

Metadata and best practices for open learner corpora

MAGALI PAQUOT (FNRS – UCLouvain)

1

Outline

Open data and FAIR principles

Why is metadata crucial in learner corpus research?

The Core Metadata Schema for Learner Corpora

OA learner corpora: where do we go from here?

2

2

Open data and FAIR principles

16. EMPHASISES that the opportunities for the optimal reuse of research data can only be realised if data are consistent with the FAIR principles (findable, accessible, interoperable and re-usable) within a secure and trustworthy environment; RECALLS the importance of storage, long term preservation and curation of research data, taking into consideration the capacity of the research group or organisation, as well as ensuring the existence of **metadata based on international standards**; ENCOURAGES Member States, the Commission and stakeholders to follow the FAIR principles in research programmes and funding mechanisms.

The transition towards an open science system – Council conclusions (adopted on 27/05/2016)

<https://data.consilium.europa.eu/doc/document/ST-8507-2016-INIT/en/pdf>

3

3

FAIR

Findable

- « Metadata and data should be easy to find for both humans and computers »
- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata

Accessible

- « Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and autorisation »

Interoperable

- « The data usually need to be integrated with other data. In addition, the data need to interoperate with applications and workflows for analysis, storage, and processing »
- (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
- (Meta)data use vocabularies that follow FAIR principles

Reusable

- « The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings »
- (Meta)data meet domain-relevant community standards

<https://www.go-fair.org/fair-principles/>

4

4

Metadata

"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information." (National Information Standards Organization, 2004)

"Metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable." (Higgins, 2007)

5

5

Why is metadata crucial in learner corpus research?

Corpus selection

- Is the learner corpus available for research purposes?
- Does the corpus include the type of data and info about the data that are needed to answer a specific research question?
- Are these two sets of data comparable?
- ...

Corpus compilation

- What variables should be collected?
- How should variables be operationalized?
- ...

Corpus analysis – study design

- Metadata = independent and/or dependent variables, explanatory variables, predictors

Interpretation of the results and cumulative research

- To what extent are results comparable?
- To what extent could the results be due to uncontrolled variables?
- ...

6

6

Metadata in LCR suffers from lack of standardisation

Major differences in quality and quantity of metadata

Different labels for the same variable

- First language vs. native language

Different categories

- Years studying English vs. Years of English at school + years of English at university

Different definitions and operationalizations

- Proficiency

Missing metadata

- Cognitive and affective variables, exposure to L2, ...

Metadata representation: heterogeneity of practices

- File headers, separate file, corpus manual

Baker et al. (2015), Granger & Paquot (2017) 7

7

Several calls in the field ...

« Probably as important as the refinement of certain variables, however, is the urgent need to standardise the metadata that come with learner corpora. » (Gilquin, 2015)

« there is increasing convergence, however, on the need of one relatively stable set of recommendable or obligatory metadata for learner characteristics, and on another set for corpus information. » (Volodina et al., 2018)

« There is a wide range of metadata that can be imagined as desirable in learner corpora, and in the best of all worlds, the community would have already agreed on a metadata schema that defines mandatory and optional values. So we would know what to collect and how, and we would have a shared understanding of what the data meant. » (Stemle et al., 2019)

« One important future step for the field is therefore to work together towards the development of some kind of shared protocol for collecting metadata that would include a minimal list of required variables and recommendations as to how best operationalize them. » (Tracy-Ventura et al., 2021)

8

8








Alex König




Egon Stemle



Jennifer Carmen Frey

Core Metadata Schema V2

[HTTPS://TINYURL.COM/L2METADATAV2](https://tinyurl.com/l2metadatav2)

9

Core

Used to describe different types of learner corpora

E.g. for corpora collected in instructional settings (primary schools, secondary schools, etc.)

Core metadata	Non-core → additional module?
corpus_mode	study_level
target_language	course_level
learner_L1	teacher_id
proficiency_level	school_id

10

10

From LCR state of the art to future practices

Available learner corpora

Desiderata as stated in LCR and L2 research more generally

- Feedback from LCR community
 - Learner_impairment, learner_region, annotation_evaluation, annotation_corrected,
- Publications

Wulff (2023)

- Age of onset and proficiency
- Multilingualism
 - The order of acquisition of acquired languages
 - The degree of use of all languages from onset to current age
 - The context of language use for all languages spoken
- Measures of language attitude and motivation

e.g. Gilquin (2015), Granger (2015), Paquot & Plonsky (2017), Tracy-Ventura et al. (2021)

11

FAIR data in LCR

Administrative metadata

- Persistent identifier
- Version
- Contact_email
- Availability
- Access requirements
- Licence
- Corpus documentation
- Link to other versions
- Corpus reference article
- Related research projects

12

12

Eight main components (159 metadata fields)

1. Administrative metadata (22)
2. Corpus design metadata (34)
3. Learner metadata (49)
4. Text metadata (15)
5. Situation and task-related metadata (21)
6. Annotation metadata (8)
7. Annotator metadata (5)
8. Transcriber metadata (5)

13

13

Situational and task characteristics

Situation	Task
Situation_participants_...	Task_prompt
Situation_medium	Task_material
Situation_purpose	Task_duration
Situation_setting	Task_duration_minutes
Situation_topic_domain	Task_ref_tools_allowed
Situation_register	Task_assessed

REGISTER STUDIES (cf. Biber's work)

14

14

Annotator/transcriber

annotator_id	a unique identifier for the annotator
annotator_L1	the first language of the annotator
annotator_target_language_competence	the competence of the annotator in the target language
annotator_type	Indicate type/experience of annotator (expert, trained, student, teacher, crowdsourcing...)
annotator_L2	other languages spoken by annotator if relevant

15

15

Inter-related components

Text (language sample)

- Learner_id
- Situation_id
- Annotation_id
- Annotator_id
- Transcriber_id

16

16

category	metadata field	description	condition	fixed attributes	example	obligatory /optional	repeat able	cardi nality
Learner metadata								
	learner_id	a unique identifier of the author of the text				obligatory	no	1
Socio-demographic differences								
socio-demographic differences	learner_age	Indicate age at text production				obligatory	no	1
	learner_gender	the gender of the learner		female/male/non-binary		obligatory	no	1
	learner_educational_background	Max level of education achieved		No schooling / Early childhood education / Primary education /		optional	no	0-1
	learner_socioeconomic_background	Informal description of a person's perceived social or economic status, operationalization should be explained in additional material			low, medium, high	optional	no	0-1
	learner_region	Use this field to provide geographical details about the learner if relevant to their language use			Spain, Catalonia, South Tyrol	optional	no	0-1
	learner_environment	The general environment of the learner (urban or rural)			urban/rural	optional	no	0-1
	learner_impairment	Does the learner have an impairment (e.g. deaf or hard-of-hearing learner)?		yes/no		optional	no	0-1
	learner_impairment_details	Please preferably choose among our list of examples. You can use tuples to provide more info about the conditions (e.g. (hearing impairment, permanent))	if "learner_impairment = yes"		hearing impairment/speech impairment/visual impairment/cognitive impairment/learning impairment/other impairments	optional	yes	0-N
Target language: learning context, exposure and proficiency								
	learner_target_language	language of specific production (if various target languages in the corpus)		[ISO language list]		obligatory	no	1
	learner_target_language_instructed	Was the learner in an instructed learning context at the time of data collection?		yes/no		obligatory	no	1
	learner_target_language_immersive_instruction	Was the learner in an immersive instructed learning context at the time of data collection? [Language immersion is a method used to		yes/no		obligatory	no	1

+ Administrative metadata Corpus design metadata **Learner** Text (language sample) Situational and task characteristics Annotation Annotator Trans

17

Obligatory/optional fields

Learner metadata

Obligatory	Optional
Learner_id	Learner_educational_background
Learner_age	Learner_socioeconomic_background
Learner_gender	Learner_region
	Learner_environment
	Learner_impairment
Learner_target_language	Learner_target_language_onset
Learner_target_language_proficiency	Learner_target_language_proficiency_CE
	FR_conversion
	Months_spent_target_language_country
Learner_L1	Learner_parent_L1
	Learner_home_language
	Learner_language_friends

18

18

Repeatable fields

L1 differences							
learner_L1	one L1 or bi-(multi)-lingual		(ISO language list)		obligatory	yes	1-N
learner_parent_L1	The first language(s) of each parent		(ISO language list)		optional	yes	0-N
learner_home_language	language(s) spoken at home (use tuple of (language ISO code, percentage) if you want to add more info)		(ISO language list)		optional	yes	0-N
learner_language_friends	the language(s) used with friends		(ISO language list)		optional	yes	0-N
learner_language_education_kindergarten	the language(s) in which the student was educated at kindergarten		(ISO language list)		optional	yes	0-N
learner_language_education_primary	the language(s) in which the student was educated at primary school		(ISO language list)		optional	yes	0-N
learner_language_education_secondary	the language(s) in which the student was educated at secondary school		(ISO language list)		optional	yes	0-N
learner_language_education_higher	the language(s) in which the student was educated in higher education		(ISO language list)		optional	yes	0-N
learner_L2	Does the learner speak other foreign languages?		(ISO language list)		obligatory	yes	1-N
Other languages: learning context, exposure and proficiency (optional and repeatable set of variables if you want to describe the other languages known in more details)							
learner_L2_language_onset	Age of onset of acquisition	for each "learner_L2"			optional	yes	0-N
learner_L2_language_learning_context	The learning context from onset to current age	for each "learner_L2"		(12-20: instructed)	optional	yes	0-N
learner_L2_language_use_context	The context of use from onset to current age (if needed, use tuples to add extra info related to time spent on these different contexts of use)	for each "learner_L2"	family/friends/leisure/public/work/education	(12-20: education, leisure)	optional	yes	0-N
learner_L2_language_listening_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_speaking_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_reading_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_writing_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_CMC	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_proficiency	the proficiency (score or level) of the learner (describe assessment method in corpus documentation)	for each "learner_L2"			optional	no	0-1
learner_L2_language_proficiency_CEFR_conversion	if available, provide conversion into CEFR	for each	A1/A2/B1/B2/C1/C2		optional	no	0-1

19

19

Set vocabulary or open fields

A thorny issue!

Target language: learning context, exposure and proficiency							
learner_target_language	language of specific production (if various target languages in the corpus)		(ISO language list)		obligatory	no	1
learner_target_language_instructed	Was the learner in an instructed learning context at the time of data collection?		yes/no		obligatory	no	1
learner_target_language_immersive_instruction	Was the learner in an immersive instructed learning context at the time of data collection? (Language immersion is a method used to teach students a second language where the language being taught is primarily used during teaching.)		yes/no		obligatory	no	1
learner_target_language_onset	Age of onset of acquisition				optional	no	0-1
learner_target_language_learning_context	The learning context of the target language from onset to current age (choose among instructed, immersive and naturalistic)			(12-20: instructed)	optional	yes	0-N
learner_target_language_use_context	The context of use of the target language from onset to current age (use tuples also if you want to add extra info related to time spent on these different contexts of use)		family/friends/leisure/public/work/education	(12-20: education, leisure)	optional	yes	0-N
learner_target_language_listening_experience	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_speaking_experience	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_reading_experience	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_writing_experience	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_use_CMC	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_proficiency	the proficiency (score or level) of the learner (describe assessment method in corpus documentation)	if proficiency_assignment_method = "learner-centred"			obligatory	yes	1-N
learner_target_language_proficiency_CEFR_conversion	if available, provide conversion into CEFR	if proficiency_assignment_method = "learner-centred"	A1/A2/B1/B2/C1/C2	(results of test/self)	optional	no	0-1
learner_target_language_official_language_testing_score	score - official language test	if corpus:data_production_setting = "language_testing_setting"			obligatory	no	1

20

20

Set vocabulary or open fields

A thorny issue!

Target language: learning context, exposure and proficiency							
learner_target_language	language of specific production (if various target languages in the corpus)	(ISO language list)		obligatory	no		1
learner_target_language_instructed	Was the learner in an instructed learning context at the time of data collection?	yes/no		obligatory	no		1
learner_target_language_immersive_instruction	Was the learner in an immersive instructed learning context at the time of data collection? (Language immersion is a method used to teach students a second language where the language being taught is primarily used during teaching.)	yes/no		obligatory	no		1
learner_target_language_onset	Age of onset of acquisition			optional	no	0-1	

« There is a need for standards which do not limit choice, but rather facilitate an accurate presentation of the choices made » (Burnard, 2017)

learner_target_language_writing_experience	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_use_CMC	Recommended scale: Never, rarely, once a month, once a week, daily				optional	no	0-1
learner_target_language_proficiency	the proficiency (score or level) of the learner (describe assessment method in corpus documentation)	if proficiency_assignment_method = "learner-centred"		(results of test/self)	obligatory	yes	1-N
learner_target_language_proficiency_CEFR_conversion	if available, provide conversion into CEFR	if proficiency_assignment_method = "learner-centred"	A1/A2/B1/B2/C1/C2		optional	no	0-1
learner_target_language_official_language_testing_score	score - official language test	if corpus:data_production_setting = "language_testing_setting"			obligatory	no	1

21

21

Feedback from the LCR community at each revision stage

Ln differences							
learner_L1	one L1 or bi-(multi)-lingual	(ISO language list)		obligatory	yes	1-N	
learner_parent_L1	The first language(s) of each parent	(ISO language list)		optional	yes	0-N	
learner_home_language	language(s) spoken at home (use tuple of (language ISO code, percentage) if you want to add more info)	(ISO language list)		optional	yes	0-N	
learner_language_friends	the language(s) used with friends	(ISO language list)		optional	yes	0-N	
learner_language_education_kindergarten	the language(s) in which the student was educated at kindergarten	(ISO language list)		optional	yes	0-N	
learner_language_education_primary	the language(s) in which the student was educated at primary school	(ISO language list)		optional	yes	0-N	
learner_language_education_secondary	the language(s) in which the student was educated at secondary school	(ISO language list)		optional	yes	0-N	
learner_language_education_higher	the language(s) in which the student was educated in higher education	(ISO language list)		optional	yes	0-N	
learner_L2	Does the learner speak other foreign languages?	(ISO language list)		obligatory	yes	1-N	
Other languages: learning context, exposure and proficiency (optional and repeatable set of variables if you want to describe the other languages known in more details)							
learner_L2_language_onset	Age of onset of acquisition	for each "learner_L2"		optional	yes	0-N	
learner_L2_language_learning_context	The learning context from onset to current age	for each "learner_L2"		(12-20: instructed)	optional	yes	0-N
learner_L2_language_use_context	The context of use from onset to current age (if needed, use tuples to add extra info related to time spent on these different contexts of use)	for each "learner_L2"	family/friends/leisure/public/work/education	(12-20: education/leisure)	optional	yes	0-N
learner_L2_language_listening_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_speaking_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_reading_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_writing_experience	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_CMC	Recommended scale: Never, rarely, once a month, once a week, daily	for each "learner_L2"			optional	no	0-1
learner_L2_language_proficiency	the proficiency (score or level) of the learner (describe assessment method in corpus documentation)	for each "learner_L2"			optional	no	0-1
learner_L2_language_proficiency_CEFR_conversion	if available, provide conversion into CEFR	for each "learner_L2"	A1/A2/B1/B2/C1/C2		optional	no	0-1

22

22

Release of version 2 planned for 2024

Final checks on the « descriptions »

- Set vocabulary

Documentation

- Essential to have a good readme file/manual to define all the variables and provide some background info

23

23

What's next?

TOWARDS MORE COLLABORATIVE EFFORT IN THE
FIELD

24

24

Maintenance and further development

« Metadata schemas develop in response to a community need and often gain wide acceptance, or are widely used while still in development. Maintenance by nationally or internationally recognised centres of excellence (...) or support from a professional body increases both visibility and take-up so that they become a community's standard schema » (Higgins, 2007).



25

25

The CMS for metadata collection, sharing and use

« the extent to which established standards are used is determined by the availability of tools that work with this standardized data » (Lehmber & Wörner, 2008)

Several possible developments:

- CMDI profile in CLARIN component registry
- Open-source interface and/or form to fill in and export metadata in various formats

26

26

Domain-specific modules provided by experts in the field

Integration of specialized metadata

- Multimodal learner corpora
 - Info on modality (gestures, eye-gaze, facial expressions)
- Translation learner corpora
 - Source texts, translation tasks
- CLIL learner corpora
- Process learner corpora
-

... by experts who work with such corpora !

27

27

Conclusion

28

28

Learner corpus research

Field in rapid expansion

Time to work on standardisation

- Metadata
- Data format
- Transcription guidelines
- Linguistic annotations
- Error annotation

Granger & Paquot (2017) 29

29

A collaborative effort

Collaborative response from the LCR community

1. Collaborative work across several institutions and centres
 - EURAC, CECL, CLARIN, CLARIN K centre for learner corpus research
2. Feedback collection phases from the LCR community

→ Acceptance from the research community is key and has been part of the work flow since the beginning of the collaboration

30

30

Around 160 variables ... you must be kidding!

Many variables will only need to be filled in once

- Administrative
- Corpus design
- All learner/text/situation variables shared across the learner corpus

Only a limited number of variables are obligatory

- Learner variables [10]
- Text [15]

31

31

(Metadata) standardisation

FAIR metadata standards in LCR

Study quality (Paquot & Plonsky, 2017)

LCR contribution to cumulative research

- “Looking at all this from the perspective of comparability of different corpora and different research studies, which should be objective, repeatable, and reproducible, it is striking how difficult it is to compare the results from one corpus with results from another corpus or even with results from the same corpus at another time.” (Stemle et al, 2019)

Guidelines and best practices

32

32

OA learner corpora

Data collection

- Metadata
- Ethics
 - Informed consent
 - Anonymization
 - Data Management Plan



Data curation

- Anonymization > pseudonymization (cf. work by E. Volodina & colleagues)
- Computational Approaches to Language Data Pseudonymization
CALD-pseudo workshop at EACL 2024
<https://mormor-karl.github.io/events/CALD-pseudo/>

Data sharing

- CLARIN, OSF, Dataverse, IRIS
- Persistent identifier

<https://uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora.html>

33

33



Thank you
very much
for your
attention

34

34

Core Metadata Schema for Learner Corpora

Version 1

Paquot, M., König, A., Stemle, E. & Frey, J.-C (2023). Core Metadata Schema for Learner Corpora, <https://doi.org/10.14428/DVN/4CDX3P>, Open Data @ UCLouvain, V1, UNF:6:WhLZTg+knFe2FjjgxGg3Uw== [fileUNF]

Version 2 (draft version)

<https://tinyurl.com/L2metadataV2>

35

35

Corpus metadata	Corpus_proficiency_level: beginner/intermediate/advanced Corpus_proficiency_assignment_available: yes/no Corpus_proficiency_assignment_method: learner-centred or text-centred Corpus_[<i>learner/text</i>].assignment_method Corpus_[<i>learner/text</i>].assignment_instrument Corpus_[<i>learner/text</i>].assignment_documtation
Text metadata	text_proficiency text_official_language_testing_score text_proficiency_CEFR_conversion
Learner metadata	learner_target_language_proficiency learner_target_language_proficiency_cefr_conversion

36

36

Please let us know what you think ...

<https://tinyurl.com/L2metadataV2>

Feedback from the community by November 30

<https://tinyurl.com/LCRmetafeedback>

- Give your name
- Remember our focus on CORE metadata ;-)

37