Aufbau einer Korpusinfrastruktur für die Beobachtung des Schreibgebrauchs

Fischer, Peter M.
peter.fischer@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Diewald, Nils
diewald@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Kupietz, Marc
kupietz@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Witt, Andreas
witt@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Mit dem Ziel, eine systematische Beobachtung des Schreibgebrauchs unter Verwendung computerlinguistischer Methoden zu ermöglichen, wurde 2013 das vom BMBF geförderte Forschungsprojekt *Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen* ins Leben gerufen. An diesem beteiligen sich gemeinschaftlich das Institut für Deutsche Sprache, das Institut für Computerlinguistik der Universität des Saarlandes, sowie die Wörterbuchverlage Bibliographisches Institut GmbH (Dudenverlag) und Wahrig bei Brockhaus. Das Projekt hat sich u.a. zur Aufgabe gemacht, eine zweckdienliche Datengrundlage (Fischer i.E.) und ein dazugehöriges Methodeninventar (Scholze-Stubenrecht 2013) aufzubauen.

Für die Erstellung von Korpusanalysen mit Auswertung nach eigens erarbeiteten Bewertungskriterien (Krome 2013) ist das Projekt auf eine geeignete Korpusinfrastruktur angewiesen, die es den beteiligten Partnern erlaubt, entsprechende Suchanfragen auf den einerseits umfangreichen (über 10 Mrd. Tokens), andererseits aus datenschutz- und urheberrechtlichen Gründen mitunter verteilt liegenden Ressourcen effizient und zuverlässig durchzuführen. Dabei wird entsprechend Jim Grays (Gray 2003) Maxime "put the computation near the data" (Kupietz et al. 2014) der Ansatz verteilter virtueller Korpora bzw. Kollektionen (van Uytvanck 2010) verfolgt, der darauf abzielt, dedizierte, auf die spezifischen Suchanfragen ausgerichtete Subkorpora zu definieren und auf diesen rechtskonform zu operieren.

KorAP (Ba#ski et al. 2013) ist eine Such- und Analyseplattform, die eine solche Infrastruktur zur Verfügung stellt. Sie wurde als Nachfolgesystem von COSMAS-II (Bodmer 1996) am Institut für Deutsche Sprache als primäre Schnittstelle für den Zugriff auf DeReKo (Kupietz / Lüngen 2014), das Deutsche Referenzkorpus, entwickelt. KorAP ermöglicht die Suche in sehr großen, mehrfach annotierten, und heterogen lizenzierten Korpora über eine Vielzahl von Suchoperatoren verschiedener Anfragesprachen. Die dynamische Erstellung virtueller Korpora wird dabei durch Kombination von Metadatenkriterien realisiert. Dies steht auch im Einklang mit dem Projektdesiderat, die Anbindung an die europäische Sprachressourceninfrastruktur CLARIN, die bereits eine Fülle von Werkzeugen anbietet, zu intensivieren und damit die Sichtbarkeit der Ressourcen auch im internationalen Kontext zu erhöhen.

Dieses Paper beleuchtet damit jene Arbeiten, die sich mit dem Prozess des Aufbaus der Korpusinfrastruktur, d.h. der Aufbereitung, Organisation und Bereitstellung der Datengrundlage befassen.

Als empirische Basis des Projektes dient die parallele Beobachtung und Auswertung von drei Zielgruppen und Ebenen der Textproduktion, nämlich die der professionellen Schreiber (in Zeitungen, Zeitschriften usw.), die den Schreibgebrauch der Schreibgemeinschaft heute entscheidend mitbestimmen, die der Schüler (in Klassenarbeiten, Abituraufsätzen, Literaturwettbewerben usw.), die als Repräsentanten der jungen Generation im schulischen Kontext an die amtlichen Regeln zur Rechtschreibung gebunden sind, und die der Internetnutzer (in E-Mails, sozialen Netzwerken, Meinungsportalen usw.), die in einer im Vergleich zu Druckerzeugnissen weniger kontrollierten Umgebung Entwicklungs- und Fehlertendenzen viel früher und deutlicher wiedergeben können als das beispielsweise in Zeitungstexten oder belletristischen Korpora der Fall ist. Dementsprechend steuern diese drei heterogenen Quellen auch unterschiedliche Informationen bei und stellen den Aufbau der Korpusinfrastruktur vor individuelle Herausforderungen.

Aus korpustechnologischer Sicht konnte das Projekt in Teilen auf bereits vorhandene, wohlstrukturierte und linguistisch aufbereitete Ressourcen wie das Deutsche Referenzkorpus DeReKo (Kupietz / Lüngen 2014), das WAHRIG Textkorpus [Digital] (Krome 2010) oder das Dudenkorpus (Münzberg 2011) zurückgreifen, während andere erst akquiriert, für eine maschinelle Verarbeitung vorbereitet und mit linguistischen Informationen angereichert werden mussten. Da entsprechende sprachtechnologische Verfahren (Tokenisierung, Lemmatisierung, Wortart-Tagging, flache syntaktische Analyse) jedoch überwiegend für stärker kontrollierte Texte entwickelt wurden und daher nicht auf alle diese drei Quellen gleichermaßen anwendbar sind, mussten überdies zunächst geeignete Werkzeuge (weiter-)entwickelt werden (Horbach et al. 2015), um einen

für Vergleichsanalysen ausgewogenen Aufbereitungsstand zu erreichen.

Neben diesen linguistischen Merkmalen verfügen die Texte auch über gewisse Metadaten, die aber in Struktur und Ausprägung stark an den Ressourcenkontext gebunden sind und deshalb mitunter entsprechend heterogen ausfallen. Das Zurückgreifen auf diese Informationen stellt jedoch bei synchronen wie diachronen Auswertungen ein für die systematische Beobachtung des Schreibgebrauchs zentrales Nutzungsszenario dar, das eine ordentliche Zusammenstellung solcher Zusatzinformationen erfordert. Folglich ist für die Erstellung virtueller Korpora und damit für ihre anfrageoptimierte Bereitstellung innerhalb der Analyseinfrastruktur die Erfassung von Metadaten unerlässlich. Die folgende Aufstellung zeigt eine Übersicht der Ressourcentypen und ihrer Metadaten.

Texte professioneller Schreiber (am Beispiel Zeitschriftenkorpus)

- Name der Zeitung
- Nummer der Ausgabe
- Titel des Artikels
- Untertitel des Artikels
- Name des Autors
- Ort der Veröffentlichung
- Tag der Veröffentlichung
- Textklasse (z.B. Wirtschaft oder Sport)
- Textsorte (z.B. Gerichtsurteil oder Satire)

Schülertexte (am Beispiel Literaturwettbewerbskorpus)

- Name des Wettbewerbs
- Jahrgang (Einsendeschluss)
- Titel des Textes
- Altersklasse des Autors
- Geschlecht des Autors

Internettexte (am Beispiel Zeitungsleserkommentarkorpus)

- Name der Zeitung
- Titel des Artikels
- Teaser des Artikels
- Schlagwörter zum Artikel
- Tag der Artikelveröffentlichung
- Pseudonym des Kommentarautors
- Tag der Kommentarveröffentlichung
- Titel des Kommentars

Die Grundstrukturierung der Datenbasis samt aller Annotationen und Metadaten erfolgt einheitlich gemäß den Vorgaben von TEI P5 (TEI Consortium 2007), das als auf das Kodieren von Textkorpora ausgerichtetes und auf XML aufbauendes Datenformat einen langjährig etablierten Standard zur Strukturierung linguistischer Daten darstellt. Zur Auszeichnung der Wortartinformationen (POS) wurde das Stuttgart-Tübingen-Tagset STTS (Schiller et al. 1999) herangezogen, bzw. im Falle der nicht-professionellen Textsubstanzen um Elemente aus STTS 2.0 (Bartz et al. 2014), einer abwärtskompatiblen Weiterentwicklung, die speziell auf die Anwendung auf Ressourcen aus internetbasierter Kommunikation optimiert wurde, ergänzt. Die TEI-kodierten Daten werden daraufhin in die interne KorAP-Repräsentation überführt und indiziert. Für vorhandene Metadaten werden optimierte Indizierungsstrategien gewählt, um beispielsweise eine Kriterienwahl über reguläre Ausdrücke oder Zahlenbereiche zu ermöglichen.

Leider dürfen die von den Projektpartnern separat aufgebauten bzw. dort bereits vorliegenden Korpora aus datenschutz- und urheberrechtlichen Gründen jedoch nicht als solche an die jeweils anderen Partner weitergegeben, damit also auch nicht an einem Ort zentral zusammengetragen werden. Dieser Umstand verteilt liegender Ressourcen erfordert die Schaffung einer Möglichkeit, zentrale Anfragen parallel an die einzelnen real existierenden Korpora zu stellen und in einem zweiten Schritt die Resultate der jeweiligen Standorte konzertiert zusammenzuführen.

Dafür wurde die KorAP-Architektur um das Konzept entfernter, selbstverwalteter Knoten erweitert. Hierbei sind Korpuseigner für die technische Bereitstellung von Daten selbst verantwortlich. Auf diese Weise behalten sie die uneingeschränkte Kontrolle über den Zugriff auf ihre Daten, während gleichzeitig der zentrale Abruf über eine Web-Schnittstelle erhalten bleibt. Die Lokalität der Daten für die Suche und die Erstellung virtueller Korpora ist dabei ohne Bedeutung. Für die Aggregation der Suchresultate müssen bereitgestellte Daten lediglich zuvor mit ihren Metadaten an der zentralen Schnittstelle registriert werden. Dieses Vorgehen ist effizient, zuverlässig und rechtskonform durchführbar.

Bibliographie

Ba#ski, Piotr / Bingel, Joachim / Diewald, Nils / Frick, Elena / Hanl, Michael / Kupietz, Marc / P#zik, Piotr / Schnober, Carsten / Witt, Andreas (2013): "KorAP: the new corpus analysis platform at IDS Mannheim." Präsentiert auf der *6th Conference on Language and Technology (LTC-2013), Poznan, Polen, Dezember 2013*.

Bartz, Thomas / Beißwenger, Michael / Storrer, Angelika (2014): "Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge", in: *Zeitschrift für germanistische Linguistik* 28, 1: 157-198.

Bodmer, Franck (1996): "Aspekte der Abfragekompononente von COSMAS-II", in: *LDV-INFO* 8. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung 112-122.

Fischer, Peter M. (i.E.): *Eine Datenbasis zur Beobachtung des Schreibgebrauchs im Deutschen*

Gray, Jim (2003): *Distributed Computing Economics*. Technical Report MSR-TR-2003-24. San Francisco: Microsoft Research.

Horbach, Andrea / Thater, Stefan / Steffen, Diana / Fischer, Peter M. / Witt, Andreas / Pinkal, Manfred (2015): "Internet Corpora: A Challenge for Linguistic Processing", in: *Datenbank-Spektrum*15, 1: 41-47 http://link.springer.com/article/10.1007/s13222-014-0172-z [letzter Zugriff 26. Februar 2016].

Krome, Sabine (2010): "Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus Digital", in: Kratochvílová, Iva / Wolf, Norbert Richard (eds.): *Kompendium Korpuslinguistik*. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Heidelberg: Universitätsverlag Winter 117-134.

Krome, Sabine (2013): "Digitale Datenflut: Chancen und Tücken eines Textkorpus zur deutschen Gegenwartssprache. Anforderungsprofil, Methoden und Instrumentarien zur Beobachtung des aktuellen Sprach- und Schreibgebrauchs", in: Kratochvílová, Iva / Wolf, Norbert Richard (eds.): *Grundlagen einer sprachwissenschaftlichen Quellenkunde*. Tübingen: Narr Verlag 49-66.

Kupietz, Marc / Lüngen, Harald (2014): "Recent Developments in DeReKo", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* 2378-2385.

Kupietz, Marc / Lüngen, Harald / Ba#ski, Piotr / Belica, Cyril (2014): "Maximizing the Potential of Very Large Corpora", in: *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2)* 1-6.

Münzberg, Franziska (2011): "Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht", in: Konopka, Marek / Kubczak, Jacqueline / Mair, Christian / Šticha, František / Waßner, Ulrich H. (eds.): *Grammatik und Korpora 2009*. Tübingen: Narr Francke Attempto 181–197.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Technical report. Tübingen / Stuttgart: Universität Stuttgart / Universität Tübingen http://www.ims.unistuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf [letzter Zugriff 26. Februar 2016].

Scholze-Stubenrecht, Werner (2013): "The World Wide Web as a resource for lexicography", in: Gouws, Rufus H. / Heid, Ulrich / Schweickard, Wolfgang /Wiegand, Herbert Ernst (Hrsg.): *Dictionaries*. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography (= HSK 5.4) 1365-1374. Berlin / New York: Mouton de Gruyter.

TEI Consortium (2007): *Guidelines for Electronic Text Encoding and Interchange (TEI P5)*. The TEI Consortium http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 26. Februar 2016].

van Uytvanck, Dieter (2010): *CLARIN Short Guide on Virtual Collections*. Technical report. CLARIN http://www.clarin.eu/files/virtual_collections-CLARIN-ShortGuide.pdf [letzter Zugriff 26. Februar 2016].