrum von akademischen Nutzern zugänglich machen.

Doch warum nutzen die Sozialwissenschaften die vorhandenen neuen Möglichkeiten computer- oder korpuslinguistischer Verfahren der Textanalyse noch so wenig? Warum dominieren immer noch die qualitativ-interpretativen, small-n-Forschungsdesigns über solche, die zumindest auch die Chancen computer- und korpuslinguistischer Methoden für large-scale, big-n, long-time-Forschungsdesigns in den Sozialwissenschaften erschließen? Liegt es an mangelnder Anwenderfreundlichkeit? Oder an einer schlechten Werbung? Wir behaupten, dass die vorhandenen Potentiale aufgrund beträchtlicher Hindernisse nicht ausgeschöpft werden können. Erst wenn diese abgebaut werden, können ebenfalls notwendige Schritte der Verbesserung der Anwenderfreundlichkeit und der Bekanntheit neuer Methoden aus den Digital Humanities den gewünschten Erfolg zeitigen.

Unserer Erfahrung nach scheitert die interdisziplinäre Zusammenarbeit in den Digital Humanities oft an zwei typischen neuralgischen Punkten:

1. Der extreme Aufwand bei der individuellen Forschungsfragen folgenden Korpuserstellung und der damit verbundenen Materialaufbereitung sowie dem

[3] Für einen Überblick: Krippendorff 2004, S. 281-307; Alexa / Zuell 2000.

Datenmanagement großer Textmengen aus heterogenen und auch in Zukunft nicht standardisierbaren Quellen lässt Sozialwissenschaftler meist schon im Vorfeld der eigentlichen Anwendung computer- und korpuslinguistischer Ansätze scheitern.
2. Sozialwissenschaftler sind daran interessiert, über die Analyse manifester Textinhalte komplexe Sinnzusammenhänge zu rekonstruieren. Sie suchen meist nach *abstrakten* Konzepten, die in der Alltagssprache kaum direkt geäußert werden. Gängige Anwendungen und wörterbuchbasierte Tools werden der Heterogenität der Anforderungen und der Notwendigkeit, von Forschungsprojekt zu Forschungsprojekt neue *Operationalisierungen komplexer fachwissenschaftlicher Begriffe* vorzunehmen, nicht gerecht.

(ad 1) : Die Verfügbarkeit digitaler Textarchive (insbesondere von Zeitungstexten über LexisNexis oder Factiva, politischen Dokumenten und Rechtstexten z.B. über EU-Lex) führte zu einem Boom von textanalytischen Forschungsprojekten in den Sozialwissenschaften – oft mit ländervergleichendem Interesse. Doch meist werden kleine Korpora intensiv mit qualitativen Verfahren analysiert. Korpus- und computerlinguistische Verfahren, mit denen sehr viel größere Fallzahlen bewältigt werden könnten, sind die seltene Ausnahme. Wenige Sozialwissenschaftler nutzen diese Möglichkeiten bisher.[4] Sie tun dies auch in der Hoffnung, so Daten zu generieren, die Anspruch auf Repräsentativität erheben können und sich mit den gleichen anspruchsvollen statistischen Verfahren auswerten lassen wie die ›harten‹ Daten der quantitativen Sozialforschung, um sich zu diesen in Beziehung zu setzen.[5]

Leider erweist sich die Aussicht auf leichten Zugang zu großen Textmengen oft als Falle. Schon im Vorfeld der Analyse großer Textmengen stellen sich Sozialwissenschaftlern oft unüberwindliche Probleme bei der Erstellung und Aufbereitung des für ihre Fragestellung relevanten Korpus. Manche Archive erlauben das schnelle Download von bis zu 200 Texten *en bloc* (z.B. LexisNexis), andere nur Text für Text. *En bloc* geladene Texte müssen vor der Bearbeitung jedoch wieder in Einzeltexte zerlegt werden. Verschiedene Quellen bieten Texte in unterschiedlichen Text- und Zeichenkodierungen an, gehen mit Sonderzeichen unterschiedlich um und kennzeichnen Metadaten (Datum, Quelle, Autoren etc.) auf unterschiedliche Weise. All dies erschwert das Einlesen des Textmaterials in ein einheitliches Format und erfordert eine Vielzahl von Arbeitsschritten, die nur unter Zuhilfenahme etlicher passgenau programmierter Software-Skripte zu bewältigen sind.

Die großen, leicht zugänglichen Textarchive lassen sich nur mit relativ einfachen Schlagworten und booleschen Schlagwortkombinationen durchsuchen, was zu einer Fülle von semantischen Doppeldeutigkeiten und damit Samplingfehlern[6] oder – bei zu spezifischer

[4] Große Textmengen wurden für ländervergleichende Untersuchungen politischer Kommunikation z.B. von den folgenden Autoren genutzt: Baker / McEnery 2005; Kantner 2006b; Kantner 2009; Kantner et al. 2008; Koenig et al. 2006; Kutter 2007; Liebert 2007; Renfordt 2009.

[5] Herrmann 2002, S. 125.

[6] Texte werden im Prozess des Sampling irrtümlich ausgewählt, wenn die Suchworte im Text metaphorisch gebraucht werden. Wird wie in unserem Fall mit Suchworten aus dem Themenfeld Krieg, Militär, Friedenstruppen sowie einer Liste von Krisenstaaten im Untersuchungszeitraum gesucht, können sich durch kriegerische Metaphern z.B. in der Sportberichterstattung Samplingfehler ergeben (z.B. könnte darüber berichtet werden, dass in einem Fußballspiel der Schiedsrichter in der 56. Minute intervenierte ..., ein Spieler aus Bosnien sich verletzte oder auf den Zuschauerrängen teilweise bürgerkriegsartige Zustände herrschten). Sport-Vokabeln auszuschließen, ist jedoch kein geeignetes Gegenmittel, da umgekehrt relevante Artikel oft

Fassung – zu Auslassung von Teilen des relevanten Materials führt. Bei der Auswahl der inhaltlich relevanten Texte (Sampling) aus Archiven fallen zudem Dubletten[7] an. Dieses ›weiße Rauschen‹ (*noise*)[8] muss entfernt werden, wenn valide und reliable Forschungsergebnisse erzielt werden sollen.

Diese Probleme sind keine ›Einstiegsprobleme‹ der Umstellung auf digitales Arbeiten, die im Laufe der Zeit verschwinden werden. Ein wesentlicher Zug der Originalität sozialwissenschaftlicher Forschung besteht darin, beständig neue Quellen zu erschließen. Selbst wenn eines Tages ganze Bibliotheken, Zeitungsarchive und Dokumentensammlungen von Rechtstexten in standardisierter elektronischer Form vorlägen, stellten sie doch immer nur einen Teil des möglicherweise relevanten Materials dar. Zur Lösung der genannten Probleme gibt es jedoch noch keine kommerziellen Softwarepakete, die die notwendigen Arbeitsschritte benutzerfreundlich integrieren. Jedes Team schreibt sich daher die nötigen Skripte selbst – was den Teilnehmerkreis automatisch auf solche mit Informatikern beschränkt, Insellösungen hervorbringt und eine erhöhte Gefahr von methodischen Artefakten birgt –, oder aber man kapituliert vor allzu großen Textmengen. Im letzteren Falle sind Sozialwissenschaftler in der Regel auf solche Textmengen verwiesen, die sie noch vollständig lesen, inhaltlich überblicken und manuell bereinigen können. Ein Korpus, das jedoch so klein ist, dass es sich manuell bereinigen lässt, lässt sich in der Folge auch qualitativ-interpretativ auswerten. Computer- und korpuslinguistische Verfahren kommen erst gar nicht zur Anwendung.

Zur Schließung dieser fundamentalen Lücke entwickelt e-Identity eine *Explorationswerkbank* für Korpuserstellung, -bereinigung und -management, welche die Nutzer unterschiedlichsten Quellen, Textformaten und Sprachen individuell anpassen können (Abschnitt 3.1 und 3.2).

(ad 2) : Schwierigkeiten bei der Operationalisierung komplexer fachwissenschaftlicher Begriffe resultieren einerseits aus den Besonderheiten des Gegenstandsbereichs der Sozialwissenschaften. Die Gegenstände dieser Fächer sind zum großen Teil keine materiellen Faktizitäten, die durch die Beobachtung von einem unbeteiligten, ›objektiven‹ Beobachter gemessen werden könnten.[9] Gerade die interessantesten und am intensivsten debattierten Forschungsgegenstände dieser Fächer sind in den Worten eines der französischen Gründerväter der Soziologie »soziale Fakten«.[10] Abstrakte Gegenstände wie z.B. »Identitäten«, »Werte«, »Staat«, »Macht«, »soziale Gerechtigkeit« kann man nicht anfassen. Sie existieren nicht außerhalb einer sprachlich verfassten Praxis der Kommunikation und Interaktion, in der die Menschen sich Gedanken über ihr Zusammenleben machen, darüber streiten und in

sportliche Metaphern gebrauchen oder über sportliche Ereignisse berichten (z.B. wenn die multinationalen Truppen in Kundus ein Freundschaftsspiel organisierten, bei dem einheimische, niederländische und deutsche Soldaten gegeneinander antraten).

[7] Typische Ursachen dafür sind z.B. Dopplungen in den Datenbanken (z.B. A-, B- und Online-Versionen von Zeitungstexten) oder unterkomplexe Suchmasken, die eine Zerlegung einer komplexeren Schlagwortkombination nötig machen.

[8] Gabrielatos 2007, S. 6; Kantner et al. 2011.

[9] Solche ›harten‹ Fakten gibt es natürlich auch. Sie sind z.B. der statistischen Datenerhebung zugänglich. Beispiele hierfür wären demographische und sozio-ökonomische Daten (z.B. Verteilung von Bildung, Einkommen, Armut, Arbeitslosigkeit etc.) oder Daten über politische Präferenzen (z.B. Umfragen, Sonntagsfrage, Wahlergebnisse).

[10] Durkheim 1964 [1895].

Diskursen gemeinschaftlich deuten, sozusagen ›sozial konstruieren‹ und dabei immer wieder neu interpretieren.

Zum anderen resultieren notorische Probleme der Operationalisierung aus der Kluft zwischen theoretischen Konzepten und der Alltagssprache. Sozialwissenschaftler suchen in verschiedenen Arten von Texten Antworten auf Fragen, die vor dem Hintergrund ihrer wissenschaftlichen Theorien relevant sind. Sie wollen nicht einfach beschreiben, *wie* über etwas sprachlich kommuniziert wird, sondern sie beobachten Kommunikationsprozesse durch die Brille theoretischer Begriffe, die gerade *nicht* direkt in der Alltagssprache verwendet werden. Der begriffliche Gehalt ist dabei selbst immer Gegenstand wissenschaftlicher Kontroversen, weil er abhängig ist von der Art und Weise, wie ein Gegenstand oder Gegenstandsbereich im Hinblick auf eine leitende Fragestellung konzeptualisiert wird. Dies hat einen notorischen Pluralismus der sozialwissenschaftlichen Begrifflichkeiten zur Folge. Ein exzellentes Beispiel für diese Schwierigkeiten stellt die Analyse kollektiver Identitäten dar. Kollektive Identität interessiert viele Sozialwissenschaftler. Der Historiker Lutz Niederhammer und andere beklagen eine ausufernde Proliferation von Identitätskonzepten,[11] dennoch kommt man ohne diesen Begriff nicht aus,[12] mit dem u.a. erfasst werden soll, was denn eigentlich Gemeinschaften zusammenhält, was den sozialen und kulturellen Kitt ausmacht. Was wir begrifflich unter »kollektiver Identität« verstehen und wie wir es empirisch operationalisieren, ist wesentlich abhängig davon, welchen Identitätsbegriff wir verwenden. Derzeit sind recht unterschiedliche und wohl auch sich wechselseitig ausschließende Identitätsbegriffe auf dem Theoriemarkt vorhanden: zwischen dem sozialpsychologischen, einem hermeneutisch-pragmatischen oder einem poststrukturalistischen Identitätsbegriff liegen Welten. Mit unterschiedlichen theoretischen Brillen achten Forscher auf ganz unterschiedliche Ausdrücke: Aus der Perspektive differenztheoretischer Ansätze (so von Carl Schmidt, Niklas Luhmann, Jacques Derrida und Michel Foucault) zeigt sich der Diskurs einer Ingroup an der Art, wie sie Outgroups (*the other*) benennt, stigmatisiert und sich in Abgrenzung zu ihnen definiert. Sozialpsychologische Ansätze würden nach auffälligen positiven emotionalen Ausdrücken für die eigene Gruppe als Held oder Opfer im untersuchten Textmaterial suchen. Hermeneutisch-pragmatische Ansätze würden nach problemorientierter Kommunikation über das, was aufs Ganze gesehen für das jeweilige Kollektiv – aus Sicht seiner Mitglieder – gut ist, suchen.[13]

Empirisch folgt daraus, dass gerade die interessantesten theoretischen Konzepte in den Sozialwissenschaften von verschiedenen Fachwissenschaftlern höchst unterschiedlich operationalisiert werden. In e-Identity versuchen wir das Mit- oder Gegeneinander von verschiedenen kollektiven Identitäten (z.B. ethnische, nationale, europäische, transatlantische, religiöse Identitäten) in Zeitungsdebatten über Kriege und humanitäre militärische Interventionen zu analysieren. Sprachliche Ausdrücke, die vor dem Hintergrund der einen Theorie als Ausdruck »nationaler Identität« gelten, mögen jedoch vor dem Hintergrund einer anderen Theorie irrelevant sein oder etwas ganz anderes messen. Die Hoffnung, ein für alle Mal universelle linguistische Muster zu identifizieren oder Fachbegriffe über kontextunabhängige Lexika zu operationalisieren, beurteilen wir daher skeptisch. Wir suchen

[11] Niethammer 2000; Brubaker / Cooper 2000.
[12] Kantner 2006a; Risse 2010.
[13] Kantner 2004; Kantner 2006a; Tietz 2002a; Tietz 2002b; Tietz 2010.

nicht nach einfachen sprachlichen Ausdrücken (z.B. Deutschland = Hamburg, Berlin, Elbe, Bodensee, Schwaben etc.), sondern nach Bedeutungen, die in der Alltagssprache selten direkt geäußert werden. Typische Ausdrücke für »deutsche nationale Identität« wären eher Umschreibungen (z.B. »unsere historische Verantwortung«, »Berlin muss endlich…«, »Deutschland in Europa«, »Bündnisverpflichtungen«), die nur in bestimmten Kontexten das Gesuchte meinen, während jedes einzelne Wort in anderen Kontexten womöglich nicht über seine wörtliche Bedeutung hinaus verweist. Denn das, was diesen abstrakten theoretischen Begriffen in der Lebenswelt der Menschen entspricht, ist stark kontextabhängig: nationale Identität äußert sich höchst unterschiedlich in verschiedenen Ländern zu verschiedenen Zeiten, ja selbst zur gleichen Zeit in Bezug auf verschiedene politische Themen – man vergleiche z.B. die Diskussion um deutsche Identität in Fragen der Integration von Migranten mit der Diskussion um deutsche Identität in außenpolitischen Fragen. Sprecher verschiedener sozialer Schichten verleihen ihren kollektiven Zugehörigkeiten auf sehr unterschiedliche Weise Ausdruck. Weitere textanalytische Probleme ergeben sich daraus, dass sich Sprecher sich im gleichen Text als Mitglied verschiedener Gruppen bekennen können oder sich von einer nahen Ebene der Identifikation zu abstrakteren bewegen. »Wir Deutschen«, wird beispielsweise häufig sinngemäß argumentiert, haben aufgrund unserer Geschichte eine besondere historische Verantwortung und sollten uns darum besonders stark dafür einsetzen, dass die »internationale Gemeinschaft« den Konflikt im Land X auf diese oder jene Weise löst. Moderne Menschen haben »multiple Identitäten«[14]: gute Europäer zu sein, wurde Bestandteil der deutschen nationalen Identität nach 1945.

Die aus sozialwissenschaftlicher Perspektive interessanten und forschungsleitenden Konzepte sind folglich nicht standardisierbar. Computer- und korpuslinguistische Ansätze sollten daher dem einzelnen Forscherteam Raum für seine *eigene* Operationalisierung lassen und es dabei möglichst umfassend unterstützen (z.B. im Wechselspiel von manueller und automatischer Annotation in »lernenden« Anwendungen).[15] In e-Identity stellt der *Complex Concept Builder* [16] eine solche flexible und interaktive Werkzeugumgebung zur Verfügung (einige Aspekte dazu werden in Abschnitt 3.3 angesprochen).

3. Computerlinguistische Methoden in der Korpusaufbereitung

Dieser Abschnitt soll anhand von zentralen Teilprozeduren in den Abläufen der Dokumentenauswahl, -aufbereitung und der Vorbereitung einer theoretischen Inhaltsanalyse zeigen, wie durch den Einsatz von anspruchsvollen computerlinguistischen Methoden und Werkzeugen für datenintensive sozialwissenschaftliche Korpusstudien einerseits generische Lösungen für wiederkehrende Aufgaben angeboten werden können und andererseits eine sehr spezifische Anpassung des Ablaufs an die theoretischen Annahmen und technischen

[14] Risse 2010.
[15] Dies schließt natürlich nicht aus, dass bewährte Operationalisierungen für die im Umfeld dieser komplexen fachlichen Konzepte ausgedrückten Sachverhalte, Bewertungen und Beziehungen etc. wie üblich analysiert werden können.
[16] Eine Darstellung des Complex Concept Builder findet sich in Blessing et al. 2012.

Randbedingungen der jeweiligen Studie ermöglicht wird. Den Gesamtablauf der Verarbeitung von Dokumenten im *e-Identity*-Projekt skizziert Abbildung 1 schematisch.

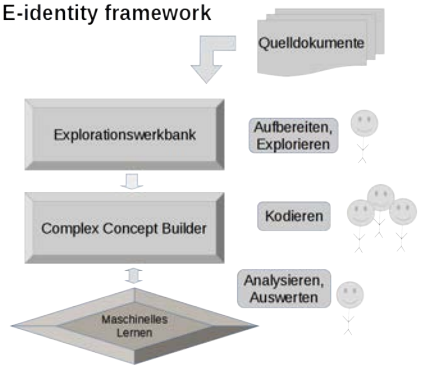

Abb. 1: Schematische Darstellung der Dokumentenverarbeitungskette (Quelle: Eigene Darstellung).

Zunächst (im Schema ganz oben) werden Rohtexte importiert, die im Rahmen der sogenannten *Explorationswerkbank* verwaltet, gesichtet und aufbereitet werden. Ziel ist ein für die weiteren Schritte geeignetes und den methodischen Ansprüchen einer späteren Analyse entsprechendes Korpus. Gleichzeitig bilden die Explorationsmöglichkeiten des entsprechend vorbereiteten Korpus eine wichtige Grundlage für die Hypothesenbildung bzw. Verfeinerung von Ausgangshypothesen. Abschnitt 3.1 widmet sich den Teilaspekten der Datensammlung und -vorbereitung im Detail[17], Abschnitt 3.2 illustriert anhand einiger Beispiele, welche Möglichkeiten sich aus einer computerlinguistischen Aufbereitung für die Arbeit mit dem Korpus ergeben.

Für den anschließenden Arbeitsschritt der Korpusannotation bzw. des Kodierens müssen die theoretischen Konzepte, auf die die späteren Analysen abheben, zunächst entwickelt und operationalisiert und anschließend in der Annotation/Kodierung umgesetzt werden. Zur Unterstützung dieser Prozesse werden in e-Identity Methoden geprüft, die im sogenannten *Complex Concept Builder* subsumiert sind, der nicht im technischen Hauptfokus des vorliegenden Beitrags – den computerlinguistischen Aspekten der Dokumentenaufbereitung – liegt. Die Frage nach den Möglichkeiten der computerlinguistischen Unterstützung des Annotations-Prozesses, der einer sehr gut koordinierten Abstimmung mit den konkreten Fragestellungen aus der Politikwissenschaft bedarf, spielt jedoch zentral in den Kodierungsschritt hinein und wird in Abschnitt 3.3 behandelt.

Im Anschluss an die Annotation/Kodierung stehen die Daten für unterschiedliche Arten der Auswertung zur Verfügung: sie können (i) direkt abgefragt, (ii) für weitere Explorationen geeignet aggregiert und visualisiert oder (iii) einer statistischen Auswertung unterzogen

[17] Aus Platzgründen kann hier nur ein Überblick zu den verwendeten Techniken gegeben werden. Eine detaillierte Darstellung der verwendeten Methoden findet sich in Kliche et al. 2014.

werden. Mit Verfahren aus dem maschinellen Lernen können schließlich (iv) auf Basis der vorliegenden Annotationen Modelle für eine automatische Analyse weiterer Daten trainiert und evaluiert werden bzw. (v) in einem interaktiven Zyklus aus vorhandenen Bausteinen weiterführende Analysen erzeugt werden. Aspekte der späteren Auswertung sind von großer Bedeutung bei der Konzeption der Aufbereitung.

3.1. Computerlinguistische Methoden bei der Datensammlung und -vorbereitung

Für die Untersuchung der kollektiven Identitäten wurde ein Sample mit Zeitungsdaten im Umfang von mehr als 800.000 Artikeln erstellt, die von fünf digitalen Medienportalen heruntergeladen wurden. Die Daten stammen aus 12 Zeitungen aus sechs Ländern und enthalten Artikel in drei Sprachen (Deutsch, Englisch, Französisch). Abbildung 2 zeigt rohe Textdaten, wie sie von Medienportalen bereitgestellt werden.[18] Die verschiedenen Medienportale präsentieren die Daten in jeweils eigenen Datenstrukturen. Neben den Fließtexten enthalten die Artikel Metadaten, die für die textanalytische Arbeit wertvoll sind. Die Metadaten liegen zum Teil strukturiert und zum Teil semi-strukturiert vor. Die Beispiele in Abbildung 2 machen deutlich, dass die rohen Textdaten erst erschlossen werden müssen, um die textlichen Inhalte und Metadaten für die Textanalyse verwenden zu können.

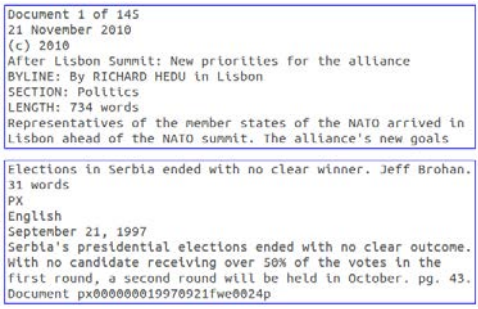

Abb. 2: Abbildung 2 zeigt Rohe Textdaten, wie sie auf digitalen Medienportalen bereitgestellt werden (Quelle: Eigene Darstellung).

Wir entwickeln mit der *Explorationswerkbank* ein generisches Werkzeug, mit dem Textwissenschaftler rohe Textdaten in unterschiedlichen Datenstrukturen erschließen können. Die Voraussetzung ist, dass die Daten in einer festen Datenstruktur vorliegen. Mit der Werkbank können die Textdaten in textstrukturelle Einheiten (im Folgenden: »Artikel«) segmentiert werden. Aus den Artikeln werden die textlichen Inhalte und Metadaten extrahiert und aus den Daten ein Korpus erstellt. Die Werkbank stellt anschließend Funktionen für die Bereinigung des Korpus und für die textanalytische Arbeit mit den Daten bereit. Ein Wizard führt die Anwender in einer intuitiv nachvollziehbaren Weise durch die Schritte der Datenerschließung. Die Arbeitsschritte des Wizards werden im Folgenden beschrieben.

[18] Die Beispiele in diesem Artikel wurden neutralisiert, um Urheberrechtsverletzungen zu vermeiden.

3.1.1 Schritt 1: Die Zerlegung der rohen Textdaten in Artikel

Der Erschließung beginnt mit dem Import der Rohdaten. Textwissenschaftler können Textdaten in verschiedenen Formaten importieren: DOCX, RTF, ODT (Open Office), HTML, TXT (Plaintext ohne Markup). Der Wizard überführt die Daten in Plaintext ohne Markup. Wir konvertieren die unterschiedlichen Datenformate in Plaintext und nehmen den Verlust textstruktureller Merkmale (Fonts, Schriftgröße usw.) in Kauf, um die Texte in einem einheitlichen Format ablegen zu können und um die Textdaten für anschließende computerlinguistische Verarbeitungsschritte zugänglich zu machen. Im speziellen Fall von Zeitungstexten hat das Fehlen von detaillierten auf die Textstruktur bezogenen Metadaten nach unseren Erfahrungen keine negativen Auswirkungen. Die Zeichenkodierungen Latin 1 und Unicode werden in Unicode vereinheitlicht.

Die Anwender lassen sich anschließend Ausschnitte der Rohdaten in einem Vorschaufenster anzeigen. Abbildung 3 zeigt das Vorschaufenster mit rohen Textdaten.

Abb. 3: Ein Ausschnitt von rohen Textdaten wird in einem Vorschaufenster angezeigt. Die Anwender definieren eine Segmentierungsregel, um die Rohdaten in textstrukturelle Einheiten (»Artikel«) zu segmentieren. Im Beispiel trennen Zeilen, die mit der Zeichenfolge kf00 beginnen, die einzelnen Artikel (Quelle: Eigene Darstellung).

Die Anwender suchen in den Rohdaten nach Hinweisen, die die Grenze zwischen Artikeln markieren, und erstellen Segmentierungsregeln. Um eine Segmentierungsregel zu prüfen, werden die von der Regel erfassten Textstellen im Vorschaufenster farblich hervorgehoben, wie es in Abbildung 1 dargestellt ist. Für die Segmentierungsregeln stehen folgende Funktionen zur Verfügung:

- Definitionen des Segments: Eine feste Zeichenfolge oder ein regulärer Ausdruck;
- Angaben zur minimalen und maximalen Länge des Segments;
- Definitionen für das vorangegangene und das nachfolgende Segment und Angaben zu deren Segmentlängen.

Die Segmentierung ist zeilenbasiert. Der Wizard sucht in den Rohdaten nach den Zeilen (»Segmenten«), die von einer Segmentierungsregel erfasst werden, um die Rohdaten in Artikel aufzutrennen. Die Anwender können wählen, ob eine Zeile, die von einer Segmentierungsregel erkannt wird, zum vorausgehenden oder zum folgenden Artikel gezählt wird. Anschließend wendet der Wizard die Regel auf die importierten Daten an und legt die segmentierten Artikel

in einem Repositorium ab. Die zugrundeliegenden Segmentierungsregeln und ein Zeitstempel werden als Prozessmetadaten festgehalten. Für jeden importierten Artikel wird außerdem die Anzahl der Segmente (d. h. Zeilen) als einfacher Importcheck festgehalten: Auffallende Abweichungen der Artikellängen können darauf hinweisen, dass Artikelgrenzen nicht richtig erkannt wurden.

3.1.2 Schritt 2: Die Extraktion der textlichen Inhalte und Metadaten aus den Artikeln

In einem zweiten Schritt werden aus den Artikeln die textlichen Inhalte und Metadaten extrahiert. Die Anwender können sich wieder in einem Vorschaufenster Beispielartikel anzeigen lassen und Regeln definieren, nach denen Metadaten und Fließtexte aus den Artikeln extrahiert werden. Die Regeln erkennen wieder nur Segmente, d. h. Zeilen in den Rohdaten der Artikel. Für jede Regel wird ein Bezeichner angegeben. Der Bezeichner ist i. d. R. der Name des extrahierten Metadatums: Datum, Autoren, usw. Für die Erstellung der Extraktionsregeln für Fließtexte und Metadaten stehen mehr Funktionen als für die Segmentierung der Rohdaten in Artikel zur Verfügung. Die Anwender können für die Extraktion von Metadaten und Fließtexten verschiedene Eigenschaften von Segmenten verwenden:

(1) Textmerkmale:

Bestimmte Textmerkmale dienen als positive bzw. negative Indikatoren für Metadaten. Ein positiver Indikator zeigt das Vorhandensein eines Metadatums an. Ein negativer Indikator verhindert die Interpretation eines Segments als Metadaten-Indikator.

(2) Kontextregeln:

Segmente können alternativ

1. im Text Vorgänger oder Nachfolger von anderen Segmenten sein, die ihrerseits Indikatoren enthalten.

Die Segmente können danach beschrieben werden, wo sie in der zu analysierenden textstrukturellen Einheit auftreten (z. B. am Anfang oder am Ende der Einheit, oder nach dem Titel-Segment eines Zeitungsartikels).

(3) Typische Beispiele für Eigenschaften von Segmenten, die solche Segmente zu Indikatoren für Metadaten machen, sind:

1. das Auftreten von bestimmten Schlüsselwörtern (»Anker«);
2. das Auftreten von Mustern von Wörtern oder Zeichenketten einer bestimmten Art (z. B. Jahreszahlen, Datumsmuster, Strukturen aus Vorname und Nachname). Damit

solche Muster leichter formuliert werden können, können Wortlisten angelegt und eingebunden werden;
3. die minimale bzw. maximale Länge der Segmente.

Die Segmente, die von einer Regel erfasst werden, können wieder im Vorschauartikel farblich hervorgehoben werden, um die Regeln zu testen. Die Anwender können festlegen, ob für eine Regel das gesamte Segment im Repositorium abgelegt werden soll, oder nur der Bereich im Segment, der von der Regel erfasst wird. Für einen Bezeichner können mehrere Regeln erstellt werden. Beispielsweise können Autorenangaben an mehreren Stellen im Artikel auftreten: Am Ende des Fließtextes, unter der Überschrift oder als Teil der strukturierten Metadaten zu Beginn eines Artikels.

Die Anwender erstellen mehrere Regeln und speichern diese als Schablone ab, die in der Folge auf eine bestimmte Datenstruktur appliziert werden kann. Für die Schablone wird ein Name vergeben. In der Schablone werden zusätzliche Prozessmetadaten abgelegt: Beim Speichern jeder Regel wird ein Zeitstempel gesetzt. Weiter wird für jede Regel ein Kommentarfeld bereitgestellt, in dem festgehalten werden kann, für welche Phänomene in den Rohdaten eine Regel erstellt wurde. Schließlich hält die Schablone fest, welche Verarbeitungsschritte die erstellten Regeln verlangen, beispielsweise ob die Tokenisierung der Daten oder die Bestimmung von Wortarten erforderlich ist.

3.1.3 Schritt 3: Die Erschließung der Textdaten

Im dritten Schritt wird die Schablone auf die Artikel gelegt, um die Regeln anzuwenden. Der Wizard prüft, welche Vorverarbeitungsschritte für die Regeln der Schablone notwendig sind, und zeigt sie den Anwendern als Voreinstellungen in einem Formularfeld an. Die Anwender können diese Voreinstellungen erweitern, beispielsweise, wenn Annotationen wie Wortarten für die Volltextsuche zur Verfügung stehen sollen. Anschließend starten die Anwender den Analyseprozess. Die Artikel werden zunächst vorverarbeitet. Um die zeitintensiven Verarbeitungsschritte zu bündeln, erfolgt ohne weitere Interaktion mit den Nutzern nach der Vorverarbeitung die Anwendung der Regeln der Schablone. Die textlichen Inhalte und Metadaten werden erkannt und im Repositorium als Datensätze abgelegt.

3.1.4 Schritt 4: Die Bereinigung des Korpus

Die *Explorationswerkbank* integriert Funktionen, um die Korpusdaten zu bereinigen. Die Anwender können nach verschiedenen Kriterien Datensätze aus dem Korpus ausschließen. Ausgeschlossene Datensätze werden nicht gelöscht, sondern im Repositorium markiert. Eine maximale Länge für leere Artikel und für kurze Artikel kann festgelegt werden. Die Werkbank enthält eine Funktion, um Dubletten und Semi-Dubletten im Korpus aufzudecken. Die Anwender wählen dazu auf einer Skala zwischen 0 und 1 einen Ähnlichkeitswert. Für Textpaare, deren Ähnlichkeit oberhalb dieses Schwellenwerts liegt, wird der kürzere Text aus dem Korpus ausgeschlossen und der längere Text beibehalten. Wenn mehr als zwei Texte identisch oder

ähnlich sind, wird der längste Text beibehalten und die übrigen Texte ausgeschlossen. Die Werkbank bestimmt den Anteil von Zahlen und nicht alphanumerischen Zeichen im Text, um Einheiten auszufiltern, die keinen Fließtext enthalten.

Durch Anwendung einer Topic-Model-Analyse kann jedem Artikel eine Topic-Verteilung zugewiesen werden. Dieser Schritt kann zwar nicht zum automatischen Ausschluss von Artikeln genutzt werden, aber die Exploration der Topics bietet eine effiziente Methode, um eine Kandidatenmenge von Off-Topic-Artikeln manuell zu sichten. Aufbauend auf den manuellen Klassifikationen kann mittels maschinellen Lernens ein Off-Topic-Filter analog zu einem Spam-Filter trainiert werden. In unserem Projekt können dadurch z.B. Artikel zu historischen Kriegen, Sportereignissen und Medienkritiken ausgefiltert werden, die nicht Gegenstand des Forschungskorpus sind.

3.2 Computerlinguistische Methoden für die Korpusexploration

Nach Bereitstellung im Korpus werden die Texte mit computerlinguistischen Werkzeugen aus der CLARIN-Infrastruktur[19] maschinell analysiert. Diese tiefe Analyse ist wichtig, um inhaltliche Untersuchungen zu ermöglichen, die über einfaches Bestimmen von Worthäufigkeiten hinausgehen.

Abbildung 4 skizziert diesen Verarbeitungsschritt. Die Daten werden zur internen Weiterverarbeitung im UIMA-Format in einer Datenbank gespeichert. Diese Art der Repräsentation hat den Vorteil, dass die Annotationsebenen dynamisch erweitert werden können (z.B. können neue Werkzeuge eingebunden werden, die am Projektanfang noch nicht bekannt sind). Des Weiteren erlaubt UIMA auch mehrfache Annotationen und überlappende Annotationen, was eine Grundlage für eine differenzierte Analyse darstellt.

Das Bewusstsein, dass computerlinguistische Verfahren nicht perfekt sind, ist ein wichtiges Kriterium, das den Endanwendern in unterschiedlichster Weise vermittelt werden muss. Der Einsatz unterschiedlicher Analysewerkzeuge auf gleicher Ebene (z.B. mehrere verschiedene Part-of-Speech-Tagger) kann helfen, das System robuster zu machen: bei schwierig zu analysierenden Daten werden unterschiedliche Vorhersagen gemacht, die Wahrscheinlichkeit einer richtigen Lösung steigt also; bei kanonischen Dateninstanzen werden die Werkzeuge übereinstimmen, die Analyse wird somit als relativ verlässlich erkennbar. Je nach Anforderung kann durch die Werkzeugkombination ein Gesamtsystem mit großer Abdeckung/Erkennung (*recall*) oder mit vordringlichem Gewicht auf Güte (*precision*) erzeugt werden.

[19]Die verwendeten Stuttgarter Werkzeuge werden detailliert in Mahlow et al. 2014 beschrieben. Eine ausführliche Darstellung, wie die Werkzeuge zu einer Applikation zusammengefasst werden, findet sich in Blessing 2014. Des Weiteren kann eine globale Übersicht der CLARIN-D-Werkzeuge online abgerufen werden.

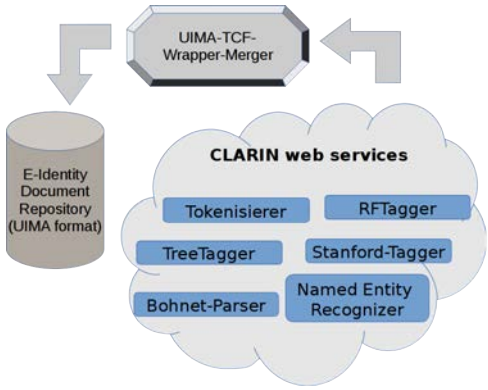

Abb. 4: Computerlinguistische Verarbeitungsschritte, die im Hintergrund der Dokumentenaufbereitung ablaufen (Quelle: Eigene Darstellung).

Folgendes Beispiel (Abbildung 5) zeigt die Analyse des Begriffs »Nato« in englischen Texten durch drei unterschiedliche Part-of-Speech Tagger (TreeTagger[20], Stanford-Tagger[21] und Bohnet-Tagger[22]). Man sieht, dass die Werkzeuge in 88% der Fälle zum gleichen Ergebnis kommen, aber in 12% der Fälle (das ist jedes achte Vorkommen) Uneinigkeit besteht. Je nach Anwendung können diese restlichen 12% unterschiedlich interpretiert werden. Die Entscheidung bleibt so dynamisch konfigurierbar und es geht keine Informationen durch ein zu frühes Vereinigen verloren.

	TreeTagger	Stanford-Tagger	Bohnet-Tagger	Frequenz
NATO	NNP	NNP	NNP	471
Nato	NN	NNP	NNP	26
Nato	NNP	IN	NNP	118
Nato	NNP	NN	JJ	18
Nato	NNP	NN	NNP	19
Nato	NNP	NNP	JJ	364
Nato	NNP	NNP	NN	309
Nato	NNP	NNP	NNP	22.840

Abb. 5: Illustration des redundanten Einsatzes mehrerer Part-of-Speech-Tagger (Quelle: Eigene Darstellung).

Die automatischen Annotationen können je nach Benutzereinstellung in den Texten visualisiert werden, um bei der Exploration Schlüsselstellen im Text schneller zu finden (bei der Kodierung inhaltlicher Kategorien darf diese Funktion nicht verwendet werden, um einen Bias zu vermeiden).

[20] Schmid 1995.
[21] Toutanova 2003.
[22] Bohnet 2010.

Die verarbeiteten Daten werden indiziert, um eine schnelle Volltextsuche zu ermöglichen. Suchkriterien der Volltextsuche können neben Wortformen auch die Annotationen sein, die in den computerlinguistischen Vorverarbeitungsschritten für die Wörter vergeben worden sind (z. B. Lemmaangaben, Wortarten, etc.); außerdem können die Einträge der Wortlisten mitbenutzt werden, die für die Indikatorensuche angelegt wurden, z. B. Listen von Vornamen, Personen, Orten, Nachrichtenagenturen usw.

Die *Explorationswerkbank* bietet Methoden zur statistischen Auswertung von Artikeln anhand des in den Metadaten hinterlegten Erscheinungsdatums an. So können in unterschiedlichen Zeitintervallen (jährlich, monatlich, wöchentlich, täglich) Zählungen gemacht werden und damit die Medienaufmerksamkeit für die Themen des Samples auf der Zeitlinie erfasst werden (»Issue Cycles«). Diese Ergebnisse werden in der folgenden inhaltlichen Analyse außerdem zur Normalisierung verwendet.

3.3. Computerlinguistische Unterstützung bei der Annotation

Die genauen textanalytischen Fragestellungen leiten sich in e-Identity aus politikwissenschaftlichen Überlegungen ab und sind mit studienspezifischen Konzepten verbunden. Somit kommen unüberwachte Verfahren, wie sie gern bei der Big-Data-Analyse eingesetzt werden, nur bedingt in Frage. Die Wahrscheinlichkeit, dass ein solches Verfahren genau die gestellte Fragestellung beantwortet, ist sehr gering. Es muss also durch manuelle Kodierung/Annotation ein grundlegendes Wissen erzeugt werden, mithilfe dessen überwachte maschinelle Verfahren arbeiten können.

Im Vorgängerprojekt zu e-Identity wurden komplette Artikel inhaltlich bewertet. Dies ist für die Auswertung der politikwissenschaftlichen Fragestellung völlig ausreichend, aber es hat sich gezeigt, dass diese grobkörnige Kodierung nicht ausreichend ist, um mittels computerlinguistischer Verfahren eine großflächige Analyse vorzunehmen. Unser Ansatz bietet einen neuen Freiheitsgrad, indem jeweils die gewünschte Textstelle zur Kodierung ausgewählt werden muss.

Je nach Projekt verschieben sich die Fragestellungen und somit auch die Variablen der Kodierungen. Dies wird in unserem System durch ein frei rekonfigurierbares Kodierschema ermöglicht. Politikwissenschaftler können dazu ein hierarchisches Kodierbuch inklusive Anmerkungen, die als Tooltipp angezeigt werden, hochladen. Das System generiert daraus eine interaktive Oberfläche, die den Kodierern angezeigt wird. Abbildung 6 zeigt einen Ausschnitt aus der aktuell generierten Schnittstelle für die Kodierung.

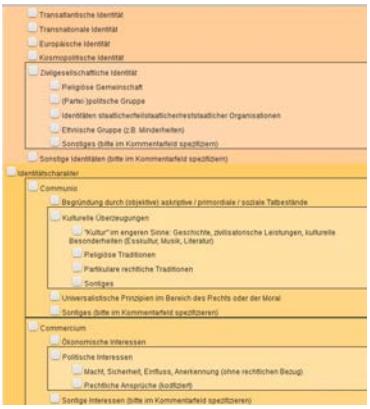

Abb. 6: Die Benutzerschnittstelle für die politikwissenschaftliche Kodierung/Annotation (Quelle: Eigene Darstellung).

3.4 Qualitätssicherung

Qualitätssicherung ist sehr wichtig, um verlässliche Ergebnisse zu erhalten. Unser System bietet durch Mehrfachkodierung ausgewählter Artikel eine Möglichkeit die Qualität der Kodierungen kontinuierlich zu messen.

Dies erlaubt gegebenenfalls ein Eingreifen und Nachschulen von Kodierern, ebenso wird hierdurch die Qualität des Kodierhandbuchs geprüft. Unser System enthält hierfür eine Kodiermanagement-Einheit, welche die Zuteilung und Verwaltung der zu kodierenden Artikel übernimmt. Außerdem kann hiermit auch der Fortschritt der einzelnen Kodierer geprüft werden und je nach Anforderung unterschiedliche Ziele gesetzt werden.

Durch die Komplexität, die durch freie Kodierung beliebiger Textstellen entsteht, ist die Definition einer einfachen Metrik zur Bestimmung der Qualität schwierig. Unser System bietet einen Zwischenschritt mittels graphischer Oberfläche, die es erlaubt, die Kodierung in jedem Artikel visuell zu vergleichen. Abbildung 7 illustriert eine solche Analyse. Im unteren Teil sind die einzelnen Kodierungen aufgetragen. Jede Zeile repräsentiert einen Kodierer bzw. eine Kodiererin und jedes farbliche Intervall zeigt eine Kodierung an. Die x-Achse beschreibt die Textposition (zeichenweise). Man sieht nun sehr schnell, welche Textstellen eine gute oder weniger gute Übereinstimmung haben. Per Mouse-over können zu jedem Intervall die kodierten Variablen angezeigt werden. Zusätzlich wird im Artikel, der in der oberen Hälfte sichtbar ist, die entsprechende Textstelle hervorgehoben.

Abb. 7: Vergleichende Darstellung der Kodierungsergebnisse verschiedener Kodierer (Quelle: Eigene Darstellung).

Der Vorteil unserer Implementierung als Webanwendung ist, dass die Benutzer keine spezielle Software installieren müssen. Außerdem kann ein Kodierer standortunabhängig auf das System zugreifen und seine Arbeit jederzeit unterbrechen und wieder aufnehmen.

Die Einbindung bestehender Systeme ist ein weiterer Aspekt unseres Systems. CQPWeb bietet z.B. ein mächtiges Abfragesystem, um mit linguistisch annotierten Daten zu interagieren. Dies kann z.B. genutzt werden, um sich Kollokationen berechnen zu lassen: Tauchen bestimmte Wortkombinationen im Korpus (oder in einem geeignet eingeschränkten Subkorpus) häufiger auf, als dies aufgrund ihrer Basis-Verteilung zu erwarten wäre? Abbildung 8 zeigt eine Wortliste, für die einfache frequenzbasierte Analysen durchgeführt und dann in CQPWeb weiterverarbeitet werden können. Kollokationsanalysen können abseits des linguistischen Interesses an idiomatischen Wortverbindungen ausgesprochen hilfreich bei der Exploration eines Korpus sein: So lässt sich beispielsweise fragen, welche Adjektive im Kontext des Lemmas ›Europe‹ in einem Subkorpus britischer Zeitungen systematisch häufiger auftauchen als im US-amerikanischen Vergleichskorpus. Das Potenzial, das sich aus den linguistisch aufbereiteten Texten für weitere Analysen ergibt, ist bei weitem noch nicht erschöpft. Für eine sinnvolle Nutzung ist jedoch eine nicht-triviale Abstimmung der technischen Analysemöglichkeiten auf konkrete nicht-linguistische Fragestellungen nötig.

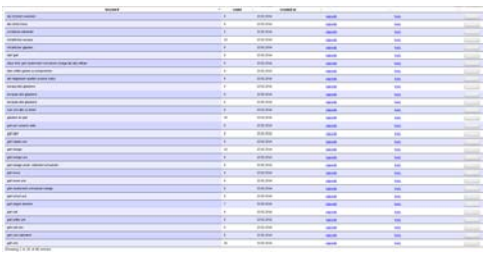

Abb. 8: Screenshot zur Kollokationsextraktion (Quelle: Eigene Darstellung).

4. Schluss

Der vorgelegte Beitrag gibt einen Überblick über eine Werkzeugarchitektur zur Unterstützung der Vorbereitung und Erschließung großer Korpora für sozialwissenschaftliche

Forschungsfragen. In den Aufbereitungsprozess werden computerlinguistische Methoden so eingebunden, dass linguistische Zwischenrepräsentationen ausgenutzt werden, ohne dass sich die beteiligten Fachwissenschaftler in ihre Details einarbeiten müssen. Der Ansatz fußt grundsätzlich auf generischen Werkzeugketten und ist damit übertragbar auf andere Projekte, unterstützt jedoch gleichzeitig eine sorgfältige Annotation bzw. Kodierung von studienspezifischen Konzepten, die u.a. als Basis für überwachte Lernverfahren genutzt werden kann.

Im Sinne eines transparenten Werkzeugangebots, das eine Sensibilität für mögliche Analysefehler erzeugt und aufrecht erhält, wird den Nutzern bei der Exploration und Hypothesenbildung immer der Bezug zu den tatsächlichen Textdaten und zwischengelagerten Analyseschritten vor Augen geführt.

Die Erfahrung der Projektkooperation zwischen den Computerlinguisten und den Politikwissenschaftlern hat gezeigt, dass Phasen des intensiven Austauschs zum jeweils fachspezifischen Verständnis zentraler Begriffe und Methoden von großer Bedeutung für die Effektivität der gemeinsamen Entwicklungsschritte sind. Ein Einsatz von vordefinierten Werkzeuglösungen kann nicht in der gleichen Weise in einen sozialwissenschaftlichen Erkenntnisprozess eingebunden werden wie die schrittweise Erarbeitung eines textanalytischen Werkzeuginventars, das auf die speziellen Schwierigkeiten und Möglichkeiten der Korpussituation und der verfolgten analytischen Fragestellung angepasst ist.

Bibliographische Angaben

Melina Alexa / Cornelia Zuell: Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review. In: Quality and Quantity 34 (2000) H. 3, S. 299–321. [Nachweis im OPAC]

Paul Baker / Tony McEnery: A Corpus-Based Approach to Discourses of Refugees and Asylum Seekers in UN and Newspaper Texts. In: Journal of Language and Politics 4 (2005) H. 2, S. 197–226. [Nachweis im OPAC]

André Blessing / Jens Stegmann / Jonas Kuhn (2012a): SOA meets Relation Extraction: Less may be more in Interaction. In: Proceedings of the Joint CLARIN-D/DARIAH Workshop at Digital Humanities Conference. Hamburg 2013, S. 6–11. [online]

André Blessing / Jonathan Sonntag / Fritz Kliche / Ulrich Heid / Jonas Kuhn / Manfred Stede (2012b): Towards a Tool for Interactive Concept Building for Large Scale Analysis in the Humanities. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities Association for Computational Linguistics. Sofia 2012, S. 55–64. [online]

André Blessing / Jonas Kuhn: Textual Emigration Analysis (TEA). In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA). Reykjavik 2014, S. 2089–2093. [Nachweis im GBV]

Bernd Bohnet: Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In: The 23rd International Conference on Computational Linguistics (COLING 2010). Beijing 2010, S. 89–97. [Nachweis im GBV]

William R. Brubaker / Frederick Cooper: Beyond »Identity«. In: Theory and Society 29 (2000) H.1, S. 1–47. [Nachweis im GBV]

Emile Durkheim: The Rules of Sociological Method. New York 1964 [zuerst 1895]. [Nachweis im GBV]

Constantinos Gabrielatos: Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database. In: ICAME Journal 31 (2007), S. 5–43. [Nachweis im GBV]

Richard K. Herrmann: Linking Theory to Evidence in International Relations. In: Handbook of International Relations. Hg. von Walter Carlsnaes / Thomas Risse / Beth A. Simmons. London 2002, S. 119–136. [Nachweis im GBV]

Cathleen Kantner: Kein Modernes Babel. Kommunikative Voraussetzungen Europäischer Öffentlichkeit. Wiesbaden 2004. [Nachweis im GBV]

Cathleen Kantner (2006a): Collective Identity as Shared Ethical Self-Understanding: The Case of the Emerging European Identity. In: European Journal of Social Theory 9 (2006) H. 4, S. 501–523. [Nachweis im GBV]

Cathleen Kantner (2006b): Die thematische Verschränkung nationaler Öffentlichkeiten in Europa und die Qualität transnationaler politischer Kommunikation. In: Demokratie in Der Mediengesellschaft. Hg. von Kurt Imhof / Roger Blum / Heinz Bonfadelli / Otfried Jarren. Wiesbaden 2006, S. 145–160. [Nachweis im GBV]

Cathleen Kantner: Transnational Identity-Discourse in the Mass Media. Humanitarian Military Interventions and the Emergence of a European Identity (1990–2006). Berlin 2009.

Cathleen Kantner / Amelie Kutter / Andreas Hildebrandt / Mark Püttcher: How to Get Rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts. Stuttgart 2011. [online]

Cathleen Kantner / Amelie Kutter / Swantje Renfordt: The Perception of the EU as an Emerging Security Actor in Media Debates on Humanitarian and Military Interventions (1990–2006). In: RECON Online Working Paper 19 (2008). Oslo 2008. [online]

Fritz Kliche / André Blessing / Ulrich Heid / Jonathan Sonntag: The *eIdentity* Text ExplorationWorkbench. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) European Language Resources Association (ELRA). Reykjavik 2014, S. 691–697. [Nachweis im GBV]

Thomas Koenig / Sabina Mihelj / John Downey / Mine Gencel Bek: Media Framings of the Issue of Turkish Accession to the EU. In: Innovation: The European Journal of Social Sciences 19 (2006) H. 2, S. 149–169. [Nachweis im GBV]

Klaus Krippendorff: Content Analysis: An Introduction to Its Methodology. London 2004. [Nachweis im GBV]

Amelie Kutter: Petitioner or Partner? Constructions of European Integration in Polish Print Media Debates on the EU Constitutional Treaty. In: Discourse and Contemporary Social Change. Hg. von Norman Fairclough / Giuseppina Cortese / Patrizia Ardizzone. Bern 2007, S. 433–457. [Nachweis im GBV]

Ulrike Liebert: Introduction: Structuring Political Conflict About Europe: National Media in Transnational Discourse Analysis. In: Perspectives on European Politics and Society 8 (2007) H. 3, S. 236–260. [Nachweis im GBV]

Cerstin Mahlow / Kerstin Eckart / Jens Stegmann / André Blessing / Gregor Thiele / Markus Gärtner / Jonas Kuhn: Resources, Tools, and Applications at the CLARIN Center Stuttgart. In: Proceedings der 12. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014). Hildesheim 2014, S. 11–21. [Nachweis im GBV]

Lutz Niethammer: »Kollektive Identität«. Heimliche Quellen Einer Unheimlichen Konjunktur. Reinbek bei Hamburg 2000. [Nachweis im GBV]

Swantje Renfordt: Framing the use of force : an international rule of law in media reporting ; a comparative analysis of western media debates about military interventions 1990 - 2005. Diss. Berlin 2009. [Nachweis im GBV]

Thomas Risse: A Community of Europeans? Transnational Identities and Public Spheres. Ithaca 2010. [Nachweis im GBV]

Helmut Schmid: Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop. Dublin 1995, S. 47–50. [online]

Udo Tietz (2002a): Die Grenzen des »Wir«. Eine Theorie Der Gemeinschaft. Frankfurt 2002. [Nachweis im GBV]

Udo Tietz (2002b): Gemeinsinn, Gemeinwohl Und Die Grenzen Des »Wir«. In: Gemeinwohl und Gemeinsinn. Hg. von Herfried Münkler / Harald Bluhm. Berlin 2002, S. 37–70. [Nachweis im GBV]

Udo Tietz: Les Limites Du »Nous« Libéral. In: Qu'est-Ce Qu'un Collectif? Du Commun À La Politique. Hg. von Laurence Kaufmann / Danny Trom. Paris 2010, S. 173–195. [online]

Kristina Toutanova / Dan Klein / Christopher Manning / Yoram Singer: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL. Edmonton, Kanada 2003, S. 252–259. [Nachweis im GBV]

Abbildungslegenden und -nachweise

Abb. 1: Schematische Darstellung der Dokumentenverarbeitungskette (Quelle: Eigene Darstellung).

Abb. 2: Rohe Textdaten, wie sie auf digitalen Medienportalen bereitgestellt werden (Quelle: Eigene Darstellung).

Abb. 3: Ein Ausschnitt von rohen Textdaten wird in einem Vorschaufenster angezeigt. Die Anwender definieren eine Segmentierungsregel, um die Rohdaten in textstrukturelle Einheiten (»Artikel«) zu segmentieren. Im Beispiel trennen Zeilen, die mit der Zeichenfolge kf00 beginnen, die einzelnen Artikel (Quelle: Eigene Darstellung).

Abb. 4: Computerlinguistische Verarbeitungsschritte, die im Hintergrund der Dokumentenaufbereitung ablaufen (Quelle: Eigene Darstellung).

Abb. 5: Illustration des redundanten Einsatzes mehrerer Part-of-Speech-Tagger (Quelle: Eigene Darstellung).

Abb. 6: Die Benutzerschnittstelle für die politikwissenschaftliche Kodierung/Annotation (Quelle: Eigene Darstellung).

Abb. 7: Vergleichende Darstellung der Kodierungsergebnisse verschiedener Kodierer (Quelle: Eigene Darstellung).

Abb. 8: Screenshot zur Kollokationsextraktion (Quelle: Eigene Darstellung).