Das Dortmunder Chat-Korpus in CLARIN-D: Modellierung und Mehrwerte

,
michael.beisswenger@tu-dortmund.de
TU Dortmund, Deutschland

,
herold@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

,
luengen@ids-mannheim.de
Institut für deutsche Sprache, Mannheim, Deutschland

,
astorrer@mail.uni-mannheim.de
Universität Mannheim, Deutschland

Einleitung und Projekthintergrund

Die Kommunikation im Internet bzw. mit sozialen Medien hat in den vergangenen zwei Jahrzehnten in den geisteswissenschaftlichen Disziplinen eine zunehmende Aufmerksamkeit erfahren. Zahlreiche sprach-, sozial- und medienwissenschaftliche Analysen haben die sprachlichen und interaktionalen Besonderheiten bei der Kommunikation in Chats, Foren, Weblogs und sozialen Netzwerken, per SMS und WhatsApp als einen neuen Gegenstand geisteswissenschaftlicher Forschung erschlossen. Durch ihre digitale Verfügbarkeit sind Sprachdaten aus solchen Genres – im Gegensatz etwa zu Aufzeichnungen von Gesprächen – einfach zu gewinnen und für Forschungszwecke speicherbar. Trotzdem gibt es bislang wenige Korpora zur Sprachverwendung in sozialen Medien, die für Analysezwecke im Bereich der Digital Humanities aufbereitet sind und die der Scientific Community zur Nutzung zur Verfügung stehen. Das hat zum einen mit unklaren rechtlichen Rahmenbedingungen in Bezug auf die Nutzung und Bereitstellung digitaler Kommunikationsdaten für Forschungszwecke zu tun, zum anderen mit dem Fehlen geeigneter Standards für die Strukturbeschreibung und linguistische Annotation von Social-Media-Genres sowie der Notwendigkeit, automatische Annotationswerkszeuge für Daten dieses Typs anzupassen.

In unserem Beitrag präsentieren wir Ergebnisse aus dem Projekt „ChatCorpus2CLARIN“, das als Kurationsprojekt der fachspezifischen Arbeitsgruppe F-AG 1 „Deutsche Philologie“. von Mai 2015 bis Februar 2016 vom BMBF gefördert wird. Ziel des Projekts ist es, das *Dortmunder Chat-Korpus*, ein existierendes Korpus zur Sprachverwendung und Sprachvariation in der deutschsprachigen Chat-Kommunikation, in die Korpus-Infrastrukturen der CLARIN-D-Zentren an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und am Institut für Deutsche Sprache (IDS) Mannheim zu integrieren. Dabei geht es insbesondere um die Herstellung einer Interoperabilität der Zielressource mit Korpora zur gesprochenen und geschriebenen Sprache (DWDS-Korpora, DeReKo, FOLK), die an der BBAW und am IDS bereits vorhanden sind. Die Bereitstellung des Chat-Korpus in CLARIN-D soll einen systematischen, korpusgestützten Vergleich der Sprachverwendung in Chats mit der Sprachverwendung in mündlichen Gesprächen und in redigierten Texten erlauben und der empirischen, sprachdatengestützten Forschung zur Sprache und Interaktion in sozialen Medien somit neue Möglichkeiten eröffnen.

Um Interoperabilität mit existierenden CLARIN-D-Ressourcen herzustellen und es Forscher_innen zu ermöglichen, die unterschiedlichen Ressourcen im Forschungsprozess vernetzt zu nutzen, wird das Chat-Korpus bei der Integration unter Rückgriff auf Standards im Bereich der Digital Humanities remodelliert und um zusätzliche linguistische Annotationen erweitert. Der Beitrag beschreibt die Modellierung der Ressource und ihre Integration in CLARIN-D und zeigt, welche Mehrwerte sich für Nutzer des Korpus durch die Integration und die zusätzlichen Annotationen ergeben.

Die Ausgangsressource

Das *Dortmunder Chat-Korpus* (Beißwenger 2013) ist eine Sammlung von Chat-Mitschnitten aus vier verschiedenen Handlungsbereichen (Freizeit, Bildung, Beratung, Medien), die ca. 140.000 Chatter-Beiträge und 1,06 Mio. Token umfasst und die 2002–2008 am Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik der TU Dortmund aufgebaut wurde. Die Daten sind in einem XML-Format repräsentiert, das zentrale Strukturelemente von protokollieren Chatverläufen (sog. ‚Logfiles‘) abbildet, unterschiedliche Typen von Chat-Beiträgen unterscheidet und ausgewählte Stilelemente internetbasierter Kommunikation erfasst. Teile des Korpus werden seit 2005 über die Website http://www.chatkorpus.tu-dortmund.de zusammen mit einem einfachen, Java-basierten Abfragewerkzeug zur Verfügung gestellt. Das Korpus wird in diversen linguistischen und computerlinguistischen Projekten sowie im Bildungskontext (Schule und Hochschule) als Ressource in Forschung und Lehre genutzt.

Interoperabilität durch Anschluss an Standards im Bereich der Digital Humanities

Strukturmodellierung und Repräsentation in TEI

Für die Repräsentation der im Korpus dokumentierten Chat-Verläufe greifen wir auf die Formate der *Text Encoding Initiative* (*TEI)* zurück. In den TEI-Guidelines (TEI-P5) gibt es bislang keine Modelle für die Darstellung von Social-Media-Genres, dafür umfangreiche Module für die Strukturrepräsentation von Textgenres und von transkribierten Gesprächen. Die in den Guidelines vorgesehene Möglichkeit der *customization* macht das Encoding-Framework aber flexibel genug, um es an die Erfordernisse auch von (neuen) Genres anzupassen.

Seit 2013 beschäftigt sich in der TEI eine Special Interest Group (SIG) „Computer-mediated communication“ mit der Entwicklung eines Standards für die Modellierung von Social-Media-Genres (Beißwenger et al. 2012; Chanier et al. 2014; Margareta / Lüngen 2014). Das Projekt greift den aktuellen Stand der in der SIG diskutierten Schemaentwürfe auf, testet diese an den Daten des Chat-Korpus sowie an Ausschnitten ausgewählter weiterer Social-Media-Genres (Wikipedia-Diskussionsseiten, WhatsApp-Dialoge, News-Diskussionen, Tweets) und entwickelt sie weiter. Das dabei entstehende TEI-Schema wird in Form eines ODDdokumentiert und bildet die Grundlage für die TEI-Modellierung des kompletten Korpus. Zugleich wird das ODD, dessen Fertigstellung für Herbst 2015 vorgesehen ist, in die weitere Arbeit der SIG eingespielt.

Linguistische Basisannotation mit „STTS 2.0“

Um die Recherchemöglichkeiten im Korpus zu verbessern, wird der Ausgangsressource eine zusätzliche Annotationsebene hinzugefügt, deren Kern Part-of-speech-Informationen (PoS) bilden. Das im Projekt verwendete PoS-Tagset („STTS 2.0“, Beißwenger et al. 2015) verwendet die Kategorien des *Stuttgart-Tübingen Tagset* (STTS, Schiller et al. 1999) und erweitert diese einerseits um Tags für typische Einheiten bei der schriftlichen Sprachverwendung in Social-Media-Genres (u. a. Emoticons, Hashtags, Adressierungen) sowie um Einheiten für die Darstellung von Phänomenen, die typisch sind für Kontexte informeller, dialogischer Kommunikation (u. a. Abtönungs- und Intensitätspartikeln, Diskursmarker). Die Erweiterungen sind abgestimmt auf Erweiterungen, die am IDS für die PoS-Annotation des FOLK-Korpus zur gesprochenen Sprache zum Einsatz kommen.

Um die Annotationen nach STTS 2.0 zu erzeugen, wurde das komplette Chat-Korpus 2015 mit einem POS-Tagger annotiert, für den im BMBF-Projekt “Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen” (IDS 2014-2016) neue Taggermodelle speziell für den Umgang mit Social-Media-Genres entwickelt wurden (Horbach et al. 2014). Um das Ergebnis der automatischen Annotation manuell nachzukorrigieren und zusätzlich einzelnen Tokens normalisierte Formen zuzuordnen, wurde das Werkzeug *OrthoNormal* (Schmidt 2012) auf die Bearbeitung von Chat-Daten angepasst.

Zielressource und Mehrwerte

Die Integration in die Infrastrukturen der beteiligten CLARIN-D-Zentren umfasst die Archivierung in den Repositorien an der BBAW und am IDS, die Aufnahme der Metadaten in das Virtual Language Observatory (VLO), die Einbindung der Daten in die korpusübergreifende Suchmaschine *CLARIN Federated Content Search* sowie die Bereitstellung über Webservices.

Die rechtlichen Bedingungen der Bereitstellung werden über ein Rechtsgutachten geklärt. Je nach Ergebnis kommen für die Ressource unterschiedliche Lizenzmodelle in Frage: Als Idealfall wird eine CLARIN-Endnutzer-Lizenz vom Typ PUB („publicly available“, Oksanen et al. 2010) angestrebt, gegebenenfalls aber auch der Lizenztyp ACA-NC (akademische, nicht-kommerzielle Nutzung zum vollständigen Kopieren / Download freigegebener Ressourcen) oder, falls erforderlich, eine Beschränkung auf eine Nutzung über eine Korpusrecherchesoftware durch bei CLARIN registrierte Nutzer (Lizenztyp QAO-NC, gemäß Vorschlag in Kupietz / Lüngen 2014).

Nach der Integration wird die Zielressource für Nutzer im Bereich der Digital Humanities gegenüber der Ausgangsressource die folgenden Mehrwerte aufweisen:

- **Erweiterung der Möglichkeiten des Zugriffs und der Durchsuchbarkeit** der Ressource.
- **Interoperabilität auf der Ebene der Dokumentstruktur (TEI):** Durch die Remodellierung in einem TEI-Format wird die Ressource interoperabel mit anderen in TEI repräsentierten Sprachressourcen und Annotations- bzw. Analysewerkzeugen.
- **Linguistische Annotation:** Die Anreicherung um zusätzliche linguistische Basisannotationen wird die Möglichkeiten zur Nutzung der Ressource für die korpusgestützte Sprachanalyse erweitern und anspruchsvollere linguistische Suchanfragen ermöglichen.
- **Interoperabilität auf der Ebene der linguistischen Annotation (STTS):** Durch die Kompatibilität der Part-of-speech-Annotationen mit STTS wird die Ressource interoperabel mit anderen nach STTS annotieren Sprachressourcen.
- **Vernetzung mit Korpusressourcen anderen Typs:** Durch die Integration in CLARIN-D und die genannten Interoperabilitätsmerkmale werden die Möglichkeiten zu einem korpusgestützten Vergleich sprachlicher Besonderheiten im Chat-Korpus mit

Korpora gesprochener Sprache und Korpora redigierter Schriftlichkeit verbessert.
- **Verbesserte Auffindbarkeit der Ressource** durch die Bereitstellung standardisierter Metadaten und die Aufnahme in das VLO.

Die Ergebnisse aus dem Projekt können zum gegenwärtigen Zeitpunkt z. T. nur perspektivisch formuliert werden. Zum Termin der Konferenz werden die Projektarbeiten abgeschlossen sein und die Ergebnisse vorliegen.

Fußnoten

1. Für weitere Informationen siehe http://www.clarin-d.de/de/wissenschaftsbereiche/germanistik
2. Sie hierzu die Webseite der TEI unter http://www.tei-c.org/Activities/SIG/CMC/ .
3. Siehe http://www.tei-c.org/Guidelines/Customization/odds.xml .

Bibliographie

Beißwenger, Michael (2013): "Das Dortmunder Chat-Korpus", in: *Zeitschrift für germanistische Linguistik* 41, 1: 161-164. Erweiterte Fassung online: http://tinyurl.com/chatkorpus [letzter Zugriff 18. September 2015]. **Beißwenger, Michael / Ermakova, Maria / Geyken, Alexander / Lemnitzer, Lothar / Storrer, Angelika** (2012): "A TEI Schema for the Representation of Computer-mediated Communication", in: *Journal of the Text Encoding Initiative (jTEI)* 3. http://jtei.revues.org/476 [letzter Zugriff 18. September 2015].

Beißwenger, Michael / Bartz, Thomas / Storrer, Angelika / Westpfahl, Swantje (2015): *Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation*. https://sites.google.com/site/empirist2015/home/annotation-guidelines [letzter Zugriff 18. September 2015].

Chanier, Thierry / Poudat, Celine / Sagot, Benoit / Antoniadis, Georges / Wigham, Ciara / Hriba, Linda / Longhi, Julien / Seddah, Djamé (2014): "The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres", in: *Journal of Language Technology and Computational Linguistics* 2: 1-30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf [letzter Zugriff 18. September 2015].

Horbach, Andrea / Steffen, Diana / Thater, Stefan / Pinkal, Manfred (2014): "Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication", in: *Proceedings of KONVENS 2014* 171-177.

IDS = Institut für Deutsche Sprache (2014-2016): *Projekt Schreibgebrauch*. Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen http://www.schreibgebrauch.de/index.html .

Kupietz, Marc / Lüngen, Harald (2014): "Recent developments in DeReKo", in: Calzolari, Nicoletta / Choukri, Khalid / Declerck, Thierry / Loftsson, Hrafn / Maegaard, Bente / Mariani, Joseph / Odijk, Jan / Piperidis, Stelios (eds): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Margaretha, Eliza / Lüngen, Harald (2014): "Building Linguistic Corpora from Wikipedia Articles and Discussions", in: *Journal of Language Technology and Computational Linguistics* 2: 59-82. http://www. jlcl.org/2014_Heft2/3MargarethaLuengen.pdf [letzter Zugriff 18. September 2015].

Oksanen, Ville / Lindén, Krister / Westerlund, Hanna (2010): "Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN", in: *Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, Malta.

Schmidt, Thomas (2012): "EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language", in: *Proceedings of LREC2012* http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf [letzter Zugriff 18. September 2015].

Schiller, Anne / Teufel, Simone / Stöckert, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart: Institut für maschinelle Sprachverarbeitung.

TEI Consortium (eds.) (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange* http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 18. September 2015].