Linguistic annotation of CMC and social media corpora:

To what extent do we have to adapt existing encoding standards and tag sets?

Michael Beißwenger

technische universität
dortmund

Slovenščina na spletu in v novih medijih

JANES

University of Ljubljana
26 November 2015

CMC and corpus linguistics

- CMC: important part of everyday communication for many people – for at least 20 years.
- Broad field of research for a range of disciplines in the Humanities (since the early 90ies).
- Nevertheless: Almost no representation of CMC in corpora of written language + only very few specialized CMC corpora.

⇨ **Reasons:**
- Unclear legal situation
- Lack of standards (for describing, annotating and processing CMC data)

⇨ To date, each corpus project on CMC has to develop their own *best practices* in representing and annotating their data.

Layers of describing data in corpora

1) **The document structure**
(text genres: scientific article; dramatic text; correspondence; manuscript; spoken lanuage transcript; …)

2) **The linguistic structure**
(tokens; parts of speech; syntactic structures; features/phenomena specific for the genre(s) or the research question)

3) **The metadata**
(e.g., data describing the nature and context of the data: author; date; source; topic; social context; techological context; …)

Layers of describing data in corpora

1) **The document structure** (“**macrostructure**”)
(text genres: scientific article; dramatic text; correspondence; manuscript; spoken lanuage transcript; …)

2) **The linguistic structure** (“**microstructure**”)
(tokens; parts of speech; syntactic structures; features/phenomena specific for the genre(s) or the research question)

3) **The metadata**
(e.g., data describing the nature and context of the data: author; date; source; topic; social context; techological context; …)

Layers of describing data in corpora

1) **The document structure** ("**macrostructure**")
   (text genres: scientific article; dramatic text; correspondence; manuscript; spoken lanuage transcript; …)

2) **The linguistic structure** ("**microstructure**")
   (tokens; parts of speech; syntactic structures; features/phenomena specific for the genre(s) or the research question)

Main focus: genres of *written* CMC

CMC ‘*macro-*’ and ‘*microstructures*’

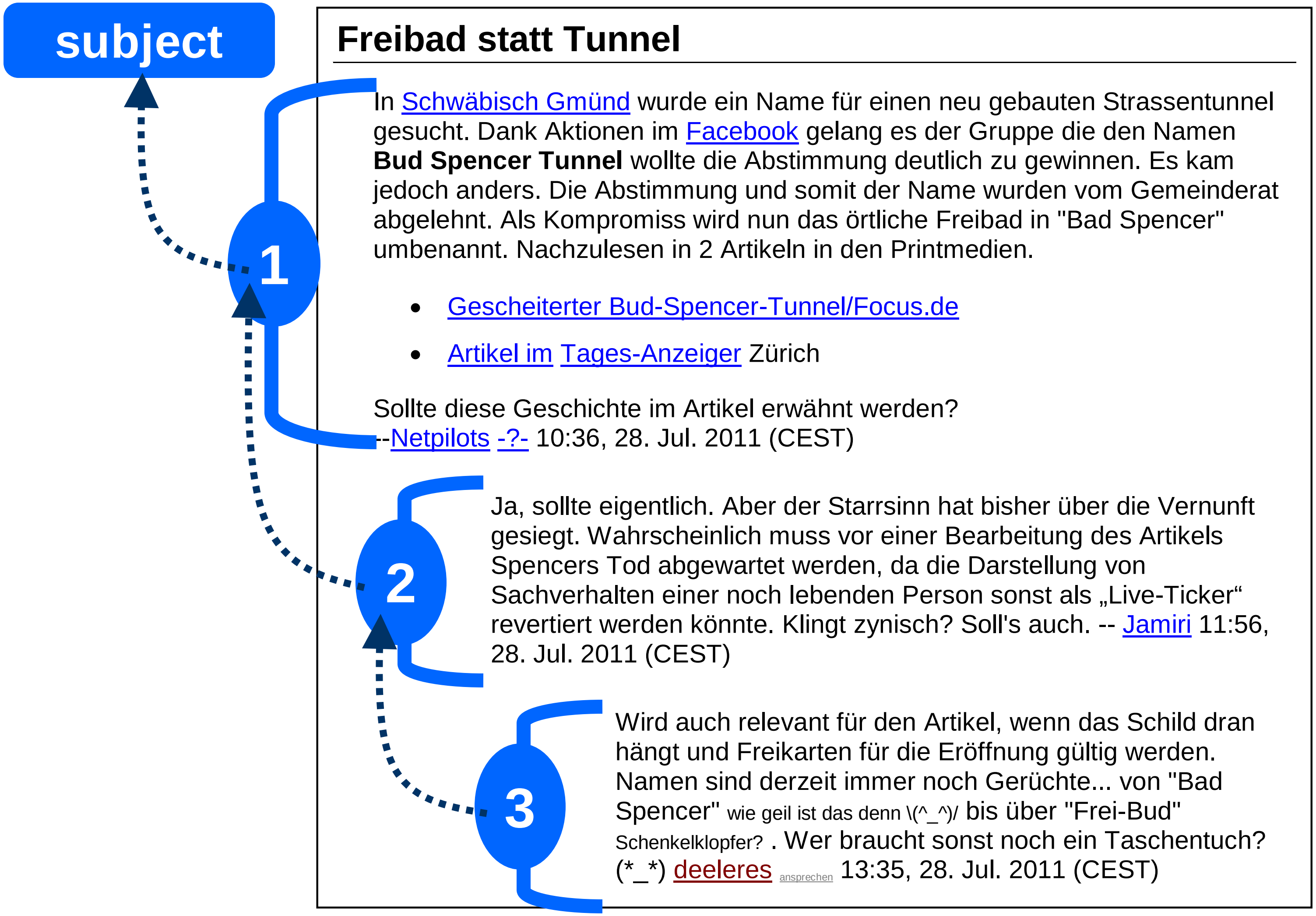
CMC macrostructures:

structure of the interaction as documented on the screen and/or in one corpus document.

The building blocks of CMC macrostructures are the individual **user posts**.

CMC macro structures are not created by one participant alone but as an interactional achievement of all participants *plus* the system.

CMC *‘micro-’* and *‘macrostructures’*



Freibad statt Tunnel

In Schwäbisch Gmünd wurde ein Name für einen neu gebauten Strassentunnel gesucht. Dank Aktionen im Facebook gelang es der Gruppe die den Namen **Bud Spencer Tunnel** wollte die Abstimmung deutlich zu gewinnen. Es kam jedoch anders. Die Abstimmung und somit der Name wurden vom Gemeinderat abgelehnt. Als Kompromiss wird nun das örtliche Freibad in "Bad Spencer" umbenannt. Nachzulesen in 2 Artikeln in den Printmedien.

- Gescheiterter Bud-Spencer-Tunnel/Focus.de
- Artikel im Tages-Anzeiger Zürich

Sollte diese Geschichte im Artikel erwähnt werden?
--Netpilots -?- 10:36, 28. Jul. 2011 (CEST)

Ja, sollte eigentlich. Aber der Starrsinn hat bisher über die Vernunft gesiegt. Wahrscheinlich muss vor einer Bearbeitung des Artikels Spencers Tod abgewartet werden, da die Darstellung von Sachverhalten einer noch lebenden Person sonst als „Live-Ticker“ revertiert werden könnte. Klingt zynisch? Soll's auch. -- Jamiri 11:56, 28. Jul. 2011 (CEST)

Wird auch relevant für den Artikel, wenn das Schild dran hängt und Freikarten für die Eröffnung gültig werden. Namen sind derzeit immer noch Gerüchte... von "Bad Spencer" wie geil ist das denn \(^_^)/ bis über "Frei-Bud" Schenkelklopfer? . Wer braucht sonst noch ein Taschentuch? (*_*) deeleres ansprechen 13:35, 28. Jul. 2011 (CEST)

CMC ‘*macro-*’ and ‘*microstructures*’

CMC microstructures:

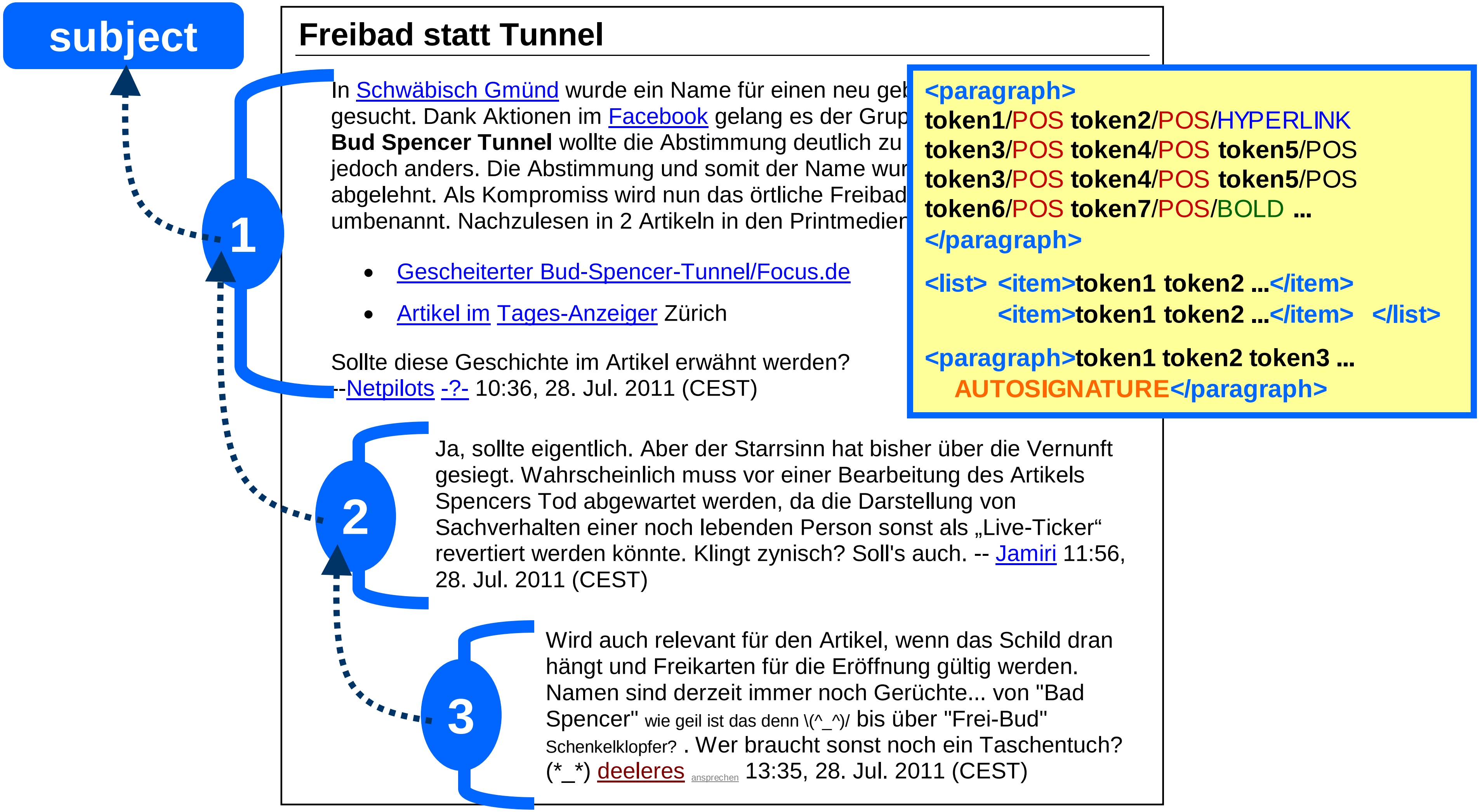
the structure of the **content of a CMC post** – i.e.:

- of the stretch of written text which one participant in the interaction has sent to the server *at once* in oder to make a new contribution

or

- of a stretch of written text created by the system which is displayed using the format of a post (e.g., “system messages” in chats).

CMC *‘micro-’* and *‘macrostructures’*



Freibad statt Tunnel

In Schwäbisch Gmünd wurde ein Name für einen neu geb
gesucht. Dank Aktionen im Facebook gelang es der Grup
Bud Spencer Tunnel wollte die Abstimmung deutlich zu
jedoch anders. Die Abstimmung und somit der Name wur
abgelehnt. Als Kompromiss wird nun das örtliche Freibad
umbenannt. Nachzulesen in 2 Artikeln in den Printmedien

- Gescheiterter Bud-Spencer-Tunnel/Focus.de
- Artikel im Tages-Anzeiger Zürich

Sollte diese Geschichte im Artikel erwähnt werden?
--Netpilots -?- 10:36, 28. Jul. 2011 (CEST)

Ja, sollte eigentlich. Aber der Starrsinn hat bisher über die Vernunft gesiegt. Wahrscheinlich muss vor einer Bearbeitung des Artikels Spencers Tod abgewartet werden, da die Darstellung von Sachverhalten einer noch lebenden Person sonst als „Live-Ticker“ revertiert werden könnte. Klingt zynisch? Soll's auch. -- Jamiri 11:56, 28. Jul. 2011 (CEST)

Wird auch relevant für den Artikel, wenn das Schild dran hängt und Freikarten für die Eröffnung gültig werden. Namen sind derzeit immer noch Gerüchte... von "Bad Spencer" wie geil ist das denn \(^_^)/ bis über "Frei-Bud" Schenkelklopfer? . Wer braucht sonst noch ein Taschentuch? (*_*) deeleres ansprechen 13:35, 28. Jul. 2011 (CEST)

```
<paragraph>
token1/POS token2/POS/HYPERLINK
token3/POS token4/POS token5/POS
token3/POS token4/POS token5/POS
token6/POS token7/POS/BOLD ...
</paragraph>

<list> <item>token1 token2 ...</item>
       <item>token1 token2 ...</item>  </list>

<paragraph>token1 token2 token3 ...
   AUTOSIGNATURE</paragraph>
```

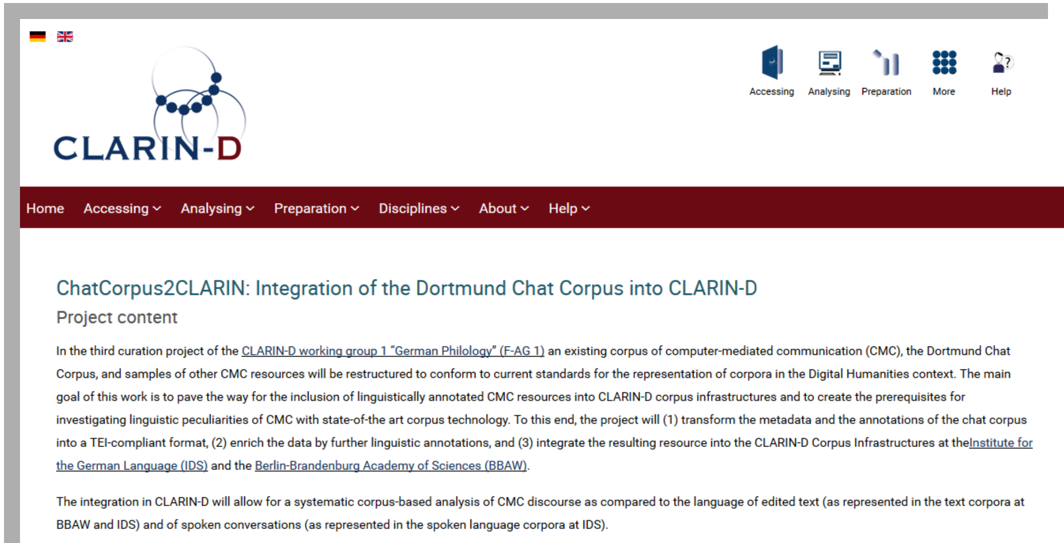
ChatCorpus2CLARIN: Project background

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.



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).

http://www.clarin-d.de/en/curation-project-1-3-german-philology

The corpus

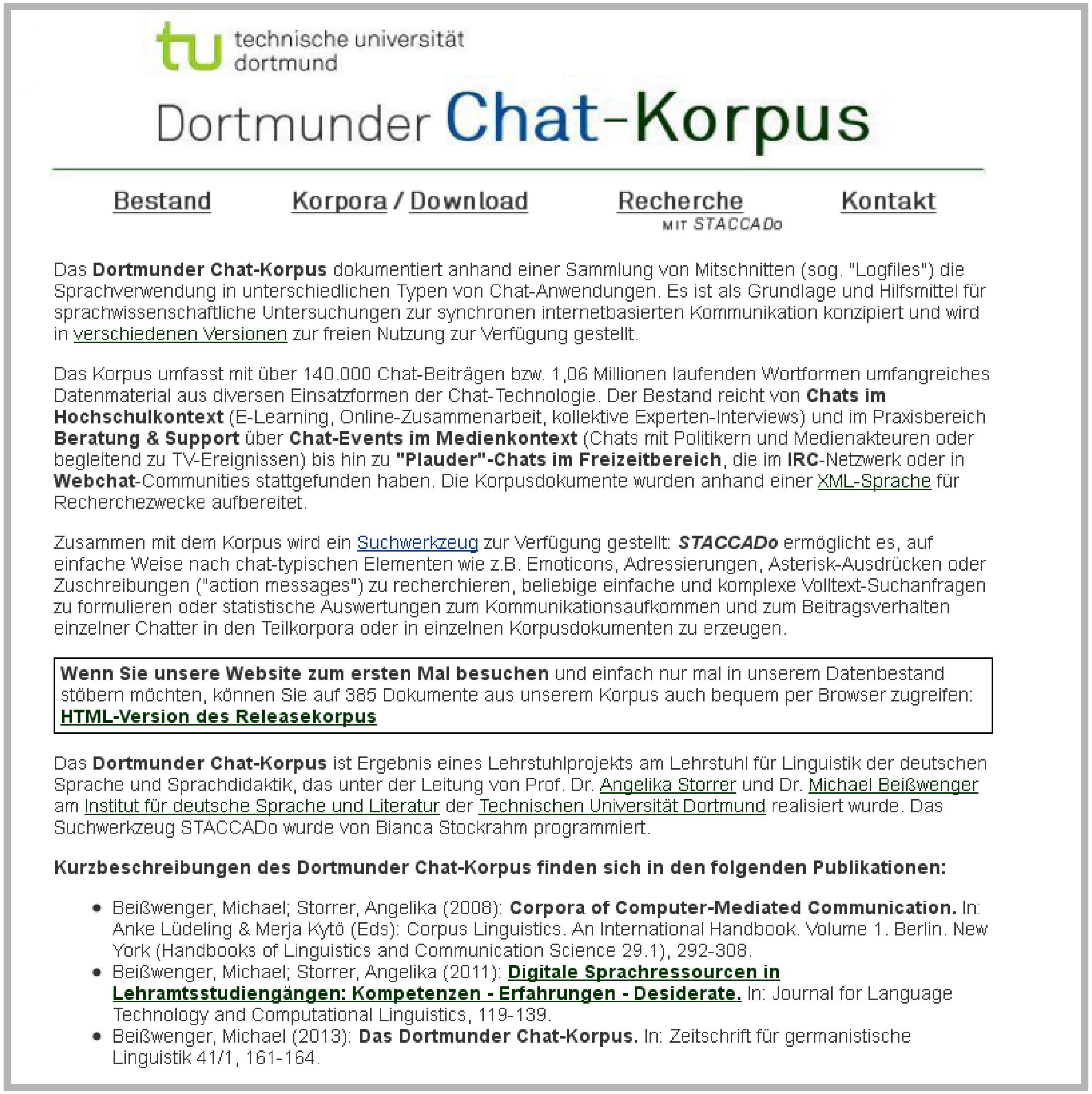
Dortmund Chat Corpus

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.



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 (Data collection „What's up, Deutschland?“ 2014/15, directed by Beat Siebenhaar/ U Leipzig)
- German Wikipedia corpus in DeReKo (IDS Mannheim)
- German News Corpus in DeReKo (IDS Mannheim)
- DWDS Blog Corpus (BBAW Berlin)

Jetzt deine WhatsApp-Nachrichten für die Wissenschaft spenden!

17x Gutscheine gewinnen!

What's up, Deutschland?

www.whatsup-deutschland.de

Sprachwissenschaftler_innen an den Universitäten Leipzig, Dortmund, Dresden, Duisburg-Essen, Hannover, Koblenz-Landau und Mannheim wollen untersuchen, wie man in Deutschland WhatsApp-Nachrichten schreibt.

Auf www.whatsup-deutschland.de findest du:

- Anleitungen zur Nachrichtenspende
- unsere Datenschutzrichtlinien
- mehr Informationen zum Gewinnspiel

Foto „Texting in the Rain“ bearbeitet von Garry Knight (by 2.0).
Foto „Teen Texting“ bearbeitet von Daniel Foster (by-nc-sa 2.0).

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 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.

Work packages (amongst others):

- Specify an XML schema for the representation of structural information on the macro- and microstructure level of CMC and convert the corpus into the target format
- Add part-of-speech information on the microstructure level

Ways to handle the lack of standards for CMC corpora

Create your own, unique XML schema or tag set

(eHumanities “1.0”)

👍 schema/tag set perfectly fits with the needs of the individual project

👎 schema/tag set is idiosyncratic, resource (corpus) is not interoperable with other resources

↕ vs

Comply with a standard (eHumanities “2.0”)

👎 compliance with an existing standard restricts the freedom to design everything in a way that perfectly fits for the peculiarities of CMC discourse

👍
- facilitates the building of corpora (availability of schemas, best practices, and tools)
- sustainability of resources
- interoperability of resources (with corpora of the same type and with corpora of other types)

⇨ Advanced opportunities for empirical research

Ways to handle the lack of standards for CMC corpora

Create your own, unique XML schema or tag set

(eHumanities “1.0”)

👍 schema/tag set perfectly fits with the needs of the individual project

👎 schema/tag set is idiosyncratic, resource (corpus) is not interoperable with other resources

vs

Comply with a standard (eHumanities “2.0”)

perspective: from best practice to standardization

👎 compliance with an existing standard restricts the freedom to design everything in a way that perfectly fits for the peculiarities of CMC discourse

👍

- facilitates the building of corpora (availability of schemas, best practices, and tools)
- sustainability of resources
- interoperability of resources (with corpora of the same type and with corpora of other types)

⇨ Advanced opportunities for empirical research

Complying with a well-established standard increases the chance that the community takes notice of the need to adapt the standard to a new subject.

⇨ extension of the standard?

Challenge I: Representation of structural information on the macro and micro level of CMC genres

Annotation framework provided by the **Text Encoding Initiative (TEI)**: *De-facto* standard in the field of Digital Humanities.

www.tei-c.org

- widely used interchange format for a variety of genres and document types (1st version of the TEI guidelines: *1990*) ⇨ interoperability of resources
- In their current version, the TEI encoding guidelines don’t include models for the representation of CMC – but the framework offers a broad sortiment of models for diverse text genres:
  - genres of edited text
  - transcriptions of spoken language
  - performance texts
  - correspondence
  - manuscript editions
  - (...)

Challenge I: Representation of structural information on the macro and micro level of CMC genres

What makes the annotation framework provided by the TEI an attractive starting point for modeling CMC genres:

Text Encoding Initiative

www.tei-c.org

- Very lively community organized in several special interest groups and workgroups which are continuously developing solutions for adapting the guidelines to new usage contexts and genres.
- **The TEI framework allows for a flexible adaptation to new genres and document types ("customization"):**

"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**."

Challenge I: Representation of structural information on the macro and micro level of CMC genres

Customizing the TEI guidelines for written CMC:

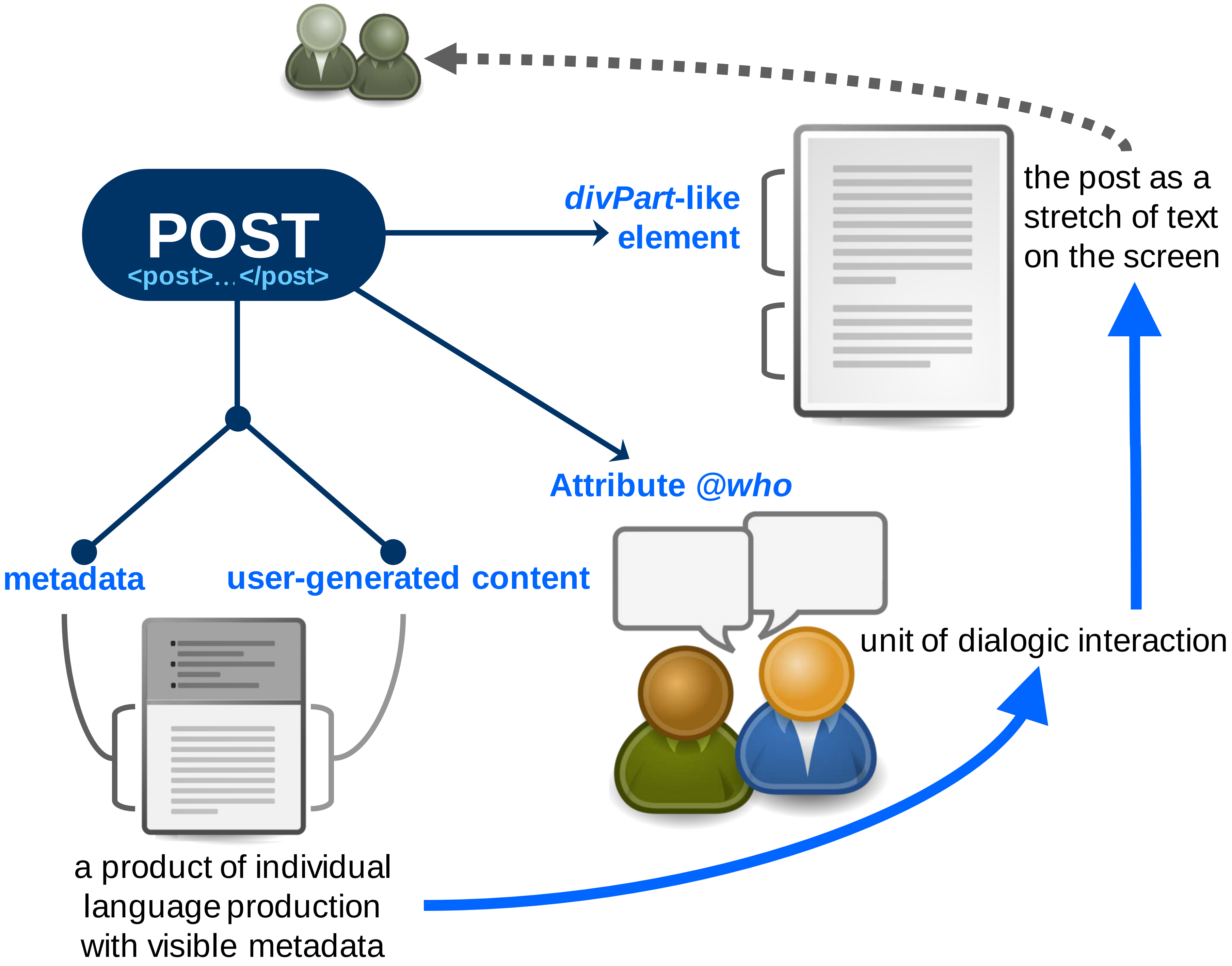
Starting points:

TEI schema drafts by Beißwenger et al. (2012) and Chanier et al. (2014)

The challenge:

The basic units of interaction on the macro level – the posts – share characteristics both with written texts and with utterances in spoken conversations.

⇨ Neither the models *pararaph* or *division* (building blocks of text structure in TEI) nor the model of the *utterance* (building blocks of transcribed speech in TEI) are useful to describe the caracteristics of CMC posts.



POST

<post>…</post>

divPart-like element

the post as a stretch of text on the screen

Attribute @*who*

metadata

user-generated content

unit of dialogic interaction

a product of individual language production with visible metadata

Modeling thread and logfile structures

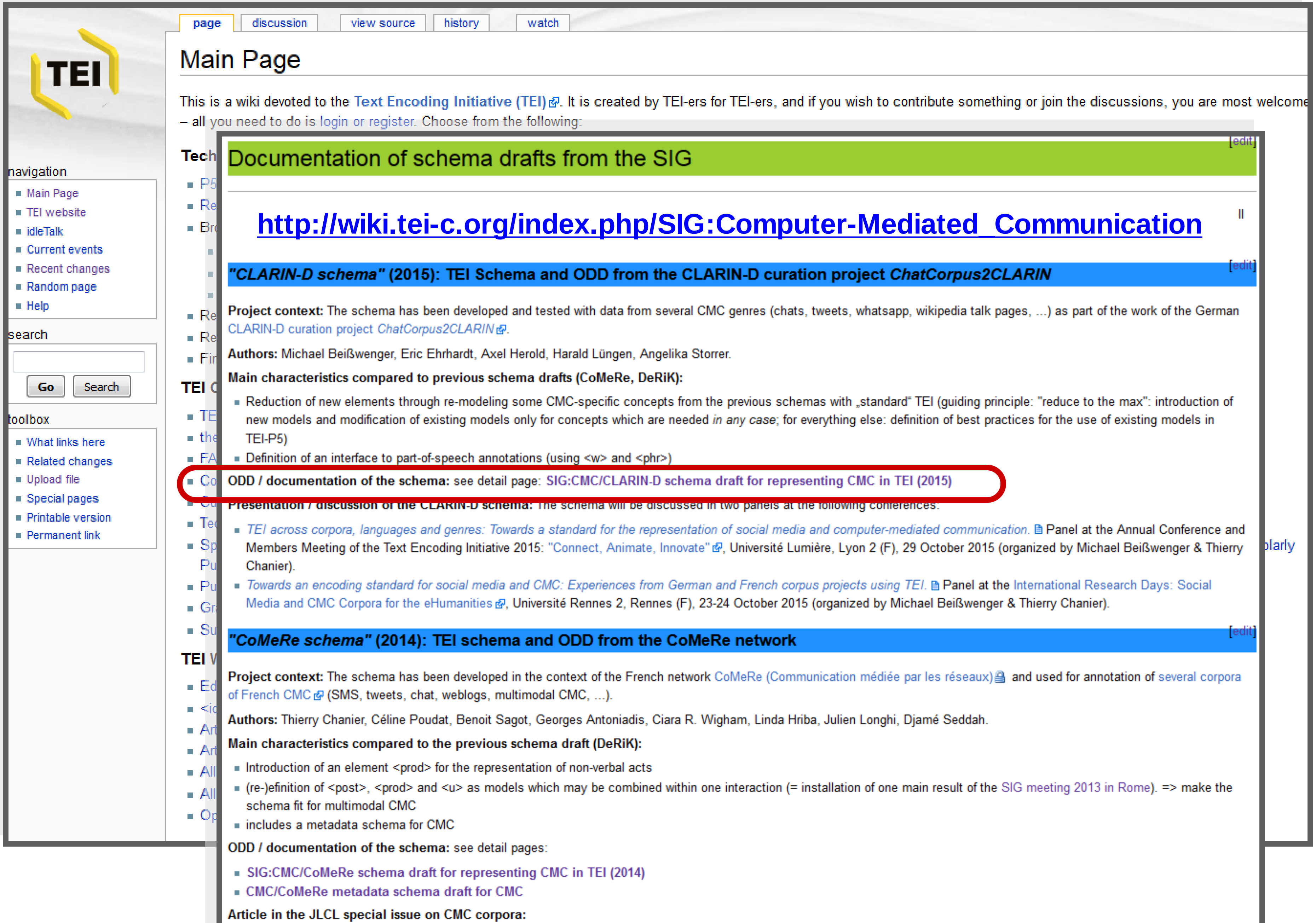
<post> attribute @replyTo:

indicates to which previous post the current post replies or refers to.

+ **Best practice for the use of the TEI standard model <div>** (typically used for the annotation of divisions in edited text):

“division in a CMC document” =
a CMC macrostructure, i.e.: a unit that consists of at least one post and typically of several posts
(types: *logfile*, *thread*, ...)

Schema drafts of the TEI-SIG on CMC



Main Page

This is a wiki devoted to the Text Encoding Initiative (TEI). It is created by TEI-ers for TEI-ers, and if you wish to contribute something or join the discussions, you are most welcome – all you need to do is login or register. Choose from the following:

Documentation of schema drafts from the SIG

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

"CLARIN-D schema" (2015): TEI Schema and ODD from the CLARIN-D curation project *ChatCorpus2CLARIN*

Project context: The schema has been developed and tested with data from several CMC genres (chats, tweets, whatsapp, wikipedia talk pages, ...) as part of the work of the German CLARIN-D curation project *ChatCorpus2CLARIN*.

Authors: Michael Beißwenger, Eric Ehrhardt, Axel Herold, Harald Lüngen, Angelika Storrer.

Main characteristics compared to previous schema drafts (CoMeRe, DeRiK):

- Reduction of new elements through re-modeling some CMC-specific concepts from the previous schemas with „standard" TEI (guiding principle: "reduce to the max": introduction of new models and modification of existing models only for concepts which are needed *in any case*; for everything else: definition of best practices for the use of existing models in TEI-P5)
- Definition of an interface to part-of-speech annotations (using <w> and <phr>)

ODD / documentation of the schema: see detail page: **SIG:CMC/CLARIN-D schema draft for representing CMC in TEI (2015)**

Presentation / discussion of the CLARIN-D schema: The schema will be discussed in two panels at the following conferences:

- *TEI across corpora, languages and genres: Towards a standard for the representation of social media and computer-mediated communication.* Panel at the Annual Conference and Members Meeting of the Text Encoding Initiative 2015: "Connect, Animate, Innovate", Université Lumière, Lyon 2 (F), 29 October 2015 (organized by Michael Beißwenger & Thierry Chanier).
- *Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI.* Panel at the International Research Days: Social Media and CMC Corpora for the eHumanities, Université Rennes 2, Rennes (F), 23-24 October 2015 (organized by Michael Beißwenger & Thierry Chanier).

"CoMeRe schema" (2014): TEI schema and ODD from the CoMeRe network

Project context: The schema has been developed in the context of the French network CoMeRe (Communication médiée par les réseaux) and used for annotation of several corpora of French CMC (SMS, tweets, chat, weblogs, multimodal CMC, ...).

Authors: Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, Djamé Seddah.

Main characteristics compared to the previous schema draft (DeRiK):

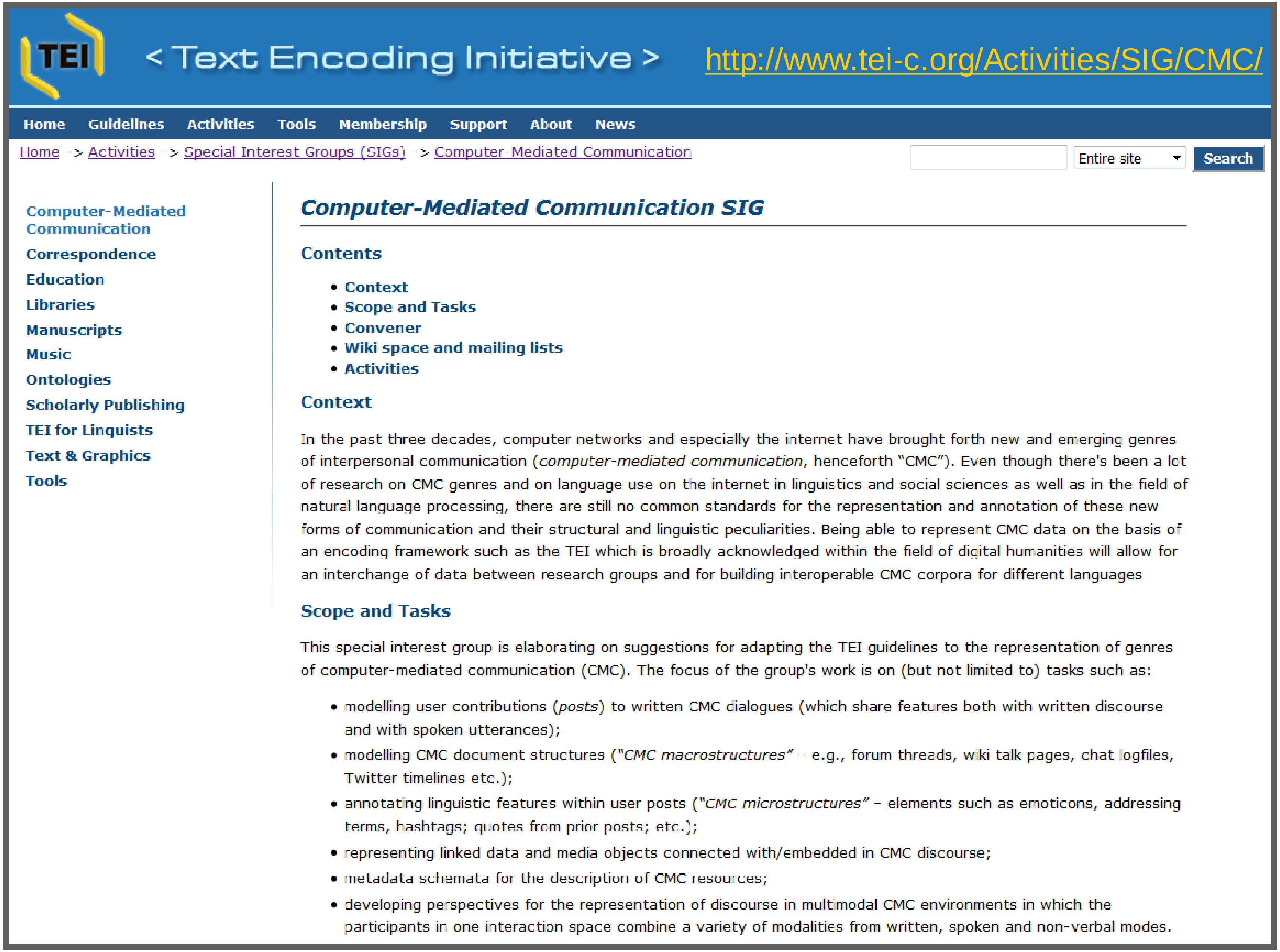
- Introduction of an element <prod> for the representation of non-verbal acts
- (re-)efinition of <post>, <prod> and <u> as models which may be combined within one interaction (= installation of one main result of the SIG meeting 2013 in Rome). => make the schema fit for multimodal CMC
- includes a metadata schema for CMC

ODD / documentation of the schema: see detail pages:

- **SIG:CMC/CoMeRe schema draft for representing CMC in TEI (2014)**
- **CMC/CoMeRe metadata schema draft for CMC**

Article in the JLCL special issue on CMC corpora:

TEI Special Interest Group (SIG) on CMC (since 2013)


TEI
< 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:

- modelling user contributions (*posts*) to written CMC dialogues (which share features both with written discourse and with spoken utterances);
- modelling CMC document structures (*"CMC macrostructures"* – e.g., forum threads, wiki talk pages, chat logfiles, Twitter timelines etc.);
- annotating linguistic features within user posts (*"CMC microstructures"* – elements such as emoticons, addressing terms, hashtags; quotes from prior posts; etc.);
- representing linked data and media objects connected with/embedded in CMC discourse;
- metadata schemata for the description of CMC resources;
- developing perspectives for the representation of discourse in multimodal CMC environments in which the participants in one interaction space combine a variety of modalities from written, spoken and non-verbal modes.

Challenge II: Part-of-speech annotations for the microlevel of CMC posts (using NLP tools & tag sets)

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);
- speedwriting phenomena (typos, omission of characters, norm-deviating use of whitespace);
- colloquial (*Wazzup?*) and creative spellings (*nyce2meetU*)

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)

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)

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

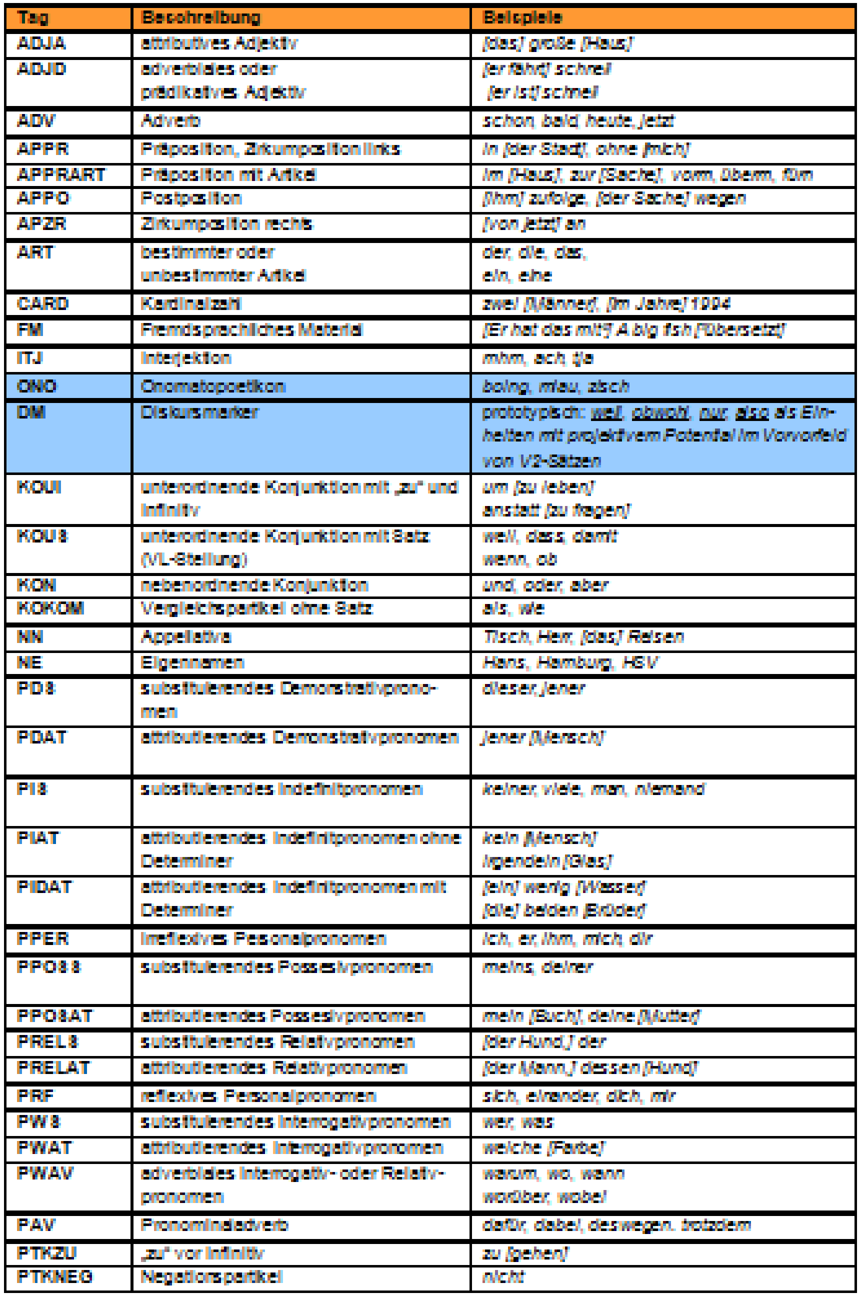
⇨ https://sites.google.com/site/empirist2015/home (supported by GSCL)

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

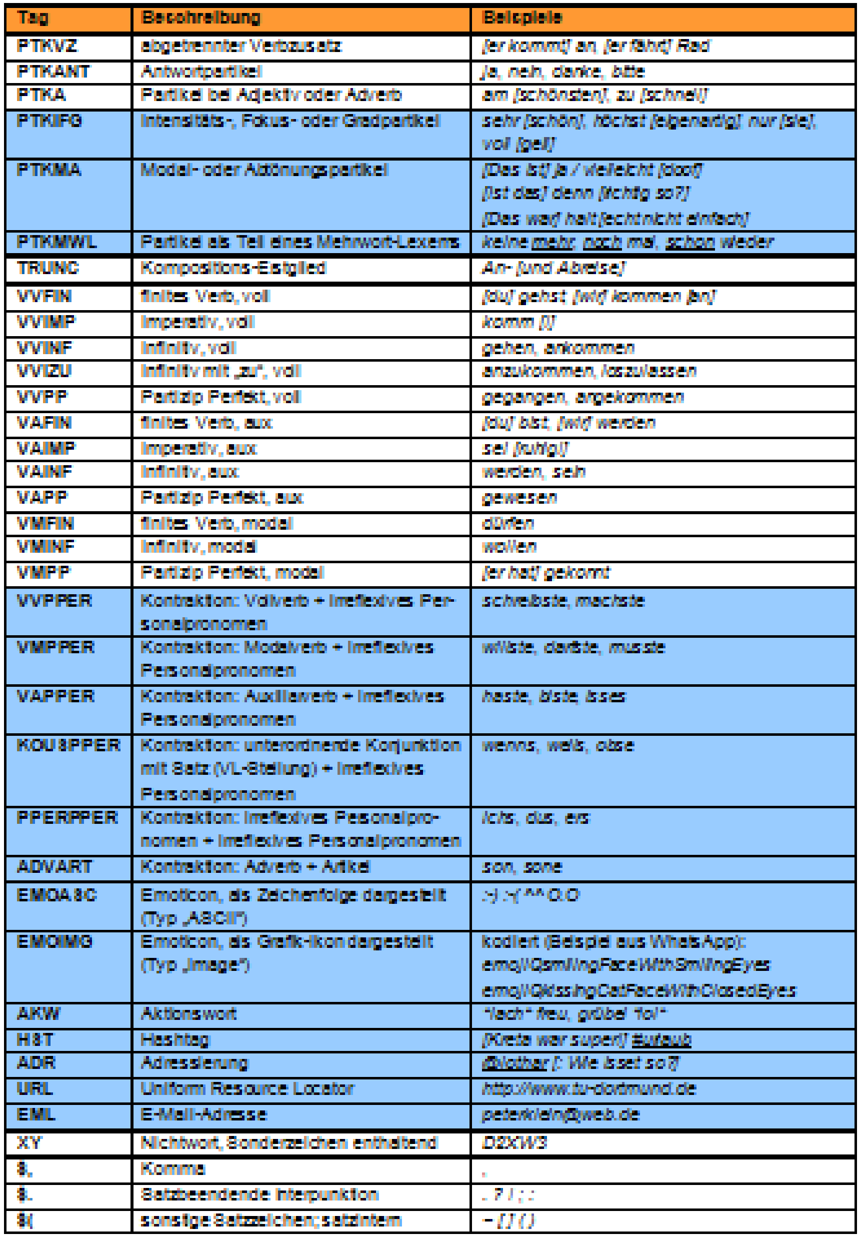
- **Basis:** The “Stuttgart Tübingen Tagset” (STTS): de-facto standard for German (focused on PoS tags for the language occuring in edited text / newspaper texts) (Schiller et al. 1999)
- **“STTS 2.0”:** canonical STTS extended with new categories, but still downward-compatible with STTS (1999)
- Compatible 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], vorm, überm, 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	Appellativa	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 Possesivpronomen	meins, deiner
PPOSAT	attribuierendes Possesivpronomen	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 [geil]
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	willste, 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 *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
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:

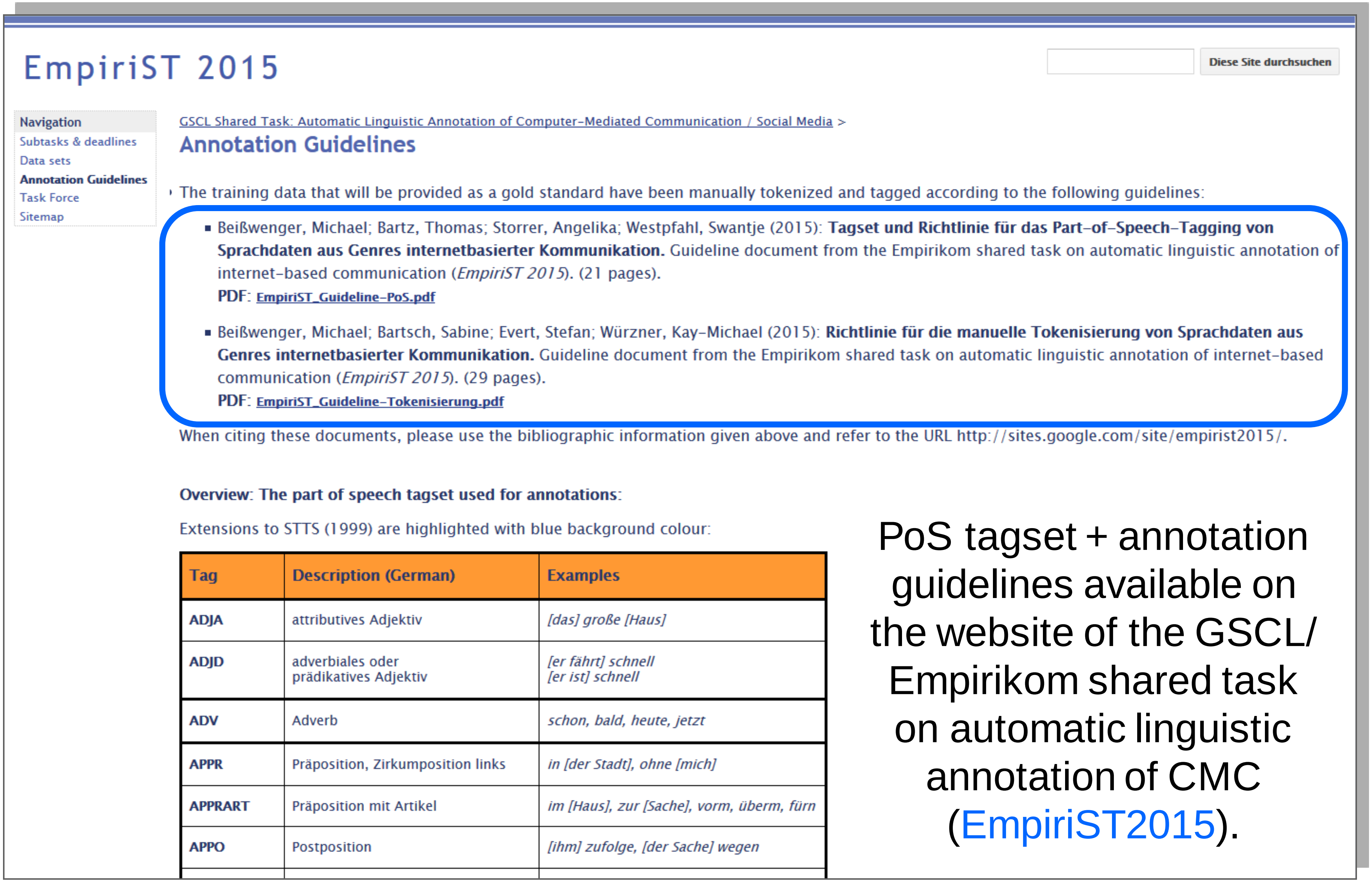
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART is already existing in STTS 1999)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone

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

Tag set and annotation guidelines *@EmpiriST2015*



EmpiriST 2015

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/

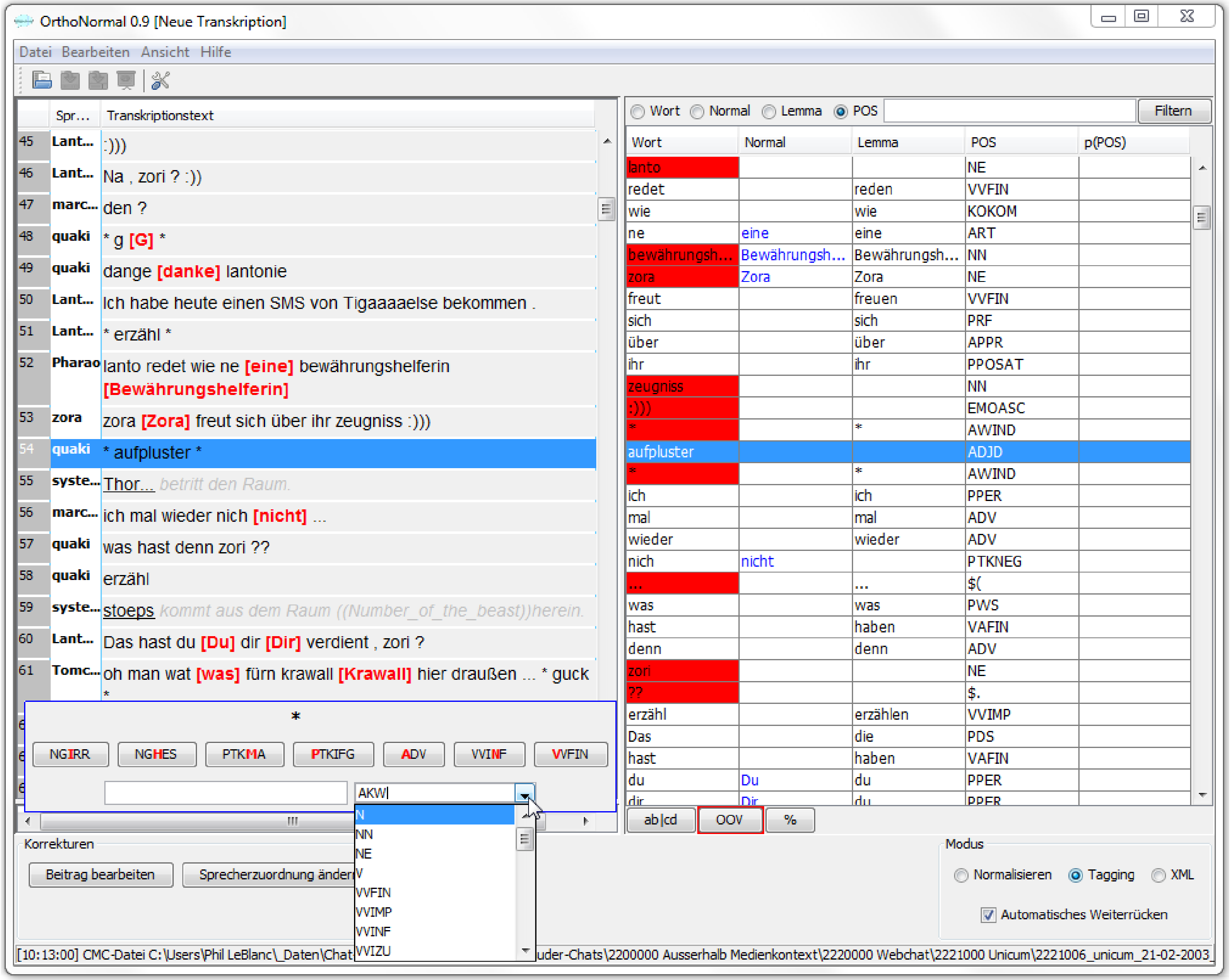
PoS annotation of the CLARIN-D project: workflow

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 used: previous version of “STTS 2.0” (Bartz et al. 2014)

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

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



	Spr…	Transkriptionstext
45	Lant...	:)))
46	Lant...	Na , zori ? :))
47	marc...	den ?
48	quaki	* g [G] *
49	quaki	dange [danke] lantonie
50	Lant...	Ich habe heute einen SMS von Tigaaaaelse bekommen .
51	Lant...	* erzähl *
52	Pharao	lanto redet wie ne [eine] bewährungshelferin [Bewährungshelferin]
53	zora	zora [Zora] freut sich über ihr zeugniss :)))
54	quaki	* aufpluster *
55	syste...	Thor... betritt den Raum.
56	marc...	ich mal wieder nich [nicht] ...
57	quaki	was hast denn zori ??
58	quaki	erzähl
59	syste...	stoeps kommt aus dem Raum ((Number_of_the_beast))herein.
60	Lant...	Das hast du [Du] dir [Dir] verdient , zori ?
61	Tomc...	oh man wat [was] fürn krawall [Krawall] hier draußen ... * guck *

○ Wort ○ Normal ○ Lemma ● POS

Wort	Normal	Lemma	POS	p(POS)
lanto			NE	
redet		reden	VVFIN	
wie		wie	KOKOM	
ne	eine	eine	ART	
bewährungsh...	Bewährungsh...	Bewährungsh...	NN	
zora	Zora	Zora	NE	
freut		freuen	VVFIN	
sich		sich	PRF	
über		über	APPR	
ihr		ihr	PPOSAT	
zeugniss			NN	
:)))			EMOASC	
*		*	AWIND	
aufpluster			ADJD	
*		*	AWIND	
ich		ich	PPER	
mal		mal	ADV	
wieder		wieder	ADV	
nich	nicht		PTKNEG	
...		...	$(	
was		was	PWS	
hast		haben	VAFIN	
denn		denn	ADV	
zori			NE	
??			$.	
erzähl		erzählen	VVIMP	
Das		die	PDS	
hast		haben	VAFIN	
du	Du	du	PPER	
dir	Dir	du	PPER	

Modus

○ Normalisieren ● Tagging ○ XML

☑ Automatisches Weiterrücken

[10:13:00] CMC-Datei C:\Users\Phil LeBlanc_Daten\Chat uder-Chats\2200000 Ausserhalb Medienkontext\2220000 Webchat\2221000 Unicum\2221006_unicum_21-02-2003

(Overview of the FOLK tools: Schmidt 2012)

Vision (1): The CLARIN-D chat corpus as a showcase

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

- 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;
- *((may probably serve as a model – or at least an example – for building and representing other German CMC corpora…..))*

Vision (2): Building a “community of best pactices”

Network of corpus projects which are developing “best practices” for *their* data and *their* research questions:

- to learn from each other through exchange of expertise, experiences, resources and tools,
- as an opportunity to find out where projects could agree upon using similar practices for similar challenges / where solutions from one project can be adopted by other projects and for other languages,
- to create show cases for what one can gain through an interoperability between CMC corpora *for different languages and genres* (⇨ new research options).

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

Thierry Chanier, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi, Djamé Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Journal of Language Technology and Computational Linguistics 28 (2).

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

Linguistic annotation of social media corpora:

To what extent do we have to adapt existing encoding standards and tag sets?

michael.beisswenger@tu-dortmund.de

technische universität
dortmund

Thank you very much for your attention! :-)

JANES

University of Ljubljana
26. November 2015

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):

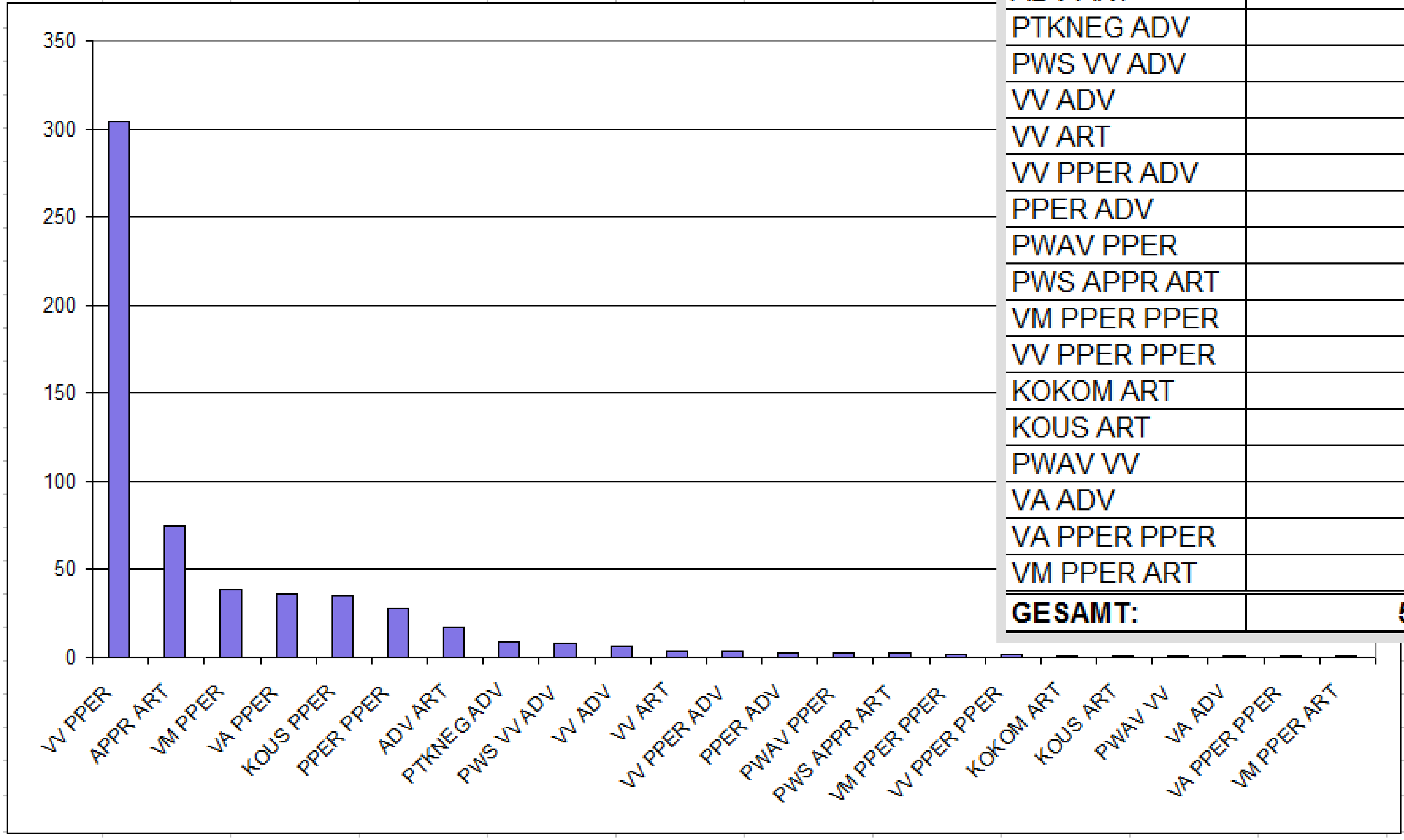
http://wiki.tei-c.org/index.php/SIG:CMC/CLARIN-D_schema_draft_for_representing_CMC_in_TEI_(2015)

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

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**


VV PPER
APPR ART
VM PPER
VA PPER
KOUS PPER
PPER PPER
ADV ART
PTKNEG ADV
PWS VV ADV
VV ADV
VV ART
VV PPER ADV
PPER ADV
PWAV PPER
PWS APPR ART
VM PPER PPER
VV PPER PPER
KOKOM ART
KOUS ART
PWAV VV
VA ADV
VA PPER PPER
VM PPER ART