AUFBAU EINES KORPUS ZUR BEOBACHTUNG DES SCHREIBGEBRAUCHS IM DEUTSCHEN

PETER M. FISCHER
INSTITUT FÜR DEUTSCHE SPRACHE
PETER.FISCHER@IDS-MANNHEIM.DE

JENS KIRSTEN
THÜRINGER BUCHLÖWE – SCHREIBWETTBEWERB DER
LITERARISCHEN GESELLSCHAFT THÜRINGEN
JENSW.KIRSTEN@GMX.DE

ANDREAS WITT
INSTITUT FÜR DEUTSCHE SPRACHE
WITT@IDS-MANNHEIM.DE

Abstract

Mitte 2013 startete das BMBF-geförderte Forschungsprojekt „Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen“. Das Gemeinschaftsprojekt zwischen dem Institut für Deutsche Sprache, dem Seminar für Computerlinguistik an der Universität des Saarlandes, der Duden Sprachtechnologie am Bibliographischen Institut und der Redaktion Brockhaus-Wahrig bei wissenmedia in der inmediaONE] GmbH hat zum Ziel, korpusbasierte Instrumentarien für eine systematische Beobachtung des Schreibgebrauchs im deutschsprachigen Raum zu entwickeln. Die Beobachtung erstreckt sich dabei nicht nur auf den Schreibusus der sog. *professionellen Schreiber* in gedruckten oder elektronischen Medien sondern auch auf den von Schülern und privaten Internetnutzern. Während die beteiligten Projektpartner bereits über umfassende, aktuelle Textkorpora verfügen, wie beispielsweise das Deutsche Refernzkorpus DEREKO (IDS, 2010), das WAHRIG Textkorpus^{digital} (Krome, 2010) oder das Dudenkorpus

(Münzberg, 2011), welche allesamt Zeitungs- und Zeitschriftentexte sowie auch fiktionale und wissenschaftliche Texte aus dem gesamten deutschsprachigen Raum beinhalten, konzentriert sich der weitere Korpusaufbau auf die Miteinbeziehung von Schüleraufsätzen und Internetbeiträgen, die bislang noch gar nicht abgedeckt sind. Für den Aufbau des Gesamtkorpus ergeben sich damit drei große Bereiche bzw. (virtuelle) Subkorpora, die ihre jeweiligen Herausforderungen mit sich bringen und nun näher betrachtet werden sollen.

Professionelle Schreiber

In diesem Subkorpus sollen die bereits vorhandenen Ressourcen, unter ihnen die o.g. drei größten Korpora der deutschen Gegenwartssprache im deutschen Sprachraum, vernetzt und, wo erforderlich, weiter aufbereitet werden. Wir werden zeigen, wie wir zu diesem Zweck sämtliche Primärdaten in ein einheitliches, für die Vergleichsforschung geeignetes Datenformat überführen und durch Annotation mehrerer linguistischer Ebenen (wie POS, Lemmatisierung, partielle Konstituentenstrukturen, orthografische Differenzierung) sowie durch metalinguistische Informationen auf eine differenzierte linguistische Auswertung vorbereiten müssen.

Internetnutzer

Zwar verfügen die Partner auch über Korpora aus internetbasierten Textquellen (z.B. Wikipedia), jedoch liegt der Fokus für diesen Bereich auf weniger kontrollierten Textsorten wie e-Mails, Weblogs (inklusive Mikroblogs wie Twitter) sowie Texte aus Diskussionsforen. Hierzu ist zum einen die rechtliche Situation zu klären, d.h., ob und welche Art von Lizenzen für welche Textsorte von wem (z.B. nur Forenbetreiber oder auch Autoren?) zu erwerben sind, ob und inwieweit die Texte (möglicherweise nicht nur ihre Metadaten) zu anonymisieren sind. Zum zweiten müssen die Betreiber von Foren, Blogs, e-Mail-Archiven und Mailinglisten (darunter Privatpersonen, Verlage und Firmen) kontaktiert werden, um Lizenzen je nach den rechtlichen Vorgaben sowie nach Möglichkeit Archivversionen fortlaufend erwerben zu können. Zum dritten ergeben sich für die Basisannotation (Struktur und Metadaten) der Texte besondere Anforderungen, die gängige Korpus-Textmodelle, die

vorwiegend auf Zeitungstexte, Belletristik und Fachtexte abzielen, noch nicht abdecken, und daher umgesetzt werden müssen.

Schüler

Für diesen Bereich werden Schülertexte unterschiedlicher Art akquiriert und in die Korpora integriert. Dafür werden zunächst Kooperationen mit Schulen eingegangen, um Kopien von Schülertexten, Klausurarbeiten und Abituraufsätzen zu bekommen. Automatische Verfahren unterstützen die Digitalisierung dieser Texte. Desweiteren werden Textdaten in digitaler Form von Schülern in Kooperation mit Schulen direkt erhoben. Schließlich werden Kooperationen mit Literaturwettbewerben für Schüler eingegangen, um Texte, die in digitaler Form eingereicht wurden, zu erwerben. Dies werden wir an der Zusammenarbeit mit der Literarischen Gesellschaft Thüringen e.V. exemplifizieren, die jährlich den Wettbewerb „Thüringer Buchlöwe – Schreibwettbewerb der Literarischen Gesellschaft Thüringen“ ausrichtet.

Gesamtkorpus

Die akquirierten und digitalisierten Texte werden konvertiert und nach einem TEI-basierten Textmodell aufbereitet und annotiert. Im Falle der Schülertexte sind rechtliche Fragen bzgl. Möglichkeiten der Herausgabe von z.B. Abituraufsätzen durch Schulen und bzgl. Anonymisierungspflicht im Vorfeld zu klären. Der Gesamtaufbau besteht im Ergebnis also in der Zusammentragung eines außerordentlich großen virtuellen Korpus zur deutschen Gegenwartssprache, das aus zu unterschiedlichen Bedingungen lizenziertem Material besteht. Das virtuelle Korpus wird für ausgesuchte Zwecke, insbesondere die Beobachtung der Schreibverwendung, genutzt werden können; eine Erweiterung der Nutzungsszenarien unter Wahrung der Lizenzanforderungen ist auch möglich.

Referenzen

Belica, C., Kupietz, M., Lüngen, H., Witt, A. (2010): The morphosyntactic annotation of DEREKO: Interpretation, opportunities and pitfalls. In M. Konopka, J. Kubczak, C. Mair, F. Šticha, and U.Wassner, editors, Selected

contributions from the conference Grammar and Corpora 2009, im Druck in Tübingen. Gunter Narr Verlag.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M. (2006): The SALSA Corpus: a German corpus resource for lexical semantics. LREC 2006

IDS (2010): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I. Institut für Deutsche Sprache. Mannheim. http://www.ids-mannheim.de/kl/projekte/archiv.html

Krome, S. (2010): Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus^{digital}. In: I. Kratochvílová, N. R. Wolf (Hgg.): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Universitätsverlag Winter, Heidelberg 2010, S. 117-134.

Kupietz, M., Witt, A., Belica, C., Keibel, H. (2010): The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In: LREC 2010 Main Conference Proceedings. Malta.

Lemnitzer, L., Geyken, A., Beißwenger, M., Storrer, A. (2011): CMC as a component of a balanced, TEI-encoded corpus representing contemporary German: goals, motivation, design issues. Abstract eines Papers zur Präsentation auf dem TEI Members' Meeting, Würzburg, Oktober 2011.

Münzberg, Franziska (2011): Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht. In: Marek Konopka et al. (Hg.): Grammatik und Korpora 2009. Tübingen: Narr Francke Attempto 2011, 181–197.

The TEI Consortium (2007): Guidelines for Electronic Text Encoding and Interchange (TEI P5). The TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Walter, S., Pinkal, M. (2005): Computational Linguistic Support for Legal Ontology Construction. In Proceedings of ICAIL 2005

Witt, A., Kupietz, M., Keibel, H. (2009): DEREKO goes P5: Customizing TEI P5 for the Mannheim German Reference Corpus, Micropaper IN: Text encoding in the era of mass digitization, Online-Konferenz-Proceedings des TEI Members Meetings 2009, http://www.lib.umich.edu/spo/teimeeting09/