CHAPTER 7

The FAIR Index of CMC Corpora

Jennifer-Carmen Frey[1], Alexander König[1, 2], Egon Stemle[1],
Achille Falaise[3], Darja Fišer[4], Harald Lüngen[5]
[1]Eurac Research, [2]CLARIN ERIC (Netherlands)
[3]University of Paris, LLF – CNRS (France)
[4]University of Ljubljana and Jožef Stefan Institute (Slovenia)
[5]Leibniz-Institut für Deutsche Sprache (Germany)

7.1. INTRODUCTION

The last few years have brought forward a significant amount of language corpora dedicated to computer-mediated communication (CMC). In Europe alone, more than 30 CMC corpora can be identified and found via a simple web search, for example through their listing in the CLARIN CMC Resource Family[1]. This development is evidence of a vibrant field of research as well as a general aim towards making existing research resources known. A substantial number of these corpora claim to become or stay available as research data resources for further exploration, replication, or referencing. However, whether research data can be made available and reusable often depends not only on the willingness of the data collector. Especially in the domain of CMC data, a highly debated realm in terms of privacy and data protection, targeted steps are needed to allow for any kind of use or reuse of data for research or dissemination purposes, including the retrieval of user consent, anonymisation, and high-quality processing. Apart from copyright issues and problematic terms of use/terms of service of certain platforms (e.g. Twitter, which does not allow the redistribution of Twitter content), other factors such as data formats, missing long-term preservation strategies, (lack of) persistent identifiers, access protocols or lacking documentation may

[1] https://www.clarin.eu/resource-families/cmc-corpora

prevent or at least hamper the dissemination and accessibility of CMC corpora.

In this article, we focus on these other factors and examine the current situation of data dissemination and provision for CMC corpora. By that we aim to give a guiding grid for future projects that will improve the transparency and replicability of research results as well as the reusability of the created resources. Based on the FAIR guiding principles for research data management (Wilkinson et al., 2016)[2], we evaluate the 20 European CMC corpora listed in the CLARIN CMC Resource family, individuate successful strategies among the existing corpora and establish best practices for future projects. We give an overview of existing approaches to data referencing, dissemination and provision in European CMC corpora, and discuss the methods, formats and strategies used. Furthermore, we discuss the need for community standards and offer recommendations for best practices when creating a new CMC corpus.

7.2. The FAIR Guiding Principles for Data Management and Stewardship

The FAIR Guiding Principles for Data Management and Stewardship, published by Wilkinson et al. (2016), provide a universal framework for data management based on the principles of Findability, Accessibility, Interoperability and Reusability. They have received international support, for example, at the G20 Summit in Hangzhou3, have helped with the establishment of community-standards for research data management (Mons et al., 2017) in various domains (e.g. Boeckhout et al., 2018) and have been incorpo-rated into relevant funding schemes like Horizon 2020 (European Commission, 2016).

In short, the FAIR principles aim to describe those characteristics of research data that encourage data reuse within the scientific community. They give added value to the scientific community by facilitating the discovery of existing data and, at the same time, provide a foundation for the transparency and reproducibility of

[2] https://www.go-fair.org/fair-principles/

[3] https://www.consilium.europa.eu/media/23621/leaders_communique hangzhousummit-final.pdf

research results by making available the original data. Overall, this also helps to ensure the long-term preservation of funded research.

7.3. THE FAIRNESS OF EUROPEAN CMC CORPORA

Over the last 20 years, a considerable number of CMC corpora have been created in Europe. The corpora display a broad range of languages, research aims, genres and topics and provide a wide variety of possibly reusable research data. However, the findability, accessibility, interoperability and reusability of the corpora differ considerably. In this article, we analyse the compliance with the FAIR principles for each of the 20 CMC corpora listed in the CLARIN CMC Resource Family[4] in order to gain an overview of the current state of affairs regarding the dissemination and provision of the available corpora. Table 1 below lists the investigated corpora and describes their general characteristics. In addition, the resources and their web references are cited in a dedicated section in the bibliography. In the remainder of this section, we present the results of our investigation. The acronyms introduced below refer to the official guidelines listed in the FAIR principles. Table 2 at the end of the article shows the results of our evaluation.

[4] We bundled together the Janes corpora, originally listed individually in the CMC Resource Family, as the individual corpora do not differ with regards to FAIR, similar to the CoMeRe corpora.

corpus name	language	data type	data format*	corpus description	availability and licensing	size in tokens
Corpus of contemporary blogs (CoCB)	Czech	blogs	XML tags for sentence mark-up	https://nlp.fi.muni.cz/projekty/cocb/	free download deposited at CLARIN CC-BY-NC-ND	1 M
SoNaR New Media (SoNaR)	Dutch	tweets, chat, SMS	FoLiA XML	(Sanders, 2012)	free download[1] deposited at CLARIN ACA-BY-NC-ND	35 M
The DiDi Corpus of South Tyrolean CMC (DIDI)	German, Italian, English, others	Facebook posts, comments, chat (instant messaging)	XML RelAnnis, JSON (Facebook)	(Frey et al., 2015, 2016)	free download[1] deposited at CLARIN ACA-BY-NC-ND	600 K
The Mixed Corpus: New Media (Mixed)	Estonian	chat (room), forum posts, news comments	TEI P5, TEI P3	http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/	on request, partly downloadable on institutional website license unknown	25 M
Suomi 24 Corpus (Suomi)	Finnish	forum posts	VRT	(Lagus et al., 2016)	free download[1] deposited at CLARIN ACA-BY-NC	2.6 B

[1] CLARIN federated identity needed (https://www.clarin.eu/content/federated-identity)

[2] Twitter data for Janes Corpora contains only annotations, text not available for download due to Twitter policy.

* Data formats assigned here comply with the formats declared in corpus descriptions, whenever this information was given. Corpora marked as TEI-CMC are modelled in customized TEI P5 formats that represent pre-versions of the new official CMC-core specification (Beißwenger & Lüngen, 2020).

CoMeRe repository (CoMeRe)	French	mails, forum posts, chat (room), tweets, and SMS	TEI-CMC	(Chanier et al., 2014)	free download deposited at Ortolang CC-BY	80 M
Dortmund Chat Corpus (Dortm. Chat)	German	chat (room)	ChatXML, TEI-CMC	(Beißwenger, 2013; Beißwenger et al., 2015)	free download deposited at CLARIN CC-BY	1 M
LITIS v.1	Lithuanian	forum posts	TXT, metadata inside file	-	free download deposited at CLARIN ACA-BY-NC-ND	190 K
Janes Corpora 1.0	Slovenian	blogs, forum posts, news comments, tweets, Wikipedia talk	TEI, vertical format	(Fišer et al., 2018)	free download[2] deposited at CLARIN CC-BY-SA	240 M
Flemish Online Teenage Talk (TeenTalk)	Dutch	Facebook posts, WhatsApp	-	-	no	2.9 M
Dereko – News and Wikipedia subcorpus	German	Newsgroup posts and Wikipedia talk	WikiXML, I5-TEI P5	(Margaretha et.al., 2014)	free download institutional website CC-BY-SA	670 M
DWDS – Blogs	German	blogs	TEI P5	(Barbaresi et al., 2014)	only corpus query	102 M
Monitor corpus of tweets from Austrian users (TAC)	German, English	Twitter	JSON (Twitter)	(Barbaresi, 2016)	on request (however, restricted by Twitter policy)	40 M

					licence unknown	
FORUMAS_INDV corpus (FOR INDV)	Lithuanian	forum posts	TXT, metadata inside file	http://dangus.vdu.lt/~jkd/eng/?page_id=16	free download on institutional website license unknown	600 K
INT_KOMETARAI_INDV2 corpus (KOM INDV)	Lithuanian	comments	TXT, metadata inside file	http://dangus.vdu.lt/~jkd/eng/?page_id=16	free download on institutional website license unknown	4 M
NTAP climate change blog corpus	Norwegian English, French	blogs	-	(Salway et al., 2016)	no	21 M
Corpus of Highly Emotive Internet Discussions (HEID)	Polish	Twitter	JSON, SQLite database	(Sobkowicz, 2016)	on request	160 M
sms4science	German, Italian, French, Romansh	SMS	-	(Dürscheid & Stark, 2011; Stähli et al., 2011)	only corpus query	0.5 M
What's up, Switzerland?	German, Italian, French, Romansh	WhatsApp	-	https://www.whatsup-switzerland.ch/	on request	5 M
The Corpus of Welsh Language Tweets (Welsh Tweets)	Welsh	Twitter	CSV	http://techiaith.cymru/data/corpora/twitter/	on request	7 M

Table 1: Corpora investigated and their basic characteristics

7.3.1. Findability of CMC Corpora

Many European CMC corpora provide a persistent identifier (PID) and can be reliably found by accessing the corpus through their PID. However, about half of the corpora we observed did not provide such a PID, and in some cases, URLs to corpora (e.g., those published in scientific articles) were no longer valid, which severely decreases corpus findability. **Assigning a persistent identifier** (FAIR: F1) to a corpus is thus of utmost importance for the identification and findability of these data resources. Some well-known examples of persistent identifiers (PID) are Cool URIs[1], Handles[2] and the well-known DOIs[3] which are technically also based on the handle protocol. Research data repositories for language corpora such as CLARIN centres (Hinrichs & Krauwer, 2014) or other data repositories such as META-SHARE[4], Zenodo[5], and figshare[6] usually provide this service for deposited resources.

CMC corpora are frequently described in dedicated corpus description articles or on project web pages. Often these articles contain the main corpus metadata regarding its author, data collection, processing, and annotation of the data. However, those descriptions are composed in prose and are therefore not machine-actionable. In order to ensure findability of data resources, **metadata should be provided in a machine-actionable, structured and searchable format** (FAIR: F2). Metadata standards like Dublin Core[7]/OLAC[8] or CLARIN's CMDI[9] can provide such a format for corpora. In any case, a minimal set of metadata fields, for example, regarding authorship, provenance and versioning of the data must be present (see also section 7.3.4). As most research data repositories do not enforce the provision of this type of metadata, it is the responsibility of the publisher to make sure that rich metadata describes all the necessary information for reproducibility and re-usability of CMC corpora. About half of the European CMC corpora we investigated are

[1] https://www.w3.org/TR/cooluris/

[2] https://handle.net/

[3] https://www.doi.org/

[4] http://www.meta-share.org/

[5] https://zenodo.org/

[6] https://figshare.com/

[7] https://www.dublincore.org/

[8] http://www.language-archives.org/OLAC/metadata.html

[9] https://www.clarin.eu/content/component-metadata

presented to the public through a dedicated research paper. In addition, corpus or project web pages describing the data are not uncommon (16/20). However, machine-actionable data is only provided for the corpora deposited in a research data repository, as those usually provide the infrastructure to store and distribute metadata in one or multiple specific metadata formats. Metadata that is exclusively stored within data files and not marked as such, as observed with some of the corpora, is not recommended. For two corpora, the NTAP corpus and Flemish Online Teenage Talk, neither machine-actionable nor other types of data descriptions were available.

It is furthermore essential that the metadata describing the corpus **explicitly references the actual data** (FAIR: F3**).** In the case of deposited European CMC corpora, this has been ensured by the direct use of the PID in the metadata file. Non-persistent hyperlinks to corpus data provided on corpus web pages or in research papers are often outdated and might therefore not reliably link to the resources. For better findability of the resources, such links should consequently be avoided, and persistent identifiers should be used instead.

To ensure findability, corpora need to be **indexed in searchable registries** (FAIR: F4). This can be either general-purpose search engines like Google or Bing or domain-specific registries that index language resources such as the CLARIN Virtual Language Observatory (VLO)[10] or the Catalogue of the Open Language Archives Community (OLAC)[11]. Also, the inclusion of a corpus in domain-relevant lists like the CLARIN CMC resource family can increase findability. For the corpora we investigated, a Google search for the corpus name did not always lead to practical corpus information like a research paper describing the corpus, a corpus or project web page or a metadata file for the corpus. This was only given for 16 of the 20 corpora. Searches for the remaining corpora lead to no results at all or to archived news events mentioning the corpus within event descriptions. Concerning the findability of corpora via search interfaces, we noticed the use of a data repository considerably increased findability because most of them automatically add the information to specialised search engines like the VLO or OLAC. The non-deposited corpora were not findable via these search engines.

[10] https://vlo.clarin.eu/

[11] http://search.language-archives.org

7.3.2. Accessibility of CMC Corpora

If one aims to make corpora accessible, access modality and access rights must be handled in a standardised and transparent way. This means that metadata and data can be **retrieved autonomously via their identifier using a standardised communication protocol** (FAIR: A1), where the **protocol is open, free and universally implementable** (FAIR: A1.1) and **allows for authentication and authorisation where necessary** (FAIR: A1.2). For all intents and purposes, this will usually be the HTTP protocol, which is always used when corpora are accessible via the web. Thus, access via an identifier (e.g. URL, ideally a PID) and retrieval via a standardised protocol are both given. As data accessibility does not necessarily mean that the data are open and freely available for the public, authentication and authorisation mechanisms should be in place as well. We found all these requirements present for the European CMC corpora that are deposited in a research data repository (8/20)[12]. Other corpora that were available as free downloads on corpus web pages (4/20) did not always provide authorisation and authentication mechanisms. Non-standardised access rights that depend on individual, personal communication (e.g. mail requests) are relatively common among the CMC corpora investigated (4/20) but should be avoided to comply with the FAIR principles, mainly because it is often not clear how and under which conditions the corpora can be accessed and reused.

Finally, accessibility includes also that **metadata are accessible even though the data themselves are not available (anymore)** (FAIR: A2). This is particularly relevant for data that have been available previously, or for data that have restricted access. Conscious steps should be taken to secure the long-term preservation of the metadata. Note that, although we could not evaluate this in our investigation, this point can also be addressed by depositing data in a research data repository because most of them have a strict policy regarding this problem.

[12] For example, the CLARIN infrastructure offers the possibility of federated identification that allows any institutional user account valid for the CLARIN infrastructure to retrieve data from CLARIN repositories (see https://www.clarin.eu/content/federated-identity).

7.3.3. Interoperability of CMC Corpora

In order to make corpora interoperable, the **language used for knowledge representation in both metadata and data should be formal, accessible and broadly applicable** (FAIR: I1). That concerns, on the one hand, the use of standardised, platform-independent formats, on the other hand, the provision of commentary and additional information in widely understood languages like English. While there is no widely agreed-upon standard for knowledge representation used in the majority of available resources, some steps have been made in recent years to establish such a standard. Three of the investigated CMC corpora use customisations of the TEI P5 Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2007) developed by members of the TEI CMC SIG[13]. However, also other typical formats for corpora, such as various XML document types, JSON or CSV for corpus data are used and fulfil the needs for formal, accessible, shared and broadly applicable knowledge representation. In terms of metadata, possible formats include Dublin Core/OLAC or the CMDI format (Broeder et al. 2012) that is usually used for corpora deposited in CLARIN repositories (8/20). Unstructured metadata, proprietary or undocumented formats and commentary or additional information that is only available in non-widely understood languages should be avoided to ensure interoperability.

To further enhance interoperability, the **use of findable, accessible, interoperable and reusable vocabularies for data and metadata representation** is recommended (FAIR: I2). To date, there are no such vocabularies known to us that would fit the needs for representing CMC corpora. The vocabularies used for data and metadata in the existing European CMC corpora are furthermore rarely standardised or even documented and therefore do not comply with FAIR. We consider this a lacuna that should be addressed in the future.

Cross-references between different data (and metadata) are not always necessary but become relevant in the presence of similarly named corpora, related projects, different versions of a corpus, or the publication of different sub-corpora (FAIR: I3). Although the need for appropriate cross-references is thus a rather subjective matter, we have

[13] https://wiki.tei-c.org/index.php?title=SIG:Computer-Mediated_Communication, see also Beißwenger & Lüngen (2020).

found some corpora (three had no or insufficient cross-references, for five corpora the references made were not entirely clear) that would benefit from clear cross-references to other projects, different versions, or related corpora.

7.3.4. Reusability of CMC Corpora

Ultimately, even findable, accessible and interoperable data are not reusable, if the specifics of data collection, authorship and pre-processing (i.e. its provenance), and the terms and conditions for its reuse (i.e. licensing) are not clear. According to the FAIR principles, data have to be **described extensively with accurate and relevant attributes** (FAIR: R1).

Many CMC corpora provide a **clear and accessible usage license** (FAIR: R1.1) (e.g. 9/20 have been licensed explicitly, 5/9 use a Creative Commons license[14], and 4/9 use an academic license). The licensing for the other corpora was, however, less explicit, sometimes being restricted to simple statements in corpus description articles (e.g. stating that the corpus is “openly available”), sometimes being left out altogether.

To **specify metadata and data provenance sufficiently** (FAIR: R1.2), CMC corpora should aim to provide the type of communication (e.g. chat, blog, forum), the origin of the data (e.g. Twitter, Facebook), the time of data creation and data collection as well as the corpus creator(s), possible updates and version numbers. However, not all corpora we investigated provided enough information to make the data reusable. Vague or blanket statements often obfuscate the data source or conditions of pre-processing as well as the versioning of the data and limit thus its reusability substantially.

Finally, **domain-relevant community standards for metadata and data** (FAIR: R1.3), still need to be established for CMC corpora, as can be seen by our investigation. This regards standardised vocabularies, minimal sets of metadata as well as data formats for CMC corpora.

[14] https://creativecommons.org/

7.4. BEST PRACTICES FOR CREATING CMC CORPORA

First and foremost, it is advisable to think about how to handle the data during and after a research project. In fact, many funding agencies already require researchers to prepare a data management plan (DMP) at the project proposal stage to formalise these thoughts, and even if it is not required, we recommend it as a reasonable first step. For preparing a DMP, good guidelines already exist, for example from the Digital Curation Centre in the UK[15], and many research institutes and universities have set up dedicated research data management offices to help researchers.

Regarding the FAIR principles, we want to emphasise that both the Findability and Accessibility principle can be realised by merely depositing the corpus in a research data repository, for example, a CLARIN Centre, which communicates the existence of the corpus to domain-relevant search engines, assigns a persistent identifier and allows the download of the data that may be restricted in access. It is also important to define a license for the corpus, which ideally does not prevent reuse, and to display this license in a prominent position. Most research data repositories prefer well known licenses, but also allow user-defined ones, and enforce an explicit choice.

Compared to Findability and Accessibility, the principles of Interoperability and Reusability are not immediately solved by depositing the corpus in a research data repository but are indeed specific to the community. First, research data must be stored in open and well documented formats. Here, the CMC community is responsible for developing and documenting common standard formats for CMC data. One important step has already been taken with the CMC core extension to the TEI P5 Guidelines, which was recently submitted to the TEI consortium as a feature request (Beißwenger & Lüngen, 2020) and will hopefully be adopted by more corpora in the future. Secondly, research data must have extensive metadata. In the case of CMC data, we consider it particularly important to provide information about the data provenance, that is, when the data were collected, what kind of data were collected and where it came from

[15] http://www.dcc.ac.uk/resources/data-management-plans

(e.g. Facebook, Twitter, blogs). In short, these best practices can be summed up by the following questions.

- Are my data findable through a search engine, the VLO, OLAC?
- Does the corpus have a Persistent Identifier?
- Is there a clear license attached (that ideally permits reuse)?
- Are the data stored in an open and well-documented format?
- Do the metadata describe the data correctly and comprehensively, also covering the provenance of it?

The CMC community is in the fortunate position that work has already been undertaken and that the community as a whole sees the need for and benefits of common standards for data formats and procedures for data documentation. With further targeted work, where the CMC community continues to discuss these issues openly, the process can serve as a model for other corpus-based disciplines. The planned establishment of a CLARIN Knowledge Centre[16] for CMC (König, 2018) will support this path by formalising and centralising existing know-how and thus acting as a catalyst for the further development of standardisation of CMC corpora. We wish and hope that the centre and this article will help to bring the CMC corpora closer to the ideal of FAIR research data management.

[16] https://www.clarin.eu/content/knowledge-centres

		CoCB	SoNaR	DIDI	Mixed	Suomi 24	CoMeRe	Dortm. Chat	LITIS	Janes	TeenTalk	Dereko	DWDS Blogs	TAC	FOR_INDV	KOM_INDV	NTAP	HEID	sms4science	What's up	Welsh Tweet
Findability																					
F1: PID		x	x	x		x	x	x	x	x		x									
F2: Machine-actionable metadata		x	x	x		x	x	x	x	x		x									
F3: Refers to PID		x	x	x		x	x	x	x	x		x			x[1]	x[1]					
F4: Indexed in search engines	Metadata	x	x	x	~	x	x	x	x	x		~	~	~	~	~		~	~	~	~
	Data	x	x	x	~	x	x	x	x	x		~			x	x					
Accessibility																					
A1: Retrieval via standardised protocol	Metadata	x	x	x		x	x	x	x	x		x		~	~	~		~	~	~	~
	Data	x	~	x		x	x	x	x	x		~			x	x					
A1.1: Open, free, universally implementable protocol	Metadata	x	x	x		x	x	x	x	x		x			~	~		~	~	~	~
	Data	x	x	x		x	x	x	x	x		~			x	x					
A1.2: Protocol allows for authentication	Metadata	x	x	x		x	x	x	x	x		/	~					~	~	~	~
	Data	x	x	x		x	x	x	x	x		/			x	x					
Interoperability																					
I1: Language for knowledge representation	Metadata	~	x	x	x	x	x	x	x	x		x									
	Data	x	x	x	x	x	x	x	x	x		x									

		CoCB	SoNaR	DIDI	Mixed	Suomi 24	CoMeRe	Dortm. Chat	LITIS	Janes	TeenTalk	Dereko	DWDS Blogs	TAC	FOR_INDV	KOM_INDV	NTAP	HEID	sms4science	What's up	Welsh Tweet
I2: FAIR vocabularies	Metadata																				
	Data																				
I3: Cross-referencing		/	~	/	/	x	x	x	/	x		~	~	/	~	~	/	/		/	
Reusability																					
R1: Extensive & machine-actionable metadata:	Author information	x	x	x	x	x	x	x	x	x				x	x	x		x	x	x	x
	Version information		x	x		x	x	x	x	x		x							x		
	Source information	x	x	x	x	x	x	x	x	x		x	~	x				x	x	x	x
	Year information	x	x	x	x	x	x	x	x	x		x	~				x	x	x	x	
R1.1: Licensing		x	x	x		x	x	x	x	x		x	x								
R1.2: Data provenance	Metadata	x	x	x	x	x	x	x	x	x		x	~	~	~	~		~	~	~	~
	Data	~	x	x	x	x	x	x	x	x		x	~	~				~	x	x	~
R1.3: Community standards	Metadata	~	x	x	x	x	x	x	x	x		x									
	Data	x	x	x	x	x	x	x	x	x		x		~							

(x) fulfilled, (~) partly fulfilled, (/) not applicable, not relevant

Table 2: Compliance of the investigated corpora with the FAIR principles.

REFERENCES

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus, *Zeitschrift für germanistische Linguistik*, 41(1), pp. 161–164.

Beißwenger, M., Ehrhardt, E., Horbach, A., Lüngen, H., Steffen, D., & Storrer, A. (2015). Adding Value to CMC Corpora: CLARINification and Part-of-Speech Annotation of the Dortmund Chat Corpus. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*, Duisburg-Essen, September 28 at GSCL. GSCL, pp. 12–16.

Beißwenger, M., & Lüngen, H. (2020). CMC-core: a schema for the representation of CMC corpora in TEI. In *Corpus 20*, Special issue on "Corpus complexes : Traitements, standardisation et analyse des corpus de communication médiée par les réseaux", C. Poudat, C. R. Wigham & L. Liégeois (eds.). https://journals.openedition.org/corpus/

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? *European Journal* of Human Genetics, 26(7), pp. 931–936.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 1387–1390.

Dürscheid, C., & Stark, E. (2011). sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland, *Digital Discourse: Language in the New Media*. Oxford University Press, p. 299.

European Commission: Directorate-General for Research & Innovation. (2016). H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 (nº 3).

Fišer, D., Ljubešić, N., & Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content, *Language Resources and Evaluation*. Springer, pp. 1–24.

Frey, J.-C., Glaznieks, A., & Stemle, E. W. (2015). The DiDi Corpus of South Tyrolean CMC Data. In Beißwenger, M. & Zesch, T. (eds) *Proceedings of the 2nd Workshop on Natural Language Processing for*

Computer-Mediated Communication / Social Media, Duisburg-Essen, September 28 at GSCL. GSCL, pp. 1–6.

Frey, J.-C., Glaznieks, A., & Stemle, E. W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. Naples, Italy.

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 1525–1531.

König, A (2018). Towards a CLARIN Knowledge Centre for CMC. *6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*. 17/18 September 2018. Antwerp.

Margaretha, E., & Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. *In Journal of Language Technology and Computational Linguistics*, 29(2), pp. 59-82.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. In *Information Services & Use*, 37 (1), pp. 49–56.

TEI Consortium (eds.). (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific data*, 3.

DATA CITATIONS

Below, all the investigated corpora are cited in the same order as presented in the tables. Where no PID was available we cited the project website instead. Website states refer to the access date 07/01/2020.

Corpus of contemporary blogs
http://hdl.handle.net/11858/00-097C-0000-000E-011B-8

SoNaR New Media
http://hdl.handle.net/10032/157d6fee6134f5beab09b159dd7c710a

DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0
http://hdl.handle.net/20.500.12124/7

The Mixed Corpus: New Media
https://www.cl.ut.ee/korpused/segakorpus/uusmeedia/

Suomi 24 Corpus[1]
http://urn.fi/urn:nbn:fi:lb-2017021502

CoMeRe repository
https://hdl.handle.net/11403/comere

Dortmund Chat Corpus
http://hdl.handle.net/11858/00-203Z-0000-002D-ECC7-2

LITIS v.1
https://hdl.handle.net/20.500.11821/11

Janes Corpora 1.0
Blog posts and comments http://hdl.handle.net/11356/1138
Forum. http://hdl.handle.net/11356/1139
News comments http://hdl.handle.net/11356/1140
Twitter http://hdl.handle.net/11356/1142
Wikipedia talk http://hdl.handle.net/11356/1137

Flemish Online Teenage Talk
http://www.clips.ua.ac.be/category/projects/flemish-online-teenage-talk

Dereko – News and Wikipedia subcorpus
https://www1.ids-mannheim.de/s/corpus-linguistics/projects/corpus-development/availability.html?L=1

DWDS – Blogs
https://www.dwds.de/d/k-web#blogs

[1] No landing page, immediately asks for login.