Hannah Kermes und Elke Teich

5 Generische Infrastruktur und spezifische Forschung: Angebote und Lösungen

Abstract: Die empirische Forschung an natürlichsprachlichen Daten geht mit grundlegenden methodischen Veränderungen einher. Immer mehr Texte stehen in digitaler Form zu Verfügung. Eine rein manuelle Vorgehensweise ist nicht möglich oder extrem zeitaufwendig. Wir zeigen welche Vorteile der Einsatz von generischen Infrastrukturkomponenten für spezifische Forschung haben kann: (i) effiziente Untersuchungen auf größeren Datenmengen, (ii) reproduzierbare und übertragbare Ergebnisse. Wir zeigen an einer konkreten Studie, wie generische Infrastruktur spezifisch angepasst und durch spezifische Lösungen ergänzt werden kann.

Keywords: Annotation, Empirie, Sprachkorpora, Textanalyse, XML

1 Einleitung

Die empirische Forschung an natürlichsprachlichen Daten geht mit grundlegenden methodischen Veränderungen einher (cf. Biber, Conrad & Reppen 1998: Kapitel 1; McEnery, Xiao & Tono 2006: 3–4). Immer mehr Texte stehen in digitaler Form zu Verfügung, ob über Plattformen wie Wikisource,[1] durch Projekte wie Project Gutenberg,[2] das Deutsche Text Archiv[3] oder das Linguistic Data Consortium.[4] Es gilt, im Hinblick auf eine Forschungsfrage in einer großen Menge an Ausgangsdaten Relevantes zu finden und zu extrahieren. Das Resultat sind oft große, zumeist multidimensionale Datensätze, die analysiert

1 http://wikisource.org/ (letzter Zugriff: 22. 4. 2018).
2 http://www.gutenberg.org/ (letzter Zugriff: 22. 4. 2018).
3 http://www.deutschestextarchiv.de/ (letzter Zugriff: 22. 4. 2018).
4 https://www.ldc.upenn.edu/ (letzter Zugriff: 22. 4. 2018).

Anmerkung: Die im Artikel beschriebenen Arbeiten wurden durch das Bundesministeriums für Bildung und Forschung im Rahmen des CLARIN-D Projekts unterstützt. Besonderer Dank gilt unseren Kollegen Peter Fankhauser, Stefan Fischer und Jörg Knappen.

Hannah Kermes, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus A2.2, D-66123 Saarbrücken, E-Mail: h.kermes@mx.uni-saarland.de
Elke Teich, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus A2.2, D-66123 Saarbrücken, E-Mail: e.teich@mx.uni-saarland.de

https://doi.org/10.1515/9783110538663-006

und schließlich interpretiert werden müssen. Eine rein manuelle Vorgehensweise ist in der Regel nicht möglich oder extrem zeitaufwendig. Zudem macht die Komplexität der einzelnen Verarbeitungsschritte den Einsatz von generischen bzw. automatischen Werkzeugen notwendig (cf. McEnery, Xiao & Tono 2006: Kapitel 1; Lemnitzer & Zinsmeister 2010: Kapitel 4). Nicht zuletzt ist der Einsatz von generischer Infrastruktur auch bezüglich Reproduzierbarkeit und Übertragbarkeit sinnvoll.

Forschungsinfrastrukturen wie CLARIN-D[5] und Plattformen wie Gate[6] oder NLTK[7] haben sich darauf spezialisiert, sprachtechnologische Werkzeuge und Sprachressourcen zur linguistischen Annotation (z. B. Wortartenannotation, syntaktischen oder semantischen Annotation) und zur Textanalyse für eine breite Zielgruppe zugänglich zu machen. Die Problematik liegt darin, die Angebote generischer Infrastruktur für die speziellen Anforderungen der eigenen Forschung nutzbar zu machen. Es gilt daher zunächst generische und spezifische Komponenten einer Studie zu identifizieren. Generische Komponenten sind z. B. vorhandene Corpora, Corpusabfragewerkzeuge oder Werkzeuge für die Corpusanalyse, aber auch Werkzeuge für die linguistische Annotation oder die OCR-Fehlerkorrektur. Spezifische Komponenten sind etwa das zu untersuchende linguistische Phänomen, aber auch eine spezielle Textgrundlage. Dabei stellen sich folgende Fragen: Wo können generische Komponenten eingesetzt werden, wo müssen spezifische Lösungen gefunden werden und wie können generische und spezifische Komponenten miteinander verknüpft werden?

Im Folgenden werden wir zunächst die allgemeine Vorgehensweise bei einer corpuslinguistischen Studie im Hinblick auf das Zusammenspiel von generischer Infrastruktur und spezifischen Lösungen diskutieren. Danach werden wir anhand einer konkreten Studie beschreiben, wie dieses Zusammenspiel in der Realität aussehen kann, von der Aufbereitung der Ausgangsdaten (Vorverarbeitung und Annotation des Corpus (Abschnitt 3.1) bis zur Datenanalyse (Abschnitt 3.4).

2 Methodik, Arbeitsabläufe und Angebote

In der Corpuslinguistik wird ein sprachliches Phänomen quantitativ und qualitativ anhand einer geeigneten Textgrundlage untersucht. Dazu werden rele-

5 https://www.clarin-d.de/ (letzter Zugriff: 22. 4. 2018).
6 https://gate.ac.uk/ (letzter Zugriff: 22. 4. 2018).
7 http://www.nltk.org (letzter Zugriff: 22. 4. 2018).

vante Beobachtungen extrahiert und anschließend analysiert und interpretiert. Die Auswahl der Textgrundlage ist dabei von dem zu untersuchenden Phänomen abhängig. Existiert kein geeignetes Corpus, so muss eines erstellt werden. Corpuslinguistische Studien lassen sich daher in zwei Hauptbereiche aufteilen: (i) die Corpuserstellung und (ii) die Corpusanalyse. Im Folgenden wollen wir nun darauf eingehen, wie generische Infrastruktur diese beiden Bereiche unterstützen kann und wo spezifische Lösungen gefunden werden müssen.

2.1 Corpuserstellung

Die Corpuserstellung gliedert sich in zwei Schritte: (i) die Vorverarbeitung und (ii) die linguistische Annotation. Dabei gehen wir von bereits digitalisierten Texten[8] aus und nehmen auch an, dass die Texte bereits ausgewählt und zu einer Textsammlung (Corpus) zusammengestellt wurden (cf. Lemnitzer & Zinsmeister 2010: Kapitel 3; McEnery, Xiao & Tono 2006: Kapitel 2). Eine solche digitalisierte aber ansonsten noch nicht weiterverarbeitete Textsammlung bezeichnen wir als Ausgangsdaten.

Bei der Weiterverarbeitung sind nun zwei Aspekte zu berücksichtigen. Erstens sind in der Regel in neu zusammengestellten Corpora Texte anderer Textsorten, Register, Zeitperioden oder Autoren enthalten als in bereits existierenden Corpora. Die über eine Infrastruktur bereitgestellten Verarbeitungskomponenten, wie z. B. Wortartentagger, sind aber in der Regel auf der Allgemeinsprache (z. B. auf Zeitungstexten) trainiert. Je mehr die Texte in einem neu zusammengestellten Corpus von der Allgemeinsprache abweichen, desto wahrscheinlicher ist es, dass generische Komponenten nicht das gewünschte Ergebnis liefern (geringe Abdeckung, hohe Fehlerrate). In diesem Fall müssen die generischen Komponenten spezifisch angepasst bzw. ergänzt werden.

Im Folgenden werden die einzelnen Komponenten der Corpuserstellung anhand der schematischen Darstellung in den Abbildungen 5.1 (Vorverarbeitung) und 5.2 (linguistische Annotation) hinsichtlich des Einsatzes von generischen Komponenten und spezifischen Lösungen näher diskutiert, wobei sowohl auf automatische Werkzeuge als auch auf manuelle Methoden eingegangen wird. Dabei gehen wir zunächst auf die Vorverarbeitung ein (Abschnitt 3.2), die sich in Datenkonversion (Überführung der Ausgangsdaten in ein standardisiertes Format) und Datenbereinigung (Lösung von Layoutproblemen und OCR Fehlerkorrektur) gliedert. Auf die Anreicherung mit Metadaten,

8 Als ein Beispiel für den Arbeitsablauf bei der Digitalisierung von Texten sei auf den Digitalisierungsworkflow des Deutschen Text Archivs verwiesen, http://www.deutschestextarchiv.de/doku/workflow (letzter Zugriff: 22. 4. 2018).

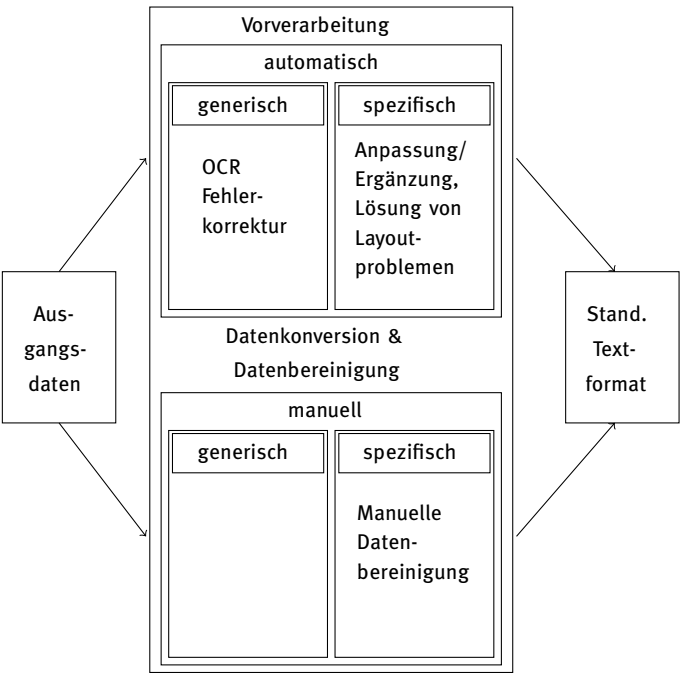

Abb. 5.1: Arbeitsablauf bei der Vorverarbeitung.

also mit das Corpus beschreibende Daten wie Autor, Titel, Erscheinungsjahr etc., die auch zur Datenkonversion gehört, wird hier nicht näher eingegangen. In Abschnitt 3.3 wird dann die linguistische Annotation beschrieben, die sich wiederum in verschiedene Annotationsebenen (Wortebene, syntaktische Ebene) und automatische und manuelle Methoden der Annotation gliedert.

2.1.1 Vorverarbeitung

Die Ausgangsdaten eines Corpus liegen nur selten in einem standardisierten Format vor, z. B. in einem TEI-konformen (*Text Encoding Initiative*[9]) XMLFormat oder einem einfachen Textdokument mit optionalem TEI-konformem XML-

9 "The Text Encoding Initiative (TEI) Guidelines are an international and interdisciplinary standard that facilitates libraries, museums, publishers, and individual scholars represent a variety of literary and linguistic texts for online research, teaching, and preservation." http://www.tei-c.org/ (letzter Zugriff: 22. 4. 2018): siehe u. a. die Kapitel 5 (*TEI Header*) und 23 (*Language Corpora*).

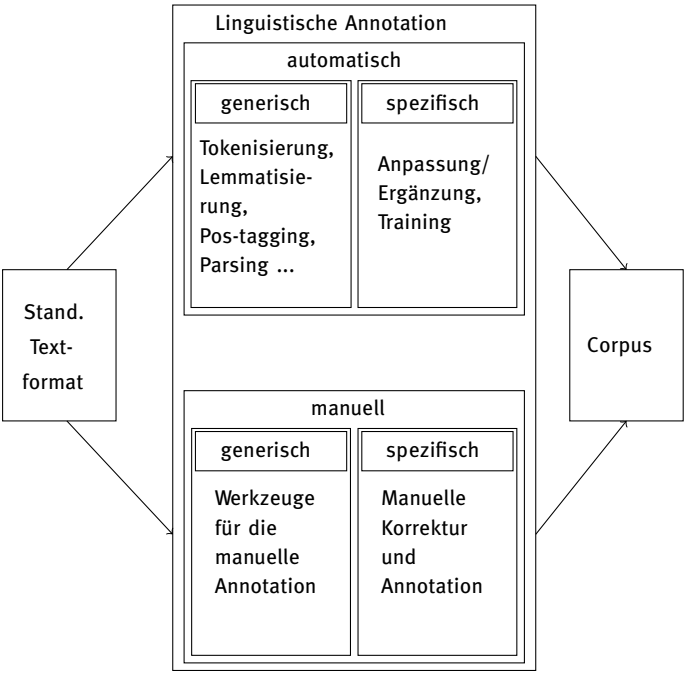

Abb. 5.2: Arbeitsablauf bei der linguistischen Annotation.

Markup. Ein wichtiger Schritt bei der Vorverarbeitung ist die Datenkonversion, also die Überführung der spezifischen Ausgangsdaten in ein standardisiertes (generisches) Format. Standardisierte Daten sind für die Weiterverarbeitung und die Wiederverwendung entscheidend.

Standardisierung bedeutet aber nicht, dass die Formate nicht auf spezifische Bedürfnisse angepasst werden können. Je nach intendiertem Verwendungszweck kann das Format daher variieren und spezifische Kriterien aufweisen.

Ein Beispiel für ein solche Spezifikation auf der Basis eines TEI Formats ist das DTA-Basisformat,[10] dessen Richtlinien auf der einen Seite eine umfassende Textaufbereitung erlauben und auf der anderen Seite die Flexibilität bei der Annotation so einschränken, dass die entstehenden Texte insgesamt kohärent sind. Es ist auf die Annotation gedruckter historischer Texte spezialisiert und dient als Basis sowohl für digitale Editionen als auch für Textcorpora.

10 http://www.deutschestextarchiv.de/doku/basisformat (letzter Zugriff: 22. 4. 2018).

Ein weiteres Beispiel ist das TCF Format,[11] ein XML Format für Textcorpora mit multidimensionaler linguistischer Annotation, das als Austauschformat für WebLicht[12] (cf. Kapitel 3.3), einer Plattform für die automatische Annotation von Textcorpora, entwickelt wurde.

Das CONLL-Format (Buchholz & Marsi 2006) und das VRT-Format der IMS Corpus Workbench (CWB)[13] sind Beispiele für textbasierte Standardformate (*one-word-per-line*). Annotationen sind entweder als TAB-getrennte Spalten kodiert oder beim VRT-Format auch zusätzlich als XML-Elemente. Durch ihren einerseits standardisierten und andererseits einfachen und flexiblen Aufbau sind sie sehr gut als Austauschformate geeignet.

Ein weiterer Aspekt der Vorverarbeitung ist die Datenbereinigung, die sich um Fehlerkorrektur und das Entfernen nicht erwünschter Elemente (Rauschen) kümmert. Die Datenbereinigung hat sowohl generische als auch spezifische Aspekte. Zu den spezifischen Aspekten gehören Layoutprobleme wie Kopf- und Fußzeilen, Graphiken und Tabellen. Obwohl diese Probleme bei vielen Corpora auftreten, sind ihre Ausprägungen und damit auch die Erkennungsmuster sehr unterschiedlich. So muss hier auf spezifische Lösungen zurückgegriffen werden, etwa auf dedizierte Skripte oder auch manuelle Korrektur bei der Entfernung von nicht-textuellen Elementen wie Graphiken oder Tabellen.

Fehler, die aufgrund einer automatischen Texterkennung (OCR-Fehler) entstanden sind, können sowohl generisch als auch spezifisch sein. Für typische OCR-Fehler gibt es eine Reihe von generischen Komponenten zur Korrektur, wobei es sich im Wesentlichen um Ersetzungslisten handelt. Ein Beispiel ist die Ersetzungsliste von Underwood & Auvil (2012) mit 50.000 Ersetzungspaaren für historische englische Texte aus den Jahren 1700–1899. Ein weiteres Beispiel ist das Projekt OCR-D,[14] das auf deutsche historische Texte spezialisiert ist. Das Projekt hat eine ganzheitliche Lösung zum Ziel und setzt bereits bei der OCR-Erkennung selbst an. Für die OCR-Korrektur soll eine Datenbank mit Ersetzungsregeln und Ersetzungsmustern aufgebaut werden, die eine möglichst breite Abdeckung hat. Die generischen Komponenten können jedoch nur die typischen OCR-Fehler abdecken. Insbesondere bei sehr spezifischen Corpora (z. B. Spezialvokabular, spezielle Renderings) kann eine Ergänzung oder Anpassung der Regeln notwendig sein, um zufriedenstellende Ergebnisse zu erzielen. Für die Evaluierung und eine etwaige Identifizierung von

11 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format (letzter Zugriff: 22. 4. 2018).

12 http://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 22. 4. 2018).

13 http://cwb.sourceforge.net (letzter Zugriff: 22. 4. 2018).

14 http://www.ocr-d.de/ (letzter Zugriff: 22. 4. 2018).

spezifischen OCR-Fehlern können wiederum generische Komponenten aus dem Language Modeling, wie etwa *Word Embeddings*, eingesetzt werden (cf. Knappen et al. 2017).

Das konvertierte und bereinigte Corpus kann dann als Basis für die Weiterverarbeitung eingesetzt werden, also z. B. für die linguistische Annotation des Corpus.

2.1.2 Linguistische Annotation

Die linguistische Annotation fügt dem Corpus weitere Abstraktionsebenen hinzu, die auf einer Interpretation der Wörter oder Wortsequenzen in ihrem Kontext basieren. So können einerseits Abfragen effizienter gestaltet werden, weil sie präziser formuliert werden können und andererseits komplexe Phänomene automatisch extrahiert werden (cf. Lemnitzer & Zinsmeister 2010: 60–62). Die unterste Ebene der Annotation ist in der Regel die Annotation auf der Wortebene, auch morphosyntaktische Annotation genannt, und besteht aus Tokenisierung, Lemmatisierung und Wortartentagging (Schmid 2008, Voutilainen 2003; Leech & Wilson 1996). Auf ihr bauen die anderen Annotationsebenen auf, wie etwa die syntaktische Annotation in Form von (partiellen) Dependenz- oder Konstituentenstrukturannotationen (Manning & Schütze 1999: Chapter 12; Langer 2001; Kermes 2008), die semantische Annotation (z. B. Named Entity Recognition, semantische Rahmen, Lesarten) und die pragmatische Annotation (Anaphernund Koreferenzauflösung, Annotation von Informationsstruktur) (cf. Lemnitzer & Zinsmeister 2010: 84–86).

Die Annotation von linguistischer Information ist in jedem Fall aufwendig. Der Einsatz von automatischen Werkzeugen ist daher sinnvoll. Für die verschiedenen Annotationsebenen gibt es eine ganze Reihe von generischen Komponenten (cf. Lemnitzer & Zinsmeister 2010; McEnery, Xiao & Tono 2006; Kübler & Zinsmeister 2015). Forschungsinfrastrukturen wie CLARIN-D stellen eine Auswahl dieser Werkzeuge webbasiert zur Verfügung mit der Möglichkeit eigene Prozessketten zusammenzustellen (cf. WebLicht; Hinrichs, Hinrichs & Zastrow 2010, Düsendi 2014). Plattformen wie Gate[15] (*General architecture for text engineering*, Cunningham, Maynard & Bontcheva 2011; Cunningham et al. 2013) oder NLTK (*Natural Language Tool Kit*)[16] stellen ähnliche Funktionalitäten zur Verfügung.

15 https://gate.ac.uk/ (letzter Zugriff: 22. 4. 2018).
16 http://www.nltk.org/ (letzter Zugriff: 22. 4. 2018).

Beim Einsatz dieser generischen Komponenten ist zu beachten, dass die Annotationen, die automatische Werkzeuge liefern nicht perfekt sind (d. h. es treten Fehler auf), da die Werkzeuge für eine bestimmte Sprachvarietät optimiert sind, zumeist für Allgemeinsprache (s. oben). Eine Evaluierung der Annotation kann die Qualität der Annotation bestimmen und Probleme aufzeigen. Die Qualität der Annotation bemisst sich nach dem Anteil der korrekten Annotation an der Anzahl aller Annotationen (der sogenannten Präzision). Ist die Qualität der Annotation zu schlecht, muss das generische Werkzeug eventuell angepasst oder neu trainiert werden. In der Regel bedeutet dies aber auch, dass eine webbasierte Prozesskette nicht mehr möglich ist, da hier nur die generischen Komponenten verwendet werden können. Die Lösung ist in diesem Fall der Aufbau einer eigenen lokalen Prozesskette mit angepassten generischen Komponenten.

Bei kleineren Corpora besteht auch die Möglichkeit manuell zu annotieren oder die Annotation manuell zu korrigieren. Auch hier gibt es generische Werkzeuge, die die manuelle Annotation unterstützen, wie etwa MMAX (Müller & Strube 2001), annotate (Brants & Plaehn 2000), SALTO für semantische Annotationen (Burchardt et al. 2006) oder EXMARaLDA[17] für Sprachcorpora. Einige davon, etwa das webbasierte Werkzeuge WebAnno,[18] unterstützen eine Vielzahl von Annotationsebenen, darunter auch selbst definierte Ebenen. WebAnno hat zudem einen automatischen Modus, der lernt und Annotationen vorschlägt (cf. Eckart de Castilho et al. 2014; Yimam et al. 2013, 2014). Die Entscheidung wieviel im Einzelfall in spezifische Lösungen investiert wird, ist immer eine Abwägung zwischen Qualität und Effizienz.

2.2 Corpusanalyse

Die spezifischste Komponente einer corpuslinguistischen Studie ist natürlich die Analyse der extrahierten Daten (z. B. eine Merkmalsverteilung). Aber auch bei der Corpusanalyse lassen sich nicht nur spezifische sondern auch generische Elemente identifizieren. Abbildung 5.3 zeigt eine schematische Darstellung des Arbeitsablaufes bei der Corpusanalyse.

Prinzipiell kann man zwei Vorgehensweisen unterscheiden, *corpus-based* (corpusbasiert oder phänomengesteuert) und *corpus-driven* (corpusgesteuert oder explorativ) (cf. Tognini-Bonelli 2001: 30; McEnery, Xiao & Tono 2006: 8–10). Bei der corpusbasierten Vorgehensweise werden zunächst die für die Un-

17 http://exmaralda.org/de/ (letzter Zugriff: 22. 4. 2018).
18 https://webanno.github.io/webanno/ (letzter Zugriff: 22. 4. 2018).

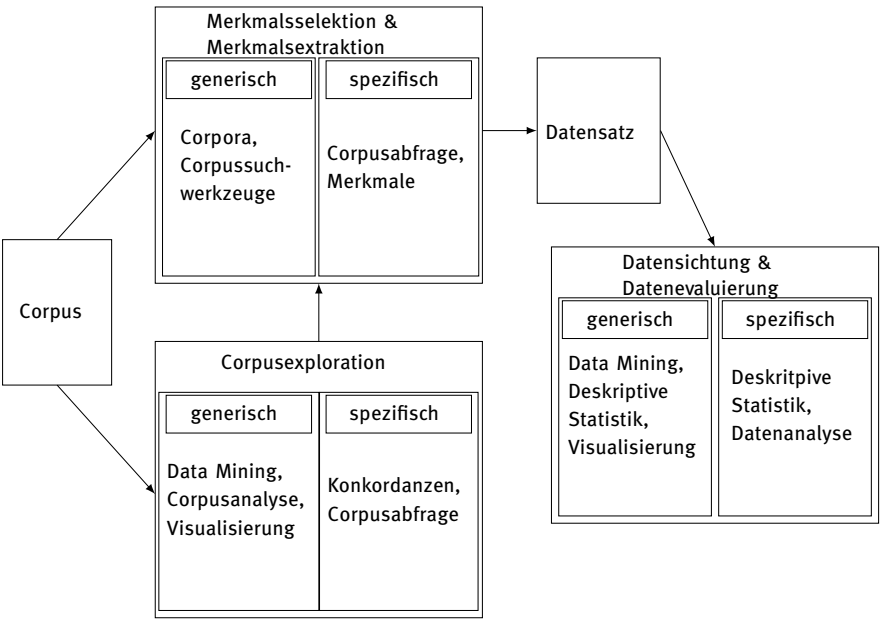

Abb. 5.3: Arbeitsablauf bei der Corpusanalyse.

tersuchung relevanten Merkmale identifiziert (Merkmalssselektion) (cf. Biber & Finegan 2014) und anschließend die entsprechenden Corpusinstanzen extrahiert (Merkmalsextraktion). Der daraus resultierende Datensatz aus Feature-Wert-Paaren wird dann gesichtet (z. B. mit Hilfe von Visualisierungen) und evaluiert (z. B. mit Methoden der deskriptiven Statistik oder Data Mining) und schließlich interpretiert.

Die corpusgesteuerten Vorgehensweise basiert auf einer explorativen Analyse des Corpus z. B. mit dem Ziel typische Merkmale für eine durch das Corpus repräsentierte Sprachvarietät zu identifizieren. Generische Komponenten der Corpusexploration beinhalten oft die einzelnen Schritte von der Merkmalsselektion, über die Merkmalsextraktion bis zur Datensichtung und Datenevaluierung. Die Corpusexploration kann aber auch als Unterstützung bei der Merkmalsselektion der corpusbasierten Methode dienen (cf. Biber & Finegan 2014). Beide Methoden ergänzen einander (McEnery, Xiao & Tono 2006).

Bei der Corpusexploration können generische Werkzeuge, z. B. aus dem Language Modeling (Word Embeddings, N-gram-Modelle) oder Corpusanalysetools wie das Voyant Tool[19] eine andere Sichtweise auf das Corpus ermög-

19 http://voyant-tools.org/ (letzter Zugriff: 22. 4. 2018).

lichen. Sie abstrahieren von der Textgrundlage durch (i) Gruppierungen in Form von Häufigkeitsverteilungen von Wörtern und Phrasen, Messungen der lexikalischen Dichte und Varianz und (ii) Hervorhebung von typischen Elementen in Form von Termen, Keywords oder grammatikalischen Einheiten. Aber auch die manuelle Sichtung eines Corpus über Corpussuche und Konkordanzen kann vor allem bei kleineren Corpora sinnvoll sein. Durch die Corpusexploration können auch noch unbekannte relevante Merkmale identifiziert werden. Für die detaillierte Analyse der Merkmale muss dann eine Datensichtung auf einer Mikroebene stattfinden, etwa durch Extraktion und/oder Datensichtung und Datenevaluierung auch auf der Textebene (Konkordanzen).

Bei der corpusbasierten Methode geht man von einem zu untersuchenden linguistischen Phänomen aus. Die Merkmalsselektion erfolgt manuell auf der Basis von vorhandenen linguistischen Studien (Merkmalskatalog als generische Ressource) aber auch über die Corpusexploration. Für die Merkmalsextraktion werden in der Regel generische Corpusabfragewerkzeuge verwendet, die Abfragen selbst sind natürlich spezifisch, aber es kann sich lohnen wiederkehrende Abfragen in einer Art Abfragebibliothek zu speichern. Inzwischen sind viele Corpora online verfügbar und abfragbar. Dabei unterscheiden sich die einzelnen Plattformen für die Corpusabfrage hinsichtlich ihrer Funktionalität. Exemplarisch seien hier für die deutsche Sprache COSMAS (*Corpus Search, Management and Analysis System*[20]) als Abfragewerkzeug für die Corpussammlung des Instituts für Deutsche Sprache, sowie das DWDS (*Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*[21]) und das Deutsche Text Archiv[22] genannt. Für englischsprachige Corpora seien die BYU Corpora[23] sowie das BNCweb[24] und die Corpora auf der CQPweb Plattform der Lancaster University[25] genannt. Schließlich sei noch das OPUS Projekt (*open parallel corpus*[26]) erwähnt, das eine große Sammlung von parallelen Übersetzungstexten als alignierte Corpora bereitstellt.

Die meisten der bereitgestellten Corpora sind lemmatisiert und wortartengetaggt. Corpusabfragen können auf diesen Annotationen aufsetzen und erlauben den Einsatz von regulären Ausdrücken bei der Abfrage. Je nach Untersuchungsgegenstand bieten diese online Plattformen unterschiedliche

20 http://www.ids-mannheim.de/cosmas2/ (letzter Zugriff: 22. 4. 2018).
21 https://www.dwds.de/ (letzter Zugriff: 22. 4. 2018).
22 http://www.deutschestextarchiv.de/ (letzter Zugriff: 22. 4. 2018).
23 http://corpus.byu.edu/ (letzter Zugriff: 22. 4. 2018).
24 http://corpora.lancs.ac.uk/BNCweb/ (letzter Zugriff: 22. 4. 2018).
25 https://cqpweb.lancs.ac.uk/ (letzter Zugriff: 22. 4. 2018).
26 http://opus.lingfil.uu.se/ (letzter Zugriff: 22. 4. 2018).

generische Lösungen für die Corpusanalyse. Als Ergebnis wird immer zumindest eine Konkordanz ausgegeben und die Anzahl der Treffer. Bei OPUS können zusätzlich noch alignierte Textstellen mit ausgegeben werden.

Einige der Plattformen bieten darüber hinaus weitere Auswertungsmöglichkeiten. Das DWDS bietet für Einzelwörter Zugriff auf lexikalische Informationen wie Bedeutung, Etymologie, Thesaurus, typische Verbindungen sowie eine Verlaufskurve der Häufigkeit des Wortes über die Zeit. Das Deutsche Text Archiv bietet u. a. die Möglichkeit Corpora mit dem Voyant Tool[27] zu analysieren (häufigste Terme und Phrasen, Häufigkeitenverteilung von Termen im Dokument, Volltextanzeige). Die Plattform der BYU Corpora bietet einige vorprozessierte Auswertungen für Einzelwörter (Kollokationen, Synonyme, Definitionen sowie Frequenzdistributionen).

Corpusplattformen, die auf der Abfragesprache CQP (Evert & Hardie 2011) und dem dafür von Andrew Hardie entwickelten webbasierten GUI CQPweb (Hardie 2012) basieren, vereinen eine effiziente und mächtige Abfragesprache mit verschiedenen Auswertungsmöglichkeiten (Häufigkeitsverteilungen, Kollokationsanalyse, Keywordanalyse). Die meisten Auswertungen sind nicht vorprozessiert, lediglich Frequenzlisten werden bereits bei der Corpusinstallation berechnet. Bei der Datenextraktion und Datenauswertung kann auf die gesamte Annotation des Corpus zugegriffen werden. Für die Extraktion von komplexen Datensätzen sind modulare Plattformen wie CQPweb besonders geeignet, da sie erlauben die Merkmale für die Extraktion sehr spezifisch zu definieren.

Ähnlich wie bei der Corpusexploration können generische Komponenten bei der Datensichtung und Datenevaluierung helfen, von den zugrundeliegenden Daten zu abstrahieren und so eine andere Sichtweise auf die extrahierten Daten ermöglichen, wodurch die Makrostruktur der Ergebnisse erst sichtbar wird. Generische Werkzeuge für die statistische Auswertung und Visualisierung wie R[28] aber auch Werkzeuge aus dem Data Mining (z. B. WeKa[29] oder Rapid Miner[30] für die Textklassifikation) bieten hier verschiedene Auswertungsmöglichkeiten, die je nach Untersuchungsgegenstand spezifisch angewandt werden können.

Durch den Einsatz von generischen Komponenten bei der Corpusanalyse kann diese effizient und schnell durchgeführt werden. Zudem gewährleistet der Einsatz von statistischen Verfahren und Methoden aus dem Data Mining ein hohes Maß an Objektivität. Es ist jedoch auch hier zu beachten, dass

27 http://voyant-tools.org/ (letzter Zugriff: 22. 4. 2018).
28 https://www.r-project.org/ (letzter Zugriff: 22. 4. 2018).
29 http://www.cs.waikato.ac.nz/~ml/weka/ (letzter Zugriff: 22. 4. 2018).
30 https://rapidminer.com/ (letzter Zugriff: 22. 4. 2018).

(i) automatische Werkzeuge keine perfekten Ergebnisse liefern, (ii) es nicht immer einfach ist, die Generalisierungen und Abstraktionen richtig zu interpretieren und (iii) die Werkzeuge per se keine spezifischen Analysen bieten. Dies bedeutet für spezifische Untersuchungen, dass bei der Interpretation der Blick auf die Mikroebene, also die Textgrundlage in Form von Textausschnitten, Konkordanzen oder Beispieldaten, nicht ausbleiben kann.

3 Zusammenspiel zwischen generischer Infrastruktur und spezifischer Forschung

In diesem Kapitel zeigen wir das Zusammenspiel zwischen generischer Infrastruktur und spezifischen Lösungen anhand eines konkreten Beispiels. Wir gehen zunächst kurz auf den linguistischen Hintergrund der Studie ein, beschreiben dann die Vorgehensweise bei der Corpuserstellung (Vorverarbeitung und linguistische Annotation) und gehen dann auf die Corpusanalyse (Merkmalsextraktion, Datensichtung und Datenevaluierung) ein.

Unser Interesse gilt der Entwicklung der englischen Wissenschaftssprache. Laut Halliday (1988) und Halliday & Martin (2005) kommt es hier aufgrund von Spezialisierung zu einer größeren Kodierungsdichte, d. h. kürzere, kompaktere sprachliche Ausdrücke werden häufiger benutzt, um dem Prinzip der Spracheffizienz zu entsprechen. Dabei zeigen sich Merkmale sprachlicher Verdichtung auf allen linguistischen Ebenen, z. B. Reduktion auf der syntaktischen Ebene (Artikelauslassung), Nominalisierungen auf der morphologischen Ebene und eine größere lexikalische Dichte auf der Wortebene (ausgeprägtere Verwendung von Inhaltswörtern).

Wir gehen nun davon aus, dass sich diese Art der Verdichtung am sprachlichen Signal als Informationsdichte messen lässt, also als Anzahl von Bits, die für die Kodierung einer gegebenen Äußerung notwendig ist (Shannon Information). Üblicherweise wird die Informationsdichte formal repräsentiert als die logarithmische Wahrscheinlichkeit einer sprachlichen Einheit gegeben einen Kontext (Crocker, Demberg & Teich 2015). Vereinfacht gesagt, je besser ein gegebener Kontext die sprachliche Einheit vorhersagen kann, desto kürzer ist die sprachliche Einheit (vgl. z. B. Variation in der Wortlänge, Mahowald et al. 2013) und desto weniger Bits werden für die Kodierung benötigt.

Für die Untersuchung brauchen wir einen Corpus des wissenschaftlichen Englisch, das eine gewisse zeitliche Ausdehnung hat, die Anfänge des wissenschaftlichen Schreibens in Englisch einschließt und eine möglichst breite Abdeckung im Bezug auf Disziplinen bietet. Vorhandene diachrone Corpora sind entweder auf eine Disziplin beschränkt oder decken nur eine bestimmte Zeit-

periode ab (z. B. das Corpus of Early Modern English Medical Texts, Taavitsainen & Pahta 2012) bzw. sind recht klein (z. B. das Coruña Corpus, Moskowich & Crespo 2007). Daher ist es hier nicht möglich auf eine vorhandene Ressource zurückzugreifen. Es musste ein Corpus neu erstellt werden. Betrachtet man die Rolle der *Royal Society of London* bei der Entwicklung der Wissenschaft ab Mitte des 17. Jahrhunderts (Atkinson 1998), so sind die *Philosophical Transactions* eine geeignete Datengrundlage. Die *Philosophical Transactions* wurden 1665 von Henry Oldenburg gegründet und sind die erste regelmäßig erscheinende Zeitschrift, die wissenschaftliche Artikel in englischer Sprache veröffentlichte. Sie enthielt ursprünglich sowohl wissenschaftliche Korrespondenz, Rezensionen und Zusammenfassungen von Büchern als auch wissenschaftliche Berichte von Beobachtungen und Experimenten. Als solches repräsentiert sie die Anfänge wissenschaftlichen Schreibens in englischer Sprache bis hin zu der Etablierung eines ersten wissenschaftlichen Standards.

3.1 Corpuserstellung

Ein weiterer Vorteil der *Philosophical Transactions* ist, dass sie bereits digitalisiert sind und von JSTOR zusammen mit Metadaten wie Autor, Texttyp und Jahr der Publikation etc. in wohlgeformtem XML bereitgestellt werden. Das Royal Society Corpus (RSC) umfasst alle Veröffentlichungen der *Philosophical Transactions* (Artikel, Buchrezensionen und Abstracts, sowie verschiedene Produktionsmodi, schriftsprachliche und gesprochensprachliche Texte) aus den ersten 200 Jahren der Zeitschrift von 1665–1869. In der aktuellen Version (2.0) hat das RSC ca. 35 Millionen Token. Bei den Digitalisaten handelt es sich um gescannte Texte aus unterschiedlichen Quellen (Tab. 5.1 zeigt einen Überblick über die Anzal der Texte pro Textkategorie und Zeitperiode), die noch weitgehend unerforscht bzw. unbekannt sind.

Tab. 5.1: Quellen des RSC.

		Buchrezen.	Artikel	Misc.	Insg.
Philosophical Transactions	1665–1678	124	641	154	919
Philosophical Transactions	1683–1775	154	3.903	338	4.395
Philosophical Transactions of the Royal Society of London	1776–1869	–	2.531	283	2.814
Abstracts of Papers Printed	1800–1842	–	1.316	15	1.331
Abstracts of Papers Communicated	1843–1861	–	429	5	434
Proceedings of RSL	1862–1869	–	1.476	38	1.528
Insgesamt		**278**	**10.296**	**833**	**11.421**

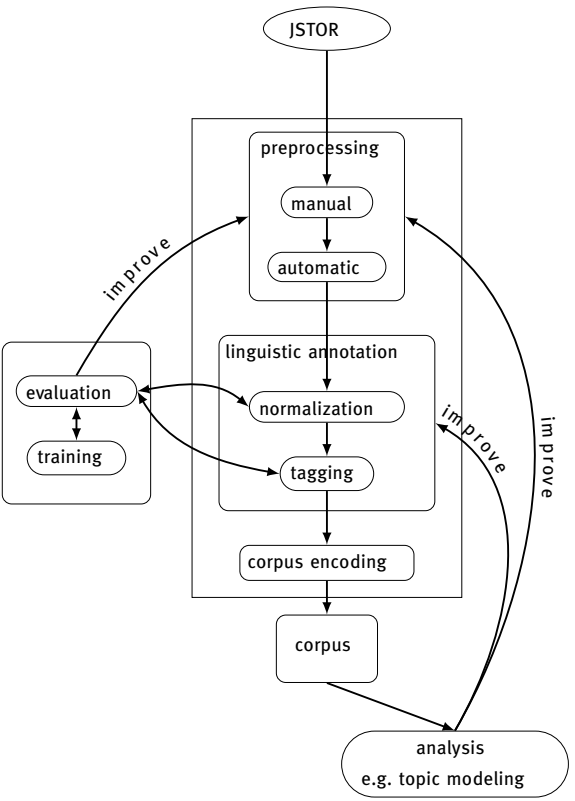

Abb. 5.4: Arbeitschritte beim Agile Corpus Building (vgl. Kermes et al. 2016a).

Wir haben uns daher für eine inkrementelle Vorgehensweise bei der Corpuserstellung entschieden, die sich an der Idee des *Agile Software Development* (Cockburn 2001) orientiert. Der gesamte Prozess der Corpuserstellung von den Ausgangsdaten bis zum abfragbaren Corpus ist weitgehend automatisiert und verwendet, wo immer möglich, generische Komponenten (etwa bei der linguistischen Annotation). Dedizierte automatische Komponenten ergänzen die generischen Werkzeuge, wenn eine generische Komponente nicht verfügbar ist. Manuelle Arbeitsschritte werden nur vorgenommen, wenn eine Automatisierung nicht möglich oder nicht sinnvoll ist und setzen direkt auf den Ausgangsdaten auf. Die Automatisierung hat den Vorteil, dass auf Probleme in der Datenqualität relativ schnell und effizient reagiert werden kann. Die entsprechenden Komponenten können angepasst oder ergänzt werden und es kann eine neue verbesserte Version des Corpus erstellt werden (cf. Kermes et al. 2016a, b). Abbildung 5.4 zeigt eine schematische Darstellung des Arbeitsablaufs.

3.2 Vorverarbeitung

Die Ausgangsdaten des RSC liegen zwar digitalisiert und in einem strukturierten XML-Format vor, eine Vorverarbeitung ist dennoch notwendig, um die Daten in ein standardisiertes Format zu überführen und zu bereinigen.

Bei der OCR-Fehlerkorrektur greifen wir auf die in Unterkapitel 2 *Methodik, Arbeitsabläufe und Angebote* beschriebene Ersetzungsliste von Underwood & Auvil (2012) zurück. Da das RSC jedoch recht spezifisch ist, zeigt sich, dass die Listen nicht einfach übernommen werden können, sondern an die speziellen Bedürfnisse angepasst werden müssen. Muster die für das RSC nicht relevant sind werden gelöscht, andere angepasst (so wird etwa ‚fhe' zu ‚the' anstatt zu ‚she') und wieder andere Muster werden ergänzt. Für die Identifizierung der spezifischen Muster werden u. a. *word embeddings* verwendet, in der Annahme, dass falsch geschriebene Wörter ähnlich verwendet werden, wie das richtig geschriebene Wort. Bisher haben wir so ca. 360 spezifische Ersetzungsmuster ergänzt (cf. Knappen et al. 2017).

Bei der Lösung der Layoutprobleme muss ebenfalls auf spezifische Lösungen zurückgegriffen werden. Dedizierte Skripte kümmern sich etwa um Kopf- und Fußzeilen im Fließtext, in der Reihenfolge vertauschte oder fehlende Seiten, uneinheitliche Seitennummern, Seitenduplikate (erste und letzte Seite) und nicht eindeutig markierte Artikelgrenzen. Ist eine automatische Lösung nicht möglich, wird manuell oder (semi-)automatisch gesichtet. Bei einigen Quellen wurden Artikelgrenzen manuell annotiert. Texte und Seiten mit großen Tabellen wurden automatisch identifiziert und anschließend manuell gesichtet.

3.3 Linguistische Annotation

Wie oben bereits diskutiert, sind die meisten linguistischen Werkzeuge für gegenwartssprachliche, allgemeinsprachliche Corpora optimiert. Das RSC ist jedoch ein historisches Corpus aus wissenschaftlichen Texten. Es weicht also sowohl zeitlich als auch bezüglich des Registers ab. Trotzdem greifen wir bei der linguistischen Annotation auf generische Werkzeuge zurück, evaluieren die Ergebnisse und passen die Werkzeuge dann gegebenenfalls an.

Für die Normalisierung (hier Modernisierung) der historischen Originalwörter verwenden wir VARD (Baron & Rayson 2008), ein regelbasiertes statistisches Werkzeug, das orthographische Varianten bzw. historische Wortformen auf gegenwartssprachliche Wörter abbildet. VARD wurde für die Zeitperiode zwischen 1450–1700 entwickelt und überschneidet sich somit mit der Zeitperiode des RSC (1665–1869). Die Evaluierung von VARD zeigt eine Präzision von 61,8 %

und einen Recall von 31,4 %. Ein speziell trainiertes Modell verbessert die Präzision um mehr als 10 % auf 72,8 %. Der Recall verdoppelt sich auf fast 57,7 %.

Für Tokenisierung, Lemmatisierung und Wortartenannotation wird der TreeTagger (Schmid 1994; 1995) verwendet. Der TreeTagger ist ein Wortartentagger der auf gegenwartssprachlichem Zeitungstext trainiert wurde. Eine Evaluierung zeigt, dass der TreeTagger mit einer Präzision von 94 % (im Gegensatz zu 97 % auf gegenwartssprachlichen Texten) auch auf dem historischen Sprachmaterial zumindest akzeptable Ergebnisse erzielt. Eine detaillierte Analyse der Taggingfehler zeigt zwei Hauptfehlerquellen: NN-NP (Verwechslung von Nomen und Eigennamen) und WP-WDT (Verwechslung von Wh-Relativpronomen mit WhArtikeln). Beide Fehlerquellen sind für viele Analysen unproblematisch. Ignoriert man NN-NP Fehler, so erhöht sich die Präzision auf 95 %. Wir verwenden den TreeTagger daher im Augenblick fast unverändert. Lediglich das Lexikon des Tokenisierers wurde um ca. 170 Abkürzungen ergänzt. Dazu wurden zunächst Abkürzungskandidaten aus dem RSC extrahiert und die Häufigsten anschließend manuell gesichtet.

Für unsere Untersuchungen benötigen wir neben den klassischen linguistischen Annotation noch andere Informationen. So annotieren wir zusätzlich *Surprisal*, also den Informationsgehalt der Wörter in Anzahl Bits, berechnet als

$$S(unit) = -\log_2 p(unit|context)$$

d. h., der (negativen logarithmischen) Wahrscheinlichkeit einer gegebenen Einheit (z. B. eines Wortes) in einem Kontext (z. B. den vorangehenden Wörtern) (vgl. Genzel & Charniak 2002). *Surprisal* (der Informationsgehalt, Levy 2008) drückt die Intuition aus, dass je unwahrscheinlicher eine sprachliche Einheit in einem bestimmten Kontext ist, desto „überraschender“ (*more surprising*) oder informativer ist diese Einheit und desto mehr Bits werden benötigt, um sie zu kodieren (und umgekehrt). *Surprisal* erlaubt es, sprachliche Einheiten auf ihren Informationsgehalt hin zu untersuchen, den Kontext der Einheit bei der Untersuchung zu berücksichtigen und so über eine rein frequenzbasierte Untersuchung hinauszugehen. Für die Annotation von *Surprisal* gibt es bisher kein generisches Werkzeug, hier musste daher eine spezifische Lösung gefunden werden. Das dedizierte Skript annotiert *Surprisal* basierend auf verschiedenen Zeitperioden und erlaubt so den schnellen Zugriff auf diese Information. Obwohl hier als spezifische Lösung entwickelt, ist das Skript insofern auch generisch, als es auch auf andere Daten angewendet werden kann.

3.4 Beispielanalyse

Für die Beispielanalyse betrachten wir den Unterschied zwischen Funktionswörtern und Inhaltswörtern sowie deren Wortarten aus informationstheoretischer Sicht (Shannon 1949). Als Maß für den Informationsgehalt einer sprachlichen Einheit verwenden wir *Surprisal* und *Average Surprisal* (AvS), den durchschnittlichen Informationsgehalt. Dabei gehen wir von der Annahme aus, dass Funktionswörter generell besser vorhersagbar sind, ein niedrigeres AvS haben, während Inhaltswörter generell weniger vorhersagbar sind, ein höheres AvS haben. Außerdem nehmen wir an, dass das AvS von Funktionswörtern über die Zeit im Wesentlichen gleich bleibt, während das AvS von Inhaltswörtern weniger konstant ist.

Unsere Annahmen stützen sich auf bereits bekannte Unterschiede zwischen Funktionswörtern und Inhaltswörtern bezüglich Vorkommenshäufigkeit, Wortlänge, Anzahl (offene vs. geschlossene Wortklasse) und Informationsgehalt (cf. Biber et al. 1999). So zeigen Piantadosi, Tily & Gibson (2011), dass der durchschnittliche Informationsgehalt die Länge eines Wortes besser vorhersagt als dessen Häufigkeit. Laut Quirk et al. (1985: 72) ist die Anzahl der möglichen Wörter in typischen Kontexten von Inhaltswörtern größer als in typischen Kontexten von Funktionswörtern. Gleichzeitig zeigen Linzen & Jaeger (2015), dass die Anzahl an Ausdrucksmöglichkeiten die Vorhersagbarkeit der folgenden syntaktischen Konstruktion beeinflusst.

Beispielhaft schauen wir uns die Entwicklung im wissenschaftlichen Englisch an. Basierend auf der Annahme, dass es hier aufgrund der sprachlichen Verdichtung (cf. Halliday 1988; Halliday & Martin 2005) zu einer verstärkten Verwendung von Inhaltswörtern kommt, die oft durch lexikalische Dichte approximiert wird, nehmen wir an, dass wir diachrone Unterschiede beim AvS von Inhaltswörtern in wissenschaftlichen Texten beobachten können und dass diese Entwicklung sich von der Entwicklung in der Allgemeinsprache unterscheidet.

Für die Untersuchung verwenden wir das RSC und als Vergleichscorpus das *Corpus of Late Modern English Texts* (CLMET, Diller, De Smet & Tyrkkö 2011). Beide Corpora sind sowohl mit CQP[31] als auch mit dessen webbasiertes GUI CQPweb – also mit generischen Werkzeugen – abfragbar. CQPweb nutzen wir zur Corpusexploration und zum Aufbau und der Evaluierung der Abfrage. Dabei nützen wir den schnellen und flexiblen Zugriff auf Konkordanzen, Häufigkeitsverteilungen sowie Sortierungen und Gruppierungen der Ergebnisse, um die Adäquatheit unserer Abfragen zu überprüfen.

31 http://www.cwb.sourceforge.net (letzter Zugriff: 22. 4. 2018).

```
tabcmd: match word, match pos, match surpr50, match text_period
tabcmd_header: word, pos, surpr, time
ex: pos_avs:: [word = "\w+" & word !=".*[0-9_]+.*" &
               pos !="SYM|FW|UH|LS|CD|V[BH].*|N.*"] |
              [pos="N.*" & word =".{3,}" &
               word != ".*[0-9].*" & word = "\w+"]
ex: pos_avs_vbhaux:: [pos="V[BH].*"][pos!="VV.*"]{4}
ex: pos_avs_vbh:: [pos="V[BH].*"][]{0,3}[pos="VV.*"]
```

Abb. 5.5: Parameterdatei für die Merkmalsextraktion.

Für die Merkmalsextraktion nutzen wir dann das zugrundeliegende generische Abfragewerkzeug CQP. Es erlaubt einen Zugriff auf das Corpus durch externe Skripte. Durch die Automatisierung wird die Merkmalsextraktion reproduzierbar und auf andere Daten übertragbar. Die spezifischen Komponenten (Corpusabfrage und Merkmale) werden in Parameterdateien gespeichert, wobei eine Parameterdatei mehrere Abfragen enthalten kann. Mit einem dedizierten Extraktionsskript können dann für diese Parameter die entsprechenden Merkmale aus beliebigen CQP-Corpora extrahiert werden.

In unserem konkreten Fall wollen wir alle englischen Wörter aus dem RSC und dem CLMET extrahieren. Die Abfrage ist so formuliert, dass keine Fremdwörter, Symbole oder Zahlen extrahiert werden. Außerdem schließen wir Nomen aus, die aus weniger als drei Buchstaben bestehen. Die Merkmale, die wir für die Corpusinstanzen extrahieren sind das Wort selbst, die Wortart, der Surprisalwert und die Zeitperiode (50-Jahre-Zeitperioden). Die Verben *be* und *have* werden separat extrahiert, um zwischen Auxiliar und Vollverb zu unterscheiden. Abbildung 5.5 zeigt die Parameterdatei für die Extraktion.

Mit `tabcmd` und `tabcmd_header` werden die Corpusattribute und die Spaltennamen für die Merkmalsextraktion definiert. Die einzelnen Abfragen werden mit ex und einem Namen gekennzeichnet. Das Ergebnis der Extraktion ist eine TAB-getrennte Feature-Wert-Tabelle, mit einer Spalte für jedes Merkmal und einer Zeile für jede Corpusinstanz, die automatisch in einer Datei mit dem Namen der Abfrage gespeichert wird (s. a. Tab. 5.2).

Für eine bessere Abstraktion fügen wir der Tabelle zwei weitere Merkmale hinzu, indem wir die Wortartentags des verwendeten Tagsets (*Penn Treebank Tagset*, Marcus, Santorini & Marcinkiewicz 1993) in übergeordnete Wortarten sowie Funktionswörter (Artikel, Präpositionen, Pronomen, Modalverben, Konjunktionen und Auxiliarverben) und Inhaltswörter (Nomen, Adjektive, Verben, Adverben) gruppieren (s. a. Tab. 5.3).

Für die statistische Auswertung und Visualisierung verwenden wir R. R bietet einerseits den Zugriff auf bereits implementierte Auswertungen und

Tab. 5.2: Feature-Wert-Tabelle als Ergebnis der Extraktion.

word	pos	avs	time
An	DT	6.51	1650
Account	NP	1.12	1650
of	IN	0.06	1650
some	DT	2.66	1650
Books	NPS	0.90	1650

Tab. 5.3: Feature-Wert-Tabelle mit Abstraktion der Wortarten.

word	pos	apos	type	avs	time
An	DT	article	fw	6.51	1650
Account	NP	noun	cw	1.12	1650
of	IN	preposition	fw	0.06	1650
some	DT	article	fw	2.66	1650
Books	NPS	noun	cw	0.90	1650

andererseits die Möglichkeit eigene Funktionen für die spezifische Datenanalyse zu schreiben. Wir verwenden auch hier dedizierte Skripte. Dabei greifen wir einerseits auf bereits implementierte Auswertungen (z. B. den Mittelwert) und Visualisierungen (hier: Graphik zur Dichteverteilung) zurück und definieren andererseits spezifische Aspekte der Datenanalyse im Skript. Durch die Automatisierung mit dedizierten R-Skripten schaffen wir auch hier eine Reproduzierbarkeit der Ergebnisse. Durch Parameter (Datendatei, Corpus, Merkmale) in den R-Skripten sind die Analysen auch auf andere Daten übertragbar, z. B. auf Daten, die aus anderen Corpora extrahiert wurden oder auf anderen Extraktionen beruhen. Das Ergebnis ist dann z. B. die Graphik einer Dichteverteilung wie in Abbildung 5.6.

Die Abbildung zeigt die diachrone Entwicklung des AvS von Wortarten im RSC als Dichteverteilung. Wir sehen u. a., dass sich das AvS von einigen Wortarten im RSC über die Zeit tatsächlich verändert. So steigt das AvS von typischen modifizierenden Wortarten wie Adjektive, Adverbien, Modalverben sowie von Pronomen über die Zeit leicht an. Für Artikel, Präpositionen und Nomen sowie für Verben und Auxiliare sinkt das AvS.

Für einen Vergleich mit dem CLMET müssen wir nun dieselbe Merkmalsextraktion und Datenanalyse auf dem CLMET durchführen. Durch den Einsatz von generischen Komponenten und die Automatisierung und Modularisierung der spezifischen Skripte, ist eine Übertragbarkeit des Prozesses auf das CLMET

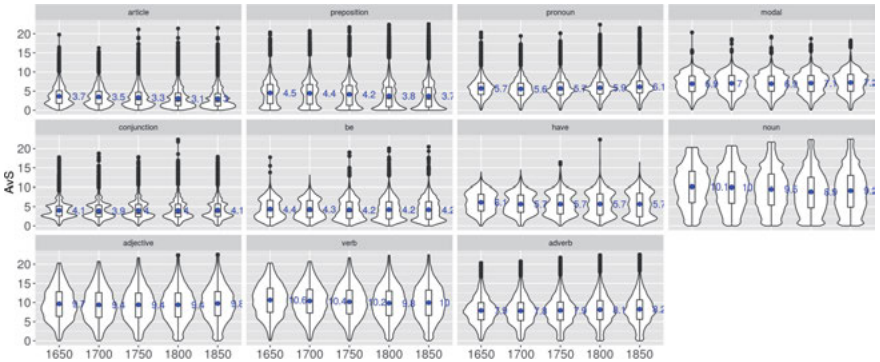

Abb. 5.6: Diachrone Entwicklung der AvS von Wortarten im RSC.

schnell und effizient möglich. Lediglich der Parameter des Corpus muss geändert werden (vgl. Kermes & Teich 2017) für eine ausführlichere Analyse des AvS von Wortarten im RSC).

4 Zusammenfassung und Schluss

Wir haben gezeigt welche Vorteile der Einsatz von generischen Infrastrukturkomponenten für spezifische Forschung haben kann: (i) Untersuchungen können auf größeren Datenmengen und effizienter durchgeführt werden und (ii) Ergebnisse können reproduziert und übertragbar gemacht werden. Dabei haben wir auch an einer konkreten Studie gezeigt, dass generische Infrastruktur auch spezifisch angepasst oder durch spezifische Lösungen ergänzt werden kann. Es zeigt sich, dass manche zunächst spezifischen Lösungen durchaus wiederverwendbar sind und so auch zu generischen Komponenten werden können bzw. diese ergänzen (z. B. erweiterte oder modifizierte Wortlisten für OCR-Korrektur, R-Skripte für komplexe Merkmalsextraktion).

Empirische Forschung an natürlichsprachlichen Daten kommt ohne den Einsatz von automatischen Verfahren, die über generische Werkzeuge (z. B. Tagger, Parser) zur Verfügung gestellt werden, nicht aus. Generische Werkzeuge unterstützen den Forschungsprozess und eröffnen neue Möglichkeiten, sie schließen aber spezifische Vorgehensweisen nicht aus und können auch manuelle Analyse und Interpretation nicht ersetzen. Dabei ist wichtig, dass man versteht, was die generischen Werkzeuge zu leisten im Stande sind und wo ihre Grenzen sind. Denn nur so ist gewährleistet, dass sie richtig eingesetzt werden und die Ergebnisse kritisch betrachtet und analysiert werden bzw. aus den Ergebnissen valide Schlussfolgerungen gezogen werden können.

Literatur

Atkinson, Dwight (1998): *Scientific discourse in sociohistorical context: The philosophical transactions of the Royal Society of London, 1675–1975*. Routledge.

Baron, Alistair & Paul Rayson (2008): VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Biber, Douglas, Susan Conrad & Randi Reppen (1998): *Corpus linguistics: Investigating language structure and use*. New York: Cambridge University Press.

Biber, Douglas & Edward Finegan (2014): On the exploitation of computerized corpora in variation studies. In Karin Aijmer & Bength Altenberg (Hrsg.), *English Corpus Linguistics*, 204. Routledge.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999): *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Brants, Thorsten & Oliver Plaehn (2000): Interactive Corpus Annotation. In *LREC*.

Buchholz, Sabine & Erwin Marsi (2006): CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164. Association for Computational Linguistics.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado & Manfred Pinkal (2006): SALTO–a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, 517–520. Citeseer.

Cockburn, Alistair (2001): *Agile Software Development*. Boston, USA: Addison-Wesley Professional.

Crocker, Matthew W., Vera Demberg & Elke Teich (2015): Information Density and Linguistic Encoding (IDeaL). *KI – Künstliche Intelligenz* doi: 10.1007/s13218-015-0391-y.

Cunningham, Hamish, Diana Maynard & Kalina Bontcheva (2011): *Text processing with gate*. Gateway Press CA.

Cunningham, Hamish, Valentin Tablan, Angus Roberts & Kalina Bontcheva (2013): Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology* 9(2), e1002854. doi: 10.1371/journal.pcbi.1002854.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö (2011): A European database of descriptors of English electronic texts. *The European English Messenger* 19, 21–35.

Düsendi, Bahadir (2014): *Erstellung annotierter Textcorpora mit WebLicht. Computerlinguistik als Sprachwissenschaft*. München: GRIN Verlag.

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych & Seid Muhie Yimam (2014): WebAnno: A flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC)*.

Evert, Stefan & Andrew Hardie (2011): Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Brimingham, UK.

Genzel, Dmitriy & Eugene Charniak (2002): Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 199–206. Association for Computational Linguistics.

Halliday, M. A. K. (1988): On the Language of Physical Science. In Mohsen Ghadessy (Hrsg.), *Registers of Written English: Situational Factors and Linguistic Features*, 162–177. London: Pinter.

Halliday, M. A. K. & J. R. Martin (2005): *Writing Science: Literacy and Discursive Power*. Taylor & Francis.

Hardie, Andrew (2012): CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics* 17(3), 380–409. doi: doi:10.1075/ijcl.17.3.04har.

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29. Association for Computational Linguistics.

Kermes, Hannah (2008): Syntactic Preprocessing. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics. An International Handbook*, Band 1 Handbücher zur Sprach- und Kommunikationswissenschaft, 598–612. de Gruyter Mouton.

Kermes, Hannah, Stefania Degaetano, Ashraf Khamis, Jörg Knappen & Elke Teich (2016a): The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the LREC 2016*, Portoroz, Slovenia.

Kermes, Hannah, Jörg Knappen, Ashraf Khamis, Stefania Degaetano-Ortlieb & Elke Teich (2016b): The Royal Society Corpus: Towards a high-quality corpus for studying diachronic variation in scientific writing. In *Proceedings of DH 2016*, Krakow, Poland.

Kermes, Hannah & Elke Teich (2017): Average surprisal of parts-of-speech. In *Proceedings of Corpus Linguistics 2017*, Birmingham.

Knappen, Jörg, Stefan Fischer, Hannah Kermes & Elke Teich (2017): The Making of the Royal Society Corpus. In *Proceedings of Nodalida 2017*, Göteborg.

Kübler, Sandra & Heike Zinsmeister (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic annotated edition Ausg.

Langer, Hagen (2001): Syntax and Parsing. In Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hrsg.), *Computerlinguistik Und Sprachtechnologie. Eine Einführung*, 203–245. Heidelberg, Berlin: Spektrum Akademischer Verlag.

Leech, Geoffrey & Andrew Wilson (1996): EAGLES recommendations for the morphosyntactic annotation of corpora. *Version of March*.

Lemnitzer, Lothar & Heike Zinsmeister (2010): *Korpuslinguistik: Eine Einführung* NarrStudienbücher. Tübingen: Narr Verlag 2. Ausg. OCLC: 643072086.

Levy, Roger (2008): A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. Honolulu.

Linzen, Tal & T. Florian Jaeger (2015): Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*. doi: 10.1111/cogs.12274.

Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi & Edward Gibson (2013): Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126(2), 313–318. doi: 10.1016/j.cognition.2012.09.010.

Manning, Christopher D. & Hinrich Schütze (1999): *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993): Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

McEnery, Tony, Richard Xiao & Yukio Tono (2006): *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Moskowich, Isabel & Begoña Crespo (2007): Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In Javier Pérez-Guerra &

Charles Jones (Hrsg.), *Of Varying Language and Opposing Creed': New Insights into Late Modern English*, 341–357. Bern: Peter Lang.

Müller, Christoph & Michael Strube (2001): MMAX: A tool for the annotation of multimodal corpora. In *In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Citeseer.

Piantadosi, S. T., H. Tily & E. Gibson (2011): Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9), 3526–3529. doi: 10.1073/pnas.1012551108.

Quirk, Randolf, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik (1985): *A comprehensive grammar of the English language*. London: Longman.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, Helmut (2008): Part-of-Speech Tagging. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics. An International Handbook*. de Gruyter Mouton.

Shannon, Claude E. (1949): *The mathematical theory of communication*. Urbana/Chicago: University of Illinois Press.

Taavitsainen, Irma & Päivi Pahta (Hrsg.) (2012): *Early Modern English Medical Texts. Corpus description and studies*. Amsterdam: John Benjamins.

Tognini-Bonelli, Elena (2001): *Corpus linguistics at work*, Band 6. John Benjamins Publishing.

Underwood, Ted & Loretta Auvil (2012): Basic OCR correction. http://usesofscale.com/gritty-details/basic-ocr-correction/ (letzter Zugriff: 12. 12. 2017).

Voutilainen, Atro (2003): Part-of-speech tagging. In Ruslan Mitkov (Hrsg.), *The Oxford Handbook of Computational Linguistics*, 219–232. Oxford University Press.

Yimam, Seid Muhie, Chris Biemann, Richard Eckart de Castilho & Iryna Gurevych (2014): Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 91–96. Baltimore, Maryland: Association for Computational Linguistics.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho & Chris Biemann (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *ACL (Conference System Demonstrations)*, 1–6.