Creation of standards for social media corpora: a digital humanities topic par excellence

Michael Beißwenger



CLARIN-PLUS Workshop:

Creation and Use of Social Media Resources

Kaunas, LT, 2017, May 18/19



Social media corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of social media data (especially when it shall be republished in a corpus)
- no standards for data collection (particularly challenging for data from the private sphere, e.g., whatsapp, SMS)
- no standards for representing/annotating the structural and linguistic peculiarities of social media genres
- no established NLP tools which can be used for automatic processing and linguistic annotation (deviation of social media discourse from the written standard)
- Only (very) few corpus resources which are available for the scientific community/the public
 - "Social media gap" in the corpus landscape

The "social media gap" in the corpus landscape



- Weak representation of language from social media environments in the corpus landscape
- ⇒ Issue for language-centered empirical research on social media
- weak representation of contemporary language in the corpus landscape
- ⇒ Issue for everybody who is interested in the analysis and description of variation and trends in contemporary language



In recent years:

 increasing number of projects in several countries which have started addressing these issues with the goal to close the gap and create social media corpora which shall be made available for the community

Overview: Beißwenger, Michael; Chanier, Thierry; Erjavec, Tomaž; Fišer, Darja; Herold, Axel; Lubešic, Nikola; Lüngen, Harald; Poudat, Céline; Stemle, Egon; Storrer, Angelika; Wigham, Ciara (2017): Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries.

Submitted for: Selected Papers from the CLARIN Annual Conference

2016, October 26–28, 2016, Aix-en-Provence, France. Linköping University Electronic Conference Proceedings.



In recent years:

- increasing number of projects in several countries which have started addressing these issues with the goal to close the gap and create social media corpora which shall be made available for the community
- Window of opportunity for joint efforts in creating standards for this new type of corpora in an bottom-up approach:
- Exchange and discussion of best practices, experiences, tools, datasets
- Workshops and other networking events on national + international level
 (<u>like this one!</u>)

A digital humanities topic par excéllence



First-wave digital humanities involved the building of infrastructure in the studying of humanities texts through digital repositories, text markup, etc., whereas **second-wave digital humanities** expands the notional limits of the archive to include digital works, and so bring to bear the humanities' own methodological toolkits to look at 'born-digital' materials [...].

(David M. Berry 2011: 3)

- ⇒ "2nd wave" DH is discovering its **interest in** researching "born-digital" materials.
- ⇒ To do empirical research, **DH needs** *resources*.
- ⇒ To create, share, connect and combine resources, **DH needs** standards.



Social media corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of social media data (especially when it shall be republished in a corpus)
- no standards for data collection (particularly challenging for data from the private sphere, e.g., whatsapp, SMS)
- no standards for representing/annotating the structural and linguistic peculiarities of social media genres
- no established NLP tools which can be used for automatic processing and linguistic annotation (deviation of social media discourse from the written standard)
- Only (very) few corpus resources which are available for the scientific community/the public
 - "Social media gap" in the corpus landscape



Social media corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of social media data (especially when it shall be republished in a corpus)
- no standards for data collection (particularly challenging for data from the private sphere, e.g., whatsapp, SMS)
- no standards for representing/annotating the structural and linguistic peculiarities of social media genres
- no established NLP tools which can be used for automatic processing and linguistic annotation (deviation of social media discourse from the written standard)
- Only (very) few corpus resources which are available for the scientific community/the public
 - "Social media gap" in the corpus landscape



Social media corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of social media data (especially when it shall be republished in a corpus)
- no Summary of a legal opinion sought for the integration of an existing corpus of German
 no chats into the CLARIN-D infrastructure.
 - d linguistic neculiarities of social media genres

Disclaimer:

- I am not a legal expert (all statements subject to correction!)
- The legal opinion refers to German law only.
- The legal opinion refers to that special corpus only.

ChatCorpus2CLARIN: Project background

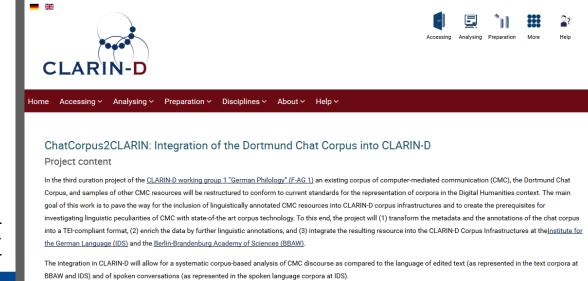
Curation project of the CLARIN-D F-AG 1 "German Philology"

Duration: May 2015 – February 2016

The task: develop a workflow and resources for the integration of an existing chat corpus into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences (http://clarin-d.de).

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

at the CLARIN-D hubs at IDS and BBAW.



CLARIN-D

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

The Dortmund Chat Corpus



- subject: the language used in German chats
- size: 1,06 million tokens
- motivation: document the range of linguistic variation in chats depending on the social domain
 ⇒ subcorpora: social chat, chats in learning and teaching, chat-based institutional advise, chats from the journalistic/media domain

sources of the data: logfiles recorded 1998-2004:

- partially recorded in publicly accessible chat rooms
- partially donated by providers of chat rooms or organisers of chat events in the context of learning and teaching (university) / institutional advise
- partially retrieved from publicly accessible chatlog archives on the web

The Dortmund Chat Corpus



- 80% of all documents, 52% of all tokens available as a free download since 2005 ("release version" of the corpus, partially anonymised), other parts of the resource only to be used in Dortmund http://www.chatkorpus.tu-dortmund.de
- Prerequisite for integration into CLARIN-D: legal clearance of conditions under which the corpus can be integrated into the CLARIN-D corpus infrastructure.
 - ⇒ legal opinion (46 pages) by *iRights.law* (John Weitzmann/Jan Schallaböck, 2016)
 - ⇒ basis: 20-page documentation of all subcorpora + data samples + several skype meetings with the legal experts



Focus: Intellectual property rights (IPR)

Basis: the regulations of the German IPR act (Urheberrecht, UrhG)

Issue:

- In order to qualify for copyright, a creation typically must meet minimal standards of originality (cf. Wikipedia).
- Not only big but also small creations can meet these standards and be considered a "work".
- Do chat posts have to be considered as "works" sensu UrhG?



Focus: Intellectual property rights (IPR)

Basis: the regulations of the German IPR act (Urheberrecht, UrhG)

Issue:

If chat posts have to be considered as "works" *sensu UrhG*, then:

- hosting of the corpus by CLARIN institutions would have to be considered as a violation of the author's copyright (= as a distribution of unauthorized copies).
- In the case of posts collected from non-public chat rooms, the provision via CLARIN could additionally be considered as a first public reproduction and, thus, as a violation of the author's publishing rights.



Focus: Intellectual property rights (IPR)

Basis: the regulations of the German IPR act (*Urheberrecht*, **UrhG**)

Legal opinion:

- The portion of "works" among the posts in the corpus is considered as fairly low (they are typically rather short and do not meet the threshold of originality).
- The chance that anybody actually claims for a IPR violation is be considered as fairly low.
- If anybody claims for an IPR violation, the economic risk is fairly low.
- Dealing with IPR is therefore rather an issue of the public image and model character of the CLARIN institutions than a legal issue.

Recommendation: To be on the safe side, make data from non-public chatrooms available for the scientific world only.



Focus: Data protection

Basis: the regulations of the German Federal data protection act (*Bundesdatenschutzgesetz*, **BDSG**)

Issue:

- How to deal with person-related data (= data that enable individual identification of the respective person)?
- To conform to the restrictions of BDSG, data should be represented in a way that the expenses that one would have to invest to individually identify a certain person are so high that it seems unrealistic that anybody would invest them.
- This is even more important since we do not have explicit (written + signed) consent of the chatters.



Focus: Data protection

Basis: the regulations of the German Federal data protection act (*Bundesdatenschutzgesetz*, **BDSG**)

Legal opinion:

- Evidence for person-related data in the corpus is low. No case in the sample in which a simple online query would have allowed for an identification of the person.
- Residual risk: Chatters may be identified based on what they state in their posts (or their individual style) by people from their individual surroundings.

Recommendation:

- 1) Anonymize all data.
- 2) Make subcorpora with data from non-public sources accessible for researchers only.

Consequences in the project



- Anonymization (partially automatically, but with a lot of manual post-editing)
 - ⇒ time-consuming extra task beyond the funding
- end of the project: May 2016; end of anonymization process: early 2017
- Availability of the corpus:
 - in the repositories of IDS and BBAW in June 2017, divided in two partial corpora (publicly available vs. scientific community only)
 - via the search engine COSMAS II (IDS) and in the DWDS portal (BBAW): summer 2017
 - licence type: CC-BY (recommended by iRights.law)



Social media corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of social media data (especially when it shall be republished in a corpus)
- no standards for data collection (particularly challenging for data from the private sphere, e.g., whatsapp, SMS)
- no standards for representing/annotating the structural and linguistic peculiarities of social media genres
- no established NLP tools which can be used for automatic processing and linguistic annotation (deviation of social media discourse from the written standard)
- Only (very) few corpus resources which are available for the scientific community/the public
 - "Social media gap" in the corpus landscape

CMC-SIG @ TEI





TEI special interest group "computer-mediated communication" http://www.tei-c.org/Activities/SIG/CMC/



mission:

suggest models for representing CMC/social media corpora in TEI format (using ongoing corpus projects as a testbed)

results so far:

three versions of a basic TEI schema: DeRiK (2012), CoMeRe (2014), CLARIN-D (2015)

available online & ready for use

Towards a basic format for representing social media resources



- How can we represent the emerging, 'born-digital' formats of interaction mediated through the online medium?
- Social media technology rapidly changes, and so do the ways how individuals adapt to the technological environments to communicate.

A basic representation format should describe basic interaction formats on a level of abstraction which is relatively robust against the permanent change of technology and communicative practices.

Why do we need a representation standard?

Interoperability



The vision: The representation of resources in compatible formats facilitates

- the combination, merging and comparative exploitation of these resources:
 - social media corpora with other social media corpora (for different languages & different genres)
 - social media corpora with corpora of other types (text corpora, spoken language corpora)

Interoperability



The vision: The representation of resources in compatible formats facilitates

- the combination, merging and comparative exploitation of these resources.
- the integration into national and pan-European common language resource infrastructures (CLARIN).

The availability of a **basic representation and exchange format** for social media corpora is a prerequisite to foster interoperability of that type.

The extension of an existing & well-established representation standard for language resources is better than the creation of a standalone solution for social media corpora.

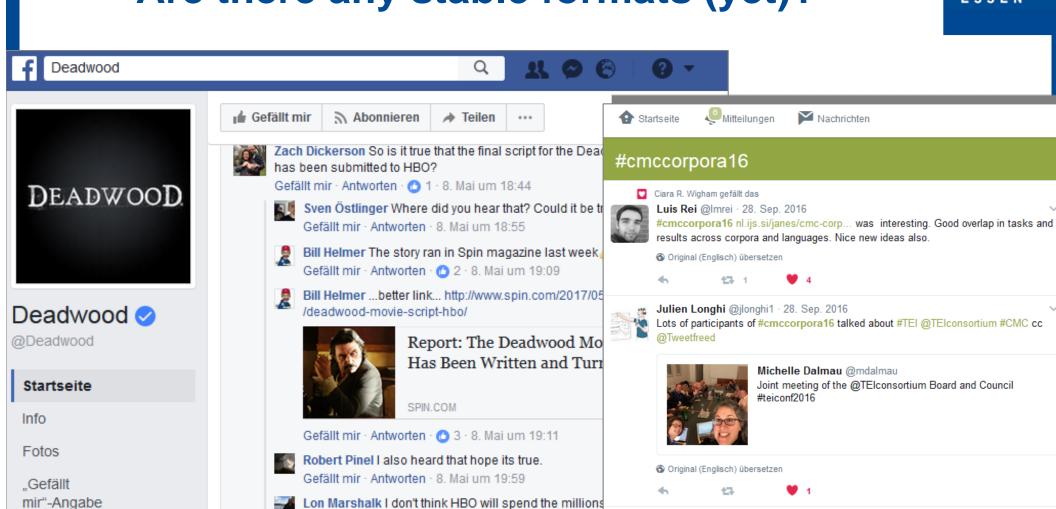
Text Encoding Initiative (TEI)



- de-facto standard for text encoding in the humanities (1st version of the guidelines: 1990); formats for the representation of a broad range of textual genres
- community-driven standard: the community of users evaluate, discuss, optimize and expand the guidelines (coordinated by an elected Council)
- TEI community provides tools for the creation and use of TEI annotation schemas
- The standard can be customised (= modified) for the specific needs of individual projects e.g. for the use in new, innovative domains



Are there any stable formats (yet)?



Teja G. @TheriaQQ 27. Sep. 2016 Great dinner @ASaperitivo #cmccorpora16



history to bring it back. Especially to get all the main ch Videos back. They scr... Mehr anzeigen Beiträge

Gefällt mir · Antworten · 6 2 · 8. Mai um 23:38

Sven Östlinger And... A completely different topic. Why notifications when someone answers here?

millions it will cost them to make a movie or another s

finale or whatever. It would be the most costly show in

Gefällt mir · Antworten · 🙆 1 · 12. Mai um 17:28

Schreibe eine Antwort

Lon Marshalk Go to the top of this page where it says "Following" and then hit click on "see first".

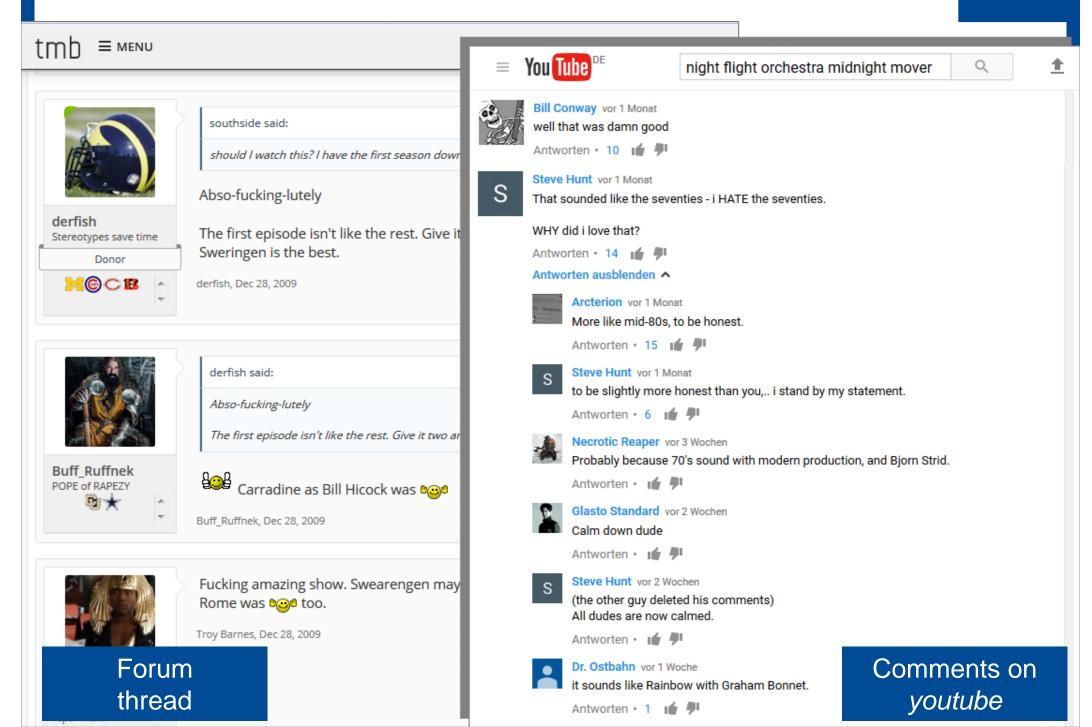
Gefällt mir · Antworten · 🙆 1 · 12. Mai um 19:57 · Bear

Facebook

Discussion on

Eine Seite erstellen

UNIVERSITÄT DUISBURG ESSEN



Are there any stable formats (yet)?

Not logged in Talk Contributions Create account Log in



Main page Contents Featured content Current events Random article Donate to Wikipedia Wikipedia store

Interaction

Help About Wikipedia Community portal Recent changes

Contact page

Tools

What links here Related changes Upload file Special pages Permanent link

Page information

Print/export

Create a book Download as PDF Printable version

Languages



Read Edit New section View history

Search Wikipedia

Talk:Deadwood (TV series)

From Wikipedia, the free encyclopedia

Contents [show]

Updated characters and provided links [edit]

Updated Joanie Stubbs and Trixie and provided links to why these characters were no based on actual people or based on several people. —The preceding unsigned comment added by 146.130.123.33 (talk) 17:05, 1 May 2007 (UTC).

I am gunna put em in alphabetical order —Preceding unsigned comment added by 174.108.8.1 15:12, 13 February 2011 (UTC)

No, do not do that without providing a legitimate reason. Alphabetical order mi give precedence to less notable characters. --- Republican Jacobite The Fortu 17:22, 13 February 2011 (UTC)

New noticeboard [edit]

A new noticeboard, Wikipedia: Fiction noticeboard, has been created. - Peregrine Fis 18:02, 1 May 2007 (UTC)

This noticeboard has been deleted per Wikipedia:Miscellany for deletion/Wikipedia:Fiction noticeboard. Please disregard the above post. ~ Jeff Q 11:26, 9 May 2007 (UTC)

historical divergence??? [edit]

Al Swearengen

... what does it matter, that he came from iowa and not from England?

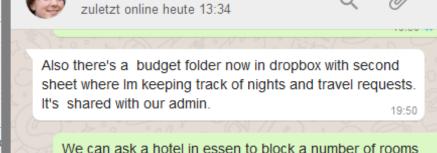
Nowhere in the series (as I remember it) this matter is stated as a fact - only Al prete sometimes to be original english ...

he any divergence since where is stated, that the historical AL never prete

Wikipedia talk page

17 August 2007 (UTC)

wood book by David Milch [edit]



Yes that's what i was thinking. Great! If we get money soon booking early will keep down costs.

Super!

for our event.

19:51

Ciara Wigham

I'll let you know the estimated prices next week, ok?

19:51 🗸

19:51 <//

19:51

Great. Have a good weekend. The embassy also sent me their logo - in dropbox. 19:51

which dropbox folder? Have u already shared it with me?

19:53 </

Yes!

19:53

Hmmm ... what's its name?

https://www.dropbox.com/sh/to /AACl8sv8PrGSwRvW2hr1y90

whatsapp interaction



They consist of **stretches of text** followed by other stretches of text which serve as contributions to an ongoing dialogue but which, different from turns at talk,

- STEP 1: are produced as a whole and in a private activity (= no access for others to the process of verbalisation),
- STEP 2: are sent to the server en bloc,
- STEP 3: are presented on the screen as a product,
- STEP 4: are read and replied by others when the producer (according to his mental representation of events) has already finished his communicative move.

Delayed communication, even in "synchronous" environments: Organisation of interaction using **posts**, not **turns**.



Delayed communication, even in "synchronous" environments: Organisation of interaction using **posts**, not **turns**.

⇒ Effects on the surface of the interaction and in the design of the posts:

'interleaved interactional patterns', 'disrupted adjacency', use of addressings, medium-specific message packaging strategies ('split turns') etc. (Garcia/Baker Jacobs 1998, 1999, Herring 1999, Zitzen/Stein 2002, Storrer 2001, Beißwenger 2003, 2007, Imo 2016, et al.)



Delayed communication, even in "synchronous" environments: Organisation of interaction using **posts**, not **turns**.

⇒ Effects behind the scenes:

Users develop strategies to deal with the temporal shift between the intention to act (= start composing a post) and its presentation to the other parties:

e.g. Beißwenger (2007, 2010):

analysis of 1,100 message production processes (screen capturing + video observation):

19% of all processes finished with complete deletion, in 71% of cases caused by new messages on the screen



"post-based written interaction":

- written language production
- interactional language: participants can switch between reader and writer role; utterances are desgined as contributions to an ongoing dialogic exchange
- logged: interactional context available in a screen protocol (permenantely or at least for the current session)
- pre-transmission composition: The production component resembles more a text-production process than the verbalisation process in spoken language
- basic unit: the post

Even though instances of this format vary over different genres, technologies and platforms, its essential features are stable for more than 25 years.

Intermediate results and next steps



Most recent schema draft ('CLARIN-D TEI schema for CMC') can be used and tested – feedback welcome! http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema

- <post> as a basic unit distinct from spoken utterances (<u>) and units of text structure (<div>,);
- set of attributes for the subclassification of posts;
- models from the TEI core and text structure which can be used for structuring the user generated content of posts;
- Posts can be grouped into cmc-specific division types (logfiles, threads);
- posts can alternate with spoken language and with nonverbal actions on the user interface (-> multimodal CMC).

Next step: "feature request" (2017) => *standardisation....*

CLARIN-ERIC UI Event

eurac research

HOME SPEAKERS CALL FOR PAPERS REGISTATION COMMITTEES VENUE

CMC-CORPORA

Posted: May 15, 2017

How to use TEI for the annotation of CMC and social media resources: a practical introduction

The goal of the event is to give a practical introduction into the annotation of language data from genres of computer-mediated communication (CMC) and social media using TEI. In an introductory section participants will learn about the general architecture of TEI encoding schemas and about rules for the creation of so-called customizations which allow for extending the use of TEI with textual genres and in domains which are not yet covered by the current version of the TEI guidelines. Examples for TEI customizations are the representation schemas for CMC/social media genres developed in the TEI special interest group "computer-mediated communication".

In a hands-on session, participants will learn how to use these customizations to create a basic TEI representation for their own CMC/social media data. For this purpose participants may bring samples from their own data/corpora or select a sample from collections of Wikipedia talk pages in several languages prepared by the instructors. Format specifications for participants' own data will be announced in advance. For the hands-on session, participants will be asked to bring a laptop computer with WLAN and a full or trial license of the oXygen XML editor.

October 04, 2017 – Eurac, Bolzano, Italy

To be announced via the CLARIN Newsflash soon..

5th Conference on CMC and Social Media Corpora for the Humanities





October 02-03, 2017 – Eurac, Bolzano, Italy https://cmc-corpora2017.eurac.edu/