Deutsches Referenzkorpus zur internetbasierten Kommunikation:

Fragen der Standardisierung und Datenerhebung

2. DTA-/CLARIN-D-Konferenz und CLARIN-D-Workshop

TEXTKORPORA IN INFRASTRUKTUREN FÜR DIE GEISTES- UND SOZIALWISSENSCHAFTEN

Michael Beißwenger (Dortmund)
Lothar Lemnitzer (Berlin)

technische universität dortmund

DEUTSCHES REFERENZKORPUS DeRiK INTERNETBASIERTE KOMMUNIKATION
DWDS
berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
technische universität dortmund

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Internetbasierte Kommunikation (IBK)

engl. ***Computer-Mediated Communication (CMC)***

Im Fokus der Erforschung internetbasierter Kommunikation stehen nicht beliebige Webtexte, sondern die sprachlichen Äußerungen in dialogischen Webgenres wie z.B. Online-Foren, Chats, Instant Messaging und Wiki-Diskussionen, in Twitter-Postings, in Kommentaren und Diskussionen in Weblogs, auf Videoplattformen (*YouTube*) und auf den Profilseiten „sozialer Netzwerke“ (*Facebook*, *MySpace*) sowie in multimodalen Kommunikationsumgebungen wie *Skype*, MMORPGs und „virtuellen Welten“ (*SecondLife* u.a.).

Wofür wird ein Referenzkorpus „IBK“ benötigt?

Aus der Perspektive der Linguistik (Medienlinguistik etc.):

Für Forschungszwecke allgemein und frei zugängliche Korpora bieten mehrere Vorteile:

(1) Der/die Forscher/in spart Zeit und muss Datensets/Korpora nicht selbst erheben und für Analysezwecke aufbereiten, bevor er/sie sich den eigentlich interessanten Fragen zuwenden kann;

(2) Untersuchungen zu bestimmten Phänomenen sind replizierbar und vergleichbar

Im Bereich der empirischen Erforschung und lexikographischen Beschreibung der deutschen Gegenwartssprache:

- Existierende Korpussammlungen zur deutschen Gegenwartssprache decken IBK nicht ab. Ein wichtiger Kommunikationsbereich ist somit in den verfügbaren Korpora nicht erfasst.

  ⇨ Kann man die deutsche Gegenwartssprache zum Stand 2014 beschreiben, wenn man den Kommunikationsbereich „IBK“ (aufgrund fehlender Abdeckung in Korpora) einfach ausspart?

Warum sind bisher so wenige IBK-Korpora allgemein verfügbar?

… weil die **rechtlichen Rahmenbedingungen** für die Erhebung, Speicherung, Bearbeitung und Bereitstellung von IBK-Daten für Forschungszwecke bislang noch großenteils unklar sind.

... weil IBK eine „moving target“ ist: schnelle Veränderlichkeit von Kommunikationstechnologien und Genres
⇨ hohe Anforderungen an **Metadaten / Dokumentation**

… weil sprachtechnologische Verfahren für die **linguistische Basisannotation** von schriftlichen Korpusdaten (Wortatenannotation, Lemmatisierung, syntaktische Strukturen) mit den Besonderheiten von IBK-Schriftlichkeit bislang nicht sinnvoll umgehen können.

… weil bislang **keine Standards für die Annotation der strukturellen Besonderheiten von IBK-Genres existieren** (Threads in Foren / sozialen Netzwerken, Chat-Logfiles, Twitter-Timelines, Wikipedia-Diskussionen, multimodale Genres usw.) ⇨ Wer ein IBK-Korpus aufbaut, kann i.d.R. keine existierenden Schemata übernehmen, sondern muss selbst Lösungen entwickeln.

⇨ Der Aufbau von IBK-Korpora ist ein eigener methodischer Teilbereich der Korpuslinguistik.

📖 Bartz et al. (2013); Beißwenger et al. (2012); Storrer (2013).

DeRiK: Referenzkorpus zu IBK (Laufendes Projekt)

Projekt: *Deutsches Referenzkorpus zur internetbasierten Kommunikation* (DeRiK)

Ziel: Aufbau einer Zusatzkomponente zu den Korpora des Projekts „Digitales Wörterbuch der deutschen Sprache“ (DWDS), die Sprachdaten aus den wichtigsten Genres internetbasierter Kommunikation umfasst und gemeinsam mit den bereits existierenden Korpusressourcen abgefragt werden kann.

(M. Beißwenger, A. Geyken, L. Lemnitzer, A. Storrer)

technische universität dortmund

UNIVERSITÄT MANNHEIM

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

DWDS

⇨ Schließung der „IBK-Lücke“ in den DWDS-Korpora zur deutschen Gegenwartssprache.

Vgl. Beißwenger et al. (2013), Beißwenger/Lemnitzer (2013).

DeRiK: eine DWDS-Zusatzkomponente zu IBK

dortmund

http://www.dwds.de

Ressourcen ▼ Erschließung ▼ Projekt ▼ Aktuelles ▼

Troll | DWDS Standardsicht | +Ressourcen

DWDS-Wörterbuch

Troll Aussprache: ▶
mask., -s/-es, -e

gespenstisches Wesen des Volksaberglaubens, besonders in der nordischen Mythologie, Unhold

Die Trolle sind die Anwälte der Tiere. Sie suchen den heim, der Tiere quält – Jahnn in Dt. Erzähler 2,89

Dazu:
- Trollblume

Version: 0.4.22 | Quelle: WDG | Artikeltyp: Vollartikel

Etymologisches Wörterbuch (nach Pfeifer)

Troll

Troll m. (in der nordischen Mythologie) 'dämonisches Wesen, Kobold', anord. *troll*, *troll*, schwed. *troll*, dän. *trold*. Die im 18. Jh. ins Dt. entlehnte Bezeichnung trifft auf einheimisches (in den Mundarten bewahrtes) *Troll* 'grober, ungeschlachter Kerl' (15. Jh.), auch 'Dämon, Unhold' (15. bis 17. Jh.). Vielleicht hervorgegangen aus germ. **truzla-* und verwandt mit trollen (s. d.).

Version: 1.0.65

DWDS-Wortprofil 3.0

Abfragewort: Troll | Vergleichswort:

Substantiv | logDice | 36

Überblick zu '*Troll*'

aussehen wie böser Dämonen Elfen erzählte von Feen Fjorde Geister Gnome Hexen Kobolde Land Reich Riesen schwarze tolle versteinerten Welt Zwerge

'*Troll*' hat Attribut

Koordination mit '*Troll*'

'*Troll*' ist Aktivsubjekt von

Version: 3.0 Einstellungen

Kernkorpus 20

Treffer: **133**, davon anzeigbar: 98 | Filter

	KWiC	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1	1999	erliegende , Chemluth und dem **Troll** machte das nichts aus , abe				
2	1999	Dann gab er dem **Troll** die Fackel und verschwand ir				
3	1999	t eine Abkürzung « , sagte der **Troll** , als hätte er meine Gedanke				
4	1999	Drache schlief , trampelte der **Troll** auf ihm herum , als sei er ein				
5	1999	schrie der **Troll** atemlos während unseres nä				
6	1999	anderen nach « , hechelte der **Troll** .				
7	1999	rief der **Troll** .				
8	1999	Der **Troll** stieg an mir hoch .				
9	1999	Chemluth und der **Troll** sahen durch das Loch zu mi				
10	1999	verkehrt gemacht « , sagte der **Troll** .				
11	1999	Chemluth und der **Troll** waren verschwunden , und ic				
12	1999	Der **Troll** zog Chemluth hoch , und be				
13	1999	rief der **Troll** .				
14	1999	» Ja , ich weiß « , sagte der **Troll** kleinlaut .				

DIE ZEIT

Treffer: **648** | Filter

	KWiC	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1	2009	Ob **Trolle** überhaupt vom Influenzaviru				
2	2009	Harry und Weasley hatten den **Troll** zur Strecke gebracht , indem				
3	2009	Weil sie dort Hexen , **Trolle** und Teufel vermuteten .				
4	2009	cht die Figur des Paladins den **Troll** besiegt : » Ich war das « ; ni				
5	2009	lf und Sauron stehen könnten : **Trolle** kämpfen gegen Krieger , Kr				
6	2009	Internet Unter **Trollen** Wo Jungen zu Männern u				
7	2009	Unter **Trollen**				
8	2009	itteln den Kopf , bevor sie sich **trollen** .				
9	2009	auf die Elfen zu achten und die **Trolle** , die seien in diesem Jahr e				
10	2008	rgischen Einläufen (von Lothar **Trolle**) auf eine Länge von fünf St				
11	2008	ndenkbar , da wird es für einen **Troll** ein Leichtes gewesen sein ,				
12	2008	entstanden sein , als ein böser **Troll** die Stadt Farum begraben w				
13	2008	en Universitätscampus voll mit **Trollen** , sprechenden Tieren und				
14	2008	or Heinrich Böll und Thaddäus **Troll** getan hatten , aus Sicht des				

DeRiK

Treffer: ??? | Filter

	KWiC	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						

... under construction ...

DeRiK: eine DWDS-Zusatzkomponente zur IBK

DeRiK ist konzipiert als

- ein **zeitlich gestaffeltes Korpus**, bei dem die Datenerhebung nicht nur einmalig, sondern mehrfach in regelmäßigen Abständen erfolgen soll, wodurch es möglich wird, auch sprachlichen Wandel *innerhalb* der internetbasierten Kommunikation darzustellen;
- ein spezialisiertes **Referenzkorpus** zur internetbasierten Kommunikation, das der Fachcommunity als Basis für korpusgestützte Untersuchungen und für die Vermittlung der sprachlichen Besonderheiten von IBK-Genres in der Lehre zur Verfügung gestellt werden soll;

Was gehört in ein Referenzkorpus?

Zusammensetzung des DWDS-Kernkorpus:

⇨ Ideal: **Ausgewogenheit** des Korpus nach Zeitabschnitten und Textsortenbereichen:

⇨ Für jede Dekade (ab 1900):

Anteil Zeitungstexte:	~27%
Anteil Werke der Literatur:	~28%
Anteil Fachtexte:	~23%
Anteil Gebrauchstexte:	~21%

Vgl. Geyken (2007), Klein/Geyken (2010)

Was gehört in ein Referenzkorpus *zur IBK*?

Offene Fragen:

⇨ Was sind relevante „IBK-Sorten“ (vergleichbar zu Textsortenbereichen)?

⇨ Wie lässt sich ein *ausgewogenes Verhältnis* festlegen?

Idee:

Kopplung des Schlüssels für die Auswahl und Zusammensetzung der Korpusdaten an die jährlich durchgeführte **ARD/ZDF-Onlinestudie**.

Idealer Schlüssel:

- Ermittlung der meistgenutzten Kommunikationsanwendungen im Internet im Untersuchungszeitraum
- Gewichtung der Präferenzen unterschiedlicher Altersgruppen nach dem Grad ihrer ‚Online-Affinität‘
- Neuberechnung des Schlüssels für die Datenaufnahme auf Basis der jeweils aktuellen Ausgabe der Studie

Was gehört in ein Referenzkorpus *zur IBK*?

Offene Fragen:

⇨ Was sind relevante „IBK-Sorten“ (vergleichbar zu Textsortenbereichen)?

⇨ Wie lässt sich ein *ausgewogenes Verhältnis* festlegen?

Pragmatischer Schlüssel:

Angesichts der unklaren Rechtslage zunächst nur Aufnahme von Daten aus solchen Anwendungen, bei denen die Nutzung juristisch unbedenklich ist (= Anwendungen, bei denen alle Inhalte explizit für eine Nutzung, Bearbeitung und Wiederbereitstellung durch Dritte lizenziert sind).

Idealer Schlüssel:

- Ermittlung der meistgenutzten Kommunikationsanwendungen im Internet im Untersuchungszeitraum
- Gewichtung der Präferenzen unterschiedlicher Altersgruppen nach dem Grad ihrer ‚Online-Affinität‘
- Neuberechnung des Schlüssels für die Datenaufnahme auf Basis der jeweils aktuellen Ausgabe der Studie

DeRiK: eine DWDS-Zusatzkomponente zur IBK

DeRiK ist konzipiert als

- ein **zeitlich gestaffeltes Korpus**, bei dem die Datenerhebung nicht nur einmalig, sondern mehrfach in regelmäßigen Abständen erfolgen soll, wodurch es möglich wird, auch sprachlichen Wandel *innerhalb* der internetbasierten Kommunikation darzustellen;
- ein **Referenzkorpus** zur internetbasierten Kommunikation, das der Fachcommunity als Basis für korpusgestützte Untersuchungen und für die Vermittlung der sprachlichen Besonderheiten von IBK-Genres in der Lehre zur Verfügung gestellt wird;
- ein **annotiertes Korpus**, das neben einer linguistischen Basisannotation auch Annotationen zu charakteristischen sprachlichen und strukturellen Besonderheiten bei der Sprachverwendung im Netz umfasst.

Problembereich: Standards für die Annotation von IBK-Korpora

- Für die Annotation von Primärdaten aus Genres internetbasierter Kommunikation gibt es derzeit noch **keine Standards**.
- Linguistische und texttechnologische Modelle, die für die Annotation von Textkorpora etabliert sind (Linguistische Basisannotation, Strukturbeschreibung von Texten), decken die strukturellen und linguistischen Besonderheiten internetbasierter Kommunikation nicht vollständig ab (Beißwenger et al. 2012).
- Sprachtechnologische Verfahren für die automatische linguistische Annotation von Textkorpora führen bei IBK-Daten zu nicht akzeptablen Ergebnissen (Phänomene „nichtstandardisierter“ Schriftlichkeit; vgl. Bartz et al. 2013).

Desiderate:

1) Anpassung texttechnologischer Standards für die Repräsentation von IBK-Genres und -Korpora (Strukturannotation, Metadaten)
   ⇨ *Interoperabilität*
2) Anpassung von Tagsets und Verfahren/Werkzeugen für die linguistische Annotation ⇨ *Suche über linguistischen Annotationen*

Desiderat: Standards für die Repräsentation von IBK-Korpora

http://www.tei-c.org/Activities/SIG/CMC/

Home Guidelines Activities Tools Membership Support About News

Home -> Activities -> Special Interest Groups (SIGs) -> Computer-Mediated Communication

Computer-Mediated Communication SIG

Contents

- Context
- Scope and Tasks
- Convener
- Wiki space and mailing lists
- Activities and cooperations

Context

In the past three decades, computer networks and especially the int
mediated communication, henceforth "CMC"). Even though there's b
sciences as well as in the field of natural language processing, there
communication and their structural and linguistic peculiarities. Being a
broadly acknowledged within the field of digital humanities will allow f
for different languages

Scope and Tasks

This special interest group is elaborating on suggestions for adapting
The focus of the group's work is on (but not limited to) tasks such as

- modelling user contributions (*posts*) to written CMC dialogues (
- modelling CMC document structures (*"CMC macrostructures"* –
- annotating linguistic features within user posts (*"CMC microstr*
etc.);
- representing linked data and media objects connected with/em
- metadata schemata for the description of CMC resources;
- developing perspectives for the representation of discourse in r
variety of modalities from written, spoken and non-verbal mode

Convener

Michael Beißwenger, TU Dortmund University

Wiki space and mailing lists

For exchange on the issues and tasks listed above, the SIG uses the talk pages in the TEI wiki and a mailing list.

Special Interest Group im Rahmen der TEI:

seit 2013: Erarbeitung eines Entwurfs zu einem TEI-Standard für die Annotation von IBK-Genres (unter Berücksichtigung sowohl schriftlicher als auch multimodaler Genres)

... auf der Grundlage eines spezialisierten TEI-Schemas, das 2011-12 im Rahmen des DeRiK-Projekts entwickelt wurde (Beißwenger et al. 2012)

... und das in 2013-14 von SIG-Mitgliedern am IDS und an der Universität Clermont-Ferrand für weitere Typen von IBK-Korpora (Deutsch und Französisch) angepasst wurde (Margaretha/Lüngen (forthc.); Chanier et al. (forthc.))

Beteiligte aus Korpusprojekten zu versch. Sprachen (CoMeRe, DeRiK, SoNaR, Web2Corpus_it, Dortmunder Chat-Korpus, Mannheimer Wikipedia-Korpus, ...)

Desiderat: Linguistische Basisannotation

Community Shared Task

zur automatischen linguistischen Annotation von IBK-Daten (2015), initiiert von Mitgliedern des Empirikom-Netzwerks und unterstützt durch die GSCL

Fokus:

- Tokenisierung
- Part-of-Speech-Annotation

Grundlage:

- Handannotierte Trainings- und Evaluationsdaten („Goldstandard“)

Ziel:

- Anpassung automatischer Verfahren an den Umgang mit IBK-Schriftlichkeit (in einem kompetitiven Szenario)

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“)	  
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.
DM	Diskursmarker	*prototypisch:* weil, obwohl, nur, also *als Einheiten mit projektivem Potenzial im Vorvorfeld von V2-Sätzen*
ONO	Onomatopoetikon	boing, miau, zisch

STTS-Tagset mit Erweiterungen für IBK und für gesprochene Sprache (Beißwenger, Bartz, Storrer, Westpfahl; forthc.)

Ausblick: Laufende und geplante Arbeiten

a) Aufgabenbereich „Repräsentation“:

- Ausarbeitung des in der TEI-SIG diskutierten Schemas (in Kooperation mit den KollegInnen aus F, IT, NL und aus dem IDS) und Anpassung für die Repräsentation des Dortmunder Chat-Korpus (→ CLARIN-D-Kurationsprojekt)
- Ziel: Bis Ende 2015 Dokumentation eines vollständigen Entwurfs, der in verschiedenen Korpora zum Deutschen und zum Französischen umgesetzt wurde und Formulierung eines Vorschlags für die Erweiterung der TEI-Guidelines um entsprechende Modelle für den Bereich IBK

b) Korpusaufbau:

- Anwendung von Verfahren zum automatischen Crawling von CC-lizenzierten IBK-Ressourcen (Blogkommentare, soziale Netzwerke, Foren, …) inkl. Lizenzerkennung und Spracherkennung
- In 2015: Erhebung eines ersten Teilkorpus für den Zeitraum 2010-2015
- Integration des „Dortmunder Chat-Korpus“ als Teilkorpus für den Zeitraum vor 2010 (Kurationsprojekt im Rahmen von CLARIN)

AusblickL Laufende und geplante Arbeiten

c) Aufgabenbereich „Linguistische Basisannotation“:

- eigene Vorarbeiten zur Anpassung sprachtechnologischer Verfahren an IBK-Schriftlichkeit an der BBAW
- Mitarbeit im GSCL-AK „Social Media / Internetbasierte Kommunikation“
- 2015: Linguistische Basisannotation für das Dortmunder Chat-Korpus (unter Nutzung aktueller Ansätze, die z. B. im Rahmen der GSCL-Shared-Task entwickelt werden; s. auch die Ansätze, die beim KONVENS-Workshop zu IBK im Okt. 2014 diskutiert wurden)
- Nutzung von Verfahren aus dem BMBF-Verbundprojekt „Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining“ (KobRA, Leitung: Angelika Storrer)

Bundesministerium für Bildung und Forschung

UNIVERSITÄT MANNHEIM

technische universität dortmund

KobRA

Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining

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

Literatur / Ressourcen

DeRiK:

- Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2013): ***DeRiK:* A German Reference Corpus of Computer-Mediated Communication**. In: Literary and Linguistic Computing (DOI: 10.1093/llc/fqt038). http://llc.oxfordjournals.org/cgi/reprint/fqt038?ijkey=GXFixqNNy0uW7cO&keytype=ref
- Beißwenger, Michael; Lemnitzer, Lothar (2013): **Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS).** In: Journal for Language Technology and Computational Linguistics 26 (2), 1-22. http://www.jlcl.org/2013_Heft2/1BeiLem.pdf

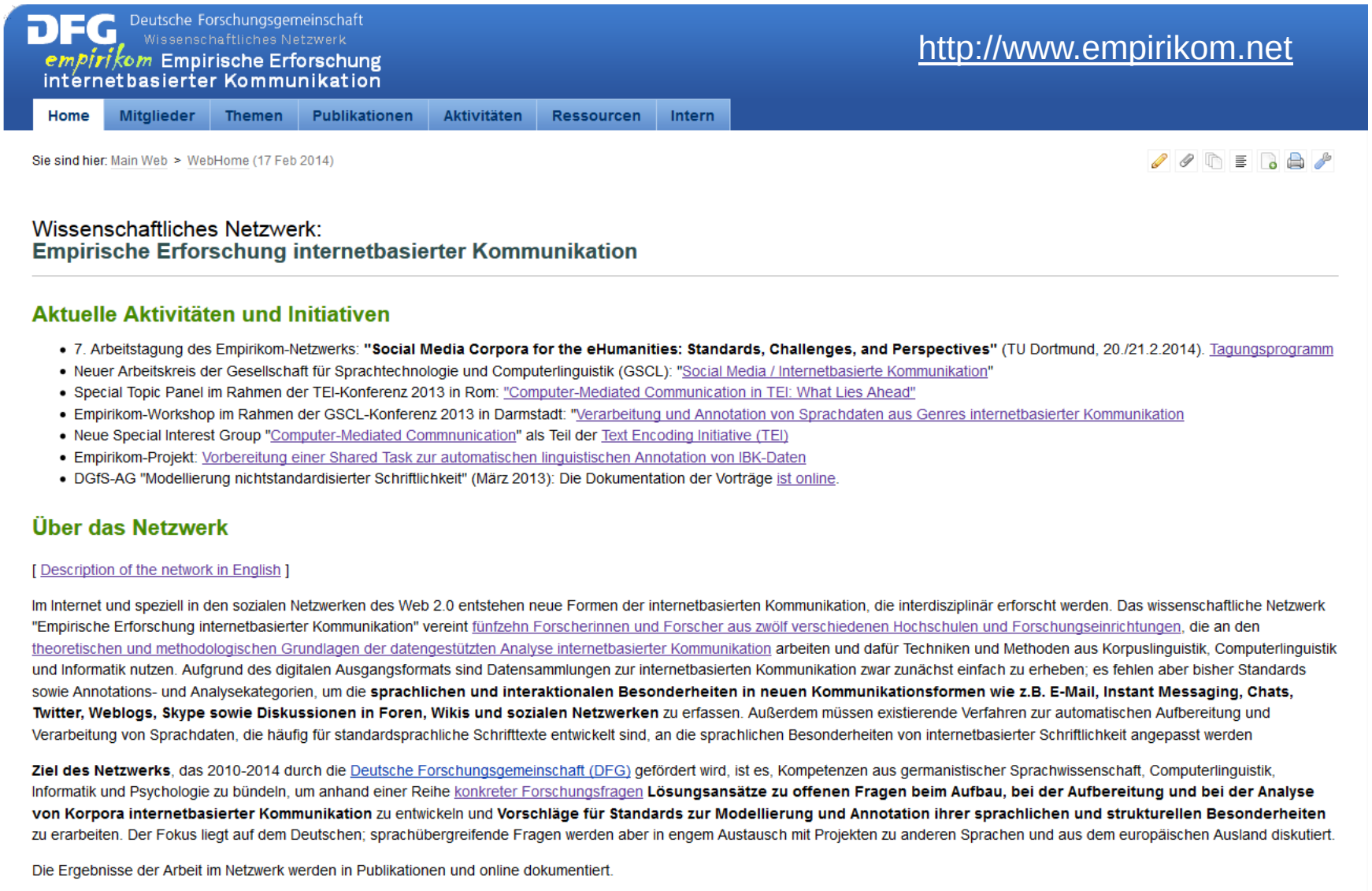
DeRiK-TEI-Schema:

- 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), Issue 3, November 2012 (DOI: 10.4000/jtei.476). http://jtei.revues.org/476

Projekte / Netzwerke:

- DFG-Netzwerk Empirikom: http://www.empirikom.net
- DWDS: http://www.dwds.de
- Dortmunder Chat-Korpus: http://www.chatkorpus.tu-dortmund.de
- TEI-SIG „Computer-Mediated Communication“: http://www.tei-c.org/Activities/SIG/CMC/
- BMBF-Projekt *KobRA*: http://www.kobra.tu-dortmund.de
- GSCL-AK „Social Media / Internetbasierte Kommunikation“: http://gscl.org/ak-ibk.html

DFG-Netzwerk *Empirikom* (2010-2014)



Desiderat: Standards für die Repräsentation von IBK-Korpora

Dortmunder Chat-Korpus

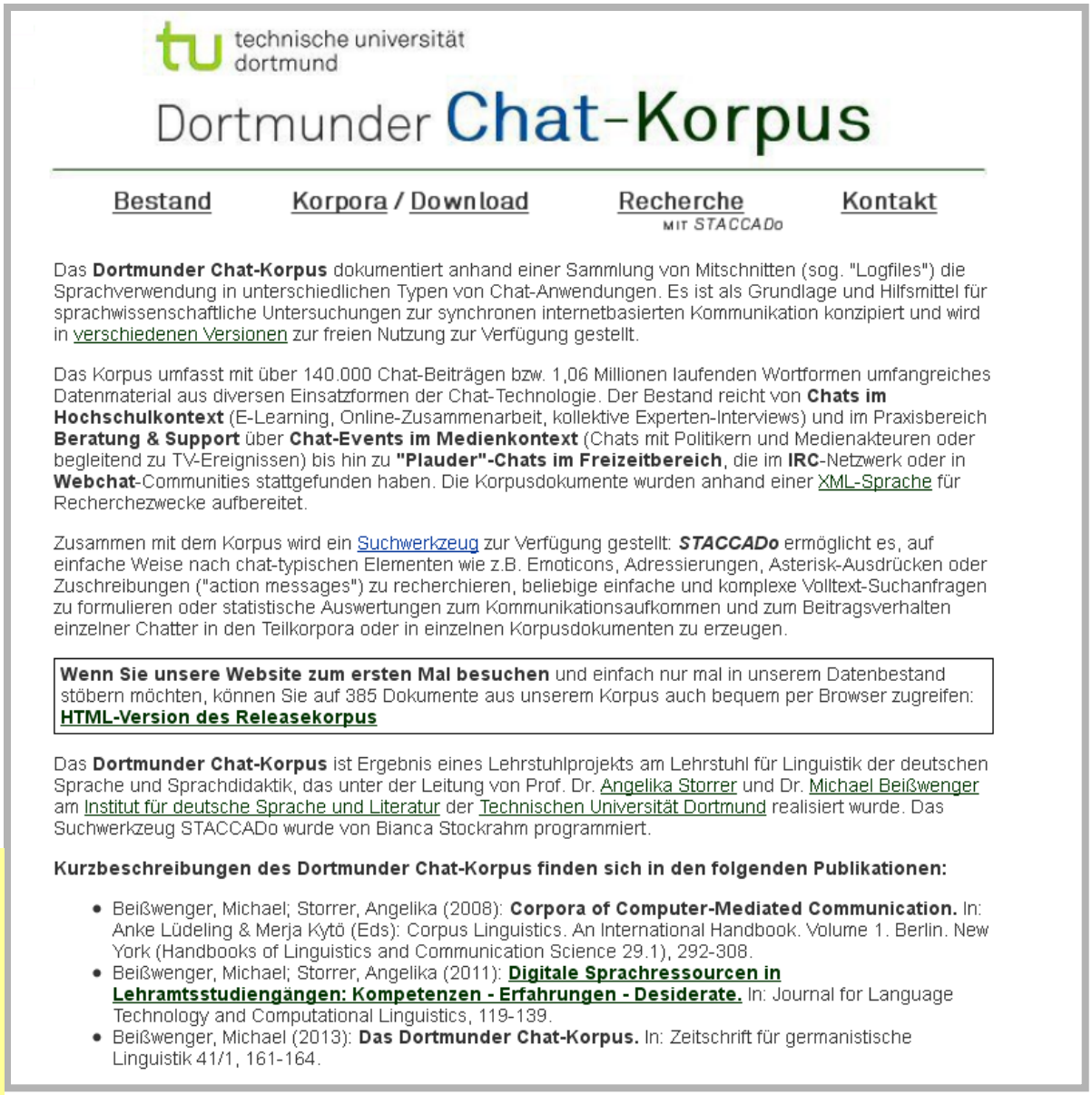
http://www.chatkorpus.tu-dortmund.de

Ergebnis eines Lehrstuhlprojekts an der TU Dortmund (2002-2008) (A. Storrer / M. Beißwenger)

CLARIN-D-Kurationsprojekt (F-AG 1 „Deutsche Philologie“, 2015):

***ChatCorpus2CLARIN*:** Integration des Dortmunder Chat-Korpus in die CLARIN-D-Korpusinfrastrukturen am Institut für deutsche Sprache (IDS) und an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW)

⇨ Anpassung des aktuellen Stands des Schemas aus der TEI-SIG + Integration der Ressource u.a. in DeRiK/DWDS



technische universität dortmund

Dortmunder Chat-Korpus

Bestand | Korpora / Download | Recherche mit STACCADo | Kontakt

Das **Dortmunder Chat-Korpus** dokumentiert anhand einer Sammlung von Mitschnitten (sog. "Logfiles") die Sprachverwendung in unterschiedlichen Typen von Chat-Anwendungen. Es ist als Grundlage und Hilfsmittel für sprachwissenschaftliche Untersuchungen zur synchronen internetbasierten Kommunikation konzipiert und wird in verschiedenen Versionen zur freien Nutzung zur Verfügung gestellt.

Das Korpus umfasst mit über 140.000 Chat-Beiträgen bzw. 1,06 Millionen laufenden Wortformen umfangreiches Datenmaterial aus diversen Einsatzformen der Chat-Technologie. Der Bestand reicht von **Chats im Hochschulkontext** (E-Learning, Online-Zusammenarbeit, kollektive Experten-Interviews) und im Praxisbereich **Beratung & Support** über **Chat-Events im Medienkontext** (Chats mit Politikern und Medienakteuren oder begleitend zu TV-Ereignissen) bis hin zu **"Plauder"-Chats im Freizeitbereich**, die im **IRC**-Netzwerk oder in **Webchat**-Communities stattgefunden haben. Die Korpusdokumente wurden anhand einer XML-Sprache für Recherchezwecke aufbereitet.

Zusammen mit dem Korpus wird ein Suchwerkzeug zur Verfügung gestellt: ***STACCADo*** ermöglicht es, auf einfache Weise nach chat-typischen Elementen wie z.B. Emoticons, Adressierungen, Asterisk-Ausdrücken oder Zuschreibungen ("action messages") zu recherchieren, beliebige einfache und komplexe Volltext-Suchanfragen zu formulieren oder statistische Auswertungen zum Kommunikationsaufkommen und zum Beitragsverhalten einzelner Chatter in den Teilkorpora oder in einzelnen Korpusdokumenten zu erzeugen.

Wenn Sie unsere Website zum ersten Mal besuchen und einfach nur mal in unserem Datenbestand stöbern möchten, können Sie auf 385 Dokumente aus unserem Korpus auch bequem per Browser zugreifen: **HTML-Version des Releasekorpus**

Das **Dortmunder Chat-Korpus** ist Ergebnis eines Lehrstuhlprojekts am Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik, das unter der Leitung von Prof. Dr. Angelika Storrer und Dr. Michael Beißwenger am Institut für deutsche Sprache und Literatur der Technischen Universität Dortmund realisiert wurde. Das Suchwerkzeug STACCADo wurde von Bianca Stockrahm programmiert.

Kurzbeschreibungen des Dortmunder Chat-Korpus finden sich in den folgenden Publikationen:

- Beißwenger, Michael; Storrer, Angelika (2008): **Corpora of Computer-Mediated Communication.** In: Anke Lüdeling & Merja Kytö (Eds): Corpus Linguistics. An International Handbook. Volume 1. Berlin. New York (Handbooks of Linguistics and Communication Science 29.1), 292-308.
- Beißwenger, Michael; Storrer, Angelika (2011): **Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen - Erfahrungen - Desiderate.** In: Journal for Language Technology and Computational Linguistics, 119-139.
- Beißwenger, Michael (2013): **Das Dortmunder Chat-Korpus.** In: Zeitschrift für germanistische Linguistik 41/1, 161-164.