Core metadata for learner corpora: draft 1.0, 7 December 2017

Sylviane Granger & Magali Paquot

Feedback most welcome: Sylviane.granger@uclouvain.be & magali.paquot@uclouvain.be

	Description	Attributes	Obligatory/optional
1. Administrative metadata			
corpus_title			obligatory
corpus_acronym			optional
distributor	*name of a person or other agency responsible for the distribution of a text*		obligatory
availability	*supplies information about the availability of a text, for example any restrictions on its use or distribution, its copyright status, any licence applying to it, etc.*	[available free of charge ; available at a cost ; available within the xx consortium/xx research group ; not available]	obligatory
licence	*contains information about a licence or other legal agreement applicable to the text.*	[academic licence/commercial licence/other]	obligatory
edition	*describes the particularities of one edition of a text*		optional
character_encoding		[UTF-8]	obligatory
markup_language		xml/other/no	obligatory
2. Corpus design metadata			
L2_target		mono_L2/multi_L2	obligatory

L2_language	describes the languages represented within a corpus	[ISO language list]	obligatory
L1_source		mono_L1/multi_L1	obligatory
L1_language	source language/s	[ISO language list]	obligatory
corpus_size	What is the size of the corpus?		obligatory
corpus_mode		written/spoken/multimodal/various	obligatory
editorial_decisions	spelling correction, normalisation, punctuation, paragraph, etc.	[readme file/corpus documentation]	obligatory if corpus_mode = written
transcription_guidelines	Link to transcription guidelines/decisions	[readme file/corpus documentation]	obligatory if corpus_mode = spoken/multimodal
written_process	Does the corpus include several versions of the same text by the same learner?	yes/no	obligatory if corpus_mode = written
longitudinal	Is the corpus longitudinal?	yes/no	obligatory
proficiency_level	What (broad) proficiency levels are represented in the corpus? (one or more options)	beginner/intermediate/advanced/unknown	obligatory
proficiency_level_type	How was proficiency assessed?	learner-based / text-based / mixed	obligatory
proficiency_level_descriptors	What proficiency descriptors were used to grade the texts or the learners?	[pointer to readme file/corpus documentation]	obligatory if proficiency_level_type ǂ unknown
field	[if element kept, the two terms will require a definition]	general language/specialized language	obligatory

official_language_testing	Was the corpus collected within the framework of official language testing (for language assessment purposes) or by academics/teachers for research/teaching purposes?	yes/no	obligatory
comp_data	Did the corpus compiler also collect other data to be used for comparison purposes? If so, what type?	learner L1 data/target language data/no	obligatory
L1_comp_data	If there is a comparable corpus of data in the first language, is the data from the same learners or different learners?	same/different	obligatory if comp_data = learner L1 data
comp_data_included	Are thecomparable texts part of this corpus?	yes/no	obligatory if comp_data = yes
comp_corpus_name	If the comparable texts are not part of this corpus, what is the name of the comparable corpus?		obligatory if comp_data_included = no
3. Corpus annotation metadata			
annotation	Are the texts annotated?	yes/no	obligatory
annotation_complete	Is the full corpus annotated?	full/partial	obligatory if annotation = yes
pos_tagged		yes/no	optional
pos_tagset	What tagset was used to pos-tag the corpus?	[PROVIDE A LIST]	obligatory if post_tagged = yes
parsed		yes/no	optional
parsing_tool	What parser was used to parse the texts?	[PROVIDE A LIST]	obligatory if parsed = yes
error_annotated		yes/no	optional

error_annotating_tool	What tool was used to error annotate the texts?	[PROVIDE A LIST]	obligatory if error_annotated = yes
annotation_other	Are the texts annotated for anything else?		optional
4. Text metadata			
text_id			obligatory
text_title	Title of the text if any		optional
topic_keywords			optional
date	Date of creation		obligatory
country	In which country was the data collected?		obligatory
institution	In which institution was the data collected?		obligatory
word_count			optional
language			obligatory if L2_target = multi_L2
mode		written/spoken/multimodal	obligatory if corpus_mode = various
written_mode		keyed-in/handwritten	obligatory if corpus_mode/mode = written
written_author_type		single author / multiple author	optional
written_multiple_author_ID	list of learner Ids		obligatory if written_author_type = multiple author
written_process_status	Version of a given text	[version 1/ version 2/version 3]	obligatory if written_process = yes
written_process_ID	Generic ID of the text for which there are several revisions/versions		obligatory if process = yes

task_type	TO BE REVISED/DESCRIBED	in-class activity / examination / home assignment / voluntary exercise / leisure activity / work activity / mixed	obligatory
task_instructions	provide task instructions (e.g. prompt) if available		optional
written_task	[CHECK FOR A MORE COMPLETE LIST (BAWE)]	argumentative essay / literary essay / narrative / letter / research report / term paper / dissertation / other	obligatory if corpus_mode/mode = written
spoken_task	[CHECK FOR A MORE COMPLETE LIST (e.g. task-based learning literature in SLA]	conversation / interview / picture description / role play / other	obligatory if corpus_mode/mode = spoken
multimodal_task	[to do]		obligatory if corpus_mode/mode = multimodal
interaction_type			obligatory if corpus_mode/mode = spoken/multimodal
sound_file	Name of the sound file attached to transcription	[file name]	obligatory if corpus_mode/mode = spoken/multimodal
sound_file_transcriber			obligatory if corpus_mode/mode = spoken/multimodal
timing		timed/untimed	obligatory
timing_duration	duration in minutes		obligatory if timing = timed

written_ref_tools	Was the learner allowed to use reference tools?	yes/no	obligatory
Written_ref_tools_set	Was the choice of reference tools open or set? (if set, provide list in readme/documentation file; if open, answer following [optional] questions)	open/set	obligatory if written_ref_tools = yes
written_ref_tools_mono_dico	Name/s of monolingual dictionary used		optional
written_ref_tools_bil_dico	Name/s of bilingual dictionary used		optional
written_ref_tools_mono_conc	Name/s of monolingual concordancer		optional
written_ref_tools_bil_conc	Name/s of bilingual concordancer		optional
written_ref_tools_checker	Name/s of spell- and grammar-checker (Word, etc.)		optional
written_ref_tools_grammar	Name/s of grammar		optional
written_ref_tools_other	Other tools		optional
task_document	Does the task require the use of other material (e.g. picture to describe, text to summarise)?	yes/no	obligatory

task_document_ID	Link to any supporting document (e.g. picture used in picture description; text used to write a summary)	[file name]	obligatory if task_document = yes
official_language_testing_type	What official language test was the learner taking?	[PROVIDE (open) LIST]	obligatory if official_language_testing = yes
official_language_testing_results	Results of the official language test	[Levels for CEFR; figures for IELTS; etc.]	obligatory if official_language_testing = yes
text_proficiency_rating_scale	What proficiency scale was used to rate the text?	TOEFL/CEFR/in-house	obligatory if proficiency_level_type = text-based
text_proficiency_rating	At what proficiency level was the text rated? (see above: provide assessment grid in readme file/documentation)		obligatory if proficiency_level_type = text-based
proficiency_level_CEFR_conversion	If available, provide conversion of the grades into CEFR (and copy conversion grid in readme file/documentation)	A1/A2/B1/B2/C1/C2	optional
5. Learner metadata			
learner_ID			obligatory
learner_status		L2 learner / trainee translator / L2 user	obligatory
age			obligatory
gender		female /male / non-binary	obligatory
L1	(one L1 or bilingual)	[ISO language list]	obligatory
home_language	Language(s) spoken at home	[ISO language list]	optional
L2_other	Does the learner speak other foreign language/s	yes / no	obligatory
L2_other_1		[ISO language list]	obligatory if L2_other = yes

L2_other_2		[ISO language list]	optional
L2_other_3		[ISO language list]	optional
learner_proficiency_level_type	Was the learner proficiency level self-assessed, externally rated (Cambridge, IELTS, etc) or internally rated (school, institute of higher education)?	no info / self-rated / externally rated / internally-rated	obligatory if proficiency_level_type = learner_based
learner_proficiency_self_rated	Self-rated competence in foreign language	[likert scale? CEFR scale?]	obligatory if learner_proficiency_level_type = self-rated
learner_proficiency_rating_scale	What proficiency scale was used to rate the learner?	TOEFL/CEFR/in-house	obligatory if learner_proficiency_level_type = externally/internally rated
learner_proficiency	What is the proficiency level of the learner?	[results of test/self-rated; draw list]	obligatory if learner_proficiency_level_type = externally/internally rated
learner_level_CEFR_conversion	If available, provide conversion into CEFR	A1/A2/B1/B2/C1/C2	optional
study_level		primary / secondary / undergraduate / graduate / post-graduate	obligatory
study_area	Current study background		optional
occupation			optional
socecStatus	*(socio-economic status) contains an informal description of a person's perceived social or economic status*	low / medium / high	optional

language_instruction_primary	Main language of instruction at primary school	[ISO language list]	optional
language_instruction_secondary	Main language of instruction at secondary school	[ISO language list]	optional
language_instruction_higher_edu	Main language of instruction in higher education	[ISO language list]	optional
L2_study_years	Total number of years studying the L2		obligatory
time_spent_L2_country	Cumulated time spent in a country where L2 is spoken (in months)		obligatory
L2_exposure_written	Exposure to L2 books, magazines, newspapers & websites	virtually never, seldom, sometimes, often, N/A	optional
L2_exposure_oral	Exposure to undubbed L2 films/TV programmes	virtually never, seldom, sometimes, often, N/A	optional
L2_interaction	Interaction with native speakers	virtually never, seldom, sometimes, often, N/A	optional
learner_aptitude	Score - aptitude test	yes/no	optional
aptitude_test_ID	Name of aptitude test		optional
aptitude_test_components	[different variables for different components of aptitude test]		
learner_intelligence	Score - intelligence test	yes/no	optional
intelligence_test_ID	Name of intelligence test		optional
intelligence_test_components	[different variables for different components of intelligence test]		
learner_motivation	Score - motivation test		optional

motivation_test_ID	Name of motivation test		optional
motivation_test_components	[different variables for different components of motivation test]		optional