

Computer-Mediated Communication in TEI: What Lies Ahead



Special Topic Panel

TEI Conference and Members Meeting 2013



Michael Beißwenger¹, Thierry Chanier², Isabella Chiari³,
Maria Ermakova⁴, Maarten van Gompel⁵, Iris Hendrickx⁵, Axel Herold⁴,
Henk van den Heuvel⁵, Lothar Lemnitzer⁴, Angelika Storrer¹

¹ TU Dortmund University (DE)

² Université Blaise Pascal, Clermont-Ferrand (F)

³ Università "La Sapienza", Rome (IT)

⁴ Berlin-Brandenburg Academy of Sciences and the Humanities (DE)

⁵ Radboud University Nijmegen (NL)



SAPIENZA
UNIVERSITÀ DI ROMA



Radboud Universiteit Nijmegen



berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

stimulate the discussion within the TEI community about

- how a standard for the representation of computer-mediated communication (CMC) in TEI should look like and
- what might be a practical and reasonable way to go about creating such a standard.

Computer-mediated communication (CMC):

Genres of interpersonal communication mediated through computer networks (the internet): chats, online forums, instant messaging, Twitter, comments on weblogs, discussions in wikis and on “social network” sites, interactions in multimodal communication environments such as Skype, MMORPGs or “virtual worlds” (e.g., SecondLife).

stimulate the discussion within the TEI community about

- how a standard for the representation of computer-mediated communication (CMC) in TEI should look like and
- what might be a practical and reasonable way to go about creating such a standard.


Goal 1: Give an outline of issues that one is faced with when designing a representation schema for CMC genres (in TEI) – as an input for the discussion

Goal 2: Receive feedback und suggestions that may be valuable for the further work of the new **TEI SIG** “**computer-mediated communication**”

⇒ <http://www.tei-c.org/Activities/SIG/CMC/>

Computer-mediated communication in TEI: What lies ahead

1. **Introduction / Paper 1** (~15 minutes):

- **Why we need a standard for the representation of CMC**
- **Modeling CMC: general requirements & issues**
- **Brief introduction of the four corpus projects** 

2. **Paper 2** (~30 minutes):

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

(Michael Beißwenger, Thierry Chanier, Maria Ermakova, Iris Hendrickx, Angelika Storrer)

3. **Discussion**, part I (15 min)

4. **Paper 3** (~15 minutes):

Metadata for CMC documents

(Axel Herold, Isabella Chiari)

5. **Discussion**, part II (15 min)

Create your own, unique XML schema (eHumanities “1.0”)

VS

Comply with a standard (eHumanities “2.0”)



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.**”

⇒ example: The TEI schema for CMC genres designed for use in the DeRiK project (cf. TEI-MM 2011 / Beißwenger et al. 2012, jTEI)

standardization

In view of the increasing importance of CMC as well as of the diverse needs to store and represent CMC data in corpora, a core framework for the representation of CMC genres should become part of the standard (which then can be customized for the specific needs of specific projects).

Who and why (we) need(s) a standard for the representation of CMC genres



A core standard for the representation of CMC would (i) facilitate building resources and (ii) foster interoperability of resources in (and between) research fields which are dealing with CMC data as a resource for research and development:

- 1) **Linguistics / Corpus Linguistics:**
building and using corpora of CMC genres as a basis for empirical research on linguistic peculiarities, linguistic variation and language change in (and through) digital media
- 2) **Language Technology / Web Mining:**
crawling and processing of web corpora as resources for applications in the field of Natural Language Processing
- 3) **other research contexts** which use CMC data as a resource for empirical (linguistic) analyses

A core standard for the representation of CMC **as part of the TEI** would – in addition – allow for an interoperability of CMC resources with other types of resources and thus **pave the way for new opportunities of empirical research** (e.g., contrastive corpus-based analyses on language use in text, speech and CMC corpora).



Requirements for a general representation framework for CMC:

- It should provide models for the annotation of the structural peculiarities of CMC genres and of typical linguistic peculiarities language use on the web.
- It should be useful for a broad range of application contexts in the Humanities and thus should take into account the requirements of projects from different research areas in which the creation of annotated CMC resources is of interest.
- 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.

General issues in designing a core framework for CMC:

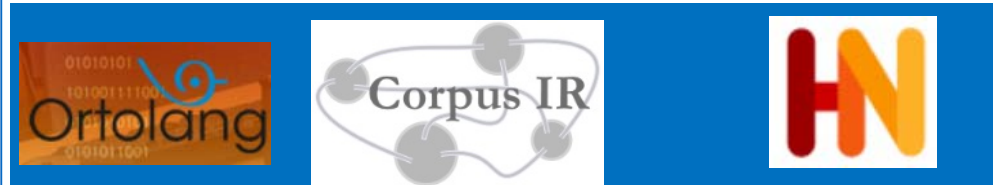
- Which linguistic and structural features of CMC should be modeled as part of the standard – and which ones are so specific that they should rather be addressed in project-specific customizations (i.e., additions to the standard)?
- How general should the “general level” be on which the standard describes the features of CMC? (Note that CMC is a moving target with emerging technologies and a continuous progress of technical frameworks and communication environments.)
- Which is a reasonable (and adequate) format for representing CMC in TEI: a collection of features in one or several model class(es)? a new module? ...
- How should a metadata schema be structured which is adapted for the representation of features of CMC interfaces/communication environments and of CMC genres/text types?

CoMeRe (Communication Médiée par les Réseaux): a reference corpus of French CMC



Project supported by the national consortium *Corpus-écrits*, sub-part of *Huma-Num*, and *Ortolang* (French correspondent to DARIAH)

Documentation and activities : <http://comere.org>



<http://corpusecrits.corpus-ir.fr/>

People: 14 research. from 8 research units. Coord: Chanier, T (Clermont), Poudat, C. & Sagot, B (Paris), Longhi, J. (Cergy), Antoniadis, G. (Grenoble)

Objective: Kernel corpus assembling existing corpora of different CMC genres and new corpora build on data extracted from the Internet. These heterogenous corpora will be structured and processed in a uniform way, complemented with metadata. CoMeRe will be released as OpenData through the national infrastructure Ortolang, following constraints which will be reused for the forthcoming “*Corpus de Référence du Français*”.



Thierry

Reffay, C., Betbeder, M.-L., Chanier, T. (2012): [Multimodal Learning and Teaching Corpora Exchange: Lessons learned in 5 years by the Mulce project](#). International Journal of Technology Enhanced Learning (IJTEL), (4) , 1/2). DOI: 10.1504/IJTEL.2012.048310

- Existing corpora of different CMC genres (Blog, forum, chat, SMS, multimodal communication systems) coming from previous research projects.
- Extension to other genres (Tweets and Wikipedia forums) extracted from Internet in 2014
- Highly heterogeneous corpora with respect to their organization (tables, simple XML, XML + schemas, etc.) which will assemble within a common TEI format for the various kinds of interactions, with a description of the Interaction Space (technological environments, participants, context).
- Shallow processing of data (described in TEI) for segmentation, token recognition and annotation of standard words, and POS.
- Superstructure for content packaging of a corpus with the TEI file, audio, video, images, when appropriate
- Correct management of ethics and rights in order to publish corpora accordingly to the OpenData requirements
- Detailed metadata with automatic correspondence between OLAC, CLARIN and TEI Header

DeRiK (Deutsches Referenzkorpus zur internetbasierten Kommunikation): a reference corpus of German CMC



Joint initiative of TU Dortmund University and the DWDS project at Berlin-Brandenburg Academy of Sciences and the Humanities (since 2010), embedded in the scientific network *Empirikom*.



People: Michael Beißwenger, Angelika Storrer (Dortmund), Maria Ermakova, Alexander Geyken, Lothar Lemnitzer (Berlin)

<http://www.empirikom.net>

Objective: Reference corpus of German CMC including data from the most prominent CMC genres (according to an annually conducted survey on internet use in Germany) – as a new component of an existing collection of corpora of written German (www.dwds.de).



Michael

For the representation of the DeRiK data, we designed a customized TEI schema (presented at TEI-MM 2011).

M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2013): [DeRIK – A German Reference Corpus of Computer-Mediated Communication](#). In: *Literary and Linguistic Computing*. <http://tinyurl.com/derik-llc>



Maria



Angelika

M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2012): [A TEI Schema for the Representation of Computer-mediated Communication](#).

In: *Journal of the Text Encoding Initiative (jTEI)*, Issue 3: <http://jtei.revues.org/476>



<http://www.empirikom.net>

Specific requirements in the DeRiK project:

- The schema should provide **elements for the annotation of units which are often regarded as “typical” for language use on the web** and which are of special interest for everybody who wants to compare linguistic features of CMC discourse with the language documented in text corpora (such as the DWDS corpora) oder in speech copora (such as the FOLK corpus).
- It should allow for an **easy (and reversible) anonymization** of CMC resources because the corpus (as part of the DWDS reference corpus) shall be made available for other researchers/to the public.
- It should allow for an **easy referencing of random samples of the corpus** (e.g., for KWIC presentation on the corpus user interface, for presentation in DWDS dictionary articles, for citation in scientific publications and in didactic materials).

SoNaR (Stevin Nederlandstalig Referentiecorpus): a reference corpus of Dutch, including CMC



SoNaR: a 500-million-word corpus for Dutch collected in 2008-2012

CLST
Centre for Language and Speech Technology
Radboud Universiteit Nijmegen



Design: contemporary written Dutch texts originating from the Netherlands and Flanders, incl. texts from conventional text types and texts from new media (CMC)

People: Maarten van Gompel, Iris Hendrickx, Henk van den Heuvel (and others)



Iris

CMC components (in million of tokens):

Discussion lists:	57 Mw	Wikipedia:	23 Mw
E-magazines:	9 Mw	Blogs:	0.1 Mw
Websites:	3 Mw	SMS:	0.7 Mw
Chats:	12 Mw	Tweets:	23 Mw

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman (2013): [The construction of a 500-million-word reference corpus of contemporary written Dutch](#). In J. Odiijk & P. Spyns (eds.), Essential Speech and Language Technology for Dutch. Springer.

Sanders, E. (2012): [Collecting and Analysing Chats and Tweets in SoNaR](#). Proceedings of LREC.

Treurniet, M., O. De Clercq, H. van den Heuvel & N. Oostdijk (2012): [Collecting a Corpus of Dutch SMS](#). Proceedings LREC-2012: 2268-2273.

van Gompel, M. (2012): [FoLiA: Format for Linguistic Annotation](#). ILK Technical Report – ILK 12-03.

SoNaR (Stevin Nederlandstalig Referentiecorpus): a reference corpus of Dutch, including CMC



- authenticity and quality of text sources
- IPR cleared
- metadata verified
- conversion to standard format

CLST
Centre for Language and Speech Technology
Radboud Universiteit Nijmegen



Iris

Annotations:

SoNaR-500: tokenization, POS tagging and lemmatization

SoNaR-1: labeling of named entities, co-reference, dependency parsing, semantic roles, and spatio-temporal relations (manually verified)

SoNaR-500 is delivered in the **FoLiA format** (van Gompel, 2012), an XML format developed for linguistic annotation.

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman (2013): **The construction of a 500-million-word reference corpus of contemporary written Dutch**. In J. Odiijk & P. Spyns (eds.), Essential Speech and Language Technology for Dutch. Springer.

Sanders, E. (2012): **Collecting and Analysing Chats and Tweets in SoNaR**. Proceedings of LREC.

Treurniet, M., O. De Clercq, H. van den Heuvel & N. Oostdijk (2012): **Collecting a Corpus of Dutch SMS**. Proceedings LREC-2012: 2268-2273.

van Gompel, M. (2012): **FoLiA: Format for Linguistic Annotation**. ILK Technical Report – ILK 12-03.

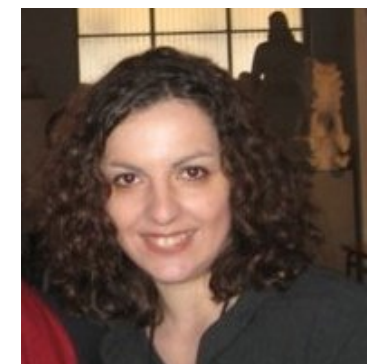
Web2Corpus_it: a balanced pilot corpus of Italian CMC



SAPIENZA
UNIVERSITÀ DI ROMA

Funded by Sapienza University of Rome

People: Isabella Chiari (coord.), Federico Albano Leoni, Sergio Bolasco, Francesco De Renzo, Sabine Koesters Gensini, Ilaria Tani, and others



Isabella

Objective: Investigate sense negotiation strategies in CMC through the development of a pilot balanced corpus of Italian CMC (1 million words) of texts belonging to the following genres:

Blog, forum, newsgroup, social network, chat.

Corpus requisites: Interactive and

dialogic exchanges, Written form, Public exchanges.

Texts are anonymized in order to hide personal data in the corpus (proper names, usernames, emails, traceable items etc.) and further annotated in a XML scheme for a general structural description. The corpus will be further POS tagged and made available for online querying.

Chiari, I e A. Canzonetti, in press: “[Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione](#)“, in Enrico Garavelli and Elina Suomela-Härmä (des.), *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*, Atti SILFI, Helsinki 18-19 June 2012), Franco Cesati Editore, Firenze.

Specific requirements in the Web2Corpus_it project:

- The anonymization procedure must mask sensible personal data while not subtracting context relevant to the understanding of text, nor introduce misleading associations and interpretations about the text.
- The metadata scheme must contain all information needed to reconstruct setting, participants, production modes, publication data and it's access properties...
- Annotation of structural textual properties should be maintained separate from detailed annotation of linguistic features of the text
- CMC specific linguistic features should have an annotation scheme and processing compatible and possibly integrated with usual and consolidated practices in linguistic annotations (integration with a reference dictionary, integration within POS tagging tools, etc.)
- The relationship between single and concatenated posts and their role within threads should be always traceable and layout choices relevant to the overall document reception annotated

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



paper 2

DeRiK project:

- 1) **Experiences and issues in representing peculiarities of CMC discourse with models from TEI-P5** – discussed on the example of <posting> and elements from *model.divWrapper*
- 2) **“Linked data” phenomena in CMC discourse** – discussed on the example of quotes in forums, hashtags and addressings in tweets

SoNaR project:

- 3) **The question of big and small modeling units** – with examples from the SoNaR project and experiences in FoLiA

CoMeRe project:

- 4) **Dealing with data from multimodal CMC environments** – with examples from LETEC (Learning & Teaching Corpora)

Experiences and issues in representing peculiarities of CMC discourse with models from TEI-P5



- In the DeRiK project, we developed a customized TEI schema for the representation of written CMC genres (2010-12).
 - **jTEI article:** M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2012): [A TEI Schema for the Representation of Computer-mediated Communication](#). In: Journal of the Text Encoding Initiative (jTEI), Issue 3: <http://jtei.revues.org/476> (DOI: 10.4000/jtei.476).
 - **ODD document and encoding examples:** <http://www.empirikom.net/bin/view/Themen/CmcTEI>
- The schema comprises a model for the representation of basic structural features of CMC genres and for a fine-grained linguistic annotation of selected “netspeak” phenomena:
 - a new element <posting> which represents the individual user’s written contributions to CMC interactions
 - elements for the annotation of CMC ‘macrostructures’ (the structural patterns above the posting level: *threads*, *logfiles*)
 - elements for the annotation of CMC-specific phenomena on the ‘microlevel’ of CMC documents (emoticons, action words, addressing terms; user signatures, postscripts, openers, closers)
 - a model for the representation of users (as authors of postings) which allows for an easy and reversible anonymization

- In the DeRiK project, we developed a **customized TEI schema** for the representation of written CMC genres (2010-12).

- **jTEI article:** M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2012): [A TEI Schema for the Representation of Computer-mediated Communication](#). In: Journal of the Text Encoding Initiative (jTEI), Issue 3: <http://jtei.revues.org/476> (DOI: 10.4000/jtei.476).
- **ODD document and encoding examples:** <http://www.empirikom.net/bin/view/Themen/CmcTEI>

Challenges we were facing when designing the schema:

find a reasonable compromise between linguistic concepts for the description of CMC and constraints set by the encoding framework – especially:

- a) when **introducing new elements** and associating them with existing model classes in TEI
- b) when **adopting existing models** from the TEI framework for the representation of similar phenomena in CMC.

Neither <u> nor <p>: postings in written CMC



1	zora freut sich über ihr zeugniss :)))
2	quaki: *aufpluster*
3	system: Thor... betritt den Raum.
4	marc30: ich mal wieder nich...
5	quaki: was hast denn zori??
6	quaki: erzähl
7	system: stoeps kommt aus dem Raum Number_of_the_beast herein.
8	Lantonie: Das hast du dir verdient, zori?
9	TomcatMJ: oh man wat fürn krawall hier draußen...*guck*
10	zora: nur einsen *brustschwell*
11	system: Emon betritt den Raum.
12	stoeps: ree :-)))
13	Emon: reee
14	system: Emon ist wieder da.
15	stoeps: r emon

chat logfile

Freibad statt Tunnel

1 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

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

3 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)

4 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" Schenkeltlopfen? . Wer braucht sonst noch ein Taschentuch? (*_*) [deeleres](#) 13:35, 28. Jul. 2011 (CEST)

5 Vorschlag zur Güte: Man läßt den Kram mit dem Freibad (zunächst) unerwähnt und schreibt lediglich ein Kapitel über die **bereits beendete (!!!)** öffentliche Wahl zur Benennung des Straßentunnels (Kurzform: Bürger sollten über Namen eines Tunnels abstimmen – „Bud-Spencer-Tunnel“ war der Sieger-Vorschlag – die Stadt Schwäbisch Gmünd hat diesen Vorschlag abgelehnt) -- [Jamiri](#) 14:23, 28. Jul. 2011 (CEST)

Ich hab grundsätzlich nichts dagegen, wenn es irgendwie erwähnt werden wird. Nur es ist immer noch nichts passiert - etabliertes Wissen ist ja vorausgesetzt und das tun wir im Moment nicht außer Tod oder vll. die [Zukunft der Erde](#). Das Echo ist zwar laut, die Welle aber auch nicht wirklich hoch. Ich würde es jetzt nicht reinschreiben wollen und das gemähte Gras wieder wachsen lassen. *Die Bud-Spencer-Statue - New York setzt auf den Koloss von Liberty Island* - (^_^) die Welle wäre wohl um einiges höher [deeleres](#) 15:43, 28. Jul. 2011 (CEST)









Wikipedia talk page

Basic units of written CMC interactions: the stretches of text that users submit to the server to make a contribution to the ongoing conversation; on the screen, these stretches of text are displayed as blocks between paragraph breaks.

Modeling options in “standard” TEI:

- <p> paragraph (*core*)
- <div> division (*default text structure*)
- <u> utterance (*transcriptions of speech*)
- <sp> speech (*core tags for drama*)

Written CMC shares characteristics both with “traditional” text *and* spoken conversation:

TEXT	CONVERSATION	
		CMC is dialogic interaction in which each communicative move creates/changes the context for follow-up moves.
		CMC discourse is organized in (and displayed on the screen as) sequences of stretches of written text which may contain internal textual structuring (use of line breaks/paragraphs, lists, varying font face, size, color and weight, embedded media objects etc.).
		Under aspects of planning and coherence, the similarity with paragraphs in monologic texts and with speech in performance texts is just a formal, not a functional one : there is no author who planned the entire dialogue in advance; instead, the dialogue is developed by the users as they go along with each posting creating the context for the partners’ next moves.
		Unlike spoken utterances, the production of people’s contributions to CMC dialogues is text-like : they are first (1) composed by their authors in their entirety, then (2) sent to the server, then (3) displayed on the screen as written messages before (4) they can be read and replied by other interlocutors.



Decision: TEI module *text structure* as the basis for our schema

⇒ allows us to describe the inner structure of postings with standard TEI elements for text structure, text design etc.

<posting>: a content unit that is being sent to the server “en bloc”.

⇒ **The posting model:**

- o a block of written text which may contain paragraphs and other types of internal textual structuring ⇒ **defined as *model.divLike***
- o each posting is assigned an author ⇒ **additional attribute *@who* from the *att.ascribed* class** which „provides attributes for elements representing speech or action that can be ascribed to a specific individual“. In the TEI standard, *att.ascribed* cannot occur with <div> (but, e.g., with <u> from the TEI module *transcribed speech*).

⇒ The concept of the element <posting> takes into account the hybrid character of written user contributions to CMC dialogues

Modeling issues: Treating <posting> as *model.divLike* implies:

- adjustments on textstructure elements (front/back, body)
- as a copy of <div>, <posting> inherits the internal structure of <div> and thus comprises a number of elements which we do not need/want



Which is the most appropriate class for <posting>: *model.divLike* or *model.divPart* (with members <u>, <p>, <sp> and <l>)?

⇒ Both solutions have their pros and cons

... or rather a new class for all CMC elements which we introduce through customization?



Subdivision classes like *model.divWrapper*, *model.divTopPart* or *model.divBottomPart* contain elements which seem useful for the encoding of elements on the microlevel of CMC postings (e.g., `<postscript>` and `<closer>`).

But: The models of these elements are optimized for the usage of salutes, signatures and postscripts in traditional letters: The distribution of `<postscript>` and `<closer>` is limited to the division top and bottom.



CMC communication, in contrast, is characterized by a less conventional style of writing than in epistolary correspondence: there are postings, for example, which consist solely of a postscript:

1

Dass die Concordia ein nicht sichtbares Heckanker hat, halte ich für ein Gerücht. Das müsste dann ja auch eine gewisse Größe, nebst Winde und Kette haben. Da kann man nicht einfach einen Klappdragen in der Hosentasche des Käptns nehmen. Heckanker können vorgeschrieben werden, wenn das aufgrund besonderer Umstände erforderlich ist, wie es oft bei schwimmenden Gerät der Fall ist. --[Wicket](#) 22:29, 22. Jan. 2012 (CET)

2

P.S.: Als *schwimmendes Gerät* bezeichnet man z.B. Arbeitspontons oder Schwimmbagger ohne eigenen Antrieb. --[Wicket](#) 22:55, 22. Jan. 2012 (CET) (Wikipedia talk page)

(similar problem with `<closer>`: user signatures in postings on Wikipedia talk pages may appear not at the end but in the middle of the posting.)

This makes it impossible to use the corresponding elements from the opener and closer parts of the standard TEI `<div>` model – even though under a functional perspective, postscripts and closers in CMC postings have similar functions.

Solution (for the time being): For the encoding of chunks of text which represent opener and closer elements, we decided to use the element `<seg>` with a corresponding attribute:

```
<posting who="#A02" synch="#t02" indentLevel="1">
  <p><seg type="opener">Servus <persName ref="#A03"/></seg>! Kennst du die
  Bearbeitung in der neuen <ref target="http://www.efloras.org/florataxon.aspx?
  flora_id=2&#65120;taxon_id=119600">Flora of China</ref>? Zwei der drei aus
  China angegebenen Arten sind zumindest nicht allgemein akzeptiert. Grundsätzlich
  muss man aber auch bei allen Arten, die aus der alten Sowjetunion beschrieben
  wurden, vorsichtig sein: Die hatten so eine Art Dogma, dass es keine Unterarten
  geben darf. So ist halt automatisch alles, was nach einer phänotypisch
  abgrenzbaren Sippe ausgesehen hat, gleich als Art beschrieben worden.
  <seg type="closer">Grüße</seg> --<autoSignature/></p>
</posting>
```

We are not satisfied with this solution, because `<seg type="opener/closer">` duplicates the components of opener and closer.

If the elements `<salute>`, `<signed>` and `<postscript>` could be made more flexible in terms of their position inside of the `<div>`, we would prefer to use them instead of creating redundant solutions:

```
<posting who="#A02" synch="t02">
  <p><opener><salute>Servus!</salute></opener>Kennst du die
  Bearbeitung...
  <closer><salute>Grüße</salute></closer></p></posting>
```

CMC environments provide their users with system functions which allow them to use hyperlinking as a device for:

- embedding media objects into their postings (⇒ YouTube links)
- linking their postings to those parts of the prior context they are referring to (⇒ quotes in forums)
- delivering their postings to their addressees (⇒ addressing terms in tweets and facebook postings)
- making their postings available as query results for certain topics/keywords (⇒ hashtags in tweets).


“Linked data” phenomena are usually processed by the system before being displayed on the screen: media links are converted into embedded media objects, quotes are introduced through automatically generated text and marked through layout features, addressings and hashtags in tweets are automatically delivered to certain users or associated with other tweets.

“Linked Data” phenomena in CMC – example 1: quotes in online forums



09-21-2010, 10:10 AM #4 (permalink)

user name
love will tear you apart



Nothing will ever be as good as The Holy Bible, Richey Edwards' parting gift. Absolutely brilliant album, I don't think anything they'll do will be as good as Generation Terrorists either but they've had some solid albums out. I'm yet to hear this one though, I'm trying to find a link as we speak. I don't like the poll options already, you can only vote Brilliant or above average. Where's the good option?!

Reply to Thread

Thread: Manic Street Preachers - Postcards from a Young Man

Title:

Message:

Rich text editor toolbar with options for Bold, Italic, Underline, Bulleted List, Numbered List, Indent, Outdent, Link, Unlink, Image, Video, and a 'Wrap [YOUTUBE] | selected text' button.

[QUOTE= **user name** ;934206]Nothing will ever be as good as The Holy Bible, Richey Edwards' parting gift. Absolutely brilliant album, I don't think anything they'll do will be as good as Generation Terrorists either but they've had some solid albums out. I'm yet to hear this one though, I'm trying to find a link as we speak. I don't like the poll options already, you can only vote Brilliant or above average. Where's the good option?![/QUOTE]

In my opinion ...|

Submit Reply


feel great.



Reply With Quote

follow-up posting with
quote included:

Quote:

Originally Posted by **user name** 

Nothing will ever be as good as The Holy Bible, Richey Edwards' parting gift. Absolutely brilliant album, I don't think anything they'll do will be as good as Generation Terrorists either but they've had some solid albums out. I'm yet to hear this one though, I'm trying to find a link as we speak. I don't like the poll options already, you can only vote Brilliant or above average. Where's the good option?!

In my opinion

“Linked Data” phenomena in CMC – example 1: quotes in online forums



- For quantitative analyses it is important to be able to filter quoted text and system-generated posting parts out.
 - For qualitative analyses quotes and system-generated posting parts are relevant content.
- ⇒ We need to annotate these phenomena and (for each platform) store information about which parts of the data are created by whom (including a distinction between user- and system-generated content).

The screenshot shows a forum interface with a 'Reply to Thread' form. The form includes a 'Title' field, a 'Message' field with a rich text editor (showing bold, italic, and underline options), and a 'Submit Reply' button. The message content is partially visible, showing a quoted section. Annotations are present:

- A red box highlights the text: "content automatically generated (= system-generated parts of the posting)". A red arrow points from this box to the "Originally Posted by user name" label in the quoted section.
- A blue box highlights the text: "content automatically reproduced (= the quoted prior posting, created by a human user of the platform)". A blue arrow points from this box to the quoted text area.
- A green box highlights the text: "content originally created by the author of the posting". A green arrow points from this box to the "In my opinion ..." text in the reply form.

“Linked Data” phenomena in CMC – example 2: hashtags and addressing terms in tweets



Open issue: How to describe phenomena in which linguistic units coincide with units of hypertext organization? (examples: hashtags and addressing terms in tweets)



Erfreulich, dass ich Vertreter der [@cornelsenverlag](#) e auf Veranstaltungen wie dem [#slml13](#) der [@werkstatt_bpb](#) antreffe. Das macht Mut.



Surprised to meet representatives of [@cornelsenverlag](#) at events such as [#slml13](#) of [@werkstatt_bpb](#). That's encouraging.

example: German tweet

In this example, < [@cornelsenverlag](#) > is:

- (i) a proper noun;
- (ii) the name (ID) of another Twitter account;
- (iii) an addressing term (indicated through <@>) used to address the tweet to the user/account „cornelsenverlag“;
- (iv) a hyperlink automatically generated by the system for tokens introduced with <@>; *target resource:* the profile page of the addressee;
- (v) a command of the author of the tweet to the system to deliver the tweet to the account „cornelsenverlag“.

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



paper 2

DeRiK project:

- 1) **Experiences and issues in representing peculiarities of CMC discourse with models from TEI-P5** – discussed on the example of <posting> and elements from *model.divWrapper*
- 2) **“Linked data” phenomena in CMC discourse** – discussed on the example of quotes in forums, hashtags and addressings in tweets

SoNaR project:

- 3) **The question of big and small modeling units** – with examples from the SoNaR project and experiences in FoLiA

CoMeRe project:

- 4) **Dealing with data from multimodal CMC environments** – with examples from LETEC (Learning & Teaching Corpora)

Natural language processing starts with **tokenisation**.

- TEI has the `<w>` element to represent single tokenised words.

What is a **token** in CMC?

- Consider data rich in abbreviations, contractions, often lacking clear boundaries to overcome character limits (twitter).
- Representation of **normalised** forms is an issue. Linguistic processing will often be based on the normalised forms.

SoNaR tweet Example:

```
<t>@Spartz just found a new website, you MUST take a  
look at it http://t.co/u20zJeiS its soooooo funny!  
#omg #roflol #sweet #totalysfw</t>
```

#omg (oh my god) #roflol (rolling on floor laughing out loud)

#totalysfw (totally safe for work)

Introduce **specific structure elements** for **specific CMC ‘tokens’**?

Twitter data: *hashtags, mentions*

Chat data: *recipient, mentions*

More general: *emoticons, URLs, email addresses*

*Several of these also often act as **hyperlinks**.*

But are those really atomic?

Hashtag examples from SoNaR:

#bff1 (best friends for life)

#OpNaarDeAH (eng: ‘lets go to the AH supermarket’)

#WhoCares

⇒ Possible solution: `<w>` inside `<hashtag>` ?

Two representation issues:

Date & time: CMC data is often strongly bound to time and sensitive to chronological order, consider chat data (IRC, IM, SMS) and forum posts.

Authorship/Actors: Representing authors/actors of/in CMC data. Not just a metadata issue!

Solutions in FoLiA, from SoNaR chat data:

```
<event xml:id="WR-P-E-L-0000000414.text.1.event.2059"
actor="twitterer414" begindatetime="2011-08-
08T17:56:11" class="Tweet">
```

```
<t>@DavidHeek Mwooooeeeah, Silence of the Lambs.</t>
```

A more generic **event** attribute is specified further by a **class**, defined in a custom **set** rather than by the format. Attributes solve time and authorship issues (*van Gompel, 2012*).

```
<event xml:id="WR-U-E-A-0000000033.text.1.event.3" class="message"
begindatetime="2011-01-11T11:27:29"
set="http://examplehost/events-chat.fsd.xml" actor="chatter201">
  <t>hallo!</t>
  <s xml:id="WR-U-E-A-0000000033.text.1.event.3.s.1">
    <w xml:id="WR-U-E-A-0000000033.text.1.event.3.s.1.w.1"
class="WORD" space="no">
      <t>hallo</t>
    </w>
    <w xml:id="WR-U-E-A-0000000033.text.1.event.3.s.1.w.2"
class="PUNCTUATION">
      <t>!</t>
    </w>
  </s>
</event>
```

What constitutes a **document** in CMC? How to group CMC data?

- Is one twitter message one document or One utterance?
- In SoNaR: Twitter data per user, SMS data per user, IRC data per room and timeframe

Representation of specific CMC non-text **events**. Consider chat data:

- *User joins a chat room*
- *User leaves a chat room*
- *User changes his/her nickname*

Solutions in **FoLiA**, from **SoNaR** chat data:

```
<event xml:id="WR-U-E-A-0000000033.text.1.event.3"
actor="chatter201"      begintatetime="2011-01-11T11:27:45"
class="nickchange">
    <feat subset="newname" class="chatter202" />
</event>
```

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



paper 2

DeRiK project:

- 1) **Experiences and issues in representing peculiarities of CMC discourse with models from TEI-P5** – discussed on the example of <posting> and elements from *model.divWrapper*
- 2) **“Linked data” phenomena in CMC discourse** – discussed on the example of quotes in forums, hashtags and addressings in tweets

SoNaR project:

- 3) **The question of big and small modeling units** – with examples from the SoNaR project and experiences in FoLiA

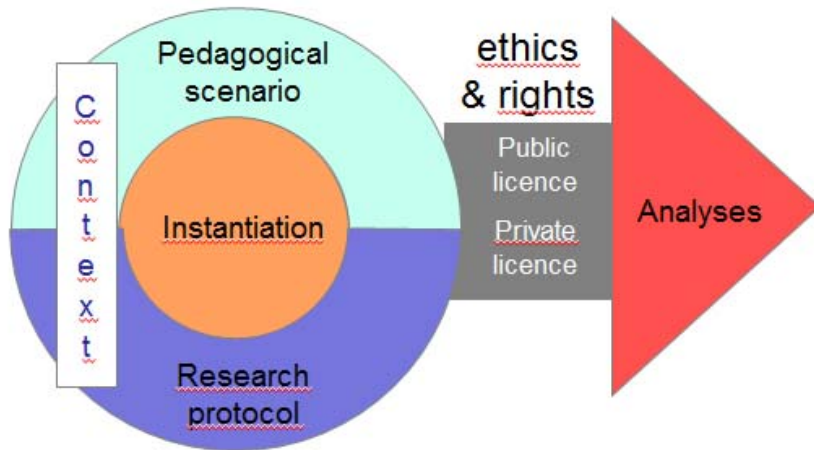
CoMeRe project:

- 4) **Dealing with data from multimodal CMC environments** – with examples from LETEC (Learning & Teaching Corpora)

Examples taken out of LETEC (Learning & Teaching Corpora), In Mulce repository

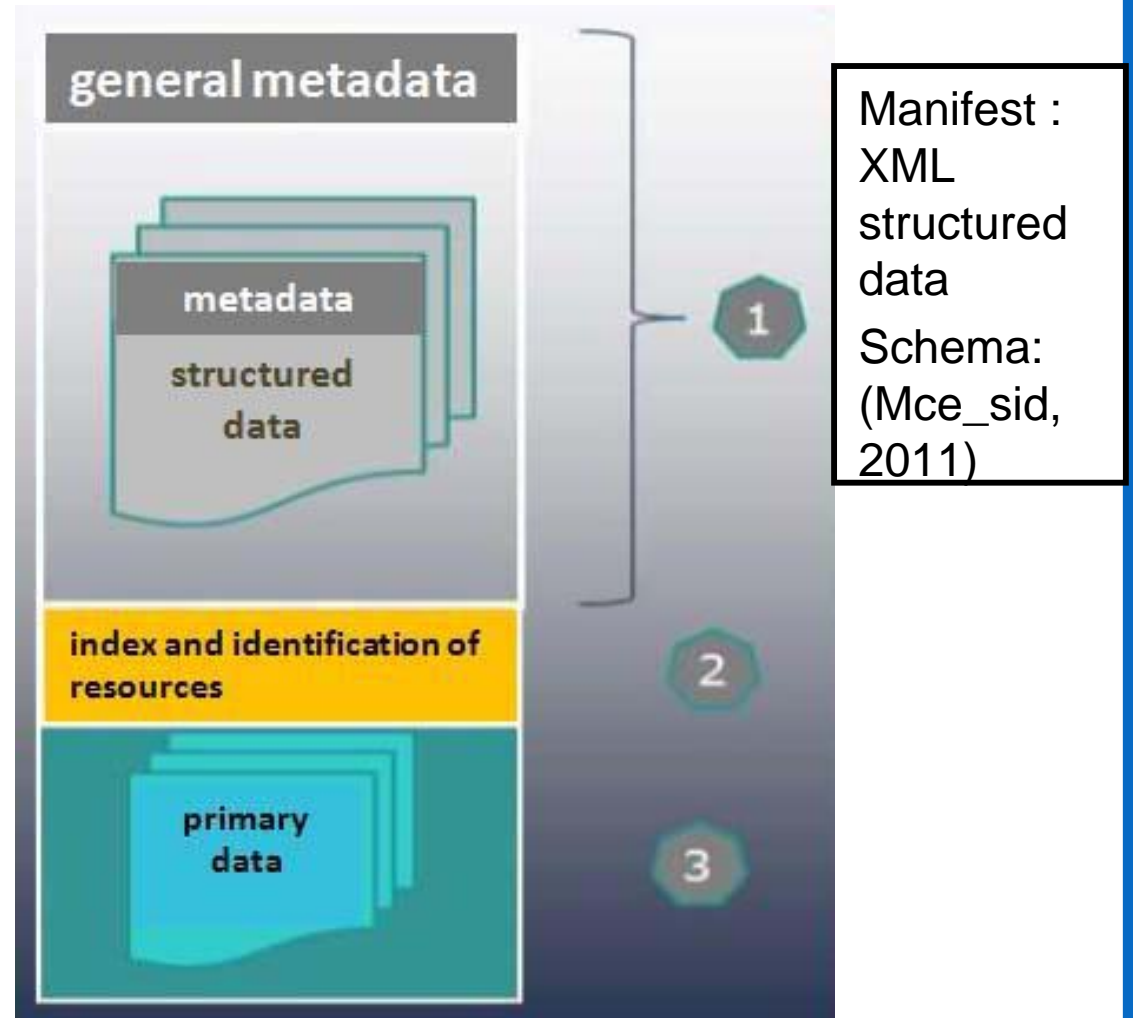


Definition & components of one corpus



"A LETEC corpus collects in a **systematic and structured way** all the data from **interactions** which occur during a course which is **partially or entirely online**. These data are enriched by technical, pedagogical and scientific information as well as information about the participants and are organized to allow **contextualized analyses** to be performed." (Mulcedocumentation, 2013)

IMS-CP content packaging for one corpus



Learning situations

Nom du LETEC	Année du projet	Langues	Domaine d'apprentissage	Institutions et participants	Environnements technologiques
Archi21	2011	Français (FLE), Anglais (LE)	CLIL / EMILE Architecture	1 Univ et 1 école d'architecture en France ; 18 part. ; 4 groupes	Monde synthétique / virtuel 3D (Second Life) ; forum audio (Voiceforum)
Favi	2006-08	Français (FLE)	Français académique	2 univ. France ; 31 parti. ; 2 sessions formation	Clavardage (MSN et WebCT)
Simuligne	2001	Français (FLE)	Langues, Interculturel	Grande-Bretagne et France ; 67 parti. ; 4 groupes	Plate-forme asynchrone (WebCT)
Copéas	2005	Anglais et Français	Langues	Grande-Bretagne et France ; 22 parti. ; 2 groupes	Plate-forme asynchrone (WebCT) et Plate-forme audio-graphique synchrone (Lyceum)
Ecofralin	2008	Espagnol et Français (FLE)	Langues, Interculturel	Colombie et France ; 24 parti. ; 4 quadrems	Blogue et Plate-forme audio-graphique synchrone (Centra)
Infral	2008	Allemand et Français (FLE)	Langues, Interculturel	Allemagne et France ; 26 parti. ; 4 quadrems	Blogue et Plate-forme audio-graphique synchrone (Centra)
Tridem06	2006	Anglais et Français (FLE)	Langues, Interculturel	Grande-Bretagne, Etats-Unis et France ; 62 parti. ; 12 quadrems	Blogue et Plate-forme audio-graphique synchrone (Lyceum)
VMT teamC	2006	Anglais	Mathématique	Etats-Unis, Singapour et Ecosse ; 12 parti. ; 1 team	Combined TextChat and Whiteboard (VMTForum2006)

- 3D : text & audio chats, non verbal
- Text chat
- Mail, forum, text chat
- text & audio chats, non verbal
- blog, audio chat, non verbal
- blog, audio chat, non verbal
- Blog, text & audio chats, non verbal
- Text chat & whiteboard

Example of multimodal environment: *Lyceum*

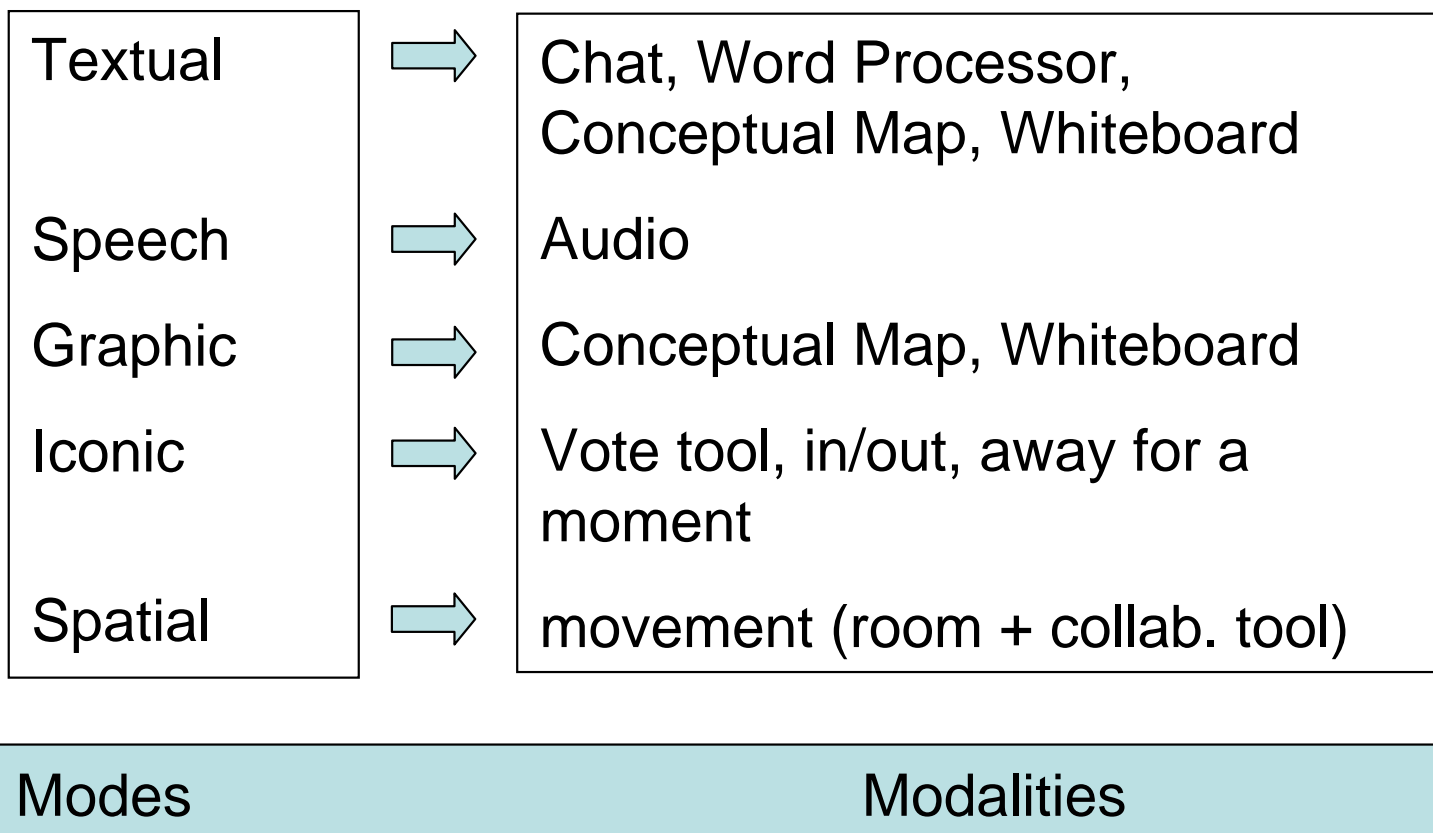
Spatial component

The screenshot displays the Lyceum multimodal environment interface. It features a central workspace with a concept map titled "Concept map: Web paid" containing nodes like "public service", "free Bruno", "pay by subscription", and "pay at time". To the left is a "Lobby" area with a grid of numbers (101-110) and a list of participants (Tim, Lucas, Bruno, etc.). The bottom section includes communication tools (Yes, No, Wipe, Talk) and a chat window showing a message: "<Tim> a version = a demo version".

Three blue numbers are overlaid on the interface: a large '1' in the lobby, a large '2' in the chat area, and a large '3' in the concept map area. Two red arrows point from the bottom text labels to these numbers: one from "Communication tools" to '2' and one from "Shared editing tools" to '3'.

Communication tools

Shared editing tools



Even more now with non verbal modes in 3D environments

Vetter, A., Chanier, T. (2006) "Supporting oral production for professional purpose, in synchronous communication with heterogeneous learners". *ReCALL*, vol. 18, 1 doi:[10.1017/S0958344006000218](https://doi.org/10.1017/S0958344006000218)

Chanier, T., Vetter, A. (2006) "Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". *Apprentissage des langues et Système d'Information et de Communication (Alsic)*, vol. 9. DOI : 10.4000/alsic.270

audio		tpa	10:43:38	10:43:46	what's mean euh + mainly ++ please what's mean mainly
		tpa	10:43:46	10:43:51	mainly is euh + usually ++
chat		tpc	10:43:49	10:43:49	principally
		tpa	10:43:51	10:43:59	but euh what euh do you write the definition of ESL it's not the + it's not the question
		tpc	10:43:58	10:43:58	generally
		tpa	10:43:59	10:44:16	no because because the + the the interactive element + on the whiteboard + euh + name ESL games + ESL help
		tpa	10:44:16	10:44:18	yes +
		tpa	10:44:18	10:44:24	yes but ++ but it's not important I think +
		tpa	10:44:24	10:44:32	I euh +++ for me euh ++ it it's important ++
		tpa	10:44:32	10:44:33	ok
		tpa	10:44:33	10:44:42	9
		tpa	10:44:42	10:44:46	do you want to answer at the third ++ euh question
		tpc	10:44:45	10:44:45	... are both equivalentents of "mainly"
		tpa	10:44:46	10:44:52	6
		tpa	10:44:52	10:44:54	euh + yes
		tpa	10:44:54	10:45:23	29
action		as	10:45:04	10:45:04	s
		prod	10:45:04	10:45:04	TT(sortir)
		tpa	10:45:23	10:45:30	I euh + for me it's a games ++ maybe + I don't know <rires>
		tpa	10:45:30	10:45:51	21
nonverbal		prod	10:45:35	10:45:41	TT(cr��er) : 'perhaps games
		tpa	10:45:51	10:46:00	euh worst euh + of them is euh ++ the mo
		tpa	10:46:00	10:46:07	yes + it's euh the more bad + but you mus
		tpa	10:46:07	10:46:09	yes <rires>
		tpa	10:46:09	10:46:14	and and ++ and help is no + no good
		tpa	10:46:14	10:46:29	15
		prod	10:46:18	10:46:27	TT(��diter) ; 'perhaps games an also help

Learner asking for help

Tutor

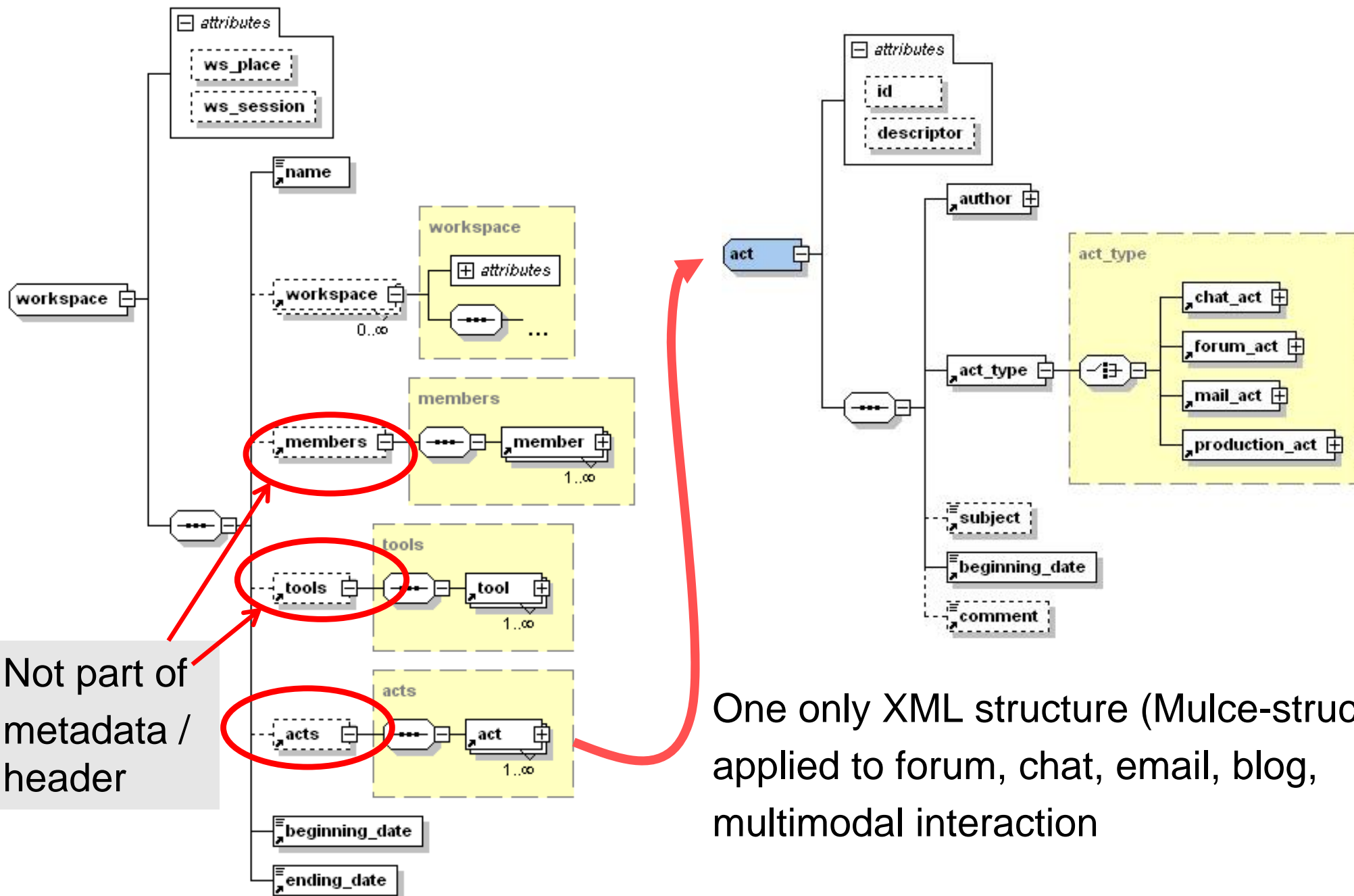
Links between acts into one space / division

chat

action

nonverbal

Converting interaction space into TEI?



Not part of metadata / header

One only XML structure (Mulce-struct) applied to forum, chat, email, blog, multimodal interaction

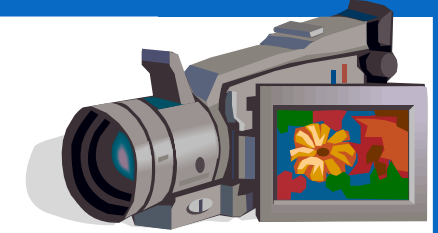
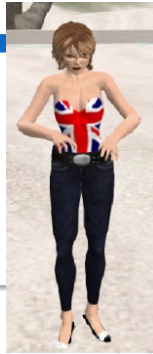
The element `<posting>` is the basic CMC-specific element in our schema. In CMC documents it represents the largest structural unit that can be assigned to one author and one point in time. **The category *posting* is defined as a content unit that has been sent to the server “en bloc”.**

TEI and CMC,
(Beißwenger et al., 2012)


- (3) **aud, tingrabu [07:20-08:48]:** ok hm for me this presentation was hm + become `<anno id="an18">too fast</anno>` because it's always the same in our architecture school euh we have not time and hm + `<anno id="an21" function="form" ntl="gram" type="cf-rpt cf-ack" ref="an19">too quickly sorry</anno>` and hm + we can't do good images because euh + euh it's xtime I don't know ++ and euh of course we whole project ++ is about motion and hm we make just some pictures hm statics pictures and hm it's + and it's it's a big matter because hm we always brought about teleportation our + motion is and hm +++ and `<anno id="an27" function="form" ntl="lex" type="rpt ack" ref="an29">everyday lack of time ok thank you</anno>` xxx and hm this is + this is hm really difficult for us because hm `<anno id="an28">we have not enough time</anno>` to do good presentation euh in + one night and I hope so tues wednesday could be better + it should be + may be I don't know `<anno id="an32" function="form" type="ack" ref="an31">[chuckles]</anno>`
- tc, <form> tfrez2, [07:32-07:33]:** `<anno id="an19" function="form" ntl="gram" type="cf-con" author="tut" ref="an18">it went too quickly?</anno>`
- tc, <form> tfrez2, [07:38-07:38]:** `<anno id="an20" function="task" type="cf-con" author="tut" ref="an18">or it was too early in the week?</anno>`
- tc, <task> romeorez [07:54-07:55]:** `<anno id="an22" ref="an20">i think it was to early</anno>`
- tc, <form> romeorez [07:59-07:59]:** `<anno id="an23" function="form" ntl="typ" type="cf-sr" author="st" ref="an22">too</anno>`
- tc, <form> tfrez2 [07:59-07:59]:** `<anno id="an24" function="form" ntl="gram" type="cf-rec" author="tut" ref="an22">too early</anno>` `<anno id="an25" function="form" type="cf-ref" author="tut" ref="an23"> ok</anno>`
- tc, <form> tfrez2 [08:08-08:10]:** `<anno id="an26" function="form" ntl="gram" type="cf-ml" author="tut" ref="an21">too quickly means that you didn't have enough time to speak</anno>`
- tc, <task form> quentinrez [08:16-08:16]:** `<anno id="an29" type="cf-pr" author="pr" ref="an28">yes, it's an everyday lack of time</anno>`
- tc, <task> romeorez [08:43-08:43]:** `<anno id="an30" ref="an28">that more that we have to show something that we don't really know </anno>`
- tc, <form> tfrez2 [08:08-08:10]:** `<anno id="an31" function="form" ntl="gram" type="cf-rec" author="tut" ref="an28 an29">you didn't have enough time</anno>`
- tc, <task> romeorez [08:43-08:44]:** `<anno id="an27" ref="an28">fore the shape</anno>`

Wigham, C.R. & T. Chanier (2013). "A study of verbal and nonverbal communication in Second Life. the ARCHI21 experience". *ReCALL* 25(1), DOI: 10.1017/S0958344012000250

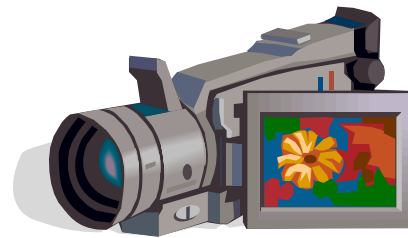
Modality interplay



1.5 mn video

tingrabu	tfrez2	romeorez	quentinrez
<p>ok for me this presentaion was become too fast because it's always the same in our architectural school we have not time and too quickly sorry and we can't do good images because it's less time uh I don't know [...] and it's a big matter because we always talk about teleportation [...] an everyday lack of time ok thank you quentinrez and this is very difficult [...]</p>	<p>it went too quickly or it was too early in the week? ok too quickly means you didn't have enough time</p>	<p>i think it was to early too</p>	<p>yes, it's an everyday lack of time</p> <p><i>* Paper: (Wigham & Chanier, 2013) CALL journal</i> <i>* Data: (Wigham, 2013) LETEC corpus</i></p> 

halshs-00878833, version 1 - 31 Oct 2013



1.5 mn video

Extract of Wigham, C.R. & Chanier, T. (2013) Pedagogical corpus: Textchat in multimodal contexts. Mulce.org : Clermont Université. [oai : mulce.org:mce-peda-textchat ; <http://repository.mulce.org>].

Metadata for CMC documents



paper 3

Why do we need metadata?

- Identification ...
- Preservation ...
- Location ...
- Evaluation ...
- Scientific validation ...
- Management ...
- Retrieval ...
- Comparison and sharing ...
- Technical dissemination ...

... of resources (and derived data) across different projects

Overall goal: being **accessible** and **useful** to others

Guiding question:

Can we provide a harmonized approach to CMC metadata, i.e. a recommended common CMC `<teiHeader>`?

Trade-off between:

- Tacit knowledge vs. explicit and formal modeling of MD
- Documentation vs. MD
- Granularity (and complexity) of MD (flat DC style vs. hierarchical TEI)

Some use cases (easy to difficult):

- Local resource/corpus management
- Indexes, catalogues (e.g. META, ELRA)
- Exploitation via research infrastructures (e.g CLARIN)

```
<Web2Corpus_itHeader>  
<recordDesc>  
<encodingDesc>  
<profileDesc>
```

```
<recordDesc>  
  <agentStmt>  
    <editor role="creator"  
type="individual">M.D.C.</editor>  
    <editor role="anonymizer"  
type="automatic">Chat Anonymizer Tool</  
editor>  
    <editor role="XML annotator"  
type="individual">M.S.</editor>  
  <revisionDesc>  
    <createdate>2012-03-01T10:37Z</  
createdate>  
    <lastmodified>2013-07-07T18:25Z</  
lastmodified >
```

Project oriented – hybrid between METS and TEI

```
    <date>2012-05-04T11:42Z </  
date>Checked manually for  
anonymisation</item>  
    <recordstatus>complete</ recordstatus>  
  </revisionDesc>  
</revisionStmt>  
<sourceStmt>  
  <URL type="root" >www.chatta.it/</URL>  
  <licence></licence>  
</sourceStmt>  
<fileStmt>  
  <saved location>...chat</saved location>  
  <filename>b_chat_3</filename>  
</fileStmt>  
</recordDesc>
```

<profileDesc>

<platformDesc>

<CMC genre="chat"/>

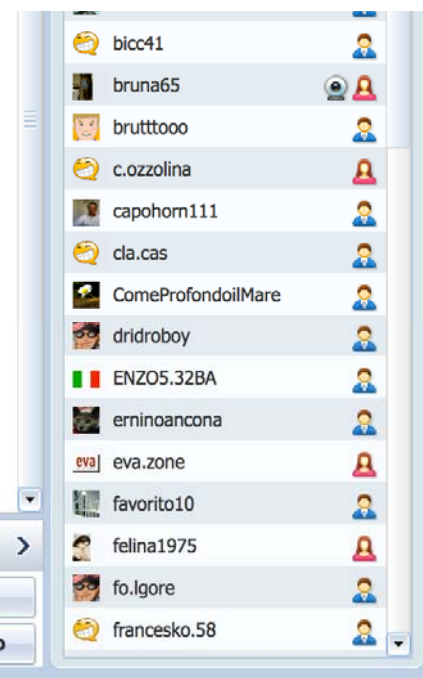
<access>public registered</access>

<channel mode="w"/>

<communication mode="synchronous"/>

<temporal dimension type="linear">up to
down</temporal dimension>

</platformDesc>



<textDesc>

<channel mode="w"/>

<access>public registered</access>

<start date>2011-02-04T10:12Z</start date>

<last contribution date>2011-12-13T11:57Z</last contribution date>

<participants number>22</participants number>

<posts in thread>134</posts in thread>

</textDesc>

</profileDesc>

* pacchetto23c è uscito dal canale

senderA: tn timer mastino

[server] * il topic può essere modificato solo da un operatore

[server] * gli utenti non presenti nel canale non possono inviare messaggi

* senderK è uscito dal canale

* senderL è entrato nel canale

* senderM è entrato nel canale

* senderN è uscito dal canale

senderB: io nn la vado a vedere

* senderO è uscito dal canale

senderB: gia domani devo fare la delice

<participantDesc>

- list of participants
- reference to anonymisation table (external)
- participant information extracted from the original page (age, sex, location, signature, etc.), only if formally present in the page template.



26-05-2011, 13:25

conogelato ◦

Candle in the wind
★★★

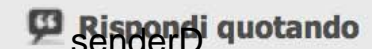


Registrato dal: 17/07/06
residenza: Empoli
Messaggi: 33,400
Post nel Blog: 1

Si, tutto vero. Ma possibile, ma dove noi legate.....

amate i vostri nemici

FILE	USERNAME	ANON.CODE
b_chat_1.txt	EmY	senderA
b_chat_1.txt	Lilly	senderAA
b_chat_1.txt	^CATGUT^	senderAB
b_chat_1.txt	MassY^	senderAC
b_chat_1.txt	Mattw27	senderAD
b_chat_1.txt	tatuatore	senderAE
b_chat_1.txt	ehi	senderAF
b_chat_1.txt	lalo	senderAG
b_chat_1.txt	J0ker	senderB
b_chat_1.txt	so`Rakkia	senderC
b_chat_1.txt	henryyy	senderD



halshts-00878833, version 1 - 31 Oct 2013

What MD needs to be recorded (in general)?

Different MD types (and different typologies for MD types)

E.g. NISO (generic view, see METS):

- **Descriptive** MD (e.g. searching, location of resources)
- **Structural** MD (e.g. technical organisation of resources)
- **Administrative** MD (e.g. legal status, preservation status)

E.g. TEI (historically tailored to digital/digitized texts):

- **fileDesc**
- **encodingDesc**
- **profileDesc**
- **revisionDesc**

Borders between MD types are somewhat blurry

What needs to be recorded (more CMC specific)?

- Genre classification (`profileDesc`, descriptive)
- Underlying text model (`encodingDesc`, structural)
- Description of interfaces/environment used for text production/communication (`profileDesc` (?), descriptive)
- Types of encoded (linguistic, communicative) features (`encodingDesc`, structural)
- Discourse participants (`profileDesc` (?), descriptive)
- Timelines (problematic, should be in `profileDesc` (?), descriptive)

... and probably a lot more

The focus should be on `profileDesc` and `encodingDesc`.

Genre classification

Maintain a finite set of genres?

Maintain a set of functional and formal features/properties:

- synchronous vs. asynchronous communication
- human—human vs. human—machine vs. ...
- written vs. spoken discourse
- 1:1 vs. 1:n vs. m:n communication
- private vs. public communication
- ...

Can we agree on a basic list of required, optional, and recommended features?

Underlying text model

CMC texts are produced by different agents (human and non-human).

Microstructure, the “building blocks”

real metadata (e.g. tagsDecl)
or rather documentation (e.g. in the ODD file)?

Macrostructure, arrangement of the “building blocks”

Typically given in the ODD and derived schemata, not in the `teiHeader`

Discourse participants

Natural place: `particDesc(/listPerson)`:

- Name
- Social description (@sex, @age, @role, occupation, ...) – **mostly of uncertain validity** (self declared, degree of certainty must be made explicit)
- Unformalized descriptions

Beware issues of anonymization/pseudonymization!

Consent of participants (implicitly given through publication of source?)

- General organization of `teiHeader` into components
- Explicit representation/description of agents that produced a text/communication
- Extension of `profileDesc` to cater for more internal and external features/properties of resources
- Agreed description of CMC genres/text types (ideally via ISOcat)
- Timeline representation as “real” metadata
- persistent reference to CMC text sources (little influence of MD creators on source persistence except full copying)



Google Nach Themen suchen ANMELDEN

Gruppen NEUES THEMA Filter

tei-cmc Öffentlich geteilt 2 von 2 Themen

- initial thoughts Von Stuart Yeates - 1 Beitrag - 2 Aufrufe 15. Sep
- Welcome to the mailing list of the TEI-CMC SIG! Von Michael Beißwenger - 1 Beitrag - 9 Aufrufe 16. Aug

Mailing list: <https://groups.google.com/d/forum/tei-cmc>

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