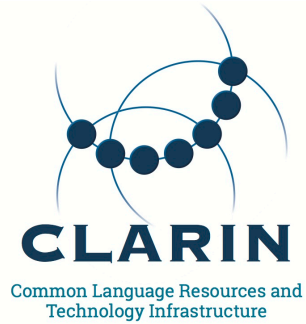
Integrating corpora of computer-mediated communication into the language resources landscape:

Initiatives and best practices from French, German, Italian and Slovenian projects

Michael Beißwenger · Thierry Chanier · Isabella Chiari
Tomaž Erjavec · Darja Fišer · Axel Herold
Nikola Ljubešić · Harald Lüngen · Céline Poudat
Egon Stemle · Angelika Storrer · Ciara Wigham



CLARIN Annual Conference
October 26–28, 2016
Aix-en-Provence

CMC corpora: a new type of resources

A new type of language resources:

corpora of computer-mediated communication / interaction in social media (CMC)

⇨ **subject:**

the interactional language use found in internet-based communication technologies –

genres such as chats, forums, tweets, social network sites (Facebook, Instagram etc.), blog comments, Wikipedia talk pages, instant messaging, whatsapp and sms dialogues – moreover in multimodal learning environments, in online (video-)conferencing systems, on youtube, and in 3D „virtual worlds“

CMC and corpus linguistics

CMC corpora: the "naughty stepchild" of text and speech corpora:

- unclear legal status of CMC 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 CMC genres
- No established NLP tools which can be used for automatic processing and linguistic annotation (deviation of CMC discourse from the written standard)

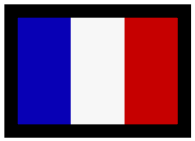
▸ Only (very) few corpus resources which are available for the scientific community/the public

☛ "CMC gap" in the corpus landscape

CMC and corpus linguistics

In recent years:

- increasing number of projects in several countries which have started addressing these issues with the goal to close the CMC gap and create CMC corpora which shall be made available as reference corpora for CMC
- Window of opportunity for joint efforts in creating best practices and standards for this new type of corpora in an bottom-up approach
- European bottom-up network of CMC corpus projects (initiated 2013, co-working on selected tasks, exchanging knowledge and practices, annual meetings, *still growing and open for new colleagues to join in*...)



Projects in four countries: FRANCE

CoMeRe

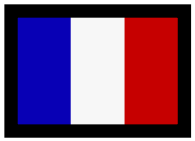
(Communication Médiée par les Réseaux)

ORTOLANG

Contact:

Thierry Chanier
Université Blaise Pascal, Clermont, France
thierry.chanier@univ-bpclermont.fr

Ciara Wigham
Université Blaise Pascal, Clermont, France
ciara.wigham@univ-bpclermont.fr




Projects in four countries: FRANCE

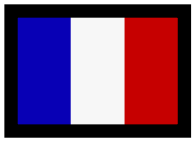
CoMeRe

(Communication Médiée par les Réseaux)

ORTOLANG

Network and resources:

- 14 researchers from different research units who had previously collected CMC corpora on a variety of CMC genres (mono- vs. multimodal; synchronous vs. asynchronous; public vs. private sphere)




Projects in four countries: FRANCE

CoMeRe

(Communication Médiée par les Réseaux)

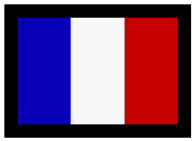
ORTOLANG

Network and resources:

- 14 researchers from different research units who had previously collected CMC corpora on a variety of CMC genres (mono- vs. multimodal; synchronous vs. asynchronous; public vs. private sphere)
- resources structured in heterogeneous formats (different XML schemas, spreadsheets, ...)

Goals of the CoMeRe project:

- design a common model for CMC discourse that fits the pre-existing corpora (⇨ CoMeRe TEI schema, Chanier et al. 2014)
- re-model the corpora in this format
- release the corpora in a common repository as open data (CC-BY-SA licence for corpus end users)




Projects in four countries: FRANCE

CoMeRe

(Communication Médiée par les Réseaux)

ORTOLANG

Current state of work:

14 CMC corpora for 9 CMC genres available as downloadable resources:

SMS	Tweets	Email	Text chat	Multimodal
- cmr-smslareunion - cmr-smsalpes - cmr-88milsms	- cmr-polititweets - cmr-intermittent	- cmr-simuligne	- cmr-getalp_org - cmr-favi	- cmr-copeas - cmr-tridem06
Wiki discussions	**Weblog**	**Discussion forum**	- cmr-favi (POS tagged)	**Multimodal + 3D**
- cmr-wikiconflits	- cmr-infral	- cmr-simuligne	- cmr-simuligne	- cmr-archi21

http://hdl.handle.net/11403/comere

Next step: Part of speech tagging for the CMC corpora (work in progress, one corpus already tagged)

Projects in four countries: GERMANY

ChatCorpus2CLARIN

CLARIN-D curation project, 2015-16

Contact:

Michael Beißwenger
University of Duisburg-Essen, Germany
michael.beisswenger@uni-due.de

Axel Herold
Berlin-Brandenburg Academy of Sciences, Berlin, Germany
herold@bbaw.de

Harald Lüngen
Institute for the German Language, Mannheim, Germany
luengen@ids-mannheim.de

Angelika Storrer
University of Mannheim, Mannheim, Germany
astorrer@mail.uni-mannheim.de

Projects in four countries: GERMANY

ChatCorpus2CLARIN

CLARIN-D curation project, 2015-16

Primary goals of the project:

- Integrate an existing 1 million token German chat corpus *(Dortmund Chat Corpus*, available for free download since 2005) into the corpus infrastructures at the CLARIN-D hubs at IDS (Mannheim) and BBAW (Berlin)
- Re-model the resource in TEI
  ⇨ CLARIN-D TEI schema (Beißwenger/Lüngen/Herold/Storrer 2015)
- Enhance the resource with additional linguistic annotations (part-of-speech tags)
  ⇨ Toolchain from U Saarbrücken (Horbach et al. 2014)
- Seek for clarification about republishing the resource as part of the CLARIN-D infrastructure
  ⇨ Legal opinion by John Weitzmann and colleagues

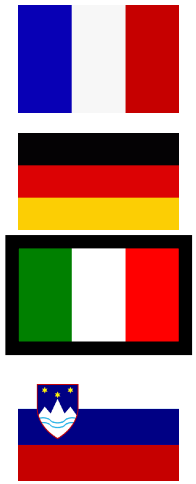
Projects in four countries: GERMANY

ChatCorpus2CLARIN

CLARIN-D curation project, 2015-16

Secondary goals of the project:

- **Create a showcase** which demonstrates what researchers can gain when CMC corpora – as part of corpus collections – can be analyzed in combination with other language resources (text and speech corpora at IDS and BBAW)
- **Intended best practice character of solutions:** The solutions developed in the project should be useful not only for the chat corpus but also for the modeling and integration of other CMC corpora into CLARIN-D (⇨ future projects)



Projects in four countries: ITALY

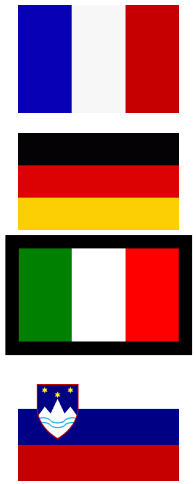
didi Digital Natives – Digital Immigrants

Project at EURAC research, Bolzano/Bozen,
06/2013 - 07/2015

EURAC research

Contact:

Egon Stemle
EURAC, Bolzano, Italy
egon.stemle@eurac.edu



Projects in four countries: ITALY

didi Digital Natives – Digital Immigrants

Project at EURAC research, Bolzano/Bozen, 06/2013 - 07/2015

EURAC research

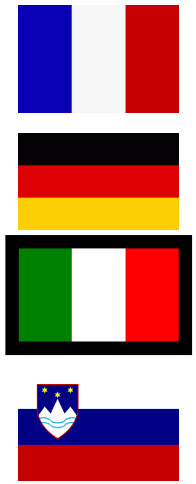
Corpus of Facebook status updates, comments, messages

Size / language(s) / data types:

- ~600.000 Tokens, ~40.000 Texts in German and Italian (often in South Tyrolean dialect) from 136 Users, incl. socio-demographic information about the users (gender, education, internet communication habits, L1(s), usage of a dialect and its micro-geographical origin)
- All German text manually normalized and anonymized

Availability:

- The corpus is accessible for querying via ANNIS or can be obtained as processable data for research purposes on http://www.eurac.edu/didi



Projects in four countries: ITALY

didi Digital Natives – Digital Immigrants

Project at EURAC research, Bolzano/Bozen, 06/2013 - 07/2015

EURAC research

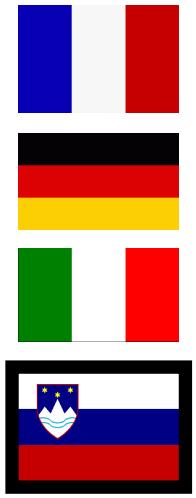
Corpus of Facebook status updates, comments, messages

Data collection from the private sphere:

- Date collected from within Facebook (users had to explicitly agree with the distribution of their data and were also able to restrict the collection of data to the publicly available part)

Further outcome:

- Recommendations for collecting private, non-public CMC data



Projects in four countries: SLOVENIA

The JANES corpus of Slovene CMC

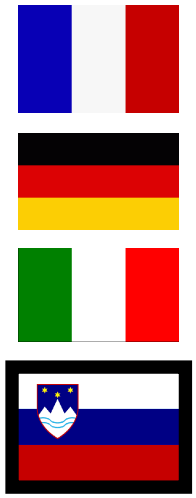
Jezikoslovna analiza nestandardne slovenščine
(Linguistic Analysis of Nonstandard Slovene)

Contact:

Darja Fišer
University of Ljubljana, Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Tomaž Erjavec
Jožef Stefan Institute, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana, Slovenia
nikola.ljubesic@ffzg.hr



Projects in four countries: SLOVENIA

The JANES corpus of Slovene CMC

Jezikoslovna analiza nestandardne slovenščine
(Linguistic Analysis of Nonstandard Slovene)

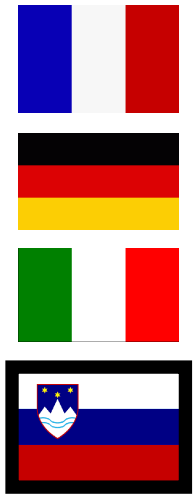
National research project *Resources, Tools and Methods for the Research of Nonstandard Internet Slovene* 2014 –2017

Goal: Building a corpus Slovene tweets, forums, blogs, news comments and Wikipedia talk pages – present version of the resource already used in a number of linguistic analyses

Workflow: crawling, cleaning, structuring, annotation

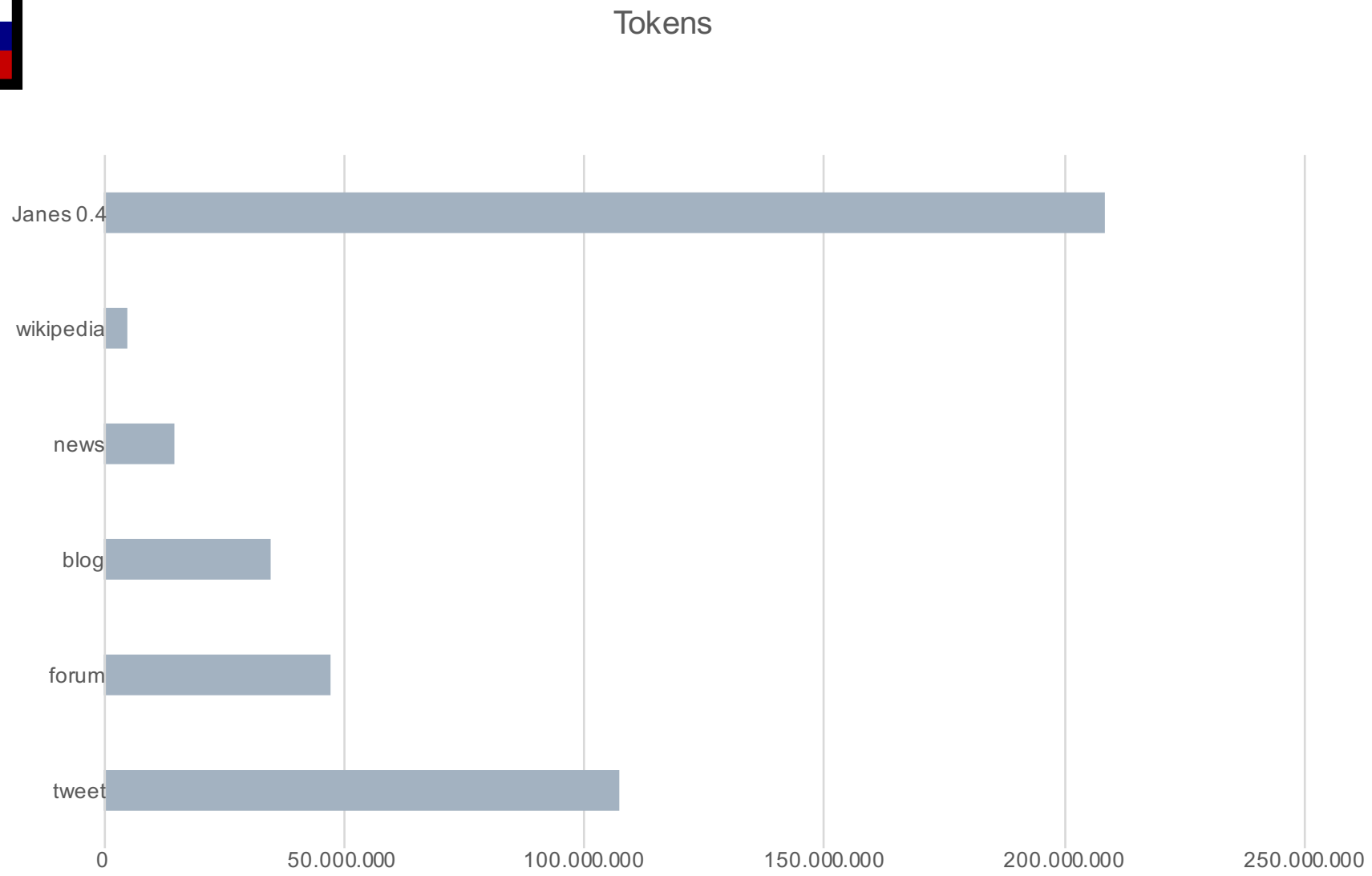
Annotation levels and strategy:

- *Word-level:* rediacritisation, normalisation, PoS tagging, lemmatisation
- *Text-level:* time, user, etc. + standardness, sentiment
- Use of manually annotated texts + machine learning methods



Projects in four countries: SLOVENIA

The JANES corpus of Slovene CMC in numbers:



Towards the creation of standards for cmc corpora

The vision: create CMC corpus resources which are

- available for the scientific community (open access)
- represented using open, non-proprietary and therefore sustainable encoding, exchange and metadata formats
- interoperable with each other as well as with state-of-the-art text and speech corpora through using acknowledged standards in the field of digital humanities
- annotated for purposes of language-centered research
- not only published as standalone resources but particularly as part of existing language resource infrastructures (CLARIN, ORTOLANG)

Activities of the network (1): The *TEI CMC-SIG*

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

mission:

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

results so far:

three customized TEI schemas for CMC:
DeRiK (2012), CoMeRe (2014), CLARIN-D (2015)
(available online & ready for use)

next step on the road map:

integration of models into the TEI standard
(feature request ⇨ standardization process ...)

Activities of the network (2): *cmc-corpora.org*

Conference on CMC and Social Media Corpora for the Humanities (*cmccorpora*)

http://www.cmc-corpora.org

- **conference series**
- **4 international events since 2013**; 5th ed. in prep. for 2017
- Latest edition (Ljubljana, Sept. 2016):
  22 contributions by 40 researchers from 24 research institutions in 11 countries, addressing key issues and current trends in corpus-based CMC research on data from 8 languages
- **Proceedings of *cmccorpora16*:**
  http://nl.ijs.si/janes/cmc-corpora2016/proceedings/

Open issues & outlook

How could CLARIN (and members of the CLARIN community) support the CMC corpus community?

- participate in the further discussion process about integrating a CMC schema into TEI – and promote the result as a CLARIN recognized best practice
- support us with seeking legal opinions on the conditions for collecting and republishing CMC data in corpora – a CLARIN guideline / best practice on how to deal with CMC and social media data in language resources would facilitate creating open access CMC resources

Don't hesitate to get in touch!

⇨ at our poster / at the CLARIN Bazaar