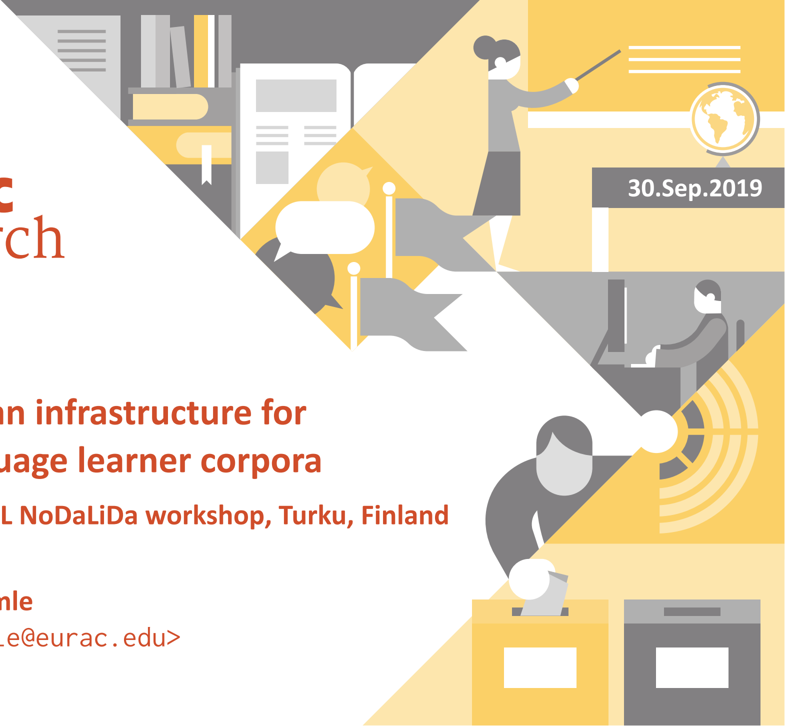
eurac
research



30.Sep.2019

Towards an infrastructure for FAIR language learner corpora

8th NLP4CALL NoDaLiDa workshop, Turku, Finland

Egon W. Stemle
<egon.stemle@eurac.edu>

Who are we?

What brought me here? – Frequently Asked Questions About Linguistics...

What *is* Linguistics, anyway?

Linguistics is the scientific study of human language.

What do you mean by ”human language”?

All humans have some language, without exception. It’s a distinguishing biological trait of Homo Sapiens. And all human languages have a lot of similarities, and they tell us quite a lot about what it means to be human, which is a big preoccupation of H. Sapiens. Here’s what Edward Sapir (considered by many the greatest American linguist of the last century) said:

> *...*
>
> *Language is the most massive and inclusive art we know, a mountainous and anonymous work of unconscious generations.*

▸ Sapir, 1921

What brought me here? – Frequently Asked Questions About Linguistics...

All right, what do you mean by ”scientific”?

Roughly, *objective*, *unbiased*, *data-oriented*, and ***reproducible***, among other meanings. Simply put, linguists are concerned with how language actually does work, rather than with how (somebody says) it ought to work. This is a fairly new approach to a very old interest, since people have always been interested in language, even though it is a hard subject to talk about.

▸ John M. Lawler’s personal web page (University of Michigan)

Background

- Recent trend for Social Sciences and Humanities (SSH) research to become **more**
  - reproducible
  - reusable
  - transparent

> *The European Commission unveiled its plans to make all data derived from EU-funded research projects findable, accessible, interoperable and reusable (FAIR). The Commission estimates that €2 billion in Horizon 2020 funding will be allocated to its so-called 'European Cloud initiative'.*

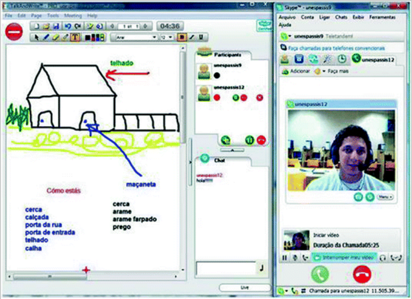
→ Research data management on the basis of FAIR (Findability, Accessibility, Interoperability, Reusability)

▸ Wilkinson et al., 2016

How applicable is FAIR to CMC corpora?

CMC corpora vs. language learner corpora

Computer-mediated communication (CMC) refers to human communication via computers and includes many different forms of synchronous, asynchronous or real-time interaction that humans have with each other using computers as tools to exchange text, images, audio and video.



unespassis12
hola!!!!

Cardoso and Matos, 2013

Findability - F

Requirements

- Rich and accurate metadata
- Machine-actionable metadata
- Unique persistent identifiers (PID) for data and metadata
- Registered/indexed in a search engine

Accessibility - A

Requirements

- Automatic retrieval of (meta)data on the basis of PID
- Allow for authentication and authorization if necessary
- Metadata should always be public

Interoperability - I

Requirements

- Use shared standards for knowledge representation
- Proper documentation
- Cross-references to other data (if necessary)

Reusability - R

Requirements

- Appropriate description of data
- Proper attribution of creators
- Clear and accessible usage license
- Extensive documentation of provenance

The CMC Corpora

The CLARIN Resource Family for CMC Corpora

- 24 corpora of computer-mediated communication of
  - various sizes (600k to 670m tokens)
  - languages (e.g. DE, LT, SI, NL, ...)
  - and sources (e.g. Twitter, Whatsapp, Blogs).

▸ CLARIN website

- Ca. 50% (13) are deposited in a CLARIN Centre or a similar repository (META-SHARE, ..., zenodo, figshare).

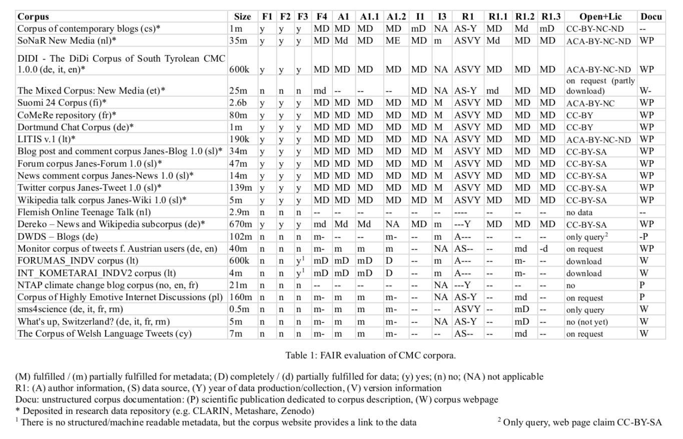
Methodology

For each corpus, we checked how well it complied with the four principles and its various subparts

1. Findability: via Google/Bing, VLO and OLAC search.
2. Accessibility: Was the data accessible?
3. Interoperability: Which format was the data in?
4. Reusability: Documentation of formats/methodology, licensing.
5. Other: Is the data openly available, is there a corpus paper or website.

The Results

Our result table



Corpus	Size	F1	F2	F3	F4	A1	A1.1	A1.2	I1	I3	R1	R1.1	R1.2	R1.3	Open+Lic	Docu
Corpus of contemporary blogs (cs)*	1m	y	y	y	MD	MD	MD	MD	mD	NA	AS-Y	MD	Md	mD	CC-BY-NC-ND	--
SoNaR New Media (nl)*	35m	y	y	y	MD	Md	MD	ME	MD	m	ASVY	Md	MD	MD	ACA-BY-NC-ND	WP
DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0 (de, it, en)*	600k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
The Mixed Corpus: New Media (et)*	25m	n	n	n	md	--	--	--	MD	NA	AS-Y	md	MD	MD	on request (partly download)	W-
Suomi 24 Corpus (fi)*	2.6b	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	ACA-BY-NC	WP
CoMeRe repository (fr)*	80m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
Dortmund Chat Corpus (de)*	1m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
LITIS v.1 (lt)*	190k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
Blog post and comment corpus Janes-Blog 1.0 (sl)*	34m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Forum corpus Janes-Forum 1.0 (sl)*	47m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
News comment corpus Janes-News 1.0 (sl)*	14m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Twitter corpus Janes-Tweet 1.0 (sl)*	139m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Wikipedia talk corpus Janes-Wiki 1.0 (sl)*	5m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Flemish Online Teenage Talk (nl)	2.9m	n	n	n	--	--	--	--	--	--	----	--	--	--	no data	--
Dereko – News and Wikipedia subcorpus (de)*	670m	y	y	y	md	Md	Md	NA	MD	m	---Y	MD	MD	MD	CC-BY-SA	WP
DWDS – Blogs (de)	102m	n	n	n	m-	--	--	m-	--	m	A---	--	--	--	only query[2]	-P
Monitor corpus of tweets f. Austrian users (de, en)	40m	n	n	n	m-	m	m	m	--	NA	AS--	--	md	-d	on request	WP
FORUMAS_INDV corpus (lt)	600k	n	n	y[1]	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
INT_KOMETARAI_INDV2 corpus (lt)	4m	n	n	y[1]	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
NTAP climate change blog corpus (no, en, fr)	21m	n	n	n	--	--	--	--	--	NA	---Y	--	--	--	no	P
Corpus of Highly Emotive Internet Discussions (pl)	160m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	md	--	on request	P
sms4science (de, it, fr, rm)	0.5m	n	n	n	m-	m	m	m-	--	--	ASVY	--	mD	--	only query	W
What's up, Switzerland? (de, it, fr, rm)	5m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	mD	--	no (not yet)	W
The Corpus of Welsh Language Tweets (cy)	7m	n	n	n	m-	m	m	m-	--	--	AS--	--	md	--	on request	W

Table 1: FAIR evaluation of CMC corpora.

(M) fulfilled / (m) partially fulfilled for metadata; (D) completely / (d) partially fulfilled for data; (y) yes; (n) no; (NA) not applicable
R1: (A) author information, (S) data source, (Y) year of data production/collection, (V) version information
Docu: unstructured corpus documentation: (P) scientific publication dedicated to corpus description, (W) corpus webpage
* Deposited in research data repository (e.g. CLARIN, Metashare, Zenodo)

[1] There is no structured/machine readable metadata, but the corpus website provides a link to the data

[2] Only query, web page claim CC-BY-SA

If you want to study the table in detail: ▸ Frey et al., 2019

The Results: Findability

Clear distinction

- Deposited Corpora
  - Provided a PID and machine-actionable metadata
  - Were indexed in domain-specific search engines (CLARIN VLO, OLAC)
- Non-Deposited Corpora
  - Metadata most of the time only available via web pages or corpus papers
  - No use of PIDs
  - Some links to data on web pages/in corpus papers were outdated
  - Some of the corpora could not be found at all

The Results: Accessibility

Again, clear distinction

- Deposited Corpora
  - All provide a clear usage license (mostly Creative Commons)
  - Data is always retrievable (if authorized) through the repository
- Non-Deposited Corpora
  - Best case: download link on corpus web page, sometimes researchers need to be contacted directly first
  - Often no or unclear usage license

The Results: Interoperability

No clear distinction between deposited and non-deposited corpora

- Deposited corpora use structured metadata, but also here often crucial information (data source, retrieval period) is missing
- Data comes in a wide range of formats (TEI, XML, JSON), but the specific format is often not well-documented
- Some corpora would have benefited from clearer version information and cross-references to related corpora

The Results: Reusability

No clear distinction between deposited and non-deposited corpora

- Extensive metadata is needed to reuse corpora, but there is no common understanding of which metadata fields need to be present
- Repositories do not demand to fill a certain set of metadata fields
- A lot of non-deposited corpora were lacking a clear license, which is crucial to reuse their data
- A large part of the corpora was missing provenance (collection period, data source) and version information

Some Conclusions

- High variability of how the corpora comply or do not comply with the FAIR principles
- In general, deposited corpora provided much better Findability and Accessibility
- Depositing did not improve Interoperability and Reusability a lot
- Provenance (source, time, description of the social media site) is crucial for reusing corpus data, but often not described in enough detail
- Notably, the community would benefit from commonly accepted guidelines on how to make corpora FAIR as well as from more agreement on used standards

On Linguistic Tools

An example: stop word lists

Open-source software packages for language processing often include stop word lists. Users may apply them without awareness of their surprising omissions (e.g. ”hasn’t” but not ”hadn’t”) and inclusions (”computer”), or their incompatibility with a particular tokenizer.

▸ Nothman et al., 2018

Reproducible Methods

Versioning of linguistic tools

- In linguistic research, handcrafted toolchains (a variety of separate programs) are very common
- Often, it will be difficult to rebuild such a toolchain exactly
  - Some tools might no longer be available (or cannot be found)
  - It might not be completely clear which specific version of a tool was used
  - Some manufacturers do not keep older versions of their software available for download
- One solution is to create a (Docker) container with a ”frozen” version of the complete toolchain
- Such a container can also be made available in a public container registry

Take-Home Message

F Deposit corpus/tool in a data repository

A Deposit corpus/tool in a data repository

I Community needs commonly accepted guidelines and standards

R As an individual project, described corpus/tool to facilitate replication and combination in different settings

References I

Cardoso, T., & Matos, F. (2013). Learning Foreign Languages in the Twenty-First Century: An Innovating Teletandem Experiment Through Skype. In A. Moreira, O. Benavides, & A. J. Mendes (Eds.), *Media in Education: Results from the 2011 ICEM and SIIE joint Conference* (pp. 87–95). doi:10.1007/978-1-4614-3175-6_7

Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC Corpora? In *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)* (pp. 25–30). Cergy-Pontoise University, France. Retrieved from https://cmccorpora19.sciencesconf.org/resource/page/id/15

König, A., & Stemle, E. W. (2019). Technical Solutions for Reproducible Research. In K. Simov & M. Eskevich (Eds.), *Proceedings of CLARIN Annual Conference 2019* (pp. 89–92). Leipzig, Germany: CLARIN. Retrieved from https://api.zotero.org/users/332053/publications/items/D2WMT2UL/file/view

Nothman, J., Qin, H., & Yurchak, R. (2018). Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)* (pp. 7–12). Melbourne, AU: Association for Computational Linguistics. Retrieved from https://aclweb.org/anthology/papers/W/W18/W18-2502/

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Harcourt, Brace. Retrieved from http://www.bartleby.com/186/

References II

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018–160018. doi:10.1038/sdata.2016.18