Three-paper panel:

Computer-mediated communication in TEI: What lies ahead

Proposal for: **The Linked TEI: Text Encoding in the Web.** 2013 Annual Conference and Members' Meeting of the TEI Consortium, 2–5 Oct 2013, Rome.

Panel organizers: Michael Beißwenger
TU Dortmund University (michael.beisswenger@tu-dortmund.de)
Lothar Lemnitzer
Berlin-Brandenburg Academy of Sciences and the Humanities (lemnitzer@bbaw.de)

Introduction

The social web has brought forth various genres of interpersonal communication (*computer-mediated communication*, henceforth: *cmc*) such as chats, discussion forums, wiki talk pages, Twitter, comment and discussion threads on weblogs and social network sites. These genres display linguistic and structural peculiarities which differ both from speech and from written text. Projects that want to build and exchange cmc corpora would greatly benefit from a standard that allows the user to annotate these peculiarities in TEI.

From the perspective of several corpus projects which aim at building and annotating cmc corpora for several European languages, this panel will discuss how the models provided by the TEI encoding framework may be adapted to the special requirements of cmc genres.

The basis of the discussion is a customized TEI schema presented at the TEI conference held in Würzburg 2011 (Beißwenger et al. 2012)[1]. The panel papers will elaborate on basic features that a TEI standard for cmc resources should include and outline open issues with which further work will have to deal.

The overall goal of the panel is to stimulate the discussion within the TEI community about how a standard for the representation of cmc in TEI should look like and what might be a practical and reasonable way to go about creating such a standard.

In order to push the development of a general standard for the representation of cmc genres and cmc discourse forward, the papers in the panel will present problem overviews for basic issues in representing cmc features in TEI P5 and outline perspectives as well as first suggestions for the treament of these challenges through modifications and expansions of the encoding framework. Starting from these suggestions, the group is planning to work out feature requests and load them onto the TEI projects page on sourceforge.net.

After a general introduction, paper 1 asserts that solutions for the representation of cmc in TEI should be included in the official TEI guidelines and not remain a task that research and corpus projects have to solve using individual customizations. In addition,

1 The ODD document can be found at http://www.empirikom.net/bin/view/Themen/CmcTEI

the paper formulates general requirements a framework for the representation of cmc (in TEI) should comply with as well as specific requirements from several projects which are currently building corpora of cmc discourse for four European languages (German, Dutch, French, and Italian).

Taking into account the requirements outlined in paper 1, paper 2 starts with an overview of existing suggestions for the representation of basic structural and linguistic features of cmc discourse in the TEI framework. It then presents considerations on the following open issues: (1) the modeling of different types of citations in cmc postings; (2) the modeling of hypermedia features (hyperlinks and linking structures, embedded media objects); (3) challenges related to 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.

Paper 3 examines the issue of metadata. It discusses general requirements for representing metadata of cmc resources and outlines a proposal for representing cmc metadata in the TEI framework.

The panel will include 30 minutes of discussion time (15 minutes each after paper 2 and 3).

Paper 1:

Modeling computer-mediated communication in TEI: requirements and perspectives

Michael Beißwenger[1], Thierry Chanier[2], Isabella Chiari[3], Maria Ermakova[4], Lothar Lemnitzer[4], Angelika Storrer[1], Maarten van Gompel[5], Henk van den Heuvel[5]

[1] TU Dortmund University (D) [2] Université Blaise Pascal, Clermont-Ferrand (F) [3] Università "La Sapienza", Rome (IT) [4] Berlin-Brandenburg Academy of Sciences and the Humanities (D) [5] Radboud University Nijmegen (NL)

This paper reports an ongoing work in a network of corpus projects which aim at building and annotating corpora of computer-mediated communication (cmc)[2] and asserts that a framework for the representation of cmc should become a part of the TEI guidelines. It gives an overview of research fields in the Humanities and Computer Sciences which would benefit from the availability of such a representation framework and outlines the basic requirements it will have to comply with:

- The schema should provide a general model for the description of the structural and linguistic peculiarities of cmc discourse.
- To be useful for a broad range of application contexts in the Humanities, it should not be designed with a single project in mind but it should take into account the specific requirements of several projects (and genre typologies) in which the creation of annotated cmc resources is of interest.

2 http://wiki.itmc.tu-dortmund.de/cmc/

- In order to be suitable for small data sets which are annotated manually and also for the annotation of big data (e.g., reference corpora in Linguistics, large web corpora in the field of Natural Language Processing), its basic structure should be defined in a way that favours or supports (at least partially) automatic annotation procedures.
- The schema should build on a review of models which already exist in the TEI framework (currently TEI P5) and adapt them to the peculiarities of cmc genres in a reasonable and practical way.
- It should reflect the fact that CMC shares characteristics with written text as well as with spoken conversation while at the same time it is significantly different from both in its textual form and in the mode of production and reception.
- It should allow for an easy (and reversible) anonymization of cmc resources for purposes in which they shall be made available for other researchers (e.g., in the case of reference corpora).
- It should allow for an easy referencing of random samples of the resource (e.g., for citation in scientific publications, didactic materials or dictionary articles).

Since papers 2 and 3 of the panel take into consideration the goals and needs of several projects which are currently dealing with the construction of corpora of cmc discourse in four European languages, paper 1 includes a brief presentation of the four projects and an outline of their project-specific requirements for an annotation schema:

- **DeRiK** ("Deutsches Referenzkorpus zur internetbasierten Kommunikation") is a joint project of TU Dortmund University and the Berlin-Brandenburg Academy of Sciences (BBAW) and the Humanities which is building a reference corpus of German cmc discourse including the most prominent cmc genres. The DeRiK corpus will form a new component of the reference corpora of contemporary written German collected in the BBAW project "Digitales Wörterbuch der deutschen Sprache" (DWDS). On the one hand, it is designed as a resource for corpus-based linguistic analyses of language use in German cmc as well as – in combination with the DWDS corpus – of the impact of cmc genres on contemporary written German. On the other hand it will serve as a resource for the lexicographic description of "netspeak" vocabulary and cmc-specific processes of lexical-semantic change in the dictionary component of the DWDS online lexical information system[3] (cf. Beißwenger et al. 2013). For annotation, DeRiK is currently using the customized TEI schema for cmc described in Beißwenger et al. (2012). The schema comprises, among others, an element for the description of user contributions to cmc conversations (the divLike element *posting*), a distinction of two major types of cmc macrostructures (the cmc-specific division types 'thread' and 'logfile'), a component for modeling the authors of cmc postings as well as elements for the annotation of selected "netspeak" features in individual user postings (emoticons, interaction words, interaction templates, addressing terms).

3 http://www.dwds.de

- The Dutch reference corpus **SoNaR** was intended to serve as a general reference for studies involving language and language use. The corpus should provide a balanced account of the standard language and the variation that occurs within it. In doing so, it allows researchers investigating language use in a particular domain (e.g. medicine) or register (e.g. academic writing) or by a specific group (e.g. professional translators) to relate their data and findings to the general reference corpus. The corpus was also intended to play a role in the benchmarking of tools and annotations. Collected in 2008-2012 the corpus contains 500 Mwords, including discussion lists, e-magazines, websites, Wikipedia, SMS, chats and tweets. SoNaR is delivered in the FoLiA format (van Gompel 2012). FoLiA aims to support a wide variety of linguistic annotations in a generic paradigm and has been successfully adopted by various projects in The Netherlands. To provide support for new media, a type of structure annotation called "event annotation" was added, which fits nicely in the paradigm. SoNaR incorporates support for tweets, chat logs and SMS. The former two have been encoded as events, in which each tweet or chat message constitutes an event. Within the event structure, further subdivisions can optionally be made, such as paragraphs, sentences, words (in case of tokenized data). Elements in FoLiA carry a class from a certain set. In this way flexibility is provided to the user. The sets can be formally defined. The events in SoNaR are assigned classes such as "tweet" or "chatmessage". The actors of the set are also explicitly annotated, and further metadata on the annotation is also supported.
- **LETEC** ("Learning & Teaching Corpora"). Mulce repository[4] is a databank of LETEC corpora built upon online learning situations (Reffay, Betbeder & Chanier, 2012). All interactions among participants have been collected and structured before their analysis. It assembles a large variety of cmc types: email, forums, chat, blogs, 3D environments with audio and text chats, etc. One of the main components of its XML structure (Mulce-struct)[5] is the *workspace*. It includes descriptions of its *members* as references to the participants registered in the learning activity, *starting* and *ending dates*, the *tools* and the interaction tracks or *acts* that occurred using these tools. Each cmc tool has a detailed and specific structure. Large subparts of the LETEC databank will be integrated in 2013-14 into a nationwide cmc corpus in French where other cmc types, such as SMS, tweets, Wikipedia forums, will be added. The cmc SIG group leading the project belongs to the national consortium "*IR corpus-écrits*" in charge of building a reference corpus in French. The cmc SIG has designed a working package which will take care of the cmc TEI structure[6] of the whole corpus and work jointly with the European colleagues gathered in this panel.
- **Web2Corpus_it** ("Corpus italiano di comunicazione mediata dal computer") is a project funded by Sapienza University of Rome in 2010 aimed at investigating meaning negotiation strategies in cmc. It focuses on conversational, interactive,

4 http://repository.mulce.org
5 Schema for the instantiation component of a LETEC corpus. http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/mce_sid.xsd
6 https://groupes.renater.fr/wiki/corpus-ecrits-nouvcom/public/proj-tei/index

public, written communication in order to build a genre-balanced cmc corpus of Italian language to be investigated both qualitatively and quantitatively. The genres included are: forum, blog, newsgroup, social network and chat (cf. Chiari and Canzonetti, in press).[7] The collected corpus comprises one million words and has been fully anonymized (by masking), in order to avoid personal details of participants being disclosed, and xml-annotated both for macro-structural properties (thread, post, sender details - avatar | signature | nickname | senderplace – subject, date, time, links and embedded media, web action elements and cmc-specific emoticons and tags and addressing terms). At present the corpus is being processed linguistically with a statistical POS tagger and lemmatizer, including a reference machine dictionary (Common Lexicon of Italian) developed in order to include cmc specific lexical items, and will be subsequently manually checked and is planned to be released in late 2013.

These four corpus projects will provide the test bed for an evaluation of the models under construction with cmc discourse from different languages.

Paper 2:

Expanding the TEI encoding framework to genres of computer-mediated communication: considerations and suggestions

Michael Beißwenger[1], Thierry Chanier[2], Isabella Chiari[3], Maria Ermakova[4], Lothar Lemnitzer[4], Angelika Storrer[1], Maarten van Gompel[5], Henk van den Heuvel[5]

[1] TU Dortmund University (D) [2] Université Blaise Pascal, Clermont-Ferrand (F) [3] Università "La Sapienza", Rome (IT)
[4] Berlin-Brandenburg Academy of Sciences and the Humanities (D) [5] Radboud University Nijmegen (NL)

The first section of this paper presents some basic suggestions for the expansion of the TEI encoding framework to the structural and linguistic particularities of cmc genres. It takes into account the general requirements as well as the project-specific requirements outlined in paper 1 and builds on the customized TEI schema for cmc which has been presented at the 2011 TEI members' meeting (published in Beißwenger et al. 2012). The suggestions describe features for the modeling of corpus documents with stored discourse from cmc genres such as online forums, chats, wiki talk pages, Twitter, weblogs or social network sites and (amongst others) refer to the following basic issues in the description of cmc:

- the representation of user postings in written cmc as units which share characteristics with both text and conversations: under aspects of planning and coherence, they are designed as moves in an ongoing conversation; under the aspect of production and reception they behave just like texts, which first have to be produced and then are presented to and received by the addressee(s) *en bloc*;

7 http://www.glottoweb.org/web2corpus/

- the need for models for the representation of *cmc macrostructures* (= the way how series of user postings are grouped / presented to the users, e.g., in the form of *logfiles*, different types of *threads*, *timelines* etc.);
- the need for elements for the annotation of cmc-specific structural and linguistic features on the *microlevel of cmc discourse* (= the content of the postings which comprises e.g. typical “netspeak” phenomena such as emoticons, action words, addressing terms; hashtags; speedwriting phenomena, phenomena of non-standardized writing; embedded hyperlinks and media objects etc.);

With the help of examples from the corpus projects introduced in paper 1, the second section of the paper will offer problem sketches of the following open issues in modeling cmc and outline some first ideas for their treatment in TEI:

- *Handling citations:* Especially in forums and Bulletin Boards, cmc postings often contain (simple and nested) citations which reproduce content that has originally been part of other authors' prior postings. A schema for the representation of cmc should include a model for the annotation of citations and for referencing citations with the cited prior postings and their authors.
- *Cmc data as hyperlinked data:* Many cmc resources contain hyperlinks and linking structures. A framework for the representation of cmc interactions must include models for the description of how postings are linked with each other and/or with other interaction-external resources on the internet. In some cmc applications (e.g., micro-blogging sites such as Twitter) the method of displaying one and the same user posting as part of a sequence may vary depending on the user’s choice (cf. e.g. on Twitter the timeline of one author’s tweets vs. the timeline of tweets by different authors which include the same hashtag). A general model for cmc resources must provide features for the description of these kinds of structures and of the target sources of the hyperlinks.
- *Dealing with data from multimodal cmc environments:* In some cmc environments users are communicating not only in a text-based mode but using a combination of text-, audio-, video- and/or 3D-based modalities of interaction (e.g., e-learning platforms, Skype, gaming environments, virtual worlds etc.). One of the challenges related to the representation of cmc discourse recorded in environments of that kind is that contributions created and sent in one modality may contribute to, and indeed supplement, a contribution in another modality. In audio-graphic conferencing environments such as *Skype*, written postings sent via chat may contribute to an ongoing spoken conversation in the audio modality. In collaborative writing environments, written postings in the chat may contribute to the creation of a longer stretch of text in the word-processing modality. One challenge of treating cmc discourse of that kind is thus the necessity to integrate and align user contributions made in different modalities into a representation of the overall *multimodal* interaction. Since TEI provides modules not only for written but also for (transcriptions of) spoken discourse, the different modes could be represented separately (using different TEI

modules) while the alignment of the utterances and postings in the different modalities would have to be solved in an additional representation which is connected with the different resources.

Paper 3:

Metadata for cmc documents

Axel Herold[1], Lothar Lemnitzer[1], Michael Beißwenger[2], Isabella Chiari[3]

[1] Berlin-Brandenburg Academy of Sciences and the Humanities (D) [2] TU Dortmund University (D)
[3] Università "La Sapienza", Rome (IT)

Extensive and correct metadata has been recognized to be a crucial property of every data object that is used as a primary data source in research contexts. Fine grained metadata allow for *identification*, *location* and *management* of resources (e.g., NISO, 2004) but also provide researchers with crucial information regarding the *suitability* of a given resource for their particular research interest. The TEI header recognizes all of these metadata requirements to different degrees (Burnard 2005).

Our paper will have a strong focus on the encoding of intrinsic properties of different cmc data sets, thus addressing the issue of finding resources which are suitable for a given research question. Ideally, this part of the metadata description is based on the model representing the primary data. In this respect our paper strongly relies on paper 2, which will propose such a model for cmc data.

An example of cmc-specific data types are emoticons: small iconic representations of an interlocutor's emotion or his/her attitude towards an utterance (either self produced or produced by others) or towards a communication peer, to name just some of their communicative functions. It is therefore desirable to either encode normalization and classification schemes for those entities within the metadata description or to provide pointers to such schemes in addition to a suitable markup of these entities within the primary data.

Cmc data often contains large portions of verbosely cited material from previous parts of the discourse. This creates a challenge to the measurement of the extent of a given resource. Depending on the assumed discourse status of cited (parts) of utterances it may be necessary to include or exclude cited material. This is a theory-dependent decision, and it should therefore be possible to give concurrent values for a single unit of measurement. Moreover, metadata information on (the handling of) citations may – to some extent – be derived from the primary data directly (see paper 2 for handling of citations in primary text).

Distinct typologies for cmc tools (including tools that were used to access the primary data) and cmc genres are needed to account for the broad range of different data sources, e.g., online forums, chats, wikis, Twitter, weblogs, social network sites, learning environments and others. We will suggest mechanisms of referencing a particular

typology of cmc genres from within the metadata, however, without making any regulations on which kind of typology should be used and referenced in a given project.

Special care must be taken in the metadata description of information about discourse participants to ensure privacy and/or anonymity of the speakers involved in the discourse. Moreover, specific metadata for cmc should also have the function of restoring context information about features of the communication mode of production and reception of cmc texts that are not evident in the text itself. This involves features such as the temporal structuring of the discourse (synchronous vs. asynchronous mode), conversational hierarchies among discourse participants (e. g. blog author vs. commentator), discourse topic/domain or accessibility of the discourse (e. g. private vs. closed vs. public). The availability of social and other context information varies greatly, not only in quantity but also in its quality, according to the primary data source. Therefore a cmc metadata scheme will have to account for different levels of reliability for such information.

Considering the given fourfold structure of the TEI header (file description, encoding description, text profile and revision description), we will identify and discuss different possibilities for recording metadata properties that are specific for cmc data:

- Cmc data comprise properties found in traditional written resources (such as books or newspapers) as well as properties found in resources of (transcribed) spoken language. Both types of resources have previously been provided with TEI-based metadata. Properties shared across different resource types can be expected to be *reusable* for cmc metadata, e.g., listPerson to denote discourse participants or profileDesc to describe general discourse settings.
- Some metadata properties that cannot be readily encoded using specific elements can still be recorded using the *generic feature structure representation* (fs). Embedding of feature structures is currently allowed for a limited set of header elements in the TEI such as *classCode, extent, language, scriptNote* and *typeNote*. Exploiting the semantic linking mechanism provided by *att.datcat* (via the ISOcat data category registry; note that *classCode* provides a native semantic interface via @scheme as well) would allow tailor-made semantics for the properties encoded in such a way. But obviously this adds a level of indirection and does not capture these properties within the TEI directly.
- A third possibility lies in the adaptation of the TEI element inventory or of suggested cmc-specific value sets for existing elements. For individual projects this can already be achieved by TEI customizations but it may hinder interoperability across resources using elements not found in the TEI guidelines – which is another argument for why models for the representation of cmc data in TEI should better be part of the official guidelines and not be something that each project needs to solve individually.

We will conclude the paper with a proposed metadata header for TEI documents encoding cmc data. We will also – at least for some prominent features of metadata for

cmc documents – show how the TEI header metadata are related to, and can be converted to, metadata components within the emerging CLARIN Metadata Framework (Component Metadata Infrastructure, CMDI).

References

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative, Issue 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2013): DeRiK: A German Reference Corpus of Computer-Mediated Communication. In: Literary and Linguistic Computing (LLC). http://llc.oxfordjournals.org/content/early/2013/07/03/llc.fqt038.full.pdf?keytype=ref&ijkey=GXFixqNNy0uW7cO

Burnard, Lou (2005): Metadata for corpus work. In: Martin Wynne (ed.): Developing Linguistic Corpora: A Guide to Good Practice. Oxford, 30-46.

Chiari, Isabella; Canzonetti, Alessio (in press): Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In: Enrico Garavelli & Elina Suomela-Härmä (eds.): Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua. Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI, Helsinki 18-19 June 2012), Franco Cesati Editore, Firenze.

[NISO 2004] National Information Standards Organization (2004): Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Oostdijk, Nelleke; Reynaert, Martin; Hoste, Véronique; Schuuman, Ineke (2013): The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In: Peter Spyns & Jan Odijk (eds): Essential Speech and Language Technology for Dutch. Springer. http://link.springer.com/chapter/10.1007/978-3-642-30910-6_13

Reffay, Christophe; Betbeder, Marie-Laure; Chanier, Thierry (2012): Multimodal Learning and Teaching Corpora Exchange: Lessons learned in 5 years by the Mulce project. Special Issue on dataTEL: Datasets and Data Supported Learning in Technology-Enhanced Learning. International Journal of Technology Enhanced Learning (IJTEL) 4 (1/2), 11-30. http://edutice.archives-ouvertes.fr/edutice-00718392 (DOI: 10.1504/IJTEL.2012.048310).

[TEI P5] TEI Consortium (eds) (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5/ (accessed 22 March 2013).

van Gompel, Maarten (2012). FoLiA: Format for Linguistic Annotation. Documentation. ILK Technical Report 12-03. Available from http://ilk.uvt.nl/downloads/pub/papers/ilk.1203.pdf