Aufbau von Social-Media-Korpora für die Digital Humanities: Standards und Perspektiven

technische universität dortmund
UNIVERSITÄT MANNHEIM
Michael Beißwenger · Thomas Bartz · Axel Herold · Marc Kupietz
Lothar Lemnitzer · Harald Lüngen · Angelika Storrer
INSTITUT FÜR DEUTSCHE SPRACHE
berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Warum brauchen wir Standards für den Aufbau von Social-Media-Korpora?

- **Erleichterung des Aufbaus** von Social-Media-Korpora (Verfügbarkeit von Annotationsschemata, Best Practices und Tools)
- **Nachnutzbarkeit** und **Kompatibilität mit gängigen Arbeitsumgebungen und Korpusanalysesystemen**
- **Interoperabilität** von Social-Media-Korpora (zu unterschiedlichen Genres, zu unterschiedlichen Sprachen)
- **Interoperabilität** von Social-Media-Korpora ***mit Korpora anderen Typs*** (Textkorpora, Gesprächskorpora)

⇨ **Erweiterung der Möglichkeiten für die empirische, korpusgestützte Sprachanalyse**

Desiderat: Entwicklung eines Standards für die Strukturannotation von Social-Media-Genres:

Lösungsvorschlag: TEI-Schema für Korpora im Bereich ‚Computer-Mediated Communication' (CMC)

⇨ Nutzung u.a. für die Annotation der Wikipedia- und News-Korpora in DeReKo, für das Dortmunder Chat-Korpus sowie (geplant) für das DWDS-Blog-Korpus und für die Daten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation".

Kooperation mit der TEI Special Interest Group „Computer-Mediated Communication" (http://www.tei-c.org/Activities/SIG/CMC/)

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**."*

Charakteristika des aktuellen Schema-Entwurfs:

(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.

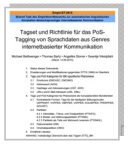
(4) **Formulierung von Best Practices für die Verwendung von Standardmodellen** – u.a. Verwendung von <w> und <phr> zur Integration von Part-of-speech-Informationen auf Ebene des user generated content in <post>-Elementen.

Das Schema wurde mit Daten aus verschiedenen CMC-Genres und -Korpora getestet (Chat, Tweets, WhatsApp, Wikipedia-Diskussionen u.a.).

Desiderat: Entwicklung von NLP-Verfahren für die linguistische Annotation von Social-Media-Genres:

Lösungsvorschlag: „STTS 2.0": Erweitertes Part-of-Speech-Tagset mit Kategorien für CMC-Phänomene auf Token-Ebene

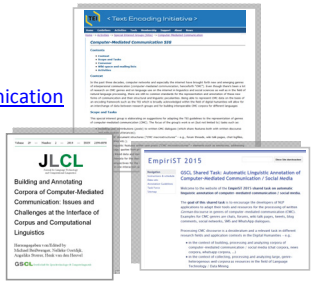
⇨ Nutzung u.a. für die Annotation des Dortmunder Chat-Korpus sowie für die Annotation der Trainings- und Evaluationsdatensets zur GSCL-Shared Task zur automatischen linguistischen Annotation deutschsprachiger CMC-/Social-Media-Daten.



Tag	Kategorie	Beispiele
I. Tags für IBK-spezifische Phänomene:		
EMO ASC	Emoticon, als Zeichenfolge dargestellt (Typ „ASCII")	:-) :-(^^ O.O
EMO IMG	Emoticon, als Grafik-Ikon dargestellt (Typ „Image")	😊 😹, *kodiert als:* emojiQsmilingFaceWithSmilingEyes emojiQkissingCatFaceWithClosedEyes
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

Initiativen und Ressourcen mit KobRA-Beteiligung:

- TEI Special Interest Group „Computer-Mediated Communication": http://www.tei-c.org/Activities/SIG/CMC/
- TEI-Schemaentwurf (ODD, Version 10/2015): http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication
- GSCL-AK „Social Media / Internetbasierte Kommunikation": http://gscl.org/ak-ibk.html
- GSCL-Shared-Task zur automatischen linguistischen Annotation deutschsprachiger CMC-/Social-Media-Daten: https://sites.google.com/site/empirist2015/
- Tagset und Guidelines für die Tokenisierung und für das Part-of-Speech-Tagging deutscher CMC-/Social-Media-Daten: https://sites.google.com/site/empirist2015/home/annotation-guidelines
- JLCL Special Issue „Building and Annotating Corpora of Computer-Mediated Communication": http://jlcl.org/



http://www.kobra.tu-dortmund.de