Community session:
Models and tools for structuring & annotating corpora of social media / computer-mediated communication

4th DARIAH-EU General VCC meeting

Rome, September 18, 2014

Michael Beißwenger (Dortmund / DE)
Thierry Chanier (Clermont / FR)
Isabella Chiari (Roma / IT)

http://www.tei-c.org/Activities/SIG/CMC/
Our subject and goals

Our subject:

- building and annotating corpora of computer-mediated communication (CMC) – as resources for empirical research on CMC phenomena in the Humanities (linguistics, communication science, language technology, …)
Computer-mediated communication (CMC):

All genres of interpersonal communication mediated through computer networks (the internet) and used via personal computers and/or mobile devices: chats, online forums, instant messaging, tweets, 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), SMS, WhatsApp, ....
Our subject and goals

Our subject:

- building and annotating corpora of computer-mediated communication (CMC) – as resources for empirical research on CMC phenomena in the Humanities (linguistics, communication science, language technology, …)

- Ongoing CMC corpus projects for several European languages (French, German, Italian, Dutch)
Our subject and goals

Our subject:

- building and annotating corpora of computer-mediated communication (CMC) – as resources for empirical research on CMC phenomena in the Humanities (linguistics, communication science, language technology, …)
- Ongoing CMC corpus projects for several European languages (French, German, Italian, Dutch)

Our vision: These corpora shall be …

- interoperable (i) with each other and (ii) with other types of linguistic corpora (text corpora, speech corpora)
- represented conformant to established encoding standards in the field of Digital Humanities
- linguistically annotated in order to allow for sophisticated queries and language-focused research
The problem / challenge:

- By now, there are no established standards for the representation of CMC genres.
- Established standards for the representation of text genres do not include models for the representation of the peculiarities of CMC.
- “Off the shelf” NLP tools for automatic linguistic analysis and annotation (tokenizers, part-of-speech taggers, lematizers, normalizers, parsers) do not perform well on CMC data (because they usually have been trained on edited text and therefore can’t handle “non-standard” phenomena and multimodal elements in CMC discourse).
Our subject and goals

Our goals:

- work on solutions for these desiderata
- develop suggestions for standards for
  - packaging and sharing (mono- and multimodal) CMC corpora,
  - modeling these types of “texts” within a framework which is conformant with the encoding framework of the Text Encoding Initiative (TEI) and thus with a widely accepted de-facto standard in the field of Digital Humanities,
  - processing and annotating these corpora (part-of-speech, normalization, ...) with NLP tools.

... in cooperation with interested experts, projects and initiatives in the field
Who belongs to our community (so far)?

Our kernel projects and founding members

- **Web2Corpus_IT**
  - Corpus italiano di comunicazione mediata dal computer
  - [http://glottoweb.org/web2corpus/](http://glottoweb.org/web2corpus/)
  - Italian CMC pilot corpus

- **French CMC corpora**
  - [CoMeRe](http://hdl.handle.net/11403/comere)
  - [ORTOLANG](http://www.ortolang.org)
  - [DARIAH-FR](http://www.dariah-eu.org)

- **German CMC corpora**
  - [Dortmund Chat Corpus](http://www.chatkorpus.tu-dortmund.de)
  - [National infrastructure for Digital Humanities](http://www.tinyurl.com/derik-llc)
  - [German Reference Corpus of CMC](http://www.tinyurl.com/derik-llc)

- **Dutch CMC corpora**
  - [CLST](http://www.clt.nl)
  - [SoNaR](http://www.sonaar.nl)

- **Scientific network**
  - "Empirical research of CMC"
  - [http://hdl.handle.net/11403/comere](http://hdl.handle.net/11403/comere)

- **National consortium on corpora**
  - [National infrastructure for Digital Humanities](http://www.tinyurl.com/derik-llc)

... YOUR projects!! 😊
Activities and initiatives (past and future)

2013, 2014
- European workshop on CMC corpora (Dortmund) ⇒ network of corpus projects
- Meetings & workshops of members of the TEI CMC-SIG (Clermont, Rome, Dortmund)
- special journal issue (JLCL)

2013 creation of the TEI-CMC SIG

2015, 2016

End of 2014
Publication of CMC corpora in open access
Several types, all TEI

2015
Submission of first model suggestions/feature requests of the TEI CMC-SIG to the TEI Community & Council

Our pathway

Activities and initiatives (past and future)

Our pathway

2013, 2014
- European workshop on CMC corpora (Dortmund) ⇒ network of corpus projects
- Meetings & workshops of members of the TEI CMC-SIG (Clermont, Rome, Dortmund)
- special journal issue (JLCL)

2013 creation of the TEI-CMC SIG

2015, 2016

End of 2014
Publication of CMC corpora in open access
Several types, all TEI

2015
Submission of first model suggestions/feature requests of the TEI CMC-SIG to the TEI Community & Council

Our pathway
We’re looking forward to get in touch!

Community Sessions

“Models and tools for structuring & annotating corpora of social media / computer-mediated communication”

PROGRAM

Pt. I: Presentations (Thu 15:00-17:00)
- Thierry Chanier (UBP Clermont): End of phase 1 of the CoMeRe project: application of the Interaction Space (TEI-CMC) model to various CMC corpora in French
- Harald Lüngen & Eliza Margareta (IDS Mannheim): Applying the TEI CMC SIG proposal to Wikipedia corpora
- Michael Beilwenger & Christian Politz (TU Dortmund): Analyzing CMC corpora using Machine Learning methods: report from the KobRA project
- Benoit Sagot (Alpago/CNRS & INRIA): Application of a general POS tagger for corpora in French to the CoMeRe-CMC corpora

Pt. II: Round table: further work on standards and joint scientific activities (Thu 17:00-18:00)
- Discussion of the current draft schema from the perspective of different projects and initiatives; ideas for annotations on the microlevel (= inner structure of postings) and about the interface between POS annotations and the microstructure in the TEI schema
- schedule for extending the standards (e.g., the next step in the TEI-CMC SIG)
- plans for scientific events in 2015-16 focused on annotations, NLP4CMC, work on wikipedia & other genres
- platform for the documentation and exchange of NLP tools and annotation experiments?

The TEI CMC-SIG on the web:
- http://www.tei-c.org/Activities/SIG/CMC/

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