An extended tag set for annotating parts of speech in CMC corpora

Thomas Bartz[1], Michael Beißwenger[1],
Eric Ehrhardt[2], Angelika Storrer[2]

1) technische universität dortmund

2) UNIVERSITÄT MANNHEIM

International Research Days:
Social Media and CMC Corpora for the eHumanities

UNIVERSITÉ RENNES 2

CLARIN-D

Journées Internationales de recherche
«Médias sociaux et corpus de communication médiée par les réseaux. Annotation, analyse, données libres»
23-24 octobre 2015

Part-of-spech tagging for CMC corpora

Without a part-of-speech (PoS) annotation:

- only very limited querying options;
- no basis for advanced processing steps which require a useful linguistic preprocessing (e.g., parse trees).

The Problem:

Part-of-speech taggers (NLP tools in general) do not perform very well on written CMC discourse:

- new elements which don’t fit into any established PoS category (emoticons, addressings, action words, hashtags);
- speedwriing phenomena (typos, omission of characters, norm-deviating use of whitespace);
- colloquial (*Wazzup?*) and creative spellings (*nyce2meetU*)

The problem

A: Rufst an wenn du köpenick bist!

B: Ja

B: Wir sehn uns ja gleich

A: Jo

B: Ersatzverkejr

B: Ich hab keine ahnubg wo der hinfährr-.-

'I have no idea, where it *[that one]* is going to-.-'

The problem

A: Rufst an wenn du köpenick bist!

B: Ja

B: Wir sehn uns ja gleich

A: Jo

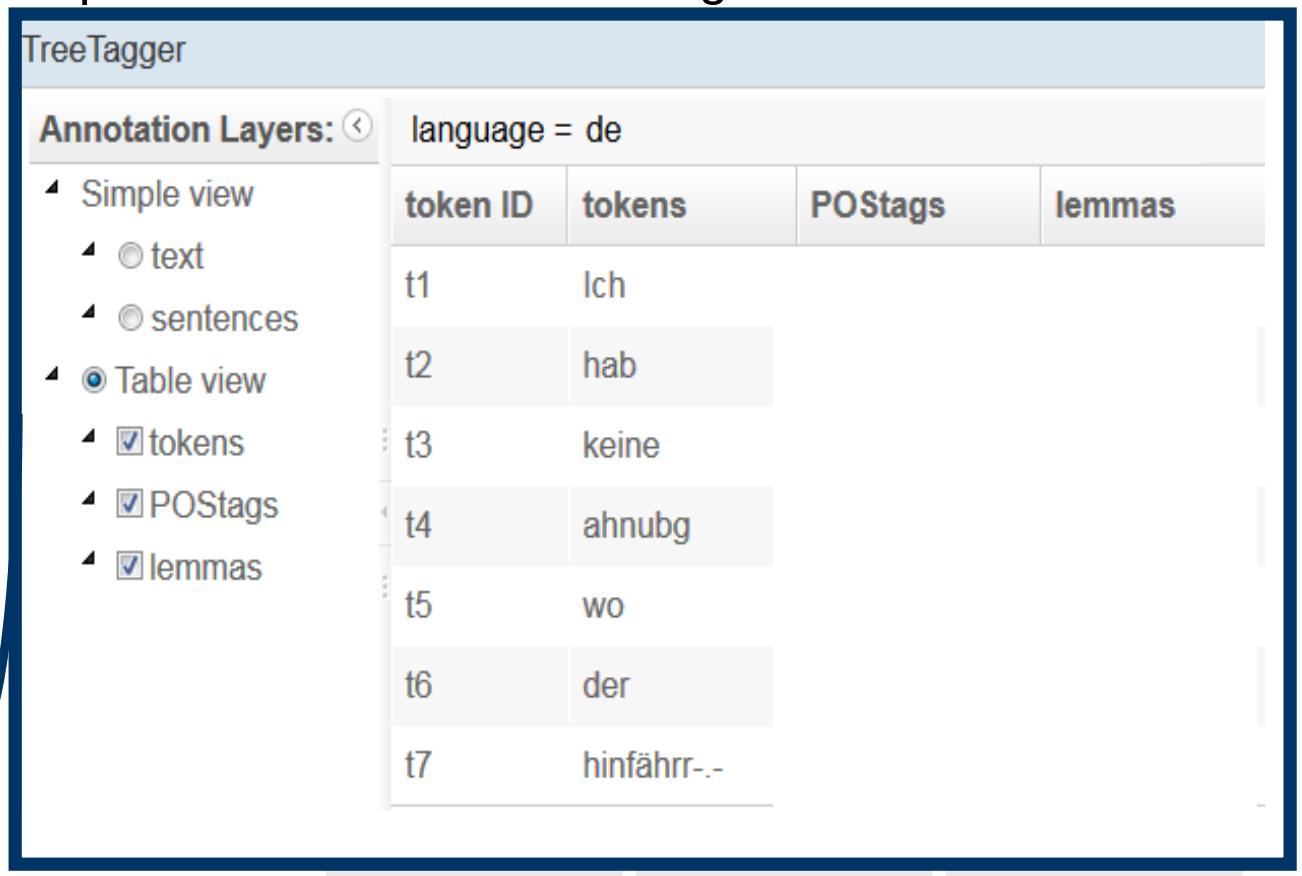
B: Ersatzverkejr

B: Ich hab keine ahnubg wo der hinfährr-.-

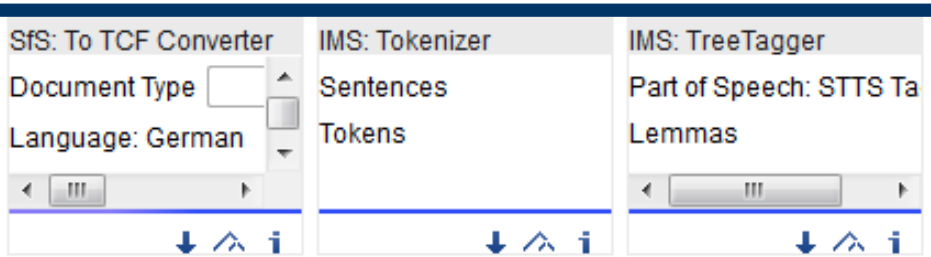
'I have no idea, where it *[that one]* is going to-.-'

WebLicht

https://weblicht.sfs.uni-tuebingen.de/weblicht/



WebLicht-Toolchain:



The problem

A: Rufst an wenn du köpenick bist!

B: Ja

B: Wir sehn uns ja gleich

A: Jo

B: Ersatzverkejr

B: Ich hab keine ahnubg wo der hinfährr-.-

'I have no idea, where it *[that one]* is going to-.-'

WebLicht

https://weblicht.sfs.uni-tuebingen.de/weblicht/

TreeTagger

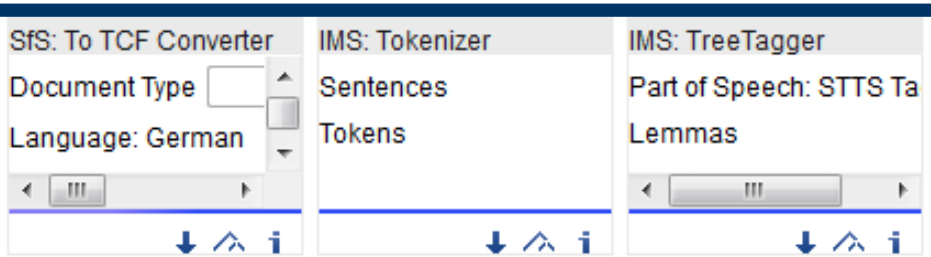
Annotation Layers:

- Simple view
  - text
  - sentences
- Table view
  - tokens
  - POStags
  - lemmas

language = de

token ID	tokens	POStags	lemmas
t1	Ich	PPER	ich
t2	hab	VAIMP	haben
t3	keine	PIAT	kein
t4	ahnubg	TRUNC	<unknown>
t5	wo	PWAV	wo
t6	der	ART	d
t7	hinfährr-.-	TRUNC	<unknown>

WebLicht-Toolchain:



The problem

A: Rufst an wenn du köpenick bist!

B: Ja

B: Wir sehn uns ja gleich

A: Jo

B: Ich hab keine ahnubg wo der hinfährr-.-

'I have no idea, where it *[that one]* is going to-.-'

WebLicht

https://weblicht.sfs.uni-tuebingen.de/weblicht/

TreeTagger

Annotation Layers:

- Simple view
  - text
  - sentences
- Table view
  - tokens
  - POStags
  - lemmas

language = de

token ID	tokens	POStags	lemmas
t1	Ich	PPER	ich
t2	hab	VAIMP	haben
t3	keine	PIAT	kein
t4	ahnubg	TRUNC	<unknown>
t5	wo	PWAV	wo
t6	der	ART	d
t7	hinfährr	ADJA	<unknown>
t8	-.-	TRUNC	<unknown>

The problem

Problems on several levels of the processing process:

- **Tokenization problems:** The tokens created in the tokenization step do not represent relevant units of the linguistic structure (e.g., due to speedwriting phenomena)
- **Categorization problems:** There's an adequate tag in the tag set but the tagger can't assign it (e.g., in the case of norm-deviating colloquial & dialect spellings)
- **Category problems:** The tagger can't assign an adequate tag because there's no adequate tag in the tag set (e.g., for emoticons, action words, addressings, hashtags, clitics which are typical of dialogical language in informal registers…)

Cf. Bartz et al. (2014)

Ways to solve the problem

Variant A:

Normalization ⇨ PoS tagging with standard tools

Open issues:

1) categories for elements that are missing in PoS tagsets for edited text

2) adapt tools for automatic normalization

Variant B:

No normalization; PoS tagging of the original data

Open issues:

1) categories for elements that are missing in PoS tagsets for edited text

2) improve tokenizers & taggers

Designing a basic PoS tag set for German CMC

- **Initiative in CLARIN-D** (2012-13) for **"updating" the canonical STTS** through adapting it for genres which its original creators didn't have in focus (Zinsmeister et al. 2014) – e.g.:

  - historical corpora
  - spoken language corpora
  - learner corpora
  - CMC

- **Discussions in the DFG network** empirikom (2010-2014, http://www.empirikom.net) on how to make NLP tools fit for automatically processing & annotating CMC corpora

⇨ Idea: Let's set up a **community shared task** on NLP for CMC in order to encourage the developers of NLP tools to adapt their tools & tagging models for CMC

⇨ https://sites.google.com/site/empirist2015/home (supported by GSCL)

What requirements should a basic PoS tag set for CMC meet?

- It should be compatible with established PoS tag sets
  ⇨ interoperability with other (types of) corpora
- For categories which occur in CMC but which are not CMC-*specific*: try to be compatible with PoS categories in other (non-CMC) genres (⇨ interoperability of corpora; interesting research questions)
- For categories which are specific to CMC: Keep it simple so that the use of the categories can easily be learned
- As long that there's no consensus in the linguistic community about how to integrate CMC elements into part-of-speech typologies: Don't try to install one (and force people to use it … because *they won't*) – instead, design your categories as theory-free as possible.

“STTS 2.0”: A basic PoS tag set for German CMC

- **Basis:** The “Stuttgart Tübingen Tagset” (STTS): de-facto standard for German (focused on PoS tags for the language occuring in edited text / newspaper texts) (Schiller et al. 1999)

1. Nomina (N)
2. Verben (V)
3. Artikel (ART)
4. Adjektive (ADJ)
5. Pronomina (P)
6. Kardinalzahlen (CARD)
7. Adverbien (ADV)
8. Konjunktionen (KO)
9. Adpositionen (AP)
10. Interjektionen (ITJ)
11. Partikeln (PTK)

“STTS 2.0”: A basic PoS tag set for German CMC

- **Basis:** The “Stuttgart Tübingen Tagset” (STTS): de-facto standard for German

PTKZU	“zu” vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
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]*
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*

Structure of STTS tags: main category > subcategory

“STTS 2.0”: A basic PoS tag set for German CMC

- **Basis:** The “Stuttgart Tübingen Tagset” (STTS): de-facto standard for German (focused on PoS tags for the language occuring in edited text / newspaper texts) (Schiller et al. 1999)
- **“STTS 2.0”:** canonical STTS extended with new categories, but still downward-compatible with STTS (1999)
- Compatible with the extended STTS for spoken language which is used for PoS tagging the FOLK corpus of spoken German at IDS Mannheim (for phenomena which are *not* in the canonical STTS and which also occur in spoken language)

“STTS 2.0”: A basic PoS tag set for German CMC

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

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*
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, wenn's, 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): *emoji1f60aSmilingFaceWithSmilingEyes* *emoji1f61aKissingCatFaceWithClosedEyes*
AKW	Aktionswort	*“lach” freu, grübel, tü!“*
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	*- [ ] ( )*

“STTS 2.0”: A basic PoS tag set for German CMC

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’	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ... *with V2 clauses*
ONO	Onomatopoeia	boing, miau, zisch

“STTS 2.0”: A basic PoS tag set for German CMC

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

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

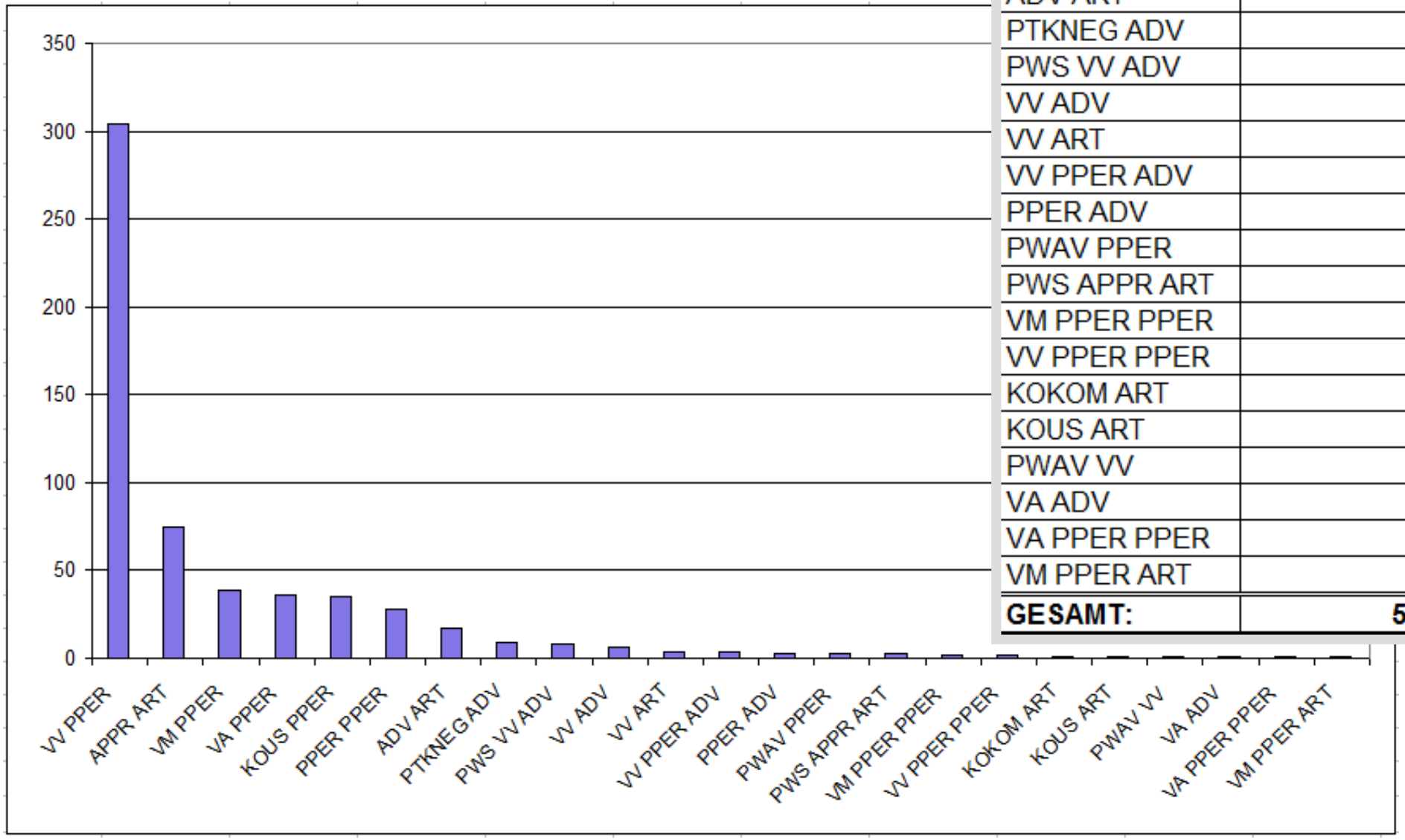
PTK IFG	'Intensitätspartikeln', 'Fokuspartikeln', 'Gradpartikeln'	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder

DM	Discourse markers	weil, obwohl, nur, also, ... *with V2 clauses*

ONO	Onomatopoeia	boing, miau, zisch

Contractions in chats

‘social chat’ subcorpus of the Dortmund chat corpus:
21 logfiles / 104.094 tokens, including 584 occurrences of colloquial contractions



Strukturtyp:	Vorkommen:
VV PPER	304
APPR ART	75
VM PPER	39
VA PPER	36
KOUS PPER	35
PPER PPER	28
ADV ART	17
PTKNEG ADV	9
PWS VV ADV	8
VV ADV	6
VV ART	4
VV PPER ADV	4
PPER ADV	3
PWAV PPER	3
PWS APPR ART	3
VM PPER PPER	2
VV PPER PPER	2
KOKOM ART	1
KOUS ART	1
PWAV VV	1
VA ADV	1
VA PPER PPER	1
VM PPER ART	1
GESAMT:	**584**

Tag set and annotation guidelines @*EmpiriST2015*

EmpiriST 2015

Diese Site durchsuchen

Navigation
- Subtasks & deadlines
- Data sets
- **Annotation Guidelines**
- Task Force
- Sitemap

GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / Social Media >

Annotation Guidelines

The training data that will be provided as a gold standard have been manually tokenized and tagged according to the following guidelines:

- Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): **Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation.** Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (*EmpiriST 2015*). (21 pages).
  PDF: EmpiriST_Guideline-PoS.pdf
- Beißwenger, Michael; Bartsch, Sabine; Evert, Stefan; Würzner, Kay-Michael (2015): **Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation.** Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (*EmpiriST 2015*). (29 pages).
  PDF: EmpiriST_Guideline-Tokenisierung.pdf

When citing these documents, please use the bibliographic information given above and refer to the URL http://sites.google.com/site/empirist2015/.

Overview: The part of speech tagset used for annotations:

Extensions to STTS (1999) are highlighted with blue background colour:

Tag	Description (German)	Examples
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], vorm, überm, fürn*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*

PoS tagset + annotation guidelines available on the website of the GSCL/ Empirikom shared task on automatic linguistic annotation of CMC (EmpiriST2015).

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

ChatCorpus2CLARIN: Project background

Curation project of the CLARIN-D F-AG 1 “German Philology”

Duration: May 2015 – February 2016

Project team: Michael Beißwenger (U Dortmund), Angelika Storrer, Eric Ehrhardt (U Mannheim), Harald Lüngen (IDS), Axel Herold (BBAW) + other colleagues at IDS and BBAW

The task: Re-modeling of the Dortmund Chat Corpus and samples of other CMC resources compliant with existing standards for the representation of corpora in the Digital Humanities. Integration into the CLARIN-D infrastructures at BBAW and IDS.

Main goal:

- Pave the way for the inclusion of linguistically annotated CMC resources into the CLARIN-D corpus infrastructures and create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology.

ChatCorpus2CLARIN: Project back

Curation project of the CLARIN-D F-AG 1 “German F

http://www.clarin-d.de/de/kurationsprojekt-1-3-germanistik

ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D

Project content

In the third curation project of the CLARIN-D working group 1 “German Philology” (F-AG 1) an existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus, and samples of other CMC resources will be restructured to conform to current standards for the representation of corpora in the Digital Humanities context. The main goal of this work is to pave the way for the inclusion of linguistically annotated CMC resources into CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology. To this end, the project will (1) transform the metadata and the annotations of the chat corpus into a TEI-compliant format, (2) enrich the data by further linguistic annotations, and (3) integrate the resulting resource into the CLARIN-D Corpus Infrastructures at theInstitute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW).

The integration in CLARIN-D will allow for a systematic corpus-based analysis of CMC discourse as compared to the language of edited text (as represented in the text corpora at BBAW and IDS) and of spoken conversations (as represented in the spoken language corpora at IDS).

The Dortmund Chat Corpus

The data for the Dortmund Chat Corpus (Beißwenger & Storrer 2008; Beißwenger 2013) was built at TU Dortmund University. The goal of the corpus project was to create a resource for researching the peculiarities and linguistic variation in written computer-mediated communication. The corpus comprises 478 logfile documents with about 140,000 postings and about 1 million tokens of German chats from different application contexts (social chats, advisory chats, chats in the context of learning and teaching, moderated chats in media context). The corpus has been annotated using an XML format (‘ChatXML’) that represents (1) the basic structure and properties of chat logfiles and postings, (2) selected “netspeak” phenomena such as emoticons, interaction words, addressing terms, nicknames and acronyms, (3) selected metadata about the chat users. Since 2005, the corpus has been made available at http://www.chatkorpus.tu-dortmund.de as an XML version for download and offline querying and as an HTML version for online browsing. It has been widely used as a resource for studying and teaching the peculiarities of German CMC discourse. In the CLARIN-D context, it has been used as one of the resources of the curation project: „ Linguistic Annotation of Non-standard Varieties – Guidelines and Best Practices“ of working group 7.

The corpus

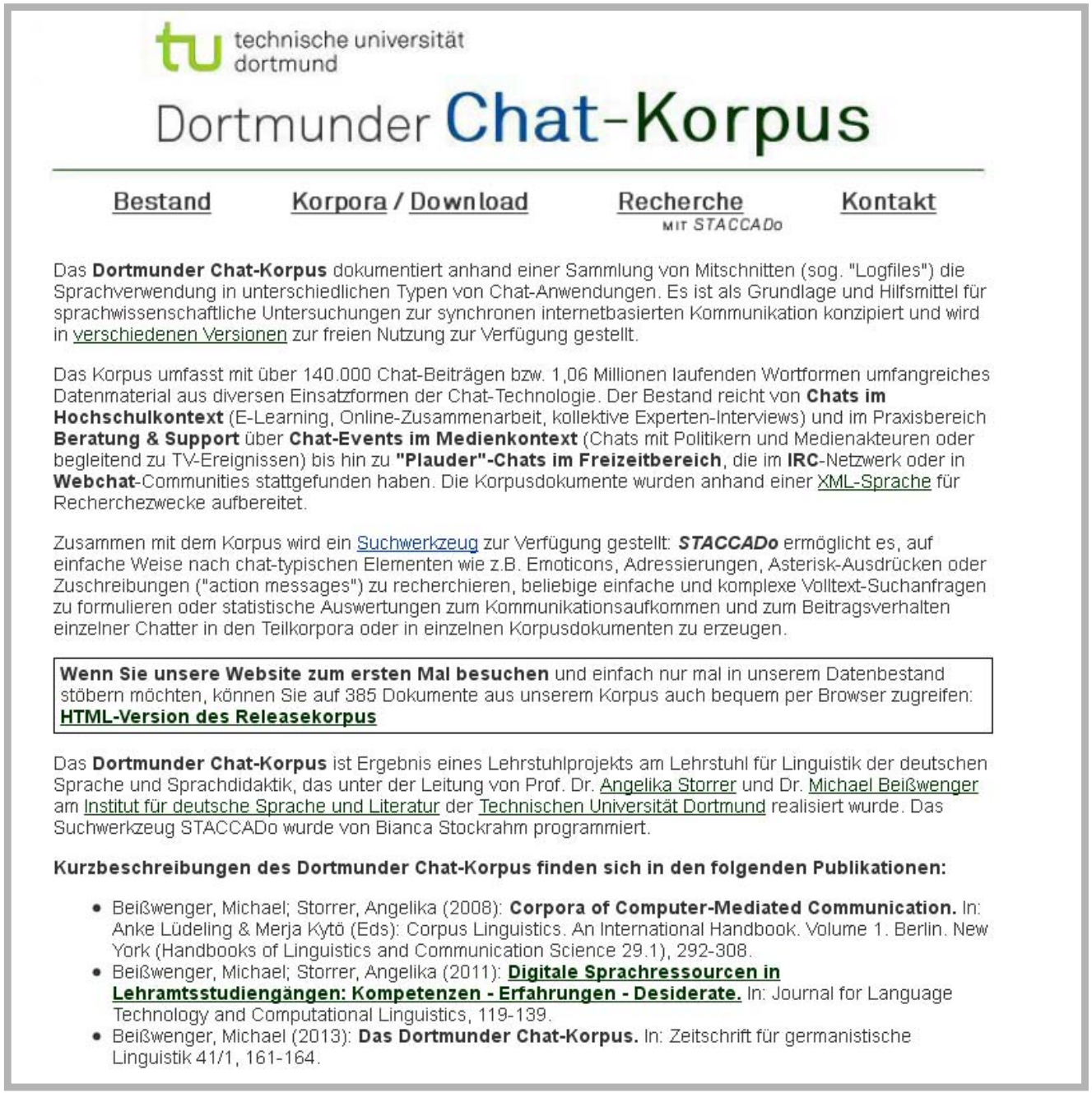
Dortmund Chat Corpus

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

478 logfile documents with 140,240 user postings or 1M words of German chat discourse.

Resource for the analysis of **linguistic variation in chats** including chats from different social/institutional contexts (social chats, advisory chats, learning and teaching, moderated chats in the media context).

Annotated in a home-grown XML format ('**ChatXML**'):
(1) basic structure of chat logfiles and postings,
(2) selected CMC phenomena, (3) selected metadata.



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.

Other corpora / data sets in the project focus

- German WhatsApp Corpus („What's up, Deutschland?“)
- German Wikipedia corpus in DeReKo
- German News Corpus in DeReKo
- DWDS Blog Corpus
- DWDS German Reference Corpus of CMC (DeRiK)

Work packages in the project

- TEI representation (⇨ “CLARIN-D schema”)
- CLARINification, legal issues + licensing
- enrich the data with additional linguistic annotations (PoS, normalised spellings, ...)

The vision

After its integration into the CLARIN-D infrastructure the resource will be characterized by the following added values:

- Advanced accessibility and retrieval options;
- interoperability with other corpus resources that are represented in TEI and with annotation and analysis tools that support the TEI format;
- advanced querying options (PoS tags, normalized spellings);
- interoperability with other corpus resources that have been tagged with STTS;
- advanced options for corpus-based analyses on the peculiarities of CMC discourse as compared to the language of edited text and of spoken language, using the text and speech corpora which are already available in the corpus infrastructures of BBAW and IDS.

PoS annotation of the corpus: workflow

1. **Automatic tokenisation, PoS annotation & lemmatisation** of the chat corpus with tools + tagging models from the BMBF project „Schreibgebrauch“ at U Saarbrücken (Horbach et al. 2014, Horbach et al. 2015) http://www.schreibgebrauch.de

   PoS tag set: previous version of “STTS 2.0” (Bartz et al. 2014)

   Representation of the tagging results as additions to the ChatXML format.

   Standard PoS taggers:

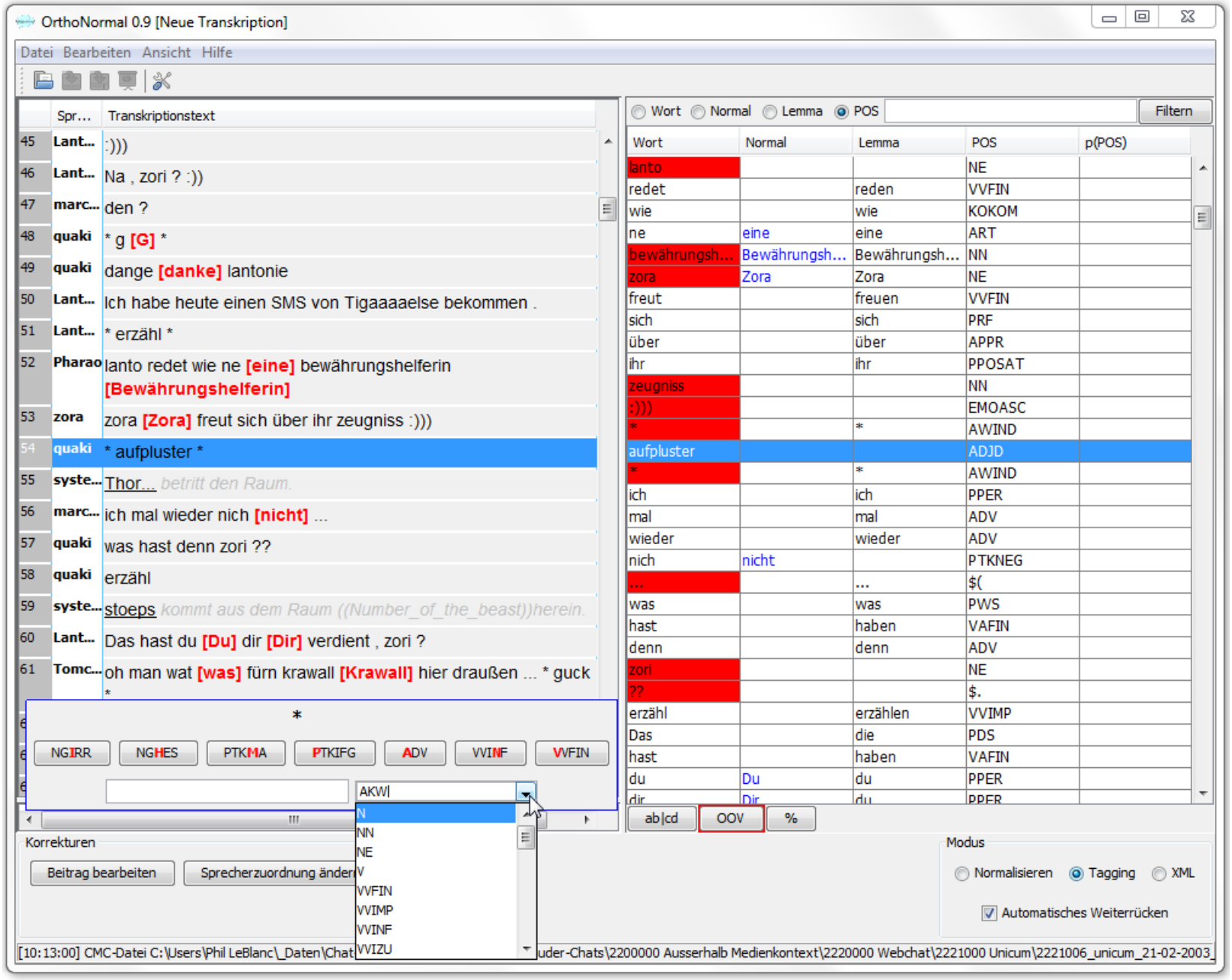
   Accuracy on Chat Corpus: ~71% (vs. 97% accurracy on Newspaper)

   Tagging models from the “Schreibgebrauch” project:

   Average accuracy on Chat Corpus: **83.5%**

2. **Manual post-processing of the tagging results** using OrthoNormal in FOLKER (preview version 1.2) with an import/export filter for PoS tagged ChatXML (defined by Thomas Schmidt/IDS)

Manual post-processing of PoS tagging results with *OrthoNormal*



(Overview of the FOLK tools: Schmidt 2012)

Using <w> for the representation of PoS information in our TEI schema

```
<post type="standard" who="#A04" auto="false" rend="color:green">
  <p>
      <w type="VVFIN">dachte</w>
      <w type="PPER">ich</w>
      < type="ADV">auch</w>
      <w type="ADV">immer</w>
      <w type="$(">,</w>
      <name type="nickname" corresp="#A09">
        <w type="NE">monk</w>
      </name>
      <w type="$.">..</w>
      <w type=„$(">*</w>
      <w type="AKW">heul</w>
      <w type=„$(">*</w>
  </p>
</post>
```

CLARIN-D TEI schema (documentation):

http://wiki.tei-c.org/index.php/SIG:CMC/CLARIN-D_schema_draft_for_representing_CMC_in_TEI_(2015)

ineli26: dachte ich auch immer, monk .. *heul*
*I was always thinking the same, monk .. *crying**

References

Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1), 157-198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41 (1), 161-164. Extended version: http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

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 (DOI: 10.4000/jtei.476).

Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline Document, Dortmund 2015. https://sites.google.com/site/empirist2015/home/annotation-guidelines

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

Horbach, Andrea; Thater, Stefan; Steffen, Diana; Fischer, Peter M.; Witt, Andreas; Pinkal, Manfred (2015): Internet Corpora: A Challenge for Linguistic Processing. In: Datenbank-Spektrum 15 (1), 41-47.

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

Schmidt, Thomas (2012): EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In: Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Zinsmeister, Heike; Heid, Ulrich; Beck, Kathrin Beck (Eds., 2014): Das STTS-Tagset für Wortartentagging - Stand und Perspektiven. Special issue of the Journal for Language Technology and Computational Linguistics. http://www.jlcl.org