Korpusbasierte Analyse internetbasierter Kommunikation: Herausforderungen und Perspektiven

Neue Wege in der Nutzung von Korpora: Data-Mining für die textorientierten Geisteswissenschaften

Fachtagung, 30. Oktober 2015

Universität Mannheim
technische universität dortmund
Bundesministerium für Bildung und Forschung
KobRA
Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining

Michael Beißwenger
Harald Lüngen
Christian Pölitz

Herausforderungen bei der korpusbasierten Analyse

[1] Werkzeuge für die automatische linguistische Annotation können mit der Schriftlichkeit in Genres internetbasierter Kommunikation nicht umgehen (“Nonstandard-Phänomene”: Abweichungen von den Normen der geschriebenen Standardsprache; fehlende Standards für die Verarbeitung netztypischer Stlelemente)

⇨ Handannotation: kostet viel Zeit (und Geld)

⇨ Volltextsuche: Hohe Zahl an unerwünschten Treffern; Trefferlisten müssen vor der Analyse intellektuell bereinigt werden (auch teuer)

[2] Zwar können IBK-Daten prinzipiell in großen Mengen aus dem Web erhoben werden (s. „web as corpus“) – die Datensets, die man bei der Analyse tatsächlich bewältigen kann, sind aber i.d.R. eher klein.

Fragestellung:

- Können Machine-Learning-Verfahren für eine Bereinigung von Trefferlisten zu Fragestellungen im Bereich IBK adaptiert werden?

Beispiel für eine Lernaufgabe: „Aktionswort-Finder“

- *Aktionswörter* basieren auf einem Wort – im Deutschen häufig einem Inflektiv –, das entweder alleine steht (*lach*, *schüttel*) oder um weitere Einheiten erweitert sein kann (*lautlach*, *kopfschüttel*).

> Ich frage mich, ob's wohl nen Fachbegriff für genau diese Art von Klangerzeugern gibt? ***grübel***
>
> ***Augenroll*** Das ist genau meine Argumentation.
>
> Deinen Kommentar finde ich ***räusper*** problematisch.

- Sie dienen zur (häufig spielerischen) Beschreibung von Emotionen, mentalen oder körperlichen Zuständen oder Aktivitäten sowie als Illokutions- und Ironiemarker.
- Sie sind typischerweise nicht syntaktisch integriert.
- Sie werden häufig (nicht immer) durch Asterisken markiert (**lach**, **freu**).

Beispiel für eine Lernaufgabe: „Aktionswort-Finder“

Daten:

z.B. (1): Trefferliste für die häufigsten Aktionswort-Formen (*lol, lach, freu, grins, wink, seufz*) (Storrer 2013)

z.B. (2): Trefferliste „Beliebige Ausdrücke zwischen Asterisken“

Aufgabe: Lerne, ausgehend von einem Sample mit manuell klassifizierten Daten, ein Modell, das es erlaubt, die Treffer automatisch in zwei Klassen zu teilen!

Beispiele für unerwünschte Treffer:

(1) Ich freu mich auf die Diskussion :)

(1) hahahahaha ich lach mich tot xD

(2) hehe, hast mich erwischt. Nein, das mit dem Zitat hatte ich im Eifer des Gefechts eingefügt und *natürlich* ist es von Watson.

(2) Go(s) = (1,8s+1)*(0,9s+1)*(0,54s+1). Die Überschwingung wird auf ca. 3 % reduziert.

KobRA-Seminar / TU-Dortmund, SS 21014

Interdisziplinäres Hauptseminar
(Germanistik / Informatik):

Korpusgestützte Analyse internetbasierter Kommunikation mit Hilfe von Data-Mining

technische universität dortmund

Korpusgestützte Analyse internetbasierter Kommunikation mit Hilfe von Data-Mining

Interdisziplinäres Projektseminar im Sommersemester 2014

PD Dr. Michael Beißwenger (Germanistik) – Dipl.-Inform. Christian Pölitz (Informatik)

Di. 10-12, R. U 331 – LABG 2009: BLS1 / BaMa 2005: F6

Seminarplan:

Termin	Sitzungsthema
Di. 08.04.	[PLENUM] **Vorstellung des Seminarkonzepts und Organisatorisches**
Di. 15.04.	[PLENUM] **Überblick: Korpusgestützte Analyse internetbasierter Kommunikation: Fragestellungen und Perspektiven (M. Beißwenger) / Methoden der Künstlichen Intelligenzforschung und des Maschinellen Lernens/Data-Mining für die Analyse von Sprachdaten (C. Pölitz)**
Di. 22.04.	[PLENUM] **Einführung in Fokusbereich 1: *Geht die Nebensatzstellung im Netz verloren?* (Untersuchung zu *Diskursmarkern* in der internetbasierten Kommunikation)**
Di. 29.04.	[PLENUM] **Einführung in Fokusbereich 2: *Wie mündlich sind „getippte Gespräche"?* (Untersuchung zu *Verschmelzungsformen* in der internetbasierten Kommunikation)**
Di. 06.05.	[GERMANISTIK] **Ausgabe der Korpusdaten und Annotationsrichtlinien (L. Cedli)**
Di.	[GERMANISTIK] **Selbstständige Analyse und Annotation der Korpusdaten (keine Sitzung)**

	non-VL	kein finV	Pseudo	Beleg
2		X		zur Redundanz von ""Die T?zai-Linie nach Nishi-Funabashi beginnt derzeit hier"": Das Wort ""redundant"" ist ""Redundanz...: Das Vorhandensein von eigentlich überflüssigen, für die Information nicht notwendigen Ele unumgänglich sind und/oder sein sollen. Das 'eigentlich' trägt dem rechnung, denn es schränkt ein: Nach 'eig nämlich 'nützlich' und 'hilfreich', 'angebracht' und 'sinnvoll'.Wenn man einen Bahnhofsartikel liest, dann mö
3	X			Wieder einmal bin ich auf einen Artikel gestoßen, der Potential hat, ich aber (aufgrund fehlenden Fachwisse allerdings vielleicht etwas kurz (Sozialverhalten könnte ergänzt werden, ansonsten fällt mir nichts ein). Wob ihn exzellent zu machen. //OT: Übrigens gibt es hier so etwas wie eine Standartgliederung für Tiere <§§>wei
4		X		Wie kommst du auf 1980? In den 1980er jahren hat sich der Maskulismus in den USA entwickelt, wurde aber Backlash auf vergleichsweise fruchtbaren Boden, auch wenn da auch viel Krakele und Skandal dabei, <§§>we abarbeiten.--
5	X			Wer meint das ich trolle kann googeln, wie gesagt Quelle is offen aber nicht unbedingt öffentlich und niema Funktechnik als selbstständiger Kaufmann auf ungewöhnliche Antennen ausgerichtet bin, mußte ich mir der sollte mann/frau eigentlich nach Antennengewinn verschieben! <§§>Weil</§§> dort fehlt das. Zur weiteren Meldung Euer unwissender suchender usw Reinhard-
				Wer kennt den Aufsatz von Gerhard Kern über "Theosophie und Anthroposophie - Eine philosophische Basis diesem befaßt. Der Text war bei Pierre Krebs angegeben. Ich hatte ihn gelöscht, <§§>weil</§§> er nicht meh die Thule-Gesellschaft einen Zusammenhang mit dem Thule-Seminar von Pierre Krebs her. Meines Erachten

akr?	Redup?	AKW:kplx?	AKW:infl?	AKW:infl:basis	match	start pos	end pos	text
		X	X	grinsen	"ganzganzfiesgrins"	3.0	28.0	<nowiki>"ganzganzfiesgrins" da ergibt sich doch eine tolle gelegenheit für ein AP, nicht wahr "diabolischgrins" – Atlan Disk. 14:43, 6. Feb. 2010 (CET)
					"kann"	184.0	190.0	Nun, für sachlich falsch halte ich zunächst die Aussage, das d in Adlershof würde wie t gesprochen. Dass "Du" d sprichst, ohne die Position der Lippen zu verändern, mag sein. Ich "kann" es _notfalls_ auch, das geht dann abe Richtung Lallen. Üblicherweise unterscheiden sich aber bei "mir" jedenfalls die Laute d und t in Lippenspannung damit –form, beim d wird "mein" Mund runder.Ebenso spreche "ich" die Kombination [Vokal/Konsonant/l oder r/Vo wenn überhaupt – nur in Komposita
		X	X	zucken	"schulterzuck"	64.0	78.0	Ja, nun ja, da AMD ja aber bei amdcompare selber von G1 spricht "schulterzuck" Heise berichtet auch nicht von Änderungen am Kern selber... Ich würd sagen, wir lassen es mal basierend auf den offiziellen Angaben von AMD

Sommersemester 2014			**Michael Beißwenger** (Germanistik) **Christian Pölitz** (Informatik)
Korpusgestützte Analyse internetbasierter Kommunikation mit Hilfe von Data-Mining			
Hauptseminar	2 SWS	Di 10-12 Uhr, R. U331	Beginn: 8.4.2014
LABG 2009: BLS1 / BaMaLa 2005: F6 / B.a./M.a. Angewandte Sprachwissenschaft: M2, M3, M7			

Kurzbeschreibung:

Ziel des Seminars ist es, anhand ausgewählter sprachwissenschaftlicher Fragestellungen den Einsatz innovativer Informatikmethoden („Data-Mining", maschinelles Lernen) für die empirische korpusgestützte Analyse internetbasierter Kommunikation zu erproben.

Um Besonderheiten der Sprachverwendung in der internetbasierten Kommunikation auf der Basis großer Datensammlungen (Korpora) quantitativ und qualitativ untersuchen zu können, bedarf es automatischer Verfahren, die die Daten vorsortieren, klassifizieren und nach potenziell interessanten Belegen durchforsten. Um solche Verfahren entwickeln zu können, bedarf es sprachwissenschaftlichen Know-Hows, das in Form sogenannter „Annotationen" in kleine Datensets eingebracht wird.

Im Seminar werden wir in kleinen Analyseprojekten solche Annnotations- und Entwicklungsprozesse durchspielen. Dabei arbeiten Studierende der Germanistik mit Studierenden der Informatik zusammen, wobei die Germanistik-Studierenden Daten auf der

KobRA-Seminar / TU-Dortmund, SS 21014

Seminarprojekt: Automatische Eliminierung von Pseudotreffern und Finden von „Nadeln im Heuhaufen“ für große Trefferlisten zu ausgewählten sprachlichen Phänomenen internetbasierter Kommunikation – *zum Beispiel:*

- **Aktionswörter:**

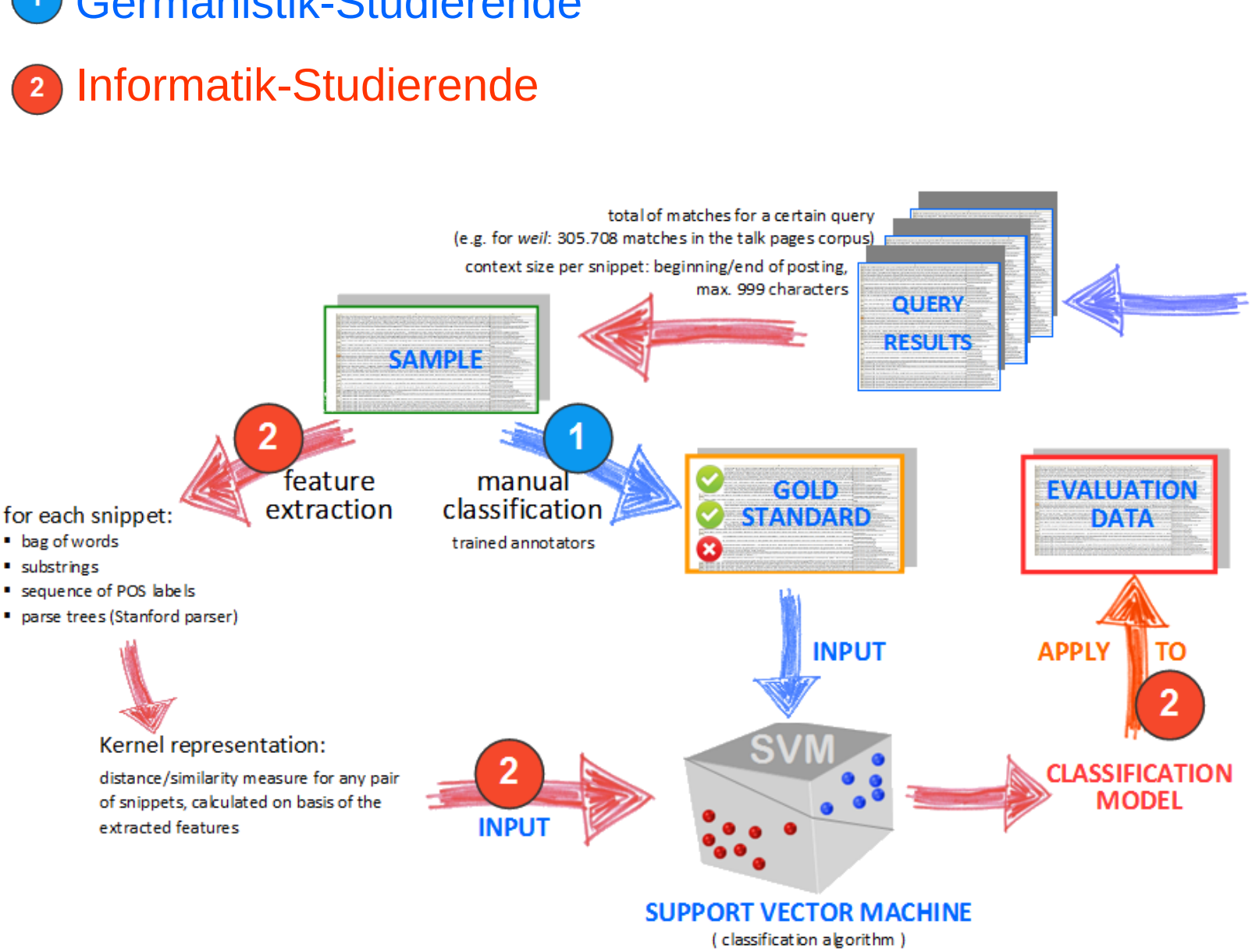
  freu, *lach*, *schmunzel*, *ganzfiesgrins*, …

- **nicht-kanonische Verwendungen von *weil* und *obwohl* (V2 anstelle von V-L):**

  ja toll aber so richtig steht es nicht drin weil damals sollten wir nämlich eine arbeit in informatik machen über das dualsystem

1 Germanistik-Studierende

2 Informatik-Studierende



Das Korpus

Wikipedia-Korpus 2013 in DeReKo
(Kupietz & Lüngen 2014)
http://www.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

	Articles	Talk pages
I5 file size	~16G	~4.8G
# pages	~1,6M	~550K
# postings	–	~5,5M
# tokens	~678M	~264M

Das Korpus ist repäsentiert in I5 (Sperberg-McQueen & Lüngen 2012), der Customization des P5-Encoding-Formats der Text Encoding Initiative (TEI, http://tei-c.org) für DeReKo. I5 integriert Modelle für die Strukturbeschreibung von Genres internetbasierter Kommunikation aus dem adaptierten TEI-Schema von Beißwenger et al. (2012).

```
<div n="2" type="thread">
  <head type="cross">
    <s>Totensonntag in der DDR</s>
  </head>
  <posting indentLevel="0" who="WU00000000">
    <p>
      <s>Hallo, weiß jemand ob es auch einen
      Totensonntag in der DDR Gab?? Danke
      </s>
    </p>
  </posting>
  <posting indentLevel="1" synch="t00121163"
  who="WU00006525">
    <p>
      <s broken="yes">Warum sollte es den dort
      nicht gegeben haben?</s>
      <s>Auch in der DDR hörte das Kirchenjahr
      mit dem Ewigkeitssonntag/Totensonntag auf
      und das neue fing mit dem 1. Advent wieder
      an.</s>
      <s>--<autoSignature/> 23:23, 5. Dez. 2006
      (CET) </s>
    </p>
  </posting>
  [...]
```

Die Grundlage für das Seminar bildet das Diskussionsseiten-Teilkorpus. Die Konvertierung nach I5 ist beschrieben in Margaretha & Lüngen (2014).

WIKIPEDIA IN DEREKO (2013)

	Artikel	Diskussionen
# WP-Seiten (Texte)	1.585.823	554.617
# Postings	--	5.469.899
# Tokens	678.444.643	272.654.185
I5-Dateigröße	16G	4,8G

- Encoding in I5 + CMC (Sperberg-McQueen/Lüngen 2012, Beißwenger et al. 2015)
- Threads, Heuristiken für Posting-Segmentierung in Diskussionen
- POS-Annotationen mit TreeTagger/ STTS 1.0 (standoff)
- COSMAS II oder Download v. http://corpora.ids-mannheim.de/pub/wikipedia-2013/

Referenz: Eliza Margaretha / Harald Lüngen (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: *Journal of Language Technology and Computational Linguistics (JLCL)* 29 (2), S. 59-82, http://www.jlcl.org/

CMC DOKUMENTSTRUKTUR IN I5

IBK-ELEMENTE ADAPTIERT VON BEIßWENGER ET AL. 2012

```
<div n="2" type="thread">
   <head type="cross">
      <s>Totensonntag in der DDR</s>
   </head>
   <posting indentLevel="0" who="WU00000000">
     <p>
       <s>Hallo, weiß jemand ob es auch einen Totensonntag in der DDR
          Gab?? Danke</s>
     </p>
   </posting>
   <posting indentLevel="1" synch="t00121163" who="WU00006525">
     <p>
       <s broken="yes">Warum sollte es den dort nicht gegeben haben?</s>
       <s>Auch in der DDR hörte das Kirchenjahr mit dem Ewigkeitssonntag/
          Totensonntag auf und das neue fing mit dem 1. Advent wieder
          an.</s>
       <s>--<autoSignature/> 23:23, 5. Dez. 2006 (CET) </s>
     </p>
   </posting>
   […]
```

WIKIPEDIA-KORPORA AM IDS (2015)
KONVERTIERUNG: ELIZA MARGARETHA

Neue Features 2015:

- Nutzerdiskussionen (*User Talk Pages*)
- Verbesserung der Posting-Segmentierung
- Language-Links in Metadaten

	Artikel #tok	Diskussionen #tok	Nutzerdiskussionen #tok
Deutsch (de)	796.638.747	309.897.027	271.441.322
Englisch (en)	2.403.943.177	1.270.217.981	2.698.338.998
Französisch (fr)	764.459.026	137.107.729	372.639.260
Ungarisch (hu)	117.987.947	8.293.799	26.215.158
Norwegisch (no)	99.014.144	5.314.362	32.481.331
Spanisch (es)	578.882.431	54.907.258	276.034.367
Kroatisch (hr)	46.641.724	2.480.966	18.731.167
Italienisch (it)	463.022.806	49.825.036	125.573.567
Polnisch (pl)	298.207.197	16.558.557	64-126.136

Ergebnisse der Seminarprojekte und Desiderate

Lernaufgaben zu Aktionswörtern:

(1) Precision: 87% Recall: 92%

(2) Precision: 74% Recall: 71%

Identifizierung nicht-kanonischer Verwendungen von *weil*:

Precision: 13% Recall: 55%

Wenn man zum reinen Bag-of-words-Ansatz Part-of-speech- und Parse-Tree-Kernels zuschaltet, werden die Ergebnisse sogar schlechter.

⇨ **Eine Verbesserung der Lernverfahren setzt eine Anpassung der genutzten Sprachverarbeitungs-werkzeuge voraus.**

Desiderat 1: NLP für CMC

Die Probleme betreffen verschiedene Ebenen des Verarbeitungsprozesses:

- **Tokenisierungsprobleme:** Der Tokenisierungsprozess erzeugt Tokens, die keine sinnvollen linguistischen Einheiten darstellen (z.B. aufgrund von *speedwriting phenomena*)
- **Kategorisierungsprobleme:** Es gibt eine passende Kategorie im verwendeten Tagset, der Tagger kann das entsprechende Tag aber nicht zuweisen (z.B. aufgrund von umgangssprachlichen Schreibungen)
- **Kategorienprobleme:** Der Tagger kann kein sinnvolles Tag zuweisen, da für die betreffende Kategorie im Tagset kein Tag existiert (z.B. im Falle von Emoticons, Emojis, Hsshtags, Aktionswörtern, konzeptionell mündlichen Verschmelzungsformen)

Cf. Bartz et al. (2014)

“STTS 2.0”: Erweitertes Part-of-speech-Tagset für IBK

Tag	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, heute, jetzt*
APPR	Präposition, Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache], vom, übern, fürn*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit?] A big fish [übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
ONO	Onomatopoetikon	*boing, miau, zisch*
DM	Diskursmarker	*prototypisch: weil, obwohl, nur, also als Einheiten mit projektivem Potential im Vorvorfeld von V2-Sätzen*
KOUI	unterordnende Konjunktion mit „zu“ und Infinitiv	*um [zu leben]* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz (VL-Stellung)	*weil, dass, damit* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel ohne Satz	*als, wie*
NN	Appellative	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativpronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativpronomen	*jener [Mensch]*
PIS	substituierendes Indefinitpronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinitpronomen ohne Determiner	*kein [Mensch]* *irgendein [Glas]*
PIDAT	attribuierendes Indefinitpronomen mit Determiner	*[ein] wenig [Wasser]* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessivpronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	substituierendes Relativpronomen	*[der Hund,] der*
PRELAT	attribuierendes Relativpronomen	*[der Mann,] dessen [Hund]*
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe]*
PWAV	adverbiales Interrogativ- oder Relativpronomen	*warum, wo, wann* *worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	„zu“ vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*

Tag	Beschreibung	Beispiele
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] Rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten], zu [schnell]*
PTKIFG	Intensitäts-, Fokus- oder Gradpartikel	*sehr [schön], höchst [eigenartig], nur [sie], voll [toll]*
PTKMA	Modal- oder Abtönungspartikel	*[Das ist] ja / vielleicht [doof]* *[Ist das] denn [richtig so?]* *[Das war halt] [echt nicht einfach]*
PTKMWL	Partikel als Teil eines Mehrwort-Lexems	*keine mehr, noch mal, schon wieder*
TRUNC	Kompositions-Erstglied	*An- [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit „zu“, voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig]!*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
VVPPER	Kontraktion: Vollverb + irreflexives Personalpronomen	*schreibste, machste*
VMPPER	Kontraktion: Modalverb + irreflexives Personalpronomen	*wilste, darfste, musste*
VAPPER	Kontraktion: Auxiliarverb + irreflexives Personalpronomen	*haste, biste, isses*
KOUSPPER	Kontraktion: unterordnende Konjunktion mit Satz (VL-Stellung) + irreflexives Personalpronomen	*wenns, weils, obse*
PPERPPER	Kontraktion: irreflexives Personalpronomen + irreflexives Personalpronomen	*ichs, dus, ers*
ADVART	Kontraktion: Adverb + Artikel	*son, sone*
EMOASC	Emoticon, als Zeichenfolge dargestellt (Typ „ASCII“)	*:-) :-(^^ O.O*
EMOIMG	Emoticon, als Grafik-Ikon dargestellt (Typ „Image“)	kodiert (Beispiel aus WhatsApp): *emoji/QsmilingFaceWithSmilingEyes* *emoji/QkissingCatFaceWithClosedEyes*
AKW	Aktionswort	*“lach” freu, grübel “toll”*
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*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	*,*
$.	Satzbeendende Interpunktion	*. ? ! ; :*
$(	sonstige Satzzeichen; satzintern	*- [ ] ( )*

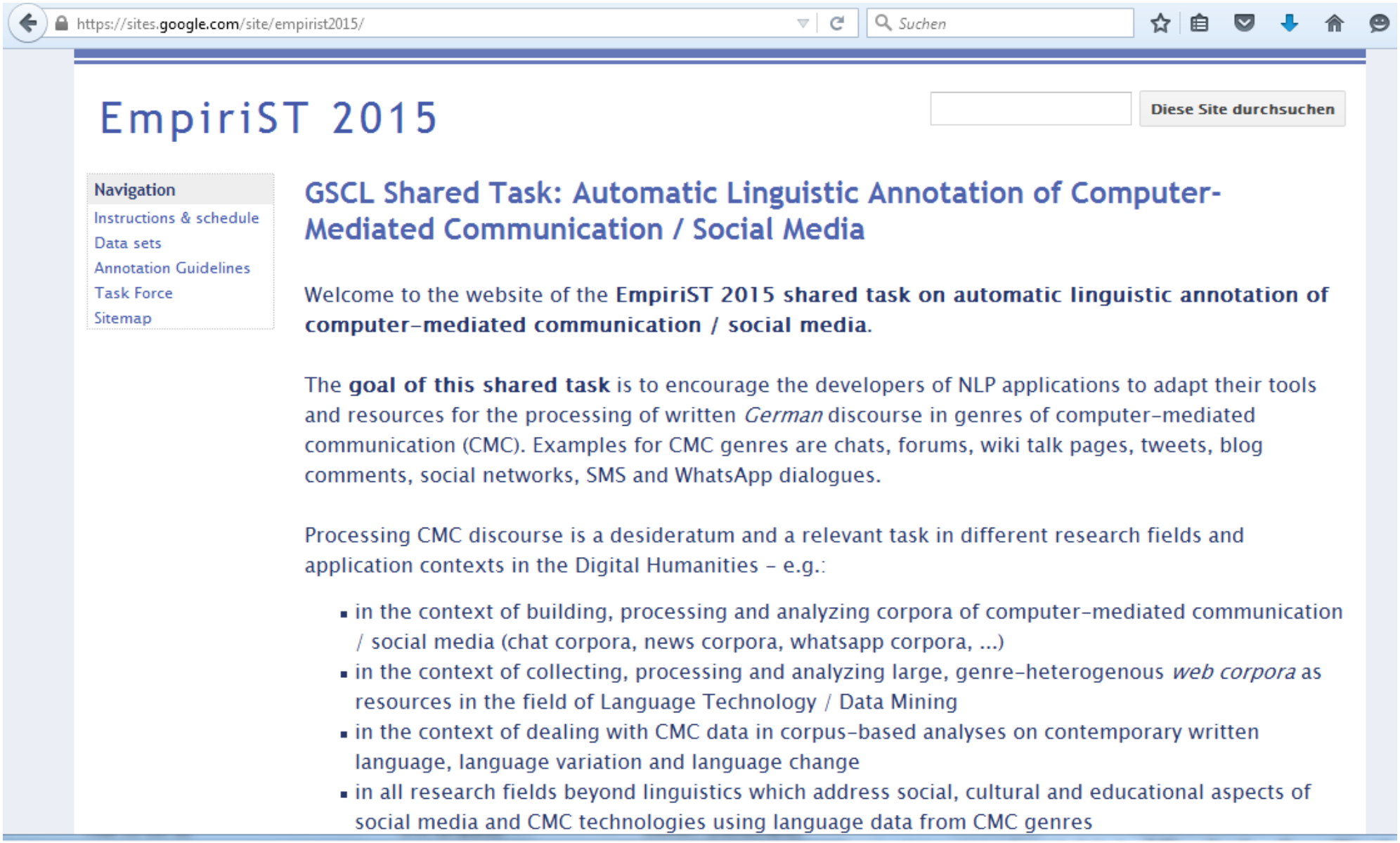
https://sites.google.com/site/empirist2015/home/annotation-guidelines

“STTS 2.0”: Erweitertes Part-of-speech-Tagset für IBK

PoS tag	Category	Examples
I. Tags for phenomena which are specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	Graphic emoticon	
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://www.tu-dortmund.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART is already existing in STTS 1999)	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	‘Intensitätspartikeln’, ‘Fokuspartikeln’, ‘Gradpartikeln’	
PTK MA	Modal particles	
PTK MWL	Particle as part of a multi-word lexeme	
DM	Discourse markers	
ONO	Onomatopoeia	

Abgestimmt auf die STTS-Erweiterungen für das Tagging gesprochener Sprache (FOLK-Korpus, IDS)

GSCL Shared Task zum PoS-Tagging für IBK



https://sites.google.com/site/empirist2015/home/

Desiderat 2: Standards für die Annotation von IBK-Korpora

Es müssen Annotationsstandards entwickelt werden, die es erlauben,

1) die Ergebnisse von Sprachverarbeitungsverfahren sinnvoll und abfragbar in Korpora zu annotieren;

2) diese Annotationen in Formaten zu repräsentieren, die interoperabel sind mit Standards, die für die Annotation von Text- und Gesprächskorpora eingesetzt werden (⇨ vergleichende korpusbasierte Analyse von IBK mit Text- und Gesprächsdaten);

3) die linguistische Annotation mit einer sinnvollen Annotation der strukturellen Besonderheiten von IBK-Genres (z.B. Threadstrukturen) und zugehörigen Metadaten zu verbinden;

4) Ergebnisse von Korpusanalysen, die diese Annotationen nutzen, wiederum als neue Annotationen (in standardisierten Formaten) den Korpora hinzuzufügen.

TEI Special Interest Group (SIG) zu IBK

< Text Encoding Initiative > 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

Entire site Search

Computer-Mediated Communication
Correspondence
Education
Libraries
Manuscripts
Music
Ontologies
Scholarly Publishing
TEI for Linguists
Text & Graphics
Tools

Computer-Mediated Communication SIG

Contents

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

Context

In the past three decades, computer networks and especially the internet have brought forth new and emerging genres of interpersonal communication (*computer-mediated communication*, henceforth "CMC"). Even though there's been a lot of research on CMC genres and on language use on the internet in linguistics and social sciences as well as in the field of natural language processing, there are still no common standards for the representation and annotation of these new forms of communication and their structural and linguistic peculiarities. Being able to represent CMC data on the basis of an encoding framework such as the TEI which is broadly acknowledged within the field of digital humanities will allow for an interchange of data between research groups and for building interoperable CMC corpora for different languages

Scope and Tasks

This special interest group is elaborating on suggestions for adapting the TEI guidelines to the representation of genres of computer-mediated communication (CMC). The focus of the group's work is on (but not limited to) tasks such as:

Dokumentation des aktuellen TEI-Schemaentwurfs für IBK (Stand Oktober 2015):

http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

Ausblick: Anwendung in Projekten

- CLARIN-D-Kurationsprojekt ChatCorpus2CLARIN
  http://www.clarin-d.de/de/kurationsprojekt-1-3-germanistik
- WhatsApp-Datensammlung
  (Projekt „What's up, Deutschland?“)
- Wikipedia-Korpus in DeReKo
- DWDS Blog-Korpus
- News-Korpus in DeReKo
- Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK)