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

technische universität dortmund · UNIVERSITÄT MANNHEIM · CLARIN-D · INSTITUT FÜR DEUTSCHE SPRACHE · berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Michael Beißwenger · Axel Herold · Harald Lüngen · Angelika Storrer

ChatCorpus2CLARIN: Kurationsprojekt der CLARIN-D-F-AG1 „Deutsche Philologie“

- **Integration einer existierenden Korpusressource zur internetbasierten Kommunikation** in die Ressourcen-Infrastrukturen der CLARIN-D-Zentren an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und am Institut für Deutsche Sprache (IDS), Mannheim (**Dortmunder Chat-Korpus**, http://www.chatkorpus.tu-dortmund.de)
- **Bereitstellung über die Korpusrecherchesysteme** am IDS und an der BBAW bis Herbst 2016
- **Remodellierung der Ressource auf Grundlage von sprach- und texttechnologischen Standards im Bereich der Digital Humanities**
- **Erweiterung um zusätzliche Metadaten und linguistische Annotationen** (Tokens, Parts of Speech, Normalisierung)
- **Bereitstellung unter klaren rechtlichen Bedingungen** (auf Grundlage eines Rechtsgutachtens)

Modellierungsaufgabe: Entwicklung eines Schemas für die Strukturannotation von Social-Media-Genres:

TEI-Schema für die Repräsentation von Genres im Bereich Computer-Mediated Communication (CMC) / Social Media

Weiterentwicklung von Schemaentwürfen aus der der TEI Special Interest Group „Computer-Mediated Communication“ und Erprobung mit Samples aus verschiedenen CMC-/Social-Media-Korpora: Chat, WhatsApp, Wikipedia-Diskussionen, Usenet-News, Tweets.

***Customizations* in TEI:** “Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to **permit usage in areas that the TEI has not yet envisioned**.”

Zentrale Charakteristika und Modellierungskonzepte:

(1) **Einführung neuer Modelle** für die Repräsentation CMC-spezifischer Äußerungsformate – *u.a.:*

- <post> (Beißwenger et al. 2012) für die Repräsentation von User-Postings, deren Eigenschaften weder mit oder <p>, noch mit <u> aus TEI-P5 angemessen erfasst werden;
- <prod> (Chanier et al. 2014) für die Repräsentation nicht-verbaler Aktivitäten in multimodalen CMC-Environments.

(2) Modellierung von <post>, <prod> und <u> als model.divPart-Elemente, die in **multimodalen CMC-Interaktionen** (z.B. in 3D-Welten, in Skype, in Lernumgebungen) frei kombiniert werden können.

(3) **Flexibilisierung existierender Modelle** (z.B. <signed>, <quote> und <p>), die in CMC-Genres variabler verwendet werden als in traditionellen Textgenres.

(4) **Formulierung von Best Practices für die Verwendung von Standardmodellen** – u.a. Verwendung von <w> und <phr> zur Integration von Part-of-speech-Informationen auf Ebene des user generated content in <post>-Elementen.

Schema (ODD- und RNG-Datei) verfügbar im TEI-Wiki:
http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

⇨ Input für die weitere Standardisierungsdiskussion (TEI-SIG); Ressource für die Annotation weiterer CMC-/Social-Media-Korpora

NLP4CMC: Nutzung sprachtechnologischer Verfahren für die linguistische Annotation:

„STTS 2.0“: Erweitertes Stuttgart-Tübingen-Tagset mit Part-of-speech-Kategorien für CMC-Phänomene auf Token-Ebene

abwärtskompatibel mit STTS (1999) und abgestimmt auf das erweiterte STTS-Tagset für Korpora gesprochener Sprache (FOLK-Projekt/IDS):

Tag	Kategorie	Beispiele
I. Tags für CMC-spezifische Phänomene:		
EMO ASC	Emoticon, als Zeichenfolge dargestellt (Typ „ASCII“)	:-) :-(^^ O.O
EMO IMG	Emoticon, als Grafik-Ikon dargestellt (Typ „Image“)	😊 🦊
AKW	Aktionswort	*lach*, freu, grübel, *lol*
HST	Hashtag	Kreta war super! #urlaub
ADR	Adressierung	@lothar: Wie isset so?
URL	Uniform Resource Locator	http://www.tu-dortmund.de
EML	E-Mail-Adresse	peterklein@web.de
II. Tags für Phänomene der konzeptionellen Mündlichkeit:		
VV PPER	Tags für die häufigsten Bildungsmuster kontraktierter Formen (APPRART ist in STTS bereits vorhanden)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone
PTK IFG	Intensitäts-, Fokus- oder Gradpartikel	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal- oder Abtönungspartikel	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Partikel als Teil eines Mehrwort-Lexems	keine mehr, noch mal, schon wieder
DM	Diskursmarker	*prototypisch:* weil, obwohl, nur, also *als Einheiten mit projektivem Potenzial im Vorvorfeld von V2-Sätzen*
ONO	Onomatopoetikon	boing, miau, zisch

Tagset und Annotationsrichtlinien der GSCL-Shared-Task EmpiriST:
https://sites.google.com/site/empirist2015/

Workflow:

(1) **Automatische Segmentierung und PoS-Annotation** mit den für CMC-Genres angepassten Taggermodellen aus dem BMBF-Projekt www.schreibgebrauch.de (Horbach et al. 2014) (Kooperation mit Computerlinguistik, U Saarbrücken)

(2) **Anpassung des Annotationswerkzeugs OrthoNormal** für die manuelle Annotation XML-strukturierter Chat-Daten (Kooperation mit Thomas Schmidt/IDS, http://agd.ids-mannheim.de/folker.shtml)

(3) **Manuelle Nachbearbeitung** der Annotation für einen Korpusausschnitt und Verfeinerung der Annotationsrichtlinien (Goldstandard für die Annotation weiterer Korpusteile)

Mehrwerte des integrierten Korpus:

- **Erweiterung der Möglichkeiten für die Recherche und Analyse**
- **Interoperabilität** mit anderen in TEI repräsentierten und in STTS annotierten Sprachressourcen und darauf bezogenen Annotations- und Analysewerkzeugen.
- **Linguistische Recherche und Analyse:** Die Anreicherung um zusätzliche linguistische Basisannotationen erweitert die Möglichkeiten zur Nutzung der Ressource für die korpusgestützte Sprachanalyse und ermöglicht anspruchsvollere linguistische Suchanfragen.
- **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.
- **Verbesserung der Einsatzmöglichkeiten in der Lehre** (Germanistik, Computerlinguistik, Sprach- und Texttechnologie).

http://de.clarin.eu/de/kurationsprojekt-1-3-germanistik

Kontakt: michael.beisswenger@uni-due.de

DHd 2016: *Die Digital Humanities als fächerübergreifendes Forschungsparadigma.* Universität Leipzig, 7.–12. März 2016.