Title	Overview of Resources and Tools for Computer-Mediated Communication
Version	2.5
Author(s)	DF, JL
Date	20-11-2017
Status	For distribution
Distribution	NCF, UI
ID	CE-2017-1064

Contents

1 Purpose and methods

In the following survey, our aim is to provide an overview of social media corpora, datasets and tools of the languages spoken in countries that are members and observers of CLARIN ERIC. Our motivation was to ascertain in how far they are accessible through the CLARIN infrastructure, thereby emphasising the aspects in which the presentation of the relevant information and accessibility of these corpora can be optimised from a User Involvement perspective.

In Section 2, we give an overview of the identified corpora, metadata (size, time span, annotation) and key publications of corpora that are part of the CLARIN infrastructure. In Section 3, we compile a list of smaller, more focused datasets and highlighted the NLP tasks for which the datasets are used. In Section 4, we provide information on the tools that are tailored to processing computer-mediated communication.

2 Corpora within the CLARIN infrastructure

We identified 12 corpora of computer-mediated communication (CMC) that are part of the CLARIN infrastructure. They cover 8 different languages: Slovene (5), Czech (1), Dutch (1),

Estonian (1), Finnish (1), German (1), French (1), Lithuanian (1). In Table 1, we give an overview of the identified corpora, including the information on the source of the texts included in the corpus, the size of the corpus, the time span of the texts in the corpus, the linguistic annotation, accessibility and licencing.

The hyperlinks were last accessed 15 November 2017.

Table 1: Overview of CMC corpora

Corpus name	Corpus description
Corpus of contemporary blogs Czech 1 million tokens Unclear annotation For download	This corpus consists of 1 million tokens and contains blogs posts in Czech from an unknown period. It is unclear how the corpus is annotated. The corpus can be found through the VLO and is available for download under CC-BY.
SoNaR New Media corpus Dutch 35 million tokens Tokenised, PoS-tagged, lemmatised Concordancer	This corpus consists of 35 million tokens and contains tweets, chats and SMS in Dutch from 2005 to 2012. The corpus is tokenised, PoS-tagged and lemmatised. It is available for searching online.[1] It is unclear under which licence the corpus is available. For the relevant publication, see Sanders (2012). The corpus can be found through the VLO.
The Mixed Corpus: New Media Estonian 25 million tokens Tokenised For download and concordancer	This corpus consists of 25 million tokens from chat rooms, forums, comments and newsgroups in Estonian from 2000 to 2008. The corpus is tokenised and available both for searching online and for download. It is unclear under which licence the corpus is available. We were unable to find a relevant publication for this corpus. The corpus can be found on the website of CLARIN Estonia.
Suomi 24 Corpus Finnish 2.6 billion tokens Tokenised, MSD-tagged For download and concordancer	This corpus consists of 2.6 billion tokens from the Suomi 24 discussion forum in Finnish from 2001 to 2016. The corpus is tokenised and MSD-tagged with the Turku Dependency Parser. The corpus is available for searching online and for download under the CLARIN_ACA licence. For the relevant publication, see Lagus et al. (2016), in Finnish. The corpus can be found through the VLO.
CoMeRe repository French 80 million tokens Unclear annotation For download	This French corpus contains 80 million tokens from e-mails, forums, chats, tweets and the French Wikipedia from various periods. It is unclear if and how the corpus is annotated. The corpus (or rather the corpora contained in this repository) is available for download under the CC-BY licence. For the relevant publication, see Chanier et al. (2014).

[1] However, the link to the OpenSONAR environment where this corpus is available is not trivially accessible via the VLO and was obtained elsewhere.

	We found the corpus in the ORTOLANG repository
Dortmund Chat Corpus German 1 million tokens Tokenised, PoS-tagged, lemmatised For download	This corpus consists of 1 million tokens taken from German chats from 2000 to 2006. The corpus is tokenised, PoS-tagged and lemmatised. The corpus is available under the licence CC-BY. The corpus is available for download. For the relevant publication, see Beißwenger (2013). We found the information about the corpus through the VLO, which contains several versions of this corpus.
LITIS v.1 Lithuanian 190,000 comments Unclear annotation For download	This corpus consists of roughly 190,000 comments taken from the Lithuanian portals *delfi.lt* and *lrytas.lt* from 2010 to 2014. It is unclear if and how the corpus is annotated. The corpus is available for download. The corpus is available under the ACA_CLARIN-LT_End-User-Licence-Agreement_EN-LT. We were unable to find a relevant publication for this corpus. The corpus can be found through the VLO.
Twitter corpus Janes-Tweet 1.0 Slovene 139 million tokens Tokenised, sentence segmented, MSD-tagged, lemmatised For download	This corpus contains 139 million tokens from Tweets in Slovene posted from 2013 to 2017. The corpus is tokenised, sentence segmented, MSD-tagged, lemmatised and annotated with named entities. The corpus is available for download under the CC-BY licence. The corpus can be found through the VLO.
Wikipedia talk corpus Janes-Wiki 1.0 Slovene 5 million tokens Tokenised, sentence segmented, MSD-tagged, lemmatised For download	This corpus contains 5 million tokens from Wikipedia in Slovene from an unknown period. The corpus is tokenised, sentence segmented, MSD-tagged, lemmatised and annotated with named entities. The corpus is available for download under the CC-BY licence. The corpus can be found through the VLO.
Forum corpus Janes-Forum 1.0 Slovene 47 million tokens Tokenised, sentence segmented, MSD-tagged, lemmatised For download	This corpus contains 47 million tokens from Slovene forums from an unknown period. The corpus is tokenised, sentence segmented, MSD-tagged, lemmatised and annotated with named entities (as well as partially anonymised). The corpus is available for download under the CC-BY licence. The corpus can be found through the VLO.
Blog post and comment corpus Janes-Blog 1.0 Slovene 34 million tokens	This corpus contains 34 million tokens from Slovene blogs and comments from an unknown period. The corpus is tokenised, sentence segmented, MSD-tagged, lemmatised and annotated with named entities (as well as partially anonymised). The corpus is available for download under the CC-BY licence.

Tokenised, sentence segmented, MSD-tagged, lemmatised For download	The corpus can be found through the VLO.
News comment corpus Janes-News 1.0 Slovene 14 million tokens Tokenised, sentence segmented, MSD-tagged, lemmatised For download	This corpus contains 14 million tokens from Slovene comments on newsposts from an unknown period. The corpus is tokenised, sentence segmented, MSD-tagged, lemmatised and annotated with named entities (as well as partially anonymised). The corpus is available for download under the CC-BY licence. The corpus can be found through the VLO.

2.1. Identification of the corpora

Of the 12 identified corpora, all can be found on the VLO except for the Estonian *Mixed Corpus: New Media* corpus, which can be found on the website of the Estonian consortium, and the *CoMeRe repository*, which can be found in the ORTOLANG repository. In the case of the *The Suomi 24 Corpus*, outdated versions are available through the VLO, but not the most recent one, to which a link is provided in table (1).

2.2. Availability

In terms of availability, the following 2 corpora are available both for download and through a concordancer:

(i) *The Mixed Corpus: New Media*
(ii) *Suomi 24 Corpus*

Corpus (i) is available through a dedicated concordancer; corpus (ii) is available through *Korp*.

The following 9 corpora are available for download:

(i) Corpus of contemporary blogs
(ii) Dortmund Chat Corpus
(iii) LITIS v.1 corpus
(iv) Twitter corpus Janes-Tweet 1.0
(v) Wikipedia talk corpus Janes-Wiki 1.0
(vi) Forum corpus Janes-Forum 1.0
(vii) Blog post and comment corpus Janes-Blog 1.0
(viii) News comment corpus Janes-News 1.0
(ix) CoMeRe repository

Corpus (i) is available through *LINDAT,* corpus (ii) is available through CLARIN-D, corpus (iii) is available through CLARIN-LT, corpora (iv)-(viii) are available through CLARIN.SI and corpus (ix) is available through ORTOLANG.

The *SoNaR New Media corpus* is available only through an online search environment (*OpenSONAR*).

2.3. Metadata

We have identified 4 issues with the metadata provided:

- The annotation for the Czech *Corpus of contemporary blogs*, *the LITIS v.1* corpus and *CoMeRe repository* is unclear.
- The timespan is unknown for the following 5 corpora:
  (i) *Corpus of contemporary blogs*
  (ii) *Wikipedia talk corpus Janes-Wiki 1.0*
  (iii) *Forum corpus Janes-Forum 1.0*
  (iv) *Blog post and comment corpus Janes-Blog 1.0*
  (v) *News comment corpus Janes-News 1.0*
- The licence is unclear for *The Mixed Corpus: New Media*
- The VLO record for *SoNaR New Media corpus* links to the LRT collection on *LINDAT*, which links to the main TST-centrale page, not directly to the *OpenSONAR concordancer* or download page.

3 Corpora not part of the CLARIN infrastructure

We identified 9 corpora of computer-mediated communication (CMC) that are not part of the CLARIN infrastructure. In Table 2, we give an overview of the identified corpora, including the information on the source of the texts included in the corpus, the size of the corpus, the time span of the texts in the corpus, the linguistic annotation, accessibility and licencing.

Table 2: Overview of CMC corpora not part of the CLARIN infrastructure

Corpus name	Corpus description
Flemish Online Teenage Talk Dutch 2.9 million tokens Tokenised Unavailable	This Flemish Dutch corpus consists of 2.9 million tokens from Facebook and WhatsApp from 2015 and 2016. The corpus is tokenised and is unavailable. For the relevant publication, see Hilte et al. (2016). The information regarding this corpus was provided by a participant at the CLARIN-PLUS workshop on Social Media data.
Dereko – News and Wikipedia subcorpus German 670 million tokens Unclear annotation	This German corpus consists of 670 million tokens taken from newsgroups and the German Wikipedia. We were unable to find information regarding the time span of the data. The corpus is tokenised and is available for searching online. It is unclear under which licence the corpus is available. We were unable to find a relevant publication for this corpus. We were unable to find the corpus through the CLARIN infrastructure but

Concordancer	in Beißwenger et al. (2016).
DWDS – Blogs subcorpus German 102 million tokens Unclear annotation Concordancer	This German corpus consists of 102 million tokens from blog posts. We were unable to find information regarding the time span of the data. It is unclear how the corpus is annotated. The corpus is accessible through for searching online. It is unclear under which licence the corpus is available. We were unable to find a relevant publication for this corpus. We found the corpus is Beißwenger et al. (2016).
Monitor corpus of tweets from Austrian users German and English 40 million tweets Tokenised, lemmatised Unavailable	This corpus, which contains data in German and English, is a compilation of 40 million tweets from 2007 to 2017. It is tokenised and lemmatised. This Austrian corpus is not publicly available and re-licensing of the data is forbidden. For the relevant publication, see Barbaresi et al. (2016). We found the corpus on Google.
FORUMAS_INDV corpus Lithuanian 600,000 tokens Unclear annotation For download	This Lithuanian corpus consists of 600,000 tokens from forum posts on the *lyrtas.lt* portal from 2014. The corpus is available for download. It is unclear under which license the corpus is available. For the relevant publication, see Kapočiūtė-Dzikienė et al. (2015). The information regarding this corpus was provided by a participant at the CLARIN-PLUS workshop on Social Media data.
INT_KOMETARAI_INDV2 corpus Lithuanian 4 million tokens Unclear annotation For download	This Lithuanian corpus consists of 4 million tokens from comments on the *delfi.lt* portal from 2015. The corpus is available for download. It is unclear under which license the corpus is available. For the relevant publication, see Kapočiūtė-Dzikienė et al. (2015). The information regarding this corpus was provided by a participant at the CLARIN-PLUS workshop on Social Media data.
NTAP climate change blog corpus Norwegian, English, French 21 million tokens Unclear annotation Unavailable	The Norwegian subcorpus contains 21 million tokens from blogs focussing on climate change from 2000 to 2014. It is unclear how the corpus can be accessed and under which licence it is distributed. For the relevant publication, see Salway et al. (2016). The corpus was found on Google.
Corpus of Highly Emotive Internet	This Polish corpus contains roughly 160 million tokens from Twitter. We were unable to discern the period that the corpus covers. The corpus is tokenised. It is available for download, though the authors need to be

Discussions Polish 160 million tokens Tokenised For download	contacted beforehand. It is unclear under which licence the corpus is distributed. For the relevant publication, see Sobkowicz (2016). The corpus was found on Google.
The Corpus of Welsh Language Tweets Welsh 7 million tokens Unclear annotation For download	This Welsh corpus consists of roughly 7 million tweets from an unknown period. It is also unclear how the corpus is annotated. The corpus is available for download, with the data restricted in accordance with Twitter Terms of Use. For the relevant publication, see Jones et al. (2015). The information regarding this corpus was provided by a participant at the CLARIN-PLUS workshop on Social Media data.

3.1 Identification of the corpora

Information on the following 4 corpora was provided to us by a participant at the CLARIN-PLUS workshop on Social Media data:

(i) *Flemish Online Teenage Talk*
(ii) *FORUMAS_INDV corpus*
(iii) *INT_KOMETARAI_INDV2 corpus*
(iv) *The Corpus of Welsh Language Tweets*

The following 2 corpora were identified through Beißwenger et al. (2016):

(i) *Dereko – News and Wikipedia subcorpus*
(ii) *DWDS – Blogs subcorpus*

The following 3 corpora were found on Google:

(i) *Monitor corpus of tweets from Austrian users*
(ii) *NTAP climate change blog corpus*
(iii) *Corpus of Highly Emotive Internet Discussions*

3.2 Availability

The following 4 corpora are available for download:

(i) *FORUMAS_INDV corpus*
(ii) *INT_KOMETARAI_INDV2 corpus*
(iii) *Corpus of Highly Emotive Internet Discussions*
(iv) *The Corpus of Welsh Language Tweets*

The following 2 corpora are available through a concordancer

(i) *Dereko – News and Wikipedia subcorpus*
(ii) *DWDS – Blogs subcorpus*

The following 3 corpora are unavailable:

(i) *Flemish Online Teenage Talk*
(ii) *Monitor corpus of tweets from Austrian users*
(iii) *NTAP climate change blog corpus*

3.3 Corpora under development

In addition to the 10 corpora, we have also identified the following corpora still under development:

- The Italian Web2Corpus_it corpus (cf. Chiari and Canzionetti 2014) contains texts from online forums, blogs, newsgroups, social networks and chats.
- The multilingual *What's up, Switzerland* corpus contains German, French, Italian and Romansh chats from *WhatsApp*.

4 Datasets

In addition to CMC corpora, we have identified 14 smaller, more specialised datasets compiled for particular NLP tasks. Among these, 13 datasets are monolingual and compile data from 6 different languages: Slovene (6), English (3), Italian (2), Czech (1), Greek (1), Swedish (1). 1 identified dataset is multilingual and contains German, Italian and Spanish data. In terms of data types, most (i.e. 10 out of 14) are from Twitter. We list them in Table 3, adding the NLP task they are intended for. 8 of the 14 identified datasets are available through the CLARIN infrastructure: all the six Slovene ones and the multilingual one, which is accessible in the repository of *Clarin.si*, and the Greek dataset, which is accessible in the repository of *CLARIN:EL*.

Table 3: Overview of CMC datasets

Language	Dataset description
Czech	The CSFD_CZ, Facebook_CZ, and Mall_CZ contain Facebook posts and comments on movie sites and have been annotated for sentiment analysis. The size and time span of the dataset are unknown. The texts are also PoS-tagged. For the relevant publication, see Habernal et al. (2013).
English	The *Broad Twitter Corpus* consists of 165,000 tokens from Twitter from 2009 to 2014 and has been annotated for Named Entity Recognition. For the relevant publication, see Derczynski et al. (2016).
English	The Twitter Entity Linking database consists of roughly 10,000 tokens from Twitter from 2010. For the relevant publication, see Derczynski et al. (2015). This dataset is used for Entity Linking.
Greek	The *Verbal Aggressiveness Database* contains 54,000 tweets from 2013 to 2016. We were unable to find a relevant publication for this dataset.
Italian	The *sentipolc* contains 10,000 Tweets from 2014 to 2016 and has been annotated for sentiment analysis and irony detection. The dataset is available for download on the webpage. For the relevant publication, see Barbieri et al. (2016).
Italian	The *Damage Assessment of Natural Disasters from Social Media Messages* database consists of 5,500 tweets from 2009 to 2014. For the relevant

	publication, see Cresci et al. (2015).
Multilingual	The xLiMe Twitter Corpus XTC 1.0.1 is a multilingual dataset consisting of German, Italian and Spanish tweets and is used for sentiment analysis and named-entity recognition. It is tokenised and PoS-tagged and consists of 370,000 tokens. For the relevant publication, see Rei et al. (2016).
Slovene	The *CMC training corpus Janes-Tag 1.2* contains texts from various social media sources and has been compiled for training morpho-syntactic (MSD) taggers and lemmatisers of non-standard language. It is manually tokenised, MSD-tagged and lemmatised, and consists of 75,000 tokens. For the relevant publication, see Erjavec et al. (2016).
Slovene	The *CMC training corpus Janes-Norm 1.2.* contains texts from various social media sources and has been manually annotated for training word-level normalisation of non-standard language. It consists of 180,000 tokens. For the relevant publication, see Erjavec et al. (2016).
Slovene	The CMC training corpus Janes – Syn 1.0 contains tweets and has been manually syntactically annotated. It consists of 4,388 tokens. For the relevant publication, see Arhar Holdt et al. (2017).
Slovene	The *Tweet comma corpus Janes-Vejica 1.0* contains tweets and has been manually annotated for (in)correct comma placement. It consists of 14,013 tokens. For the relevant publication, see Popič et al. (2016).
Slovene	The *CMC shortening corpus Janes-Kratko 1.0* contains tweets and has been manually annotated for studying text shortening strategies on Twitter. It consists of 20,000 tokens. For the relevant publication, see Goli et al. (2016).
Slovene	The *Dataset of normalised Slovene text KonvNormSi 1.0* contains manually normalised texts from historical and contemporary Slovene for training word-level normalisation of non-standard language. It consists of 427,000 tokens. For the relevant publication, see Ljubešić et al. (2016).
Swedish	The *Eukalyptus* dataset contains 20,000 tokens of texts from various social media from an unknown period and has been annotated for word-sense disambiguation. It is also tokenised, PoS-tagged and lemmatised. For the relevant publication, see Johansson et al. (2012).

5 The tools

Apart from the resources, we searched for language-processing tools that are tailored to working with corpora and datasets that compile various CMC data. The following tools were found within the CLARIN infrastructure.

Table 4: Overview of CMC tools

Language	Tool	Description
language-independent	1 *csmtiser*	text normalisation via character-level machine translation
language-independent	2 *tweetcat*	building Twitter corpora of smaller languages or specific geographical regions

South Slavic languages	3 *janes-ner*	Named Entity recognition systems for South Slavic languages
Slovene/Croatian/Serbian	4 *janes-tagger*	a tagger for non-standard Slovene, Croatian and Serbian
Slovene/Croatian/Serbian	5 *reldi-tagger*	a tagger and lemmatiser for Croatian, Serbian and Slovene
language-independent	6 *tweetgeo*	collection and visualising geographically-encoded data
Slovene/Croatian/Serbian	7 *redi*	a diacritic restoration tool for Croatian, Serbian and Slovene
language-independent	8 *GATE Twitter collector*	a language-independent Cloud-based tool for collecting tweets by keyword, author, geographical region and language
language-independent	9 *GATE tools*	a series of tools for Twitter specific Named Entity recognition, Named Entity linking, tokenisation, language identification, sentence splitting, normalisation, PoS-tagging, and sentiment analysis. These tools are suited for English, French and German data
Hungarian	10*Hunaccent*	an accentizer of Hungarian text
language-independent	11*twython*	an actively maintained, pure Python wrapper for the Twitter API
language-independent	12*dmi-tcat*	a set of tools used for the retrieval and collection of tweets from Twitter for statistical analysis
English	13Tweet NLP	a tokenizer, part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated corpora and web-based annotation tools

5.1 Tools

Tools 1-7 are available within the repository of *Clarin.si*. Tools 8-9 were found on the website of CLARIN-UK. The remaining tools 10-13 are not part of the CLARIN infrastructure and were pointed out by participants at the Kaunas workshop.