Linguistic annotation of CMC and social media corpora: To what extent do we have to adapt existing encoding standards and tag sets?

Michael Beißwenger

Slovenščina na spletu in v novih medijih

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 language 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; technological context; …)
Layers of describing data in corpora

1) The document structure
(text genres: scientific article; dramatic text; correspondence; manuscript; spoken language 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; technological 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 macrostructures are not created by one participant alone but as an interactional achievement of all participants plus the system.
Freibad statt Tunnel


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

Sollte diese Geschichte im Artikel erwähnt werden?


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 order 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).
Freibad statt Tunnel

In Schwäbisch Gmünd wurde ein Name für einen neu gebauten Strassentunnel gesucht. Dank Aktionen im Facebook gelang es der Gruppe Bud Spencer Tunnel wollte die Abstimmung deutlich zu gewinnen. Es kam jedoch anders. Die Abstimmung und somit der Name wurde 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?


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.

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.
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)
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

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

- 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

Comply with a standard (eHumanities “2.0”)

- 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)

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.

⇒ Advanced opportunities for empirical research

Comply with a standard (eHumanities “1.0”)

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

perspective: from best practice to standardization

⇒ 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.

- 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
  - (...)

www.tei-c.org
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:

- 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.”
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 *paraph* 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 characteristics of CMC posts.
The post as a stretch of text on the screen.

**POST**

```
<post>...</post>
```

- **divPart-like element**
- **Attribute @who**
- **metadata**
- **user-generated content**
- **a product of individual language production with visible metadata**

Unit of dialogic interaction.
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


"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 Beilwenger, Eric Ehhardt, Axel Herold, Harald Lungen, Angelika Storrer.

Main characteristics compared to previous schema drafts (CoMeRe, DeRiK):
- Reduction of new elements through re-modeling of 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 <x> and <phr>.


"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 médieé par les réseaux) and used for annotation of several corpora of French CMC (SMS, tweets, chat, weblogs, multimodal CMC, ...).

Authors: Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Clara R. Wigham, Linda Hraba, Julien Longhi, Djamel Seddah.

Main characteristics compared to the previous schema draft (DeRiK):
- Introduction of an element <prod> for the representation of non-verbal acts
- (re-)definition 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).
- 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)
- CoMeRe metadata schema draft for CMC

Article in the JCL special issue on CMC corpora:
Computer-Mediated Communication SIG

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


⇒ 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)
Basis: The “Stuttgart Tübingen Tagset” (STTS): de-facto standard for German (focused on PoS tags for the language occurring 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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Beschreibung</th>
<th>Beispiele</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJA</td>
<td>attributives Adjectiv</td>
<td>[das] große [Haus]</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverbolat oder prädikatives Adjectiv</td>
<td>[er] fähig [zum Sport]</td>
</tr>
<tr>
<td>APPR</td>
<td>Präposition</td>
<td>In [der Stadt] ohne [einem]</td>
</tr>
<tr>
<td>APPPR</td>
<td>Präposition mit Artikel</td>
<td>im [Haus] zu [der Bank]</td>
</tr>
<tr>
<td>APPPO</td>
<td>Postposition</td>
<td>an [der Straße] von [der Bank]</td>
</tr>
<tr>
<td>APPZR</td>
<td>Zirkumposition rechts</td>
<td>[von] ans [Unternehmen]</td>
</tr>
<tr>
<td>ART</td>
<td>bestimmter oder unbestimmter Artikel</td>
<td>der, die, das, die, ein, eine</td>
</tr>
<tr>
<td>CARD</td>
<td>Kardinalzahl</td>
<td>zwei [Männer] im [Jahr] 2004</td>
</tr>
<tr>
<td>FM</td>
<td>Fremdsprachliches Material</td>
<td>[Er] hat das mit [diesem Material]</td>
</tr>
<tr>
<td>ITJ</td>
<td>Interjektion</td>
<td>nhm, ich, aye, aye</td>
</tr>
<tr>
<td>ONO</td>
<td>Onomatopoeikon</td>
<td>sial, miu, zisch</td>
</tr>
<tr>
<td>DM</td>
<td>Diskursmarker</td>
<td>in [dem Blatt], was, ob, noch als [ein] Element mit [deinem] Potential im Vorfeld von VZ-Sätzen</td>
</tr>
<tr>
<td>KOU</td>
<td>unterordnende Konjunktion mit &quot;zu&quot; und infinitiv</td>
<td>um [zu leben], anzust [zu glauben]</td>
</tr>
<tr>
<td>KOU8</td>
<td>unterordnende Konjunktion (VL-Stellung)</td>
<td>dass, dass, dass, dass, dass, dass, dass, dass, dass, dass</td>
</tr>
<tr>
<td>KON</td>
<td>Nebenordnende Konjunktion mit Satz (VL-Stellung)</td>
<td>und, oder, aber</td>
</tr>
<tr>
<td>KROM</td>
<td>Vorgangsangaben ohne Satz</td>
<td>als, wie</td>
</tr>
<tr>
<td>NN</td>
<td>Nennung</td>
<td>Tisch, Frau, [das] Reisen</td>
</tr>
<tr>
<td>NE</td>
<td>Eigenschaft</td>
<td>Hans, Hamburg, Hessen</td>
</tr>
<tr>
<td>PD8</td>
<td>substituierendes Demonstrativpronomen</td>
<td>dieser, jener</td>
</tr>
<tr>
<td>PDAT</td>
<td>substituierendes Demonstrativpronomen</td>
<td>jener [Mensch]</td>
</tr>
<tr>
<td>PI8</td>
<td>substituierendes Indefinitpronomen</td>
<td>keiner, viele, man, niemand</td>
</tr>
<tr>
<td>PIA</td>
<td>substituierendes Indefinitpronomen ohne Determiner</td>
<td>kein [Mann], Bing [Bild]</td>
</tr>
<tr>
<td>PIDAT</td>
<td>substituierendes Indefinitpronomen mit Determiner</td>
<td>[ein] klein [Wasser], [dies] beiden Brüder</td>
</tr>
<tr>
<td>PPER</td>
<td>Personennennung</td>
<td>Ich, er, ihm, mich, dir</td>
</tr>
<tr>
<td>PPP</td>
<td>substituierendes Possessivpronomen</td>
<td>meiner, deiner, seinen, ihrer, ihres</td>
</tr>
<tr>
<td>PP0AT</td>
<td>attributives Possessivpronomen</td>
<td>mein [Buch], deine [Mutter]</td>
</tr>
<tr>
<td>PP08AT</td>
<td>attributives Possessivpronomen</td>
<td>mein [Buch], deine [Mutter]</td>
</tr>
<tr>
<td>PW8</td>
<td>attributives Possessivpronomen</td>
<td>war, was</td>
</tr>
<tr>
<td>PWAT</td>
<td>attributives Interrogativpronomen</td>
<td>welche [Farbe]</td>
</tr>
<tr>
<td>PWAV</td>
<td>attributives Interrogativpronomen</td>
<td>warum, wo, wann, worüber, welches</td>
</tr>
<tr>
<td>PAV</td>
<td>Pronominaladverb</td>
<td>dafür, dabei, deswegen, jedenfalls</td>
</tr>
<tr>
<td>PTHZU</td>
<td>&quot;zu&quot; vor infinit</td>
<td>zu [gehen]</td>
</tr>
<tr>
<td>PTHNEG</td>
<td>Negationsadverb</td>
<td>nicht</td>
</tr>
</tbody>
</table>

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


# "STTS 2.0": A basic PoS tag set for German CMC

<table>
<thead>
<tr>
<th>PoS tag</th>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EMO ASC</strong></td>
<td>ASCII emoticon</td>
<td>:-) :-( ^ ^ O.O</td>
</tr>
<tr>
<td><strong>EMO IMG</strong></td>
<td>Graphic emoticon</td>
<td>😊 😊 😊</td>
</tr>
<tr>
<td><strong>AKW</strong></td>
<td>Interaction word</td>
<td>&quot;lach&quot;, freu, grübl, &quot;lol&quot;</td>
</tr>
<tr>
<td><strong>HST</strong></td>
<td>Hash tag</td>
<td>Krela war super! #urlaub</td>
</tr>
<tr>
<td><strong>ADR</strong></td>
<td>Addressing term</td>
<td>@lothar: Wie isset so?</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td>Uniform resource locator</td>
<td><a href="http://www.tu-dortmund.de">http://www.tu-dortmund.de</a></td>
</tr>
<tr>
<td><strong>EML</strong></td>
<td>E-mail address</td>
<td><a href="mailto:peterklein@web.de">peterklein@web.de</a></td>
</tr>
</tbody>
</table>

## I. Tags for phenomena which are specific for CMC / social media discourse:

## II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:

<table>
<thead>
<tr>
<th>PoS tag</th>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VV PPER</strong></td>
<td>Tags for types of colloquial contractions which are frequent in CMC</td>
<td>schreibste, machste</td>
</tr>
<tr>
<td><strong>APPR ART</strong></td>
<td>(APPRART is already existing in STTS 1999)</td>
<td>vorm, überm, f ü m</td>
</tr>
<tr>
<td><strong>VM PPER</strong></td>
<td></td>
<td>willste, darfste, musste</td>
</tr>
<tr>
<td><strong>VA PPER</strong></td>
<td></td>
<td>haste, biste, isses</td>
</tr>
<tr>
<td><strong>KOUS PPER</strong></td>
<td></td>
<td>wennns, weils, obse</td>
</tr>
<tr>
<td><strong>PPER PPER</strong></td>
<td></td>
<td>ichs, dus, ers</td>
</tr>
<tr>
<td><strong>ADV ART</strong></td>
<td></td>
<td>son, sone</td>
</tr>
<tr>
<td><strong>PTK IFG</strong></td>
<td>‘Intensitätspartikeln’, ‘Fokuspartikeln’, ‘Gradpartikeln’</td>
<td>sehr schon, höchst eigenartig, nur sie, voll geil</td>
</tr>
<tr>
<td><strong>PTK MA</strong></td>
<td>Modal particles</td>
<td>Das ist ja ! vielleicht doof. Ist das denn richtig so? Das war half echt nicht einfach.</td>
</tr>
<tr>
<td><strong>PTK MWL</strong></td>
<td>Particle as part of a multi-word lexeme</td>
<td>keine mehr, noch mal, schon wieder</td>
</tr>
<tr>
<td><strong>DM</strong></td>
<td>Discourse markers</td>
<td>weil, obwohl, nur, also, … with V2 clauses</td>
</tr>
<tr>
<td><strong>ONO</strong></td>
<td>Onomatopoeia</td>
<td>boing, miau, zisch</td>
</tr>
</tbody>
</table>
“STTS 2.0”: A basic PoS tag set for German CMC

<table>
<thead>
<tr>
<th>PoS tag</th>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMO ASC</td>
<td>ASCII emoticon</td>
<td>:-) :( ^O.O O.O</td>
</tr>
<tr>
<td>EMO IMG</td>
<td>Graphic emoticon</td>
<td>😊😊😊</td>
</tr>
<tr>
<td>AKW</td>
<td>Interaction word</td>
<td><em>lach</em>, freu, grübel, <em>lol</em></td>
</tr>
<tr>
<td>HST</td>
<td>Hash tag</td>
<td>Kreta war super! #urlaub</td>
</tr>
<tr>
<td>ADR</td>
<td>Addressing term</td>
<td>@lothar: Wie isset so?</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform resource locator</td>
<td><a href="http://www.tu-dortmund.de">http://www.tu-dortmund.de</a></td>
</tr>
<tr>
<td>EML</td>
<td>E-mail address</td>
<td><a href="mailto:peterklein@web.de">peterklein@web.de</a></td>
</tr>
</tbody>
</table>

I. Tags for phenomena which are specific for CMC / social media discourse:

II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>VV PPER</td>
<td>Tags for types of colloquial contractions which are frequent in CMC</td>
<td>schreibste, machste</td>
</tr>
<tr>
<td>APPR ART</td>
<td>(APPRART is already existing in STTS 1999)</td>
<td>vorm, überm, fürn</td>
</tr>
<tr>
<td>VM PPER</td>
<td></td>
<td>willste, darfste, musste</td>
</tr>
<tr>
<td>VA PPER</td>
<td></td>
<td>haste, biste, isses</td>
</tr>
<tr>
<td>KOUS PPER</td>
<td></td>
<td>wenns, weils, obse</td>
</tr>
</tbody>
</table>
### II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>VV PPER</td>
<td>Tags for types of colloquial contractions which are frequent in CMC</td>
<td>schreibste, machste</td>
</tr>
<tr>
<td>APPR ART</td>
<td>(APPRART is already existing in STTS 1999)</td>
<td>vorn, übern, fürn</td>
</tr>
<tr>
<td>VM PPER</td>
<td></td>
<td>willste, darfste, musste</td>
</tr>
<tr>
<td>VA PPER</td>
<td></td>
<td>haste, biste, isses</td>
</tr>
<tr>
<td>KOUS PPER</td>
<td></td>
<td>wenns, weils, obse</td>
</tr>
<tr>
<td>PPER PPER</td>
<td></td>
<td>ichts, dus, ers</td>
</tr>
<tr>
<td>ADV ART</td>
<td></td>
<td>son, sone</td>
</tr>
<tr>
<td>PTK IFG</td>
<td>‘Intensitätspartikeln’, ‘Fokuspartikeln’, ‘Gradpartikeln’</td>
<td>sehr schön, höchst eigenartig, nur sie, voll geil</td>
</tr>
<tr>
<td>PTK MA</td>
<td>Modal particles</td>
<td>Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.</td>
</tr>
<tr>
<td>PTK MWL</td>
<td>Particle as part of a multi-word lexeme</td>
<td>keine mehr, noch mal, schon wieder</td>
</tr>
<tr>
<td>DM</td>
<td>Discourse markers</td>
<td>weil, obwohl, nur, also, ... with V2 clauses</td>
</tr>
<tr>
<td>ONO</td>
<td>Onomatopoeia</td>
<td>boing, miau, zisch</td>
</tr>
</tbody>
</table>
Tag set and annotation guidelines @EmpiriST2015

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

(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 practices”

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


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


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

Thank you very much for your attention! :-)

University of Ljubljana
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):

```
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

<table>
<thead>
<tr>
<th>Strukturtyp</th>
<th>Vorkommen</th>
</tr>
</thead>
<tbody>
<tr>
<td>VV PPER</td>
<td>304</td>
</tr>
<tr>
<td>APRR ART</td>
<td>75</td>
</tr>
<tr>
<td>VM PPER</td>
<td>39</td>
</tr>
<tr>
<td>VA PPER</td>
<td>36</td>
</tr>
<tr>
<td>KOUS PPER</td>
<td>35</td>
</tr>
<tr>
<td>PPER PPER</td>
<td>28</td>
</tr>
<tr>
<td>ADV ART</td>
<td>17</td>
</tr>
<tr>
<td>PTKNEG ADV</td>
<td>9</td>
</tr>
<tr>
<td>PWS VV ADV</td>
<td>8</td>
</tr>
<tr>
<td>VV ADV</td>
<td>6</td>
</tr>
<tr>
<td>VV ART</td>
<td>4</td>
</tr>
<tr>
<td>VV PPER ADV</td>
<td>4</td>
</tr>
<tr>
<td>PPER ADV</td>
<td>3</td>
</tr>
<tr>
<td>PWWV PPER</td>
<td>3</td>
</tr>
<tr>
<td>PWS APPR ART</td>
<td>3</td>
</tr>
<tr>
<td>VM PPER PPER</td>
<td>2</td>
</tr>
<tr>
<td>VV PPER PPER</td>
<td>2</td>
</tr>
<tr>
<td>KOOKM ART</td>
<td>1</td>
</tr>
<tr>
<td>KOUS ART</td>
<td>1</td>
</tr>
<tr>
<td>PWWV VV</td>
<td>1</td>
</tr>
<tr>
<td>VA ADV</td>
<td>1</td>
</tr>
<tr>
<td>VA PPER PPER</td>
<td>1</td>
</tr>
<tr>
<td>VM PPER ART</td>
<td>1</td>
</tr>
</tbody>
</table>

GESAMT: 584