Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries

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

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

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

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

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

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

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

Céline Poudat
Université de Nice Sophia Antipolis
France
poudat@unice.fr

Egon Stemle
Eurac Research
Bolzano
Italy
egon.stemle@eurac.edu

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

Ciara Wigham
Université Clermont Auvergne, France
ciara.wigham@uca.fr

Abstract

The paper presents best practices and results from projects dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC) from four different countries. Even though there are still many open issues related to building and annotating corpora of this type, there already exists a range of tested solutions which may serve as a starting point for a comprehensive discussion on how future standards for CMC corpora could (and should) be shaped like.

1 Introduction

The paper presents best practices and results from projects dedicated to the creation of corpora of computer-mediated communication and social media interactions (henceforth referred to as *CMC*) from four European countries. The projects are inter-related via a bottom-up network of researchers interested in fostering the transfer of expertise and solutions for handling this relatively new type of language resources and for modeling the structural and linguistic peculiarities of (written and multimodal) discourse found in chat, forum, sms and whatsapp interactions, in weblogs and wikis, on social network sites and in multimodal CMC environments. This new type of discourse exhibits features that cannot be adequately handled by the schemas and tools which have been developed for

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/ https://catalog.ldc.upenn.edu/LDC2010T05

Michael Beißwenger, Thierry Chanier, Tomaž Erjavec, Darja Fišer, Axel Herold, Nikola Ljubešić, Harald Lüngen, Céline Poudat, Egon Stemle, Angelika Storrer and Ciara Wigham 2017. Closing a gap in the language resources landscape: Groundwork and best practices from projects on computer-mediated communication in four European countries. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 1–18.

the representation, annotation and processing of discourse which conforms to the written standard and the structural conventions of established text types (e.g., newspaper articles, prose, scientific articles). In addition with the collection and redistribution of CMC data in linguistic corpora, legal and ethical issues arise which are not yet sufficiently covered by existing laws and ethical standards. What is more, there are no established standards for metadata and for the documentation of the (technological, hypermedial and social) context in which CMC data are typically embedded, produced and used.

Corpus-linguistic approaches to CMC have so far not found answers to all of these challenges. Nevertheless, existing projects in the field have proposed and tested an encouraging range of solutions and best practices. The joint goal of the projects and initiatives described in this paper is to pave the ground for standards which will allow CMC corpora to be interoperable (a) with each other and (b) with language resources for other types of discourse (text and speech corpora).

The paper is structured as follows: Section 2 gives an overview of existing CMC corpora and corpus projects. Section 3 describes two initiatives dedicated to the development of standards and to the exchange of knowledge related to the collection, annotation, representation and provision of CMC corpora. Section 4 gives an overview of the results and best practices from CMC corpus projects in four countries which may be useful for other projects in the field and which may serve as a starting point for a more comprehensive discussion on how future standards for CMC corpora could (and should) be shaped like.

2 Overview of CMC corpora and corpus projects

Even though research on CMC in linguistics and social sciences from its very beginning had a strong empirical focus, only few corpora or datasets have been made available to the scientific public. An overview of CMC corpora is given in Beißwenger and Storrer (2008). Examples of 'early-bird' CMC corpora are:

- the *NPS Chat Corpus* for English (Forsyth and Martell, 2007) with 45.000 tokens from age-specific chat rooms which have been annotated with part-of-speech information and a dialog-act classification. The corpus is available via the Linguistic Data Consortium (LDC).
- The *Dortmund Chat Corpus* for German (Beißwenger, 2013) which comprises 1 million tokens of chat discourse with annotations of selected CMC-specific phenomena. The corpus is available for free download since 2005[1] and will be released in an enhanced version as part of CLARIN-D in spring 2017 (cf. Sect. 4.2).

More recently, a range of projects has created (or is currently creating) resources which have been or will be made available to the public – for example (in alphabetical order):

- *CoMeRe*: a collection of 14 French corpora for 9 different CMC genres represented in TEI, available for download via ORTOLANG (cf. Sect. 4.1).[2]
- *CorCenCC-CMC*: The "e-language" component in the project "National Corpus of Contemporary Welsh" (CorCenCC, since 2016).[3]
- *DEREKO-News*: Corpus of German Newsgroups in DEREKO, since 2013, 98 million tokens, available for online querying via COSMAS II (Schröck and Lüngen, 2015).[4]
- *DEREKO-Wikipedia*: Wikipedia corpora in DEREKO: German language article talk and user talk (cf. Margaretha and Lüngen, 2014), 581 million tokens, available for online querying via COSMAS II; also downloadable.
- *DiDi corpus*: The CMC corpus from the DiDi project with 570.000 tokens of German, Italian and South Tyrolean Facebook posts and interactions, available for online querying via ANNIS (Frey et al., 2016; cf. Sect. 4.3).[5]
- *DWDS blog corpus*: The blog corpus in the corpus collection of the DWDS project: 103 million tokens from CC-licensed, mainly German blog entries, available for online querying.[6]

[1] http://chatkorpus.tu-dortmund.de/
[2] http://hdl.handle.net/11403/comere
[3] Project page: http://sites.cardiff.ac.uk/corcencc/
[4] https://cosmas2.ids-mannheim.de/
[5] http://www.eurac.edu/didi
[6] https://www.dwds.de

- *Janes*: The Corpus of Nonstandard Slovene comprising >200 million tokens from tweets, forum posts, blogs, comments on news articles and Wikipedia discussions (Fišer et al., 2016; cf. Sect. 4.4).[7]
- *sms4science.ch*: a donation-based corpus of 650.000 tokens of SMS messages collected in Switzerland and comprising discourse in non-dialectal German, French, Swiss German, Italian and Romansh (Dürscheid and Stark, 2011), available for online querying in a full text version (SMS Navigator) and as a partially annotated version represented in ANNIS.[8]
- *SoNaR-CMC*: the CMC component (chats, tweets and sms messages) in the Reference Corpus of Contemporary Dutch (SoNaR, Oostdijk et al., 2013) which is available for online querying via CLARIN-NL (OpenSoNaR).[9]
- *Suomi24*: a collection of 2.38 billion tokens of discourse from Finnish discussion forums with morpho-syntactic annotations, available for download.[10]
- *whatsup-switzerland.ch*: Corpus of the project "Whats's up, Switzerland?": a collection of 5 million tokens from 650 whatsapp chats donated by Swiss smart phone users.[11]
- *Web2Corpus_it*: a balanced CMC corpus for Italian (in preparation) including discourse from forums, blogs, newsgroups, social networks and chats (Chiari and Canzonetti, 2014) created in the context of a project on negotiation strategies.[12]

Even though the sheer availability of CMC corpora is already a big step ahead towards closing the "CMC gap" in the corpus landscape, the existing corpora, in their current state, are represented and provided using heterogeneous technologies, representation formats and annotation schemas. The availability of a flexible standard for the representation and exchange of CMC resources would allow researchers and corpus providers to combine, merge and connect their resources (*interoperability*), and facilitate corpus-based research across languages and CMC genres and beyond the limitations of single corpora. The creation of such a standard in compliance with the existing standards in the field of digital humanities would, additionally, allow to combine CMC corpora with corpora of other type (text corpora, speech corpora) and thus open up new perspectives also for corpus-based research on commonalities and differences between CMC discourse and monologic written language and spoken conversations. Moreover, compliance with existing standards would increase the *sustainability* and *reusability* of resources.

3 *cmc-corpora.org*: a European network of CMC corpus projects

Since 2013 a loose network of projects with a joint interest in building, annotating and analyzing CMC corpora has set up two initiatives in order to (1) strengthen the exchange of expertise and best practices between projects and (2) lead the discussion of a representation standard for CMC genres in the context of a well-acknowledged standardization initiative in the Digital Humanities:

3.1 Conference series on CMC corpora

The network has established a series of international workshops and conferences dedicated to the creation of CMC corpora with previous events held in Dortmund/DE (2013, 2014), Rennes/F (2015) and Ljubljana/SI (2016), and a next event (the *5th Conference on CMC and Social Media Corpora for the Humanities*) scheduled to be held in October 2017 at Eurac Research in Bolzano/IT. These conferences are defined as peer-reviewed events with a coordinating and a scientific committee.[13] Since 2016 the conferences are accompanied by peer-reviewed proceedings which are published online (cf. Fišer and Beißwenger, 2016).

[7] Project page: http://nl.ijs.si/janes/

[8] http://www.sms4science.ch

[9] https://portal.clarin.nl/node/4195

[10] http://urn.fi/urn:nbn:fi:lb-201412171

[11] http://www.whatsup-switzerland.ch/

[12] Project page: http://www.glottoweb.org/web2corpus/

[13] http://www.cmc-corpora.org

3.2 TEI special interest group on CMC

The network succeeded with a proposal for the creation of a special interest group (SIG) on Computer-Mediated Communication in the *Text Encoding Initiative* (*TEI*, http://tei-c.org) in 2013. The goal of this SIG is to extend the TEI framework with additions dedicated to the representation of the structural and linguistic peculiarities of CMC genres. Starting from a discussion of a first schema draft defined by Beißwenger et al. (2012) the SIG created two advanced schema drafts ('CoMeRe schema', 2014, 'CLARIN-D schema', 2015) which have been tested with French and German corpora and which are currently being adopted by other further projects. The schemas developed by the SIG are defined following the rules for *customization* described in the TEI guidelines[14]. The basic structure and CMC-specific models of the schemas have been discussed with the TEI community in several panels at the annual TEI conferences and members' meetings and will be presented to the TEI Technical Council in the form of feature requests, i.e. suggestions for the extension of the "official" TEI standard. The latest version of the schema which builds on its predecessors is described in Sect. 4.2.2.

4 Groundwork and best practices from projects in Germany, France, Italy and Slovenia

4.1 The CLARIN-D curation project *ChatCorpus2CLARIN* (Germany)

4.1.1 Project description

In the project *ChatCorpus2CLARIN*, an existing chat corpus for German (the *Dortmund Chat Corpus*, Beißwenger, 2013) served as a use case to demonstrate how an integration of CMC and social media resources into the CLARIN-D corpus infrastructures could be accomplished in a way that the target resource (1) conforms to established standards for the representation and linguistic annotation of corpora in the Digital Humanities context and (2) can be a useful resource for doing comparative analyses of CMC discourse with other types of corpus resources in CLARIN-D (text and speech corpora). The original resource has been compiled in 2002–2005 and comprises 1 million tokens of German chat discourse from various domains (social chat, chat in the context of learning and teaching, advisory chats, chats in the media context). The data is represented using a 'homegrown' XML format which describes (different types of) individual user posts, selected linguistic phenomena (such as emoticons, addressing terms, action words and acronyms) and selected metadata about the chats and their participants. The corpus has been available online for download since 2005.[15] It has been used as a resource in a broad range of research and teaching contexts in linguistics and language technology.

In the project, the original resource was remodeled building on schema drafts from the TEI CMC-SIG (Sect. 3) to increase its interoperability with other types of corpora provided via CLARIN-D. To extend the research and query options for the target resource the corpus, in addition, was enhanced with a layer of linguistic annotations (tokens, parts of speech, lemmas).

The project was headed by Michael Beißwenger (U Dortmund) and Angelika Storrer (U Mannheim). Researchers from the CLARIN-D hubs at the Institute for the German Language (IDS), Mannheim (Harald Lüngen), and from the Berlin-Brandenburg Academy of Sciences (Axel Herold) were closely involved into all work packages of the project. A visualization of the workflow and resources used in the integration process is given in Figure 1 and described in detail in Lüngen et al. (2016).

[14] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html

[15] http://chatkorpus.tu-dortmund.de/

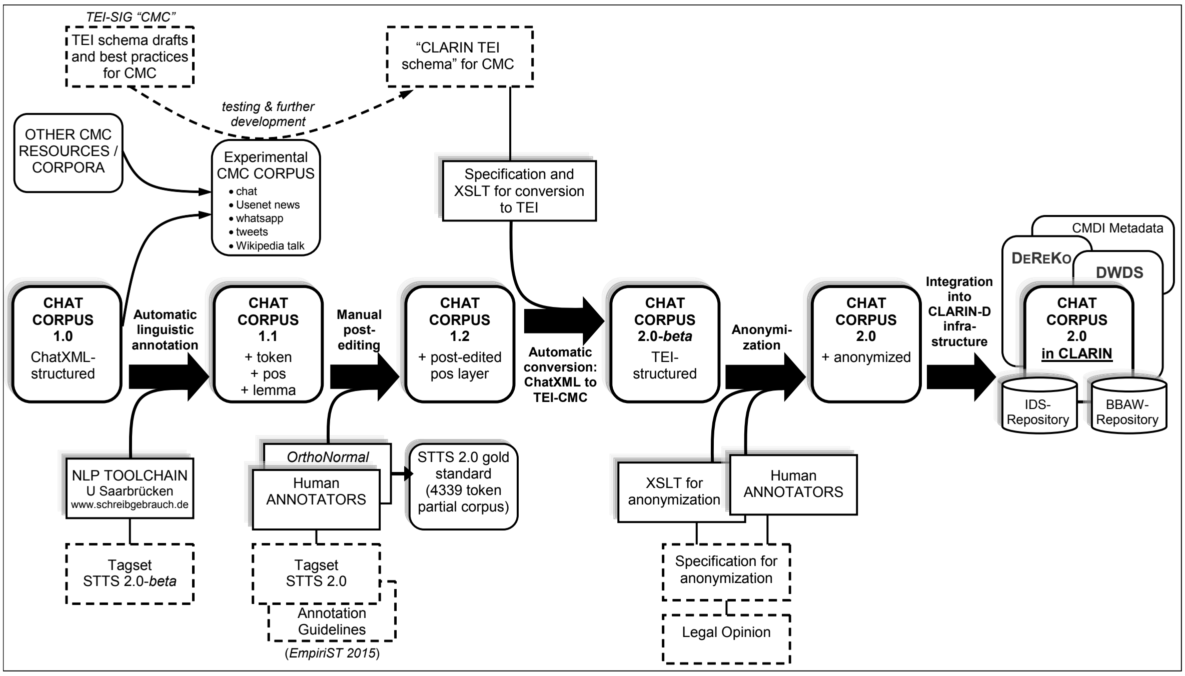


Figure 1: Integration of the Dortmund Chat Corpus into the CLARIN-D corpus infrastructure.

Main resources and work packages in the project workflow were:

- An **experimental CMC corpus:** For developing and testing the solutions developed for representing and annotating the corpus, we compiled a small experimental corpus with data from several CMC and social media genres (chat, news messages, Wikipedia talk pages, tweets, whatsapp interactions). This was done to guarantee that the annotation schema and tagset are useful not only for chat but also for a range of other types of (mainly) written CMC genres.
- **Linguistic annotation:** Tokenization, part-of-speech (PoS) tagging and lemmatization were done in two stages: (1) an automatic tagging process done at Saarland University applying the NLP toolchain described in Horbach et al. (2014) and (2) a manual post-editing phase with two trained annotators for a part of the resource (to demonstrate how a 'gold' annotation for chat data could look like).
- **The 'STTS 2.0' Part-of-Speech Tagset:** As target standard for the PoS layer, we adopted the tag set ('STTS 2.0'; Beißwenger et al., 2015) developed in the GSCL shared task on automatic linguistic annotation of CMC and social media (EmpiriST2015; Beißwenger et al., 2016)[16]. 'STTS 2.0' builds on the categories of the "Stuttgart-Tübingen Tagset" (*STTS*, Schiller et al., 1999) and introduces two types of new tags: (1) tags for phenomena which are specific for CMC and social media discourse, (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers and which can also be found in corpora of transcribed speech (e.g., in the FOLK corpus of spoken language at the IDS which uses an STTS extension which is compatible with 'STTS 2.0', Westpfahl and Schmidt, 2016). The resulting tag set is still downwardly compatible with STTS (1999) and therefore allows for interoperability with other corpora that have been tagged with STTS.
- **Legal clearance and anonymization:** Prior to the integration of the curated resource in CLARIN infrastructures, we sought a legal opinion to get a better picture of the legal conditions for republishing the material as a whole or in parts. The legal opinion (iRights.Law, 2016) carefully checked for possible restrictions arising from individual property rights, copyrights and other legal statutes. One result was that the possibility to identify individuals from their utterances (with the exception of public figures) needed to be circumvented by means of an anonymization. Only parts of the anonymization task could be done automatically; occurrences of names that had not been annotated in the original source,

[16] http://sites.google.com/site/empirist2015/

or that could not be matched to entries in the participant list automatically, had to be anonymized manually which was a very time-consuming process so that the date for the release of the integrated resource had to be postponed to spring 2017.

- **Development of a schema for remodeling the chat corpus in TEI:** To achieve interoperability with a broad range of other language resources in the digital humanities, the original resource was converted into a TEI format using customizations. Main features of the TEI schema developed in the project are outlined in Sect. 4.2.2.

The TEI schema, PoS tagset and anonymization guidelines will be reused and refined in follow-up CMC corpus projects within CLARIN-D – e.g. for representing and annotating data in the project *MoCoDa* (*Mobile Communication Database*) in which a database and web frontend for the repeated collection of data donations from whatsapp, sms and similar CMC 'apps' for mobile use will be created. The project which started in January 2017 is funded by the Ministry for Innovation, Science, Research and Technology of the German federal state North Rhine-Westfalia and by Michael Beißwenger (U Duisburg-Essen), Wolfgang Imo (U Halle-Wittenberg) and Evelyn Ziegler (U Duisburg-Essen).

4.1.2 Results beyond the resource: The 'CLARIN-D TEI schema' for CMC

The TEI schema developed in the project (the 'CLARIN-D TEI schema') is the result of a continued development based on two previous schemas for the representation of CMC discourse: the schema suggested by Beißwenger et al. (2012) and the schema developed in the CoMeRe project (Chanier et al., 2014; cf. Sect. 4.1.2). The development of the schema was fostered by extensive discussions within the TEI CMC-SIG (Sect. 3) and by discussions within the DFG scientific network *Empirikom*[17]. While aiming at providing a generic model for CMC discourse within the framework of the TEI, the CLARIN-D schema focuses on the representation of discourse captured in chat logfiles, whatsapp interactions, tweets and Wikipedia discussions. To amend the TEI guidelines (TEI-P5) for this CMC genre and reflect properties specific to logfiles, different types of customizations of the TEI guidelines had to be implemented:

(1) **New elements:** Three new elements were introduced to cater for building-blocks of computer-mediated interactions not yet covered by the TEI guidelines, namely <post> (which has first been described in the 'DeRiK TEI schema', Beißwenger et al., 2012 and see figure 2) and <prod> (originally introduced in the CoMeRe schema, Chanier et al., 2014). <post> is used to represent any written contribution to an ongoing CMC interaction which (1) has been composed by its author in its entirety as part of a private activity and, subsequently, (2) has been sent to the server *en bloc*. In contrast to <post>, <prod> represents *non-verbal* acts within a CMC environment (for details cf. Sect. 4.2.2). As another new element, we introduced <signatureContent> to allow for the unified representation of (most often automatically created) user signatures. This element may occur in meta-data descriptions of the discourse participants.

(2) **New attributes:** A new binary attribute @auto ('automatically generated') was introduced to better reflect the influence of the communication system on the discourse. In many CMC systems, non-verbal actions of participants may result in automatically generated verbal messages, e.g. the insertion of quoted material when hitting a "reply" button, the insertion of signatures into posts, or the generation of status messages. In combination with TEI's @who attribute, fine grained modeling of a message's creation context becomes possible. Because computer-assisted writing or collaborative writing in CMC may lead to parts of messages being produced by different participants, the exploitation of @who was allowed within a wider range of elements than is accepted in TEI proper (see figure 2).

[17] http://www.empirikom.net

```
<post xml:id="m645" who="#A02" synch="#t058" type="standard" auto="false">
   <note auto="true" who="#A02">for all</note>
   <anchor type="sentence_start"/>
   <ref type="addressingTerm" corresp="#A27">
      <w xml:id="m645.t1" type="ADV" lemma="nun">nun</w>
      <w xml:id="m645.t2" type="VVFIN" lemma="bitten">bitte</w>
      <w xml:id="m645.t3" type="NE" lemma="[_FEMALE-STUDENT-A27_]">[_FEMALE-STUDENT-A27_]</w>
      <w xml:id="m645.t4" type="$." lemma="!">!</w>
   </ref>
   <time> 16:48 </time>
</post>
```

Figure 2: CLARIN-D TEI snippet encoding a chat message, demonstrating the use of <post> and custom attributes. The attributes @who, @corresp, and @synch point to the list of participants and the timeline, respectively, in the TEI header.

(3) **Adaptation and extension of content models:** The content model of the generic <s> (sentence), <p> (paragraph), and <quote> elements was extended to allow for sub-elements such as <closer>, <signed>, or <postscript> to occur in a wider range of contexts than envisioned by the TEI. In CMC discourse, these types of text structure tend to be used without the rigid positional constraints found e.g. in traditional books and letters. The content model of some of the elements containing e.g. TEI's <p> and <s> elements was adapted to allow for combining these elements with the newly introduced elements as well as less restricted use of these elements in their typical contexts.

In addition to these customizations, we have defined best practices for using the TEI-P5 models <w>, <phr>, <signed>, <time>, <div>, <name> and other elements for annotating CMC phenomena and for adding part-of-speech information for every word token. Best practices have also been proposed for metadata modeling the discourse level as well as on the level of individual posts.

4.2 The *CoMeRe* project (France)

4.2.1 Project description

The *CoMeRe* project ('Communication Médiée par les Réseaux', supported in 2013-2015 by the National Written Corpora Consortium *IRCE*[18]) brought together researchers who had previously collected different types of CMC corpora in their local research teams or in previous research projects, and had structured these in a variety of formats (different XML schemas for text chat corpora, SMS corpora, and for LEarning and TEaching Corpora (LETEC)).

The primary aim of CoMeRe was to design a common model for CMC discourse that would fit the pre-existing CMC corpora, as well as new corpora collected both during the project or post-project. The secondary aim was to release these corpora in a common repository as open data, in order to provide access to a dataset with significant coverage to researchers interested in the linguistic study of CMC genres.

To address the project's primary goal, it was first necessary to develop a common document model that would fit different types of multimodal CMC data, the TEI CoMeRe schema (2014). All the 14 corpora stored in the CoMeRe repository (2016) have been structured according to this schema. To ensure open access to everyone, the data were collected from sources with appropriate licenses, anonymized, and the corpora were released under Creative Common licenses with the least possible constraints for reuse.

4.2.2 Results beyond the resource: models for representing multimodal CMC in TEI

The opportunity to collect various types of CMC corpora in different formats led us to develop a uniform format complying with the TEI-CMC SIG (Sect. 3). The CoMeRe schema had to be compatible with various genres, including sms, wiki discussions, tweets, weblogs, emails, discussion forums, text chats, oral and multimodal interactions, and multimodal interactions in 3D environments.

For a part of these genres (such as text chat or sms interactions) the users' interactions may be encoded in a way similar to the encoding suggested by Beißwenger et al. (2012), directly relying on the new <post> element (Sect. 4.1.2). For other corpora based on textual interactions, it has been

[18] http://corpusecrits.huma-num.fr/

necessary to enrich the <post> element with extra attributes such as explicit references to previous posts (email, discussion forum, weblog, wiki discussions), or to add sub-elements which describe specific structures encountered within the message contents (tweets). Since the LETEC (LEarning & TEaching) corpora had the most complex structure (Chanier and Wigham, 2016), they served as a basis to develop the CoMeRe schema. A LETEC corpus is a structured entity containing all the elements resulting from an online learning situation whose context is described by an educational scenario and a research protocol. The core data collection includes all the CMC interaction data, the course participants' productions, and the tracks, resulting from the participants' actions in the learning environment.

Indeed, LETEC participants in a course generally used several CMC tools to communicate over a period of 8 to 10 weeks. Participants resorted to various written synchronous and asynchronous communication tools, including emails, text chats, and discussion forums. The challenge was to organize the various interactions in a coherent way within the corpus structure which was the reason to develop the notion of *Interaction Space* (*IS*), with participants interacting on similar subjects using different tools within a time frame. The CoMeRe project members all agreed to adopt this concept, detailed in Chanier et al. (2014), which was generic enough to encompass the CMC genres they were dealing with.

Briefly, an Interaction Space is located within a timeframe, during which interactions occur between a set of participants within an online location. This location is defined by the properties of the set of environments used by the participants who may be either individual members or groups. The environments may be synchronous or asynchronous, mono- or multimodal, simple or complex. The traces of actions within an environment and one particular modality of a CMC tool are termed 'acts'. Working with this concept, the TEI *CoMeRe* schema was proposed. The various components of the Interaction Space are defined in the <teiHeader> of the TEI file, while the actual use of the environments by the participants interacting is described in the <body> part of the TEI file.

IS relates to the intrinsic dialogic nature of such corpora and interactions and all CoMeRe corpora were structured this way. In most cases, the structure of the dialogues could be automatically detected. Only Wikipedia discussions (and particularly the *Wikiconflits* corpus described in Poudat et al., 2017) needed further checking – because of the particularities of Wiki editing and of the fact that wikipedians do not necessarily follow Wikipedia editing recommendations.

Another best practice we worked on concerns the encoding of information on participants: this information is of course crucial for the researcher. Here again, LETEC was the type of corpus in which we had the more detailed information about participants, including information on their role (teacher, learner, domain experts), their sex, age, linguistic competence (languages studied, mastered at different levels), the institutions they belong to, etc. Another part of the information relates to the characteristics and the composition of the groups which circumscribe the space of participants' interactions: the classroom, the subgroups belonging to one or different institutions, the roles played by the participants within each group (tutor, facilitator, learner, etc.). This detailed encoding about participants was also applied to the other CoMeRe genres which did not concern learning situations. For instance, in SMS corpora, questionnaires helped researchers to collect information about participants' habits and usage of SMS, the types of phones and the writing tools they use (Panckhurst et al., 2016). All information on participants have to be encoded in a standard way, and the schema we developed will also be used in the working groups of the new national consortium *CORLI* (*Corpus, Langues and Interactions*).

Lastly, and this will be further developed within CORLI, LETEC situations not only concern environments where participants interact simultaneously within different CMC textual tools, but also CMC oral tools, and tools based on non-verbal interactions (such as collaborative word processors, concept maps, whiteboards, even interactions generated through avatars which move in 3D worlds (Wigham and Chanier, 2013)). Thanks to the speech component of the TEI, data from CMC oral tools could be encoded with the <u> element. However, a new element, currently entitled <prod>, had to be created in order to encapsulate the transcription of non-verbal acts. In the IS model, all the three elements <post>, <u>, and <prod> appear at the same level in the hierarchy. This equality reflects the fact that participants can interact at the same time through textual, oral, or nonverbal acts, each of them associated to an author, a specific duration, and a content which may provoke another participant's reaction. Studying multimodal dialogues requires an encoding of the cross references of the different acts through their head characteristic or their contents.

All the *CoMeRe* corpora were encoded according to these principles, and were deposited into ORTOLANG[19]. This infrastructure represents the most important linguistic data service at the national level. It takes care of curation and long-term archiving. It plays a role similar to other CLARIN national structures, and should in the near future become a part of the European network.

Finally, we are currently working on best practices regarding the PoS tagging of CMC corpora. Only one corpus has been processed so far (a text chat corpus, see Riou and Sagot, 2016), thanks to the MElt tagger. The CoMeRe project has a special interest in further advancing that agenda in line with the European partnership.

4.3 The *DiDi* project (Italy)

4.3.1 Project description

The goal of the regionally funded 2-year DiDi project was to build a South Tyrolean CMC corpus and document the current language use. For this purpose, we collected language data from a social networking site (SNS) and combined it with socio-demographic data about the writers, obtained from a questionnaire (Frey et al., 2016). We chose to collect data from Facebook because this SNS is well known in South Tyrol, offers a wide variety of different communication methods, and is used throughout the territory by many social groups and people of different age.

The autonomous Italian province of South Tyrol is characterized by a multilingual environment with three official languages (Italian, German, and Ladin), and an institutional bi- or trilingualism (depending on the percentage of the Ladin population). Although the project focused on the German-speaking language group, all information regarding the project, for example, the invitation to participate, the privacy agreement, the project web site, and the questionnaire for collection socio-demographic data was published in German and Italian. Consequently, speakers of both Italian and German participated in the text collection campaign.

The multilingual CMC corpus combines Facebook status updates, comments, and private messages with socio-demographic data of the writers. The corpus was enriched with linguistic annotations on thread, text and token level, and provides the following socio-demographic information about the participants: gender, education, employment, internet communication habits, communication devices in use, internet experience, first language(s) (L1), and usage of a South Tyrolean German or Italian dialect and its particular origin. On text level, the corpus was semi-automatically annotated with language code(s) and a political vs. non-political topic label. On token level, the corpus was automatically annotated with part-of-speech, lemma, and CMC phenomenon (e.g. emoticons, emojis, and iteration of graphemes and punctuation) information, and manually normalised, anonymised and annotated with information about the use of German variety.

Another focus of the project was on the users' age and on the question whether a person's age influences language use on SNS; where age is understood in two ways: as a numerical value that reflects the life span of an individual and as digital age that reflects a person's experience with the new media.

Overall, the DiDi corpus comprises public and non-public language data of 136 South Tyrolean Facebook users. The users could choose to provide either their Facebook wall communication (status updates and comments), their chat (i.e. private messages) communication or both. In the end, 50 people provided access to both types of data. 80 users only provided access to their Facebook wall and 6 users gave their chat communication. In total, the corpus consists of around 600,000 tokens that are distributed over the text categories status updates (172,66 tokens), comments (94,512 tokens) and chat messages (328,796 tokens). German language content comprises 58% of the corpus. 13% are written in Italian and 4% in English (the remainder of the messages was either classified as unidentifiable language, non-language or other language). The distribution of the languages is in line with the language backgrounds of the participants and is comparable to the multilingual community of South Tyrol.

[19] https://www.ortolang.fr/

4.3.2 Results beyond the resource: A strategy for collecting private, non-public CMC data

Although the creation and analysis of CMC corpora is currently an active research area, projects exploring private conversations have been rare (but see Dürscheid and Stark (2011) and other sms4science[20] projects, Verheijen and Stoop (2016), and also, for example, the “What's up Switzerland?” project[21]). Instead, projects often explore publicly available data from SNS (like Facebook or Twitter), or data from Wikipedia or discussion boards, where data are relatively easy to obtain. Compared to publicly available data, the acquisition of private data is considerably more difficult in terms of privacy issues, technical implementation and sampled data retrieval. Obtaining private CMC data is time-consuming for both the researchers and the participants because direct interaction between the two is needed. Additionally, the data acquisition process might involve various media discontinuities; this, in turn, causes problems in terms of consistency during data transfer and increases the risk of possible data loss.

Bolander and Locher (2014) and Beißwenger and Storrer (2008) discuss general issues and challenges for corpora of publicly available CMC data. When dealing with non-public data, the issues of data acquisition for CMC corpora become even more demanding: *legal concerns* add to *ethical issues*, and *technical demands* related to *authentic* data retrieval and the linking of *mixed resources* (for example, linking language data and socio-linguistic meta information) get more challenging. Also, for technical and legal reasons of data acquisition an interaction between the user and the researcher becomes an inevitable necessity.

The *legal* situation of using publicly available user-generated language data for research is still under debate, but the trend leans towards seeking explicit user consent. Also, the data will be bound to copyright restrictions, making every modification, (re)publication or citation, potentially problematic (Baron et al., 2012). Furthermore, ethical considerations demand that researchers acquiring private personal data should seek the user’s consent in advance and that the data is anonymised (Beißwenger and Storrer, 2008). For non-public data, this legal and ethical issues are even more critical. But also *technical constraints* make it necessary to interact with the users: most media platforms offer interfaces for third parties to explicitly request permission from the users to use their data. Finding a *representative sample of participants* for the corpus is another problem that, in fact, many corpus creation projects face. Often expensive public relation campaigns and incentives are necessary to get users to participate in projects where the requested data is personal, private and potentially intimate. Different approaches exist to gather the otherwise non-accessible private data, most of them asking for individual submissions of language data by the users.

Frey et al. (2014) considers ‘submission by the user’ to be too tedious for users and researchers, and also troublesome because of privacy concerns on the user side and authenticity doubts on the research side (the users might feel that their writing does not reflect “proper” language use, and brush it up before donating it). Instead, they suggest automatic data collection via a web appilcation: In this way, it is possible to gain user consent and socio-linguistic metadata with the highest privacy for participants (without personal interaction, no backtracking via mail addresses, etc.) and also to collect authentic language data. Additionally, it makes participating more attractive by simplifying the procedure of sharing language and metadata in an integrated, easy and time-saving way, that is also genuine in that media setting (i.e. the participation stays within the same platform, using the platform's interface and methods that are already familiar to the user). The data collection process consists of the following steps: (1) inform potential participants about the research project, the privacy policy and the data usage declaration; (2) provide options for the user to choose which content to share (private inbox and/or personal wall) and thereby increase the transparency for the user; (3) authenticate the user via the Facebook login dialogue (by using the Facebook API); (4) obtain the consent to use, save and republish the user's data (via the web application as well as via the Facebook infrastructure for privacy policies); (5) manage the registered user and the granted permissions via the Facebook login dialogue and the Facebook API; (6) request an anonymous and individual user identifier for the survey client, save permission flags, and enlist the user into an internal database; (7) redirect the user to the survey

[20] http://www.sms4science.org

[21] http://www.whatsup-switzerland.ch/

for the acquisition of the user's meta information; (8) provide dynamic feedback to the user about the current progress of the project (for example, about the amount of participants); (9) provide the possibility to share the application with Facebook friends to attract more users.

The web application and these steps keep the participation process as slim and simple as possible, and it takes users two clicks to donate their language data. There is no one-to-one interaction between an authenticated person and a researcher. Furthermore, legal and ethical constraints are met without additional effort: meta information of the questionnaire and actual language data are automatically linked with an individualised anonymous user identifier, provided by Facebook for every registered user of the web application; so, these identifiers can be used with third-party survey services without privacy problems. Moreover, the procedure facilitates the isolation of user acquisition and interaction with the actual crawling of language data. After logging in, the application grants access to the user's account for a period of 60 days, and the web application only manages registered users. Thus, using such a web application enables efficient data crawling: users do not have to wait for the language data download to complete, and the risk of data loss and other loading and saving issues decreases, as data can be retrieved in independent processes whenever capacities allow it best. Furthermore, server or system failures do not result in data loss since the data can be requested repeatedly. And finally, there are various possibilities to support the attractiveness of the research project: Dynamic feedback can be given via the application interface allowing participants to be part of a collective community project. The application can be easily shared as Facebook post, blog comment, twitter status, e-mail or any other media content, and after having finished the survey, participants can directly share the application with their friends. This workflow is genuine to social media contexts and addresses interested users wherever they happen to be. In addition, participants can be reached by Facebook via targeted advertising campaigns that address a specific user subset and are usually paid by conversions or actual reach of the advertisement.

For more details about the procedure and a discussion of problems and weaknesses see Frey et al. (2014). The anonymized corpus without the private messages is freely available for researchers, and the complete anonymized corpus is available after signing an agreement.[22]

4.4 The Janes project (Slovenia)

4.4.1 Project description

The *Janes* project[23] is compiling a corpus of Slovene user-generated content (Fišer et al., 2016) that contains five different text types of public user-generated content of varying lengths and communicative purposes: tweets, forum posts, user comments on on-line news portals (and, for completeness as well as for enabling comparative analyses, also the news articles themselves, even though they are not user-generated and will therefore not be further discussed in this paper), talk and user pages from Wikipedia, and blog posts along with user comments on these blogs. The collection of tweets and Wikipedia talk pages is comprehensive in the sense that the corpus includes all the Slovene users and their posts that we could identify at the time of harvesting. For the other text types, due to time and financial constraints, we selected only a small set of the most popular sources that at the same time offer the most textual content.

The most recent version of the corpus is v0.4 and it contains around 9 million texts comprising roughly 200 million tokens, 107 of which come from tweets, 47 from forum posts, 34 from blogs and their comments, 15 from news comments and 5 from Wikipedia. The texts in the corpus are structured according to the text types they belong to (e.g. conversation threads in forums) and contain rich metadata, which have been harvested directly during crawling and further enriched within the Janes project. The directly harvested metadata include date and time of posting, username, URL of the text, the discussion thread a text belongs to, the number of likes and retweets, etc. The enriched metadata, which have been added at either user- or text-level, are of two types: those that were added manually (account type, author's gender) and those that were added automatically (user's region, text sentiment, text standardness) (Čibej and Ljubešić, 2015, Fišer et al., 2016, Ljubešić et al., 2015). Figures 3 and 4

[22] http://www.eurac.edu/didi
[23] http://nl.ijs.si/janes/english/

show the distribution of sentiments and of levels of standardness, respectively, by account type and gender.

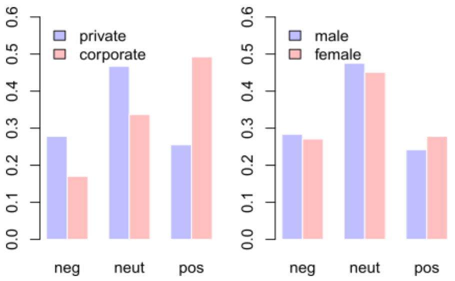


Figure 3: Sentiment of tweets by account type (left) and gender (right).

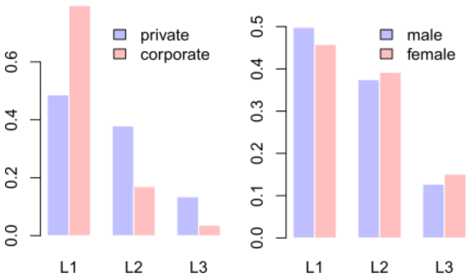
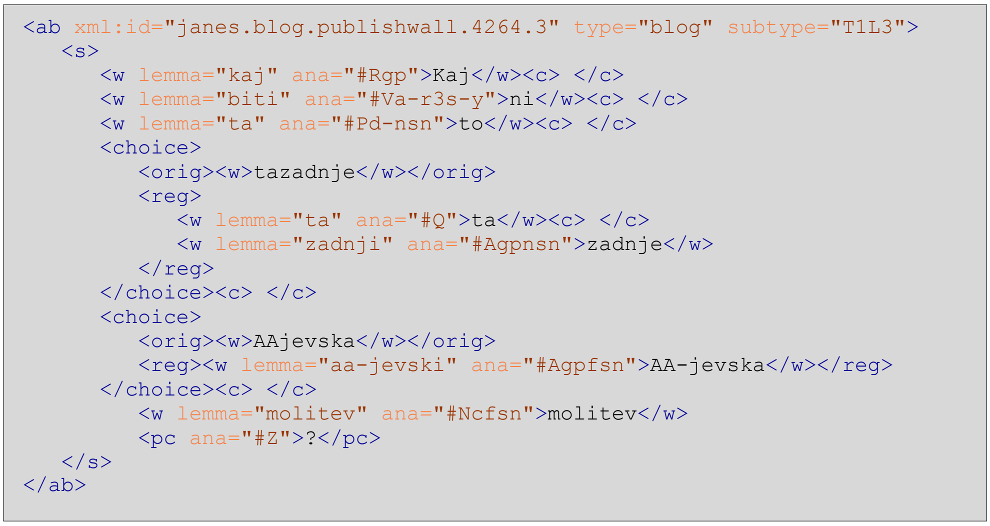


Figure 4: Standardness of tweets by account type (left) and gender (right) (L1 - completely standard, L2 - slightly non-standard, L3 - very non-standard).

For linguistic research on, as well as processing of non-standard language, the most relevant part of the metadata is the assignment of standardness scores to each text. We developed a method (Ljubešić et al., 2015) to automatically classify each text into three levels of technical and linguistic standardness. Technical standardness (T1, quite standard – T3, very non-standard) takes into account the use of spaces, punctuation, capitalisation and similar features, while linguistic standardness (L1, quite standard – L3, very non-standard) takes into account the level of adherence to the written norm and more or less conscious decisions to use non-standard language, involving spelling, lexis, morphology, and word order. On the basis of a manually labelled test set, the method has a mean error rate of 0.45 for technical and 0.54 for linguistic standardness prediction.

As further described in Section 4.5.1, the standard linguistic annotation workflow has been adapted to better tackle CMC-specific features and comprises five steps: tokenization, sentence segmentation, rediacritisation, normalization, morphosyntactic tagging, and lemmatization (Ljubešić and Erjavec, 2016, Ljubešić et al., 2016a, Ljubešić et al., 2016b) and the Janes corpus v0.4 is annotated with these levels of linguistic description.



```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
         <w lemma="molitev" ana="#Ncfsn">molitev</w>
         <pc ana="#Z">?</pc>
   </s>
</ab>
```

Figure 5: TEI encoding of a text in the JANES corpus

The corpus is encoded according to a bespoke XML schema that compactly reflects the structure of the corpus and its metadata. Version 1 will be encoded in a CMC-aware TEI (Beißwenger et al., 2012), cf. Figure 5. Apart from the XML source files, the corpus is also made available to linguists on the local installation of the noSketchEngine and SketchEngine concordancers (Kilgarriff et al., 2014), both as the entire Janes v0.4 corpus with the metadata that all the subcorpora have in common and as separate subcorpora with all the metadata available for the given subcorpus. Access to the corpus is currently restricted to project members, but steps are being taken to comply with the copyright, terms of use and privacy issues in order to make an anonymised, sampled and shuffled corpus available to other researchers as well by the end of the project (Erjavec et al., 2016a).

4.4.2 Results beyond the resource: Adaptation of NLP tools for processing Slovenian CMC language

In this section, we present the toolchain for automatic linguistic annotation of CMC we have mostly developed within the Janes project, as well as the datasets to enable its further improvements. Since most of the developed tools rely on supervised machine learning, we briefly report on the training data used and, where available, the estimated accuracy of each tool.

Tokenisation and sentence segmentation. For tokenisation and sentence segmentation, we used a new Python tool that covers Slovene, Croatian and Serbian (Ljubešić and Erjavec, 2016). Like most tokenisers, it is based on manually defined rules in the form of regular expressions and uses language-specific lexicons with, e.g. lists of abbreviations. In addition to standard rules, the tokeniser has an additional non-standard mode in which it uses less strict rules. For example, a full stop can here end a sentence even though the following word does not begin with a capital letter or is even not separated from the full stop by a space. Nevertheless, tokens that end with a full stop and are on the list of abbreviations (e.g. *prof.*) will not end a sentence. The non-standard tokeniser mode also has several additional rules, such as additional regular expressions devoted to recognising emoticons, e.g. *:-], :-PPPP,* ^_^ etc. A preliminary evaluation of the tool on tweets showed that sentence segmentation could still be significantly improved (86.3% accuracy), while tokenisation is relatively good (99.2%), taking into account that both tasks are very difficult for non-standard language.

Normalisation. Normalising non-standard word tokens to their standard form has two advantages. First, it becomes possible to search for a word without having to consider or be aware of all its spelling variants and second, normalisation makes it possible to use downstream tools for standard language processing, such as part-of-speech taggers. In the Janes corpus, the word tokens have been normalised when necessary by using a sequence of two steps. First, we use a dedicated tool (Ljubešić et al., 2016a) to restore diacritics (e.g. *krizisce → križišče*). The tool learns the rediacritisation model on a large collection of texts with diacritics paired with the same texts with the diacritics removed. The evaluation showed that the tool achieves a token accuracy of 99.62% on standard texts (Wikipedia) and 99.12% on partially non-standard texts (tweets). Second, the rediacriticised word tokens are normalised with a method that is based on character-level statistical machine translation (Ljubešić et al., 2016b). The goal of the normalisation is to translate words written in a non-standard form (e.g. *jest, jst, jas, js*) to their standard equivalent (*jaz*). The current translation model for Slovene was trained on a preliminary version of the manually normalised dataset Janes-Norm (cf. below), while the target (i.e. standard) language model was trained on the Kres balanced corpus of Slovene (Logar Berginc et al., 2012) and the tweets from the Janes corpus that were labelled as linguistically standard. It should be noted that normalisation will sometimes also span word-boundaries, i.e. there are cases where one non-standard word corresponds to two or more standard words or vice versa (e.g. *ne malo → nemalo; tamau → ta mali*).

Tagging and lemmatization. As the final step in the text annotation pipeline, the normalised tokens were annotated with their morphosyntactic description (MSD) and lemma. For this, we used a newly developed CRF-based tagger-lemmatiser that was trained for Slovene, Croatian and Serbian (Ljubešić and Erjavec, 2016). The main innovation of the tool is that it does not use its lexicon directly, as a constraint on possible MSDs of a word, but rather indirectly, as a source of features; it thus makes no distinction between known and unknown words. For Slovene, the tool was trained on the ssj500k 1.3 corpus (Krek et al., 2013) and the Sloleks 1.2 lexicon (Dobrovoljc et al., 2015). Compared to the previous best result for Slovene with the Obeliks tagger (Grčar et al., 2012), the CRF tagger reduces the relative error by almost 25% achieving a 94.3% accuracy on the test set comprising the last tenth of the ssj500k corpus. The MSD tagset used within the Janes project follows the MULTEXT Version 4 specifications (Erjavec, 2012), except that we, following Bartz et al. (2014), introduce new MSDs for the annotation of CMC-specific content, in particular Xw (e-mails, URLs), Xe (emoticons and emoji), Xh (hashtags, e.g. *#kvadogaja*) and Xa (mentions, e.g. *@dfiser3*). The lemmatisation, which is also part of the tool, takes into account the posited MSD. For pairs word-form:MSD which are already in the training lexicon, it simply retrieves the lemma, while for the rest it uses its lemmatisation model to guess the lemma.

Manually annotated datasets. To further improve our annotation tool chain, we have manually annotated two gold-standard datasets (Erjavec et al., 2016b): Janes-Norm (Erjavec et al., 2016c), which contains 7,816 texts or 184,755 tokens, is a gold-standard dataset for tokenisation, sentence segmentation and word normalisation, while Janes-Tag, (Erjavec et al., 2016d), a subset of Janes-Norm, comprises 2,958 texts or 75,276 tokens, and is a gold-standard dataset for training and evaluating morphosyntactic tagging and lemmatisation.

The annotation guidelines which were produced to guide the annotation of these two corpora to a large extent follow the guidelines for annotating standard (Holozan et al., 2008) and historical (Erjavec, 2015) Slovene, with some medium-specific modifications (e.g. the annotation of emoticons, URLs, hashtags, and mentions). At the normalisation level, special attention was paid to non-standard words with multiple spelling variants and those without a standard form (e.g. *orng, ornk, oreng, orenk* for *'very'*), foreign language elements (e.g. *updateati, updajtati, updejtati, apdejtati* for *'to update'*) and linguistic features that are not normalised (e.g. hashtags, non-standard syntax and stylistic issues). At the morphosyntactic description (MSD) and lemmatisation levels, the guidelines were designed to deal with foreign language elements, proper names and abbreviations as well as non-standard use of case and particles. All the texts were first automatically annotated, then checked and corrected manually by a team of students, with two students annotating each text and the divergent annotation checked by an experience curator. The platform used for manual annotation was WebAnno (Yimam et al., 2013).

Janes-Norm and Janes-Tag are deposited on the CLARIN.SI repository and freely available for research under the CC BY licence.

5 Outlook

In this paper, we gave an overview of results and best practices from projects in four countries dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC). The joint goal of the projects is to establish standards for the collection and representation of CMC corpora and for their integration into common resources infrastructures.

Up to now, the network has brought forward two main initiatives: a conference series dedicated to all issues related to building, annotating and analyzing CMC corpora, and a TEI-SIG focused on the integration of standards for CMC resources into the TEI framework. Both initiatives are "bottom up" with the goal to connect researchers all over Europe and to work on solutions driven by practices that have proven useful in ongoing projects. The latest edition of the conference included 22 contributions by 40 authors from 24 research institutions in 11 countries (Fišer and Beißwenger, 2016).

Nevertheless, there's still a lot of open, non-trivial issues in the field. One example is the lack of legal standards for collecting and republishing CMC data as part of language resources. Corpus builders are typically laymen when it comes to legal issues. A general legal opinion on these issues commissioned and disseminated by and via an acknowledged language resources initiative (e.g., CLARIN or its national consortia) would therefore be an important prerequisite for the further development of the CMC corpora landscape and community.

In view of the importance of CMC in everyday communication, in business, public administration, science and education, efforts in the field of establishing state-of-the-art research and resource infrastructures for the analysis of CMC phenomena are an investment in our future knowledge about how the adoption of CMC technologies affects society and how communicative practices reflect the presence of CMC as an innovative means for the organization of social interaction.

References

[Baron et al.2012] Alistair Baron, Paul Rayson, Phil Greenwood, James Walkerdine, and Awais Rashid. 2012. Children Online: A Survey of Child Language and CMC Corpora. *International Journal of Corpus Linguistics,* 17(4):443–81.

[Bartz et al.2014] Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics,* 28(1):157–198.

[Beißwenger and Storrer2008] Michael Beißwenger and Angelika Storrer. 2008. Corpora of computer-mediated communication. In: Lüdeling, Anke; Kytö, Merja (eds.). *Corpus Linguistics HSK*, vol. 29.1. Walter de Gruyter, Berlin, Germany, pp. 292–309.

[Beißwenger et al.2012] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012. A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative (Online),* (3) (doi: 10.4000/jtei.476). http://jtei.revues.org/476.

[Beißwenger2013] Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik,* 41(1):161–164.

[Beißwenger et al.2015] Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl. 2015. *Tagset and Guidelines for the PoS Tagging of Language Data from Genres of Computer-mediated Communication / Social Media*. http://sites.google.com/site/empirist2015/home/annotation-guidelines.

[Beißwenger et al.2016] Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In*: Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task.* Berlin, Germany, pp. 44–56. http://aclweb.org/anthology/W/W16/W16-2606.pdf

[Bolander and Locher2014] Brook Bolander and Miriam A. Locher. 2014. Doing Sociolinguistic Research on Computer-Mediated Data: A Review of Four Methodological Issues. *Discourse, Context & Media,* (3):14–26.

[Chanier et al.2014] Thierry Chanier, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal of language Technology and Computational Linguistics,* 29(2):1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf.

[Chanier and Wigham2016] Thierry Chanier and Ciara Wigham. 2016. Standardizing Multimodal Teaching and Learning Corpora. In: Marie-Jo, Hamel; Caws, Catherine (eds.). *Language-Learner Computer Interactions: Theory, Methodology and CALL Applications.* John Benjamins, Amsterdam, Netherlands, pp. 215-240. DOI:10.1075/lsse.2.10cha.

[Chiari and Canzonetti2014] Isabella Chiari and Alessio Canzonetti. 2014. Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In: Garavelli, Enrico; Suomela-Härmä, Elina (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, Italy, pp. 595-606.

[Čibej and Ljubešić2015] Jaka Čibej and Nikola Ljubešić. 2015. *"S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter.* Zbornik konference Slovenščina na spletu in v novih medijih, Ljubljana, Slovenia, pp. 10-14.

[CLARIN-D schema2015] CLARIN-D TEI schema for CMC corpora. 2015. http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema.

[CoMeRe repository2016] CoMeRe repository. 2016. *Corpora of Computer-Mediated Communication in French.* Ortolang.fr, Nancy, France. http://hdl.handle.net/11403/comere.

[CoMeRe schema2014] CoMeRe TEI schema for CMC corpora, version 2. 2014. https://repository.ortolang.fr/api/content/comere/v2/tei_cmr.rng and http://wiki.tei-c.org/index.php/SIG:CMC/CoMeRe_schema_draft_for_representing_CMC_in_TEI_(2014).

[Dobrovoljc et al.2015] Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. *Morphological Lexicon Sloleks 1.2.*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1039.

[Dürscheid and Stark2011] Christa Dürscheid and Elisabeth Stark. 2011. sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow, Crispin; Mroczek, Kristine (eds.): *Digital Discourse. Language in the New Media*. Oxford University Press, Oxford, UK, pp. 299-320.

[Erjavec2012] Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

[Erjavec2015] Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775.

[Erjavec et al.2016a] Tomaž Erjavec, Jaka Čibej, and Darja Fišer. 2016. Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. *Slovenščina 2.0*, 4(2):189–219.

[Erjavec et al.2016b] Tomaž Erjavec, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Darja Fišer. 2016. Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication. In*: Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processings*, Brno, the Czech Republic, pp. 29–40.

[Erjavec et al.2016c] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, and Nikola Ljubešić. 2016. *CMC Training Corpus Janes-Norm 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1084.

[Erjavec et al.2016d] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, and Nikola Ljubešić. 2016. *CMC Training Corpus Janes-Tag 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1085.

[Fišer and Beißwenger2016] Darja Fišer and Michael Beißwenger (eds.). 2016. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora2016)*. University of Ljubljana, Slovenia. http://nl.ijs.si/janes/cmc-corpora2016/proceedings/

[Fišer et al.2016] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2):67–99.

[Forsyth an Martell2007] Eric N. Forsyth and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In: *Proceedings of the First IEEE International Conference on Semantic Computing* (ICSC 2007), Irvine, USA, pp. 19-26.

[Frey et al.2014] Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2014. Collecting Language Data of Non-Public Social Media Profiles. In: *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, edited by Gertrud Faaß and Josef Ruppenhofer. Universitätsverlag Hildesheim, Hildesheim, Germany, pp. 11-15.

[Frey et. al.2016] Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2016. The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. Accepted at CLIC-it 2016.

[Grčar et al.2012] Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. *Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: a statistical morphosyntactic tagger and lemmatiser for Slovene)*. Zbornik Osme konference Jezikovne tehnologije, Ljubljana, Slovenia.

[Holozan et al.2008] Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman, and Aleš Velušćek. 2008. *Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku" (Specifications for the Training Corpus. The "Communication in Slovene" project).* http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

[Horbach et al.2014] Andrea Horbach, Diana Steffen, Steffen Thater, and Manfred Pinkal. 2014. Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In: *Proceedings of KONVENS 2014*, pp. 171–177. https://hildok.bsz-bw.de/frontdoor/index/index/docId/241.

[iRights.Law2016] iRights.Law Rechtsanwälte. 2016. *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen.* (Legal opinion for the ChatCorpus2CLARIN project, 46 pages).

[Kilgarriff et al.2014] Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

[Krek et al.2013] Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek, and Nanika Holz. 2013. *Training Corpus ssj500k 1.3*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1029.

[Ljubešić and Erjavec2016] Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In: *Proceedings of the 10th Language Resources and Evaluation Conference*, Portorož, Slovenia, pp. 1527–1531.

[Ljubešić et al.2016a] Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2016. Corpus-based diacritic restoration for South Slavic languages. In: *Proceedings of the 10th Language Resources and Evaluation Conference.* Portorož, Slovenia, pp. 3612–3616.

[Ljubešić et al.2016b] Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, pp. 146–155.

[Ljubešić et al.2015] Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, and Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. In*: Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 371–378.

[Logar Berginc et al.2012] Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba* (The Gigafida, KRES, ccGigafida and ccKRES corpora of Slovene language: compilation, content, use.) Ljubljana, Slovenia: Trojina, zavod za uporabno slovenistiko, Faculty of Social Sciences.

[Lüngen et al.2016] Harald Lüngen, Michael Beißwenger, Eric Ehrhardt, Axel Herold, and Angelika Storrer. 2016. Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016),* Bochum, Germany, pp. 156–164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf.

[Margaretha and Lüngen2014] Eliza Margaretha and Harald Lüngen. 2014. Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics,* 29(2):59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

[Oostdijk et al.2013] Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In: Spyns, Peter; Odijk, Jan (eds*). Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer Verlag, Berlin, Germany, pp. 219-247.

[Panckhurst et al.2016] Rachel Panckhurst, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, and Bertrand Verin. 2016. *88milSMS*: A corpus of authentic text messages in French. [corpus] In: Chanier, Thierry (ed*). Banque de corpus CoMeRe*. Ortolang, Nancy, France. https://hdl.handle.net/11403/comere/cmr-88milsms.

[Poudat et al.2017] Céline Poudat, Natalia Grabar, Camille Paloque-Berges, Thierry Chanier, and Kun Jin. 2017. Wikiconflits: un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In: Wigham, C.R.; Ledegen, G. (eds.). 2017. *Corpus de communication médiée par les réseaux: Construction, structuration, analyse.* Collection Humanités Numériques. L'Harmattan, Paris, France, pp. 211-222.

[Riou and Sagot2016] Stéphane Riou and Benoit Sagot. 2016. *Etiquetage morpho-syntaxique du corpus FAVI* [corpus]. D'après Yun, H. & Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. Banque de corpus CoMeRe. Ortolang, Nancy, France. http://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2

[Schiller et al.1999] Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Germany. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

[Schröck and Lüngen2015] Jasmin Schröck and Harald Lüngen. 2015. Building and Annotating a Corpus of German-Language Newsgroups. In*: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015).* Essen, Germany, pp. 17-22. https://sites.google.com/site/nlp4cmc2015/program

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

[Verheijen and Stoop2016] Lieke Verheijen and Wessel Stoop. 2016. Collecting Facebook Posts and WhatsApp Chats. In: *Proceedings. Text, Speech, and Dialogue: 19th International Conference*, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Springer International Publishing, Cham, Germany, pp. 249–58.

[Westpfahl and Schmidt2016] Swantje Westpfahl and Thomas Schmidt. 2016. *FOLK-Gold – A GOLD standard for Part-of-Speech- Tagging of Spoken German*. In: *Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC16),* Paris, France, pp. 1493-1499.

[Wigham and Chanier2013] Ciara Wigham and Thierry Chanier. 2013. Interactions Between Text Chat and Audio Modalities for L2 Communication and Feedback in the Synthetic World Second Life. *Computer Assisted Language Learning*, 28(3):260-283. DOI:10.1080/09588221.2013.851702.

[Yimam et al.2013] Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations),* Association for Computational Linguistics, Stroudsburg, USA, pp. 1–6.