The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research

Magali Paquot[1], Alexander König[2], Egon W. Stemle[3] and Jennifer-Carmen Frey[3]

[1] FNRS – Centre for English Corpus Linguistics, UCLouvain, [2] CLARIN ERIC,
[3] Institute for Applied Linguistics, Eurac Research

Abstract

Metadata is critical throughout the research process, from study design to corpus selection/compilation, result interpretability and cumulative research. To date, however, learner corpus research has not developed community standards or best practices for metadata collection and sharing. In this article, we present the results of a collaborative project aimed at addressing this issue by developing a standardised metadata schema for learner corpora. We first describe the procedure implemented to design the schema, including the ways in which we continuously involved learner corpus researchers in this initiative. We then introduce the *Core Metadata Schema for Learner Corpora* (version 2), which consists in a set of obligatory and optional variables that encapsulate crucial information about L2 data (administrative details, corpus design, text-related variables, learner-related variables, annotations, annotators, or transcribers). Finally, we discuss future developments and emphasise the importance of continued maintenance and further refinement of this schema by the research community.

Keywords: Learner corpus, metadata, community standards, FAIR principles, data discoverability, corpus compilation, study comparability

1. Introduction

Metadata can broadly be defined as ‘data about data’. When conducting a learner corpus study, metadata is crucial at every single step of the research process, from (1) study design, (2) corpus selection or compilation to (3) data analysis and interpretation of the results. First, during the study design phase, the theoretical relevance of metadata becomes evident. Metadata plays a pivotal role in identifying and defining the independent or dependent variables necessary to formulate research questions (e.g. age, proficiency, experience with studies abroad, knowledge of other languages, task effects, register differences). Second, when researchers aim to address research questions using learner corpus data, their initial step often involves exploring the availability of a learner corpus containing the necessary information. This exploration typically involves determining whether such a corpus exists and if it is accessible for research purposes. Should the existing learner corpora prove inadequate for researchers' needs, necessitating the compilation of new data, corpus compilers will need to decide what variables to collect and how to operationalize them. These decisions are extremely important, and they must be made at the start of a corpus compilation project as it is often difficult, if not impossible, to return to participants to request further information once data collection has taken place (Barker et al., 2015).

Third, during the analysis stage, high-quality metadata fields are equally essential. As put by Burnard (2004),

> it is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible. Why? Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language. A typical corpus analysis will therefore gather many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata restores and specifies that context, thus enabling us to relate the specimen to its original habitat.

Metadata not only enables researchers to effectively discuss and interpret findings within a theoretical framework, but also facilitates the assessment of result comparability and assists in identifying potential confounding variables. Accordingly, metadata plays a crucial role in fostering the progression of cumulative research endeavours.

The importance of metadata has been further heightened in the context of Open Science. Metadata is considered a key element of the FAIR principles, essential for maximising optimal reuse of research data. According to the FAIR principles, data and metadata should be **F**indable (for both humans and computers), **A**ccessible (with clear guidance on how to access the data, possibly including authentication and authorization), **I**nteroperable (with applications and workflows for analysis, storage and processing) and **R**eusable (containing any information needed by other researchers to effectively interpret and disseminate findings from subsequent use of the data) (Wilkinson et al., 2016). To put it differently, "metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable" (Higgins, 2007). Importantly, metadata should also be shared even when associated data cannot be distributed. This is because metadata provides essential context about the data, enabling researchers to understand its significance and potential applications. Additionally, well-described and comprehensive metadata provides insights into the expertise and data collection practices, thereby promoting transparency, reproducibility, and resource optimisation within the research community. To realise all these benefits, metadata should use a formal, accessible, shared and broadly applicable language for knowledge representation and meet domain-relevant community standards.[1]

To date, however, Learner Corpus Research (LCR) has not developed community standards or best practices for data collection, archiving and sharing (cf. König et al., 2021;

[1] https://www.go-fair.org/fair-principles/ (last accessed, 7 March 2024)

Stemle et al., 2019; Volodina et al., 2018). Metadata typically does not adhere to a common template or employ a unified vocabulary. In fact, there are major differences in the availability, quality and quantity of metadata in learner corpora (Barker et al., 2015; Granger & Paquot, 2017). Regrettably, metadata is also not available for a large majority of the learner corpora that remain unpublished. This is problematic, as it represents missed opportunities for cumulative knowledge building, knowledge exchange and collaboration within the LCR community. When learner corpora and/or their associated metadata are published, differences are apparent. These differences may stem from various factors, such as the specific research questions or projects for which different learner corpora were compiled, variations in available resources, and so forth. However, these differences present numerous challenges when selecting, analyzing, or comparing learner corpora. For example, it is common to observe variations in the labels assigned to the same variable (e.g. first language, mother tongue, native language). More problematically, information about the data may be organised differently so that, for example, one corpus may provide information regarding the number of years learners spent studying English, while another may present two separate variables: one for years spent studying English at school and another for years spent studying English at university. A critical issue arises with metadata that is provided with different definitions or operationalizations across corpora. This is, for example, the case for proficiency that can be represented in diverse forms such as institutional status, *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2020) levels assigned to either the learner or the text, official or unofficial exam scores, etc. (Carlsen, 2012). Across different corpora, significant gaps in metadata exist, and in the field more generally, only a limited number of learner corpora include metadata pertaining to individual differences such as cognitive competences or motivation, which nevertheless hold paramount importance for many SLA researchers (e.g. Kerz & Wiechmann, 2020; Li et al., 2022; MacWhinney, 2017).

These challenges have prompted learner corpus researchers to advocate for increased standardisation of metadata in learner corpus research (Gilquin, 2015; Granger & Paquot, 2017; Tracy-Ventura et al., 2021; Volodina et al., 2018). This call is well captured by Stemle et al. (2019), who stated that

> [t]here 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: 6)

One of the earliest efforts to address this widespread call was undertaken by Granger and Paquot (2017), who focused on core metadata, that is a set of variables or elements that are (or should be) common to all learner corpora. They recognized from the onset of the project that the heterogeneity of learner corpus types today would likely necessitate the development of additional modules to describe specific contexts (e.g. translation learner corpora, classroom learner corpora). Granger and Paquot (2017) reviewed several learner corpora to identify commonly used metadata and examined current practices in second language acquisition, corpus linguistics and digital humanities. They developed a first draft of a core metadata schema for learner corpora including about 110 elements, some obligatory and others optional, organised into five components (administrative metadata, corpus design metadata, annotation metadata, text metadata, learner metadata).

This initiative was revived in 2022 in the form of a collaborative project between the Centre for English Corpus Linguistics (UCLouvain, Belgium), the Institute for Applied Linguistics at Eurac Research (Bolzano, Italy) and CLARIN ERIC. More than two years after the beginning of this collaboration, the main objective of this article is to introduce the *Core Metadata Schema for Learner Corpora* (version 2). We hope that this schema fills a gap in the

field, gradually evolving into an established and widely used standard, maintained and enhanced by the collective efforts of our research community. In what follows, we outline the development process of the schema, emphasising its truly collaborative and community-informed nature. We then describe the schema and highlight its main characteristics. Finally, we sketch out the next steps in this collaborative effort and conclude with a few remarks.

2. Development of the *Core Metadata Schema for Learner Corpora*

As previously mentioned, Granger and Paquot (2017) initially reviewed a range of learner corpora to compile a list of commonly recorded variables. They consulted the Learner Corpora around the World webpage[2] to identify learner corpora shared with the research community, noting the types of metadata available in these corpora. Additionally, they also broadened the scope of their exploration to include neighbouring fields such as first language acquisition and corpus linguistics. They consulted the *Talkbank* website[3] and reviewed the obligatory and optional file headers in the CHAT transcription format (e.g. language, age, sex, education, interaction type, transcriber; MacWhinney, 2000). They also examined metadata descriptions in the *Text Encoding Initiative*[4] and the *Corpus Encoding Standard* (Ide, 1998). Drawing from this comprehensive survey, Granger and Paquot proposed an initial draft of a core metadata schema, which was presented at the CLARIN workshop on Interoperability of Second Language Resources and Tools held from 6th to 8th December 2017 at the University of Gothenburg, Sweden. However, despite the potential significance of this project, this work

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last accessed, 7 March 2024)
[3] https://www.talkbank.org/ (last accessed, 7 March 2024)
[4] https://tei-c.org/ (last accessed, 7 March 2024)

remained unpublished, and no further progress was made on the topic in subsequent years, despite calls from the LCR community for revisions and wider dissemination.

Eventually, this initiative was reinvigorated in 2022 when a group of learner corpus researchers attempted to use Granger and Paquot's (2017) original proposal to describe the learner corpora compiled at the Institute for Applied Linguistics, Eurac Research. Although the initial goal was modest, the collaboration quickly evolved into a more ambitious collaborative project aimed at representing the metadata recorded in currently available learner corpora while at the same time laying the ground for future research practices in the field.

A fundamental guiding principle in the development of the *Core Metadata Schema for Learner Corpora* is its collaborative and community-informed nature. From the outset, we have been committed to ensuring that the schema meets the specific needs of the LCR community by actively engaging with its members. We explicitly sought feedback from various related research communities on multiple occasions, which allowed us to enrich the schema with recommendations from learner corpus researchers and second language acquisition specialists in particular. For this purpose, we presented the first version of the schema at the *Learner Corpus Research 2022* conference, University of Padua, Italy, and asked for feedback from the community. The schema was shared online via the Dataverse of UCLouvain, Belgium[5] and we also sent requests for feedback by email via the *Learner Corpus Association* (LCA) mailing list and the *Corpora List*. Feedback was collected in a dedicated online document.

The feedback we received on the first version of the schema was immensely valuable, leading to extensive revisions that continued until Spring 2023. The revisions required by the scientific community largely focused on enhancing the description of individual characteristics of learners as well as the situational characteristics of the collected language samples. Revisions

[5] https://dataverse.uclouvain.be/dataset.xhtml?persistentId=doi:10.14428/DVN/4CDX3P (last accessed, 7 March 2024)

related to the situational characteristics align closely with recent calls for an increased emphasis on registers and tasks in learner corpus research, highlighting the significance of considering register/task in learner corpus studies (e.g., Larsson et al., 2021) (see description of the 'Situational and task characteristics' component in the next section). Revisions related to the individual learner characteristics primarily reflect advancements in the fields of applied linguistics and second language acquisition, most particularly a heightened focus on multilingualism as a central object of inquiry (Ortega, 2019) (see description of the 'learner' component below). For example, several colleagues concurred on the addition of a field called "age of onset of acquisition" to the learner metadata. Recommendations published in other research were also instrumental in this step. In particular, Wulff (2023) reports the results of a survey she conducted to gather colleagues' needs and preferences regarding future L3 and additional language (Ln) corpus compilations. Thirty-one corpus linguists and L3/Ln researchers participated in this survey. When asked about the types of metadata they would like to see included in an ideal corpus, nearly all respondents selected age of onset of acquisition and proficiency. Additionally, more than two thirds of the respondents expressed a desire for what could be referred to as a more comprehensive "multilingual" profile, which would include information on the order of acquisition of acquired languages, the degree of use of all languages from onset to current age, and the contexts of language use for all languages spoken. As will be elaborated in the next section, we have addressed these requests in the revised version of the schema by introducing a set of additional variables to describe Ln differences and the learning context, exposure and proficiency in other languages. Other revisions also reflect a broadening of the target learner populations under investigation (e.g. learners with specific learning difficulties, learners with an impairment; e.g., Kormos, 2020). For example, one colleague recommended the addition of an optional field called "learner_impairment" to accommodate

cases where the learner is deaf or hard of hearing. Another colleague proposed the inclusion of a field to account for forms of atypical learning.

We introduced the revised schema at the *32nd conference of the European Second Language Association* (EUROSLA) held in Birmingham, UK (Frey et al., 2023), and also presented it during an invited talk at the "Linguistics and Applied Linguistics Research" series of online talks organised by the University of Murcia (Paquot, 2023). Like with the first version, we also shared the updated schema via the mailing list of the LCA and the *Corpora List* in October 2023. We collected feedback using the same online document sharing method as before. Additionally, we provided colleagues with the option to contribute comments directly in the metadata schema (online spreadsheet document) or to contact us through the CLARIN Knowledge Centre for Learner Corpora contact email address. This led to a second round of edits and revisions, though this time they were minor, with the community largely approving the changes made. In what follows, we present the results of this collaborative and community-informed project. The full metadata schema is available from https://doi.org/10.14428/DVN/AAUEM2.

3. The Core Metadata Schema for Learner Corpora

The aim of the *Core Metadata Schema for Learner Corpora* (LC-meta) was to offer a comprehensive set of metadata fields that could theoretically be used to describe any type of learner corpus (e.g., corpus mode, age of the learners, target language). Elements solely applicable to specific types of learner corpora (e.g. study level, course level, teacher unique identifier, school, for a learner corpus collected in the classroom) are not included but should be described within separate components yet to be developed.

3.1. General structure of the schema: eight interrelated components

The LC-meta includes 168 elements, which may be obligatory or optional, and may be repeatable or singular. These elements are organised into eight interrelated main components:

1. Administrative metadata
2. Corpus design metadata
3. Learner metadata
4. Text metadata
5. Situational and task-related metadata
6. Annotation metadata
7. Annotator metadata
8. Transcriber metadata

Administrative metadata refers to metadata fields that provide documentary information about the learner corpus itself, including its name, version number, persistent identifier, documentation, and availability. These metadata elements are key to ensuring the adoption of the FAIR principles in LCR (König et al., 2021; Lindström Tiedemann et al., 2018). Importantly, administrative metadata fields such as "corpus_id" and "corpus_cite_as" not only contribute to establishing data management best practices, but they also facilitate proper data citation - a practice where scientists, including corpus linguists, are lagging behind (Brown, 2021).

Corpus design metadata includes metadata fields that serve to provide information about the design of the learner corpus, answering questions such as the following: What is the target language represented in the learner corpus? What are the L1 languages spoken by the learners?

Is the corpus composed of written, spoken, and/or multimodal data? Is the learner corpus longitudinal in nature? Were the language samples included in the corpus transcribed, and if so, are the transcription guidelines available? In what kinds of language production settings were the data collected? Corpus design metadata also provides details about the type of information that was collected by design. This includes recording the language proficiency levels represented in the learner corpus, specifying the approach taken to collect proficiency information (text-centred vs. learner-centred), and detailing the instrument employed for each method. Corpus design metadata also documents the efforts made by the corpus compilers to ensure data privacy protection as well as the nature of information gathered from the participants, indicating for example whether learners completed a motivation, attitude, or aptitude test as part of the data collection process.

Learner metadata provides detailed information about individual learners represented in the corpus. The learner metadata component has been extensively revised from Version 1 to Version 2, most specifically to meet the needs of researchers interested in second language acquisition, third language acquisition and multilingualism (cf. Wulff, 2023). For each learner, it is first recommended that information be collected that focuses on socio-demographic and other individual differences (e.g. age, gender, educational background, socio-economic status, region, environment, impairment, atypical learning, motivation, attitude, aptitude). Other metadata fields serve to describe the multilingual profile of the learner by focusing on (1) individual differences related to the language environment of the learner (e.g. L1 languages(s), L1 language(s) of the parents, language(s) spoken at home, language(s) spoken with friends, language(s) used at kindergarten, primary school, etc.), (2) the learning context, exposure and proficiency of the learner in the target language (e.g. type of learning context, age of onset of acquisition, context of use of the target language from onset to current age, experience with listening, speaking, reading and writing in the target

language, proficiency, months spent in a country where the target language is an official language), and (3) the learning context, exposure and proficiency in any other languages spoken by the learner (in the form of a repeatable set of variables – the same variables as used to describe experience with the target language - for every L2 language spoken by the learner).

Text metadata fields inform about the specificities of each text within a learner corpus, with "text" being used here as an umbrella term for all types of learner language productions. They include links to all associated files such as handwritten text in pdf format, sound files, and videos, information regarding the target language represented in each individual text, the text version number if applicable, time of creation, the proficiency score or level attributed to the text, etc. Additionally, several fields serve to link the text with other components: a reference to a unique identifier for the learner(s) who authored the text, as well as, if relevant, a reference to a unique identifier for the human transcriber who transcribed the text, a reference to a unique identifier for the different annotations provided for the text, and a reference to a unique identifier for the human annotator who provided these annotations.

Situational and task-related variables serve to describe the situational characteristics of a language event and/or the specificities of tasks that were used to generate the learner language samples. Based on Biber and Conrad (2019: 40), metadata fields that characterise the language production situation include the number of addressors and addressees, the personal relationship between them, the presence of on-lookers, the mode and medium of communication, situation circumstances (real time, prepared, (potentially) revised and resubmitted), the broad context (education, work, family, leisure), the communicative purpose (e.g. explanatory, descriptive, persuasive, narrative), the broad topic domain (e.g. domestic, daily activities, science, education, politics), and the register (e.g. essay, summary, translation, creative writing, conversation). In instructed language learning settings, task-related metadata fields provide further information related to the task given to learners to generate language samples. For

example, they include a link to any supporting documents provided as a prompt, information about whether the task was timed or not, whether reference tools were allowed or not, and whether the task was part of an exam or any other type of language assessment setting. While a text typically relates to one specific task or situation, a learner corpus should be described with as many separate sets of situational and task-related variables as there are situations or tasks represented in it.

While the components described above are considered obligatory for the description of all learner corpora, the remaining metadata components are optional, but they are considered essential towards the development of data management best practices in learner corpus research. Annotation metadata is meant to provide detailed information about the types of manual and automatic annotation used in a learner corpus. For each type of annotation provided with a learner corpus, corpus compilers are encouraged to describe the following elements: the type of annotation (manual vs. automatic), the name of the tool used to annotate, the version of the tool, unique identifier(s) for the human annotator(s) who provided manual annotations, and the extent to which the annotations were evaluated and/or corrected. Finally, annotator/transcriber metadata are optional components that contain the same set of five metadata fields that describe each annotator/transcriber with a unique identifier and some details about their L1s and L2s, their proficiency in the target language and their status (whether they are novice or expert annotators, trained or not, etc.). While this is not yet a common practice in learner corpus compilation projects, steps have already been taken to standardise this practice in other research contexts. For example, the CHAT transcription format offers an optional ‘transcriber’ header for two main purposes. First, its inclusion can prove very helpful when questions arise at a later stage. Second, it can also serve as a means of acknowledging the researchers who have dedicated their time to making the data available for further analysis (MacWhinney, 2024: 37).

Components are interrelated, facilitating the representation of different corpus designs and characteristics. Except for administrative metadata and corpus design metadata, components are also repeatable, allowing for the representation of different corpus structures and accommodating the increased metadata complexity observed in the field in recent years. This feature is also intended to maintain flexibility and simplicity within the schema. Thus, each text in a learner corpus is linked to one or more learner metadata components, providing detailed information about the learner(s) who produced the text. Conversely, more than one text can be linked to the same learner profile. If multiple situations or tasks are represented in the learner corpus, each text is also associated with a situation component describing the specific situation or task that served to generate the learner language sample; otherwise, this information is more usefully stored as part of the corpus design metadata. Additionally, a text will also be linked to components representing the different types of annotation available for that text (if there are different types of annotations for texts in a learner corpus). Lastly, a text can also be associated with a component that describes the annotator and/or transcriber who processed that specific language sample.

3.2. Metadata elements: design principles and key characteristics

Each metadata element follows the same organising principles. For every metadata field, there is a corresponding metadata field name accompanied by a short description (e.g. corpus_size_learners: Number of learners; corpus_transcription: Were the language samples included in the corpus transcribed?). Additional columns provide further information about conditional questions, recommended fixed attributes, examples of possible answers, the obligatory status of metadata fields, and their repeatability. More specifically, the "condition" column serves to identify metadata fields that should only be completed if another metadata

field was answered in a specific manner (for example, “learner_impairment_details” should only be filled in if a ‘yes’ response is provided to the preceding question (learner_impairment: Does the learner have an impairment?). In our future efforts to develop a user-friendly interface for collecting learner corpus metadata, this column will also be used to configure conditional questions, directing respondents to different sections of the survey based on their answers to specific questions.

The “fixed attributes” column presents a list of predetermined terms for metadata that can be described with a closed vocabulary. For example, the metadata field “text_proficiency_CEFR_conversion” is intended to capture the conversion of text proficiency scores into CEFR levels. Consequently, the field should only be populated with the CEFR levels A1, A2, B1, B2, C1 and C2. This feature will also prove beneficial when developing an online data collection interface, enabling the use of closed questions. Next, the “further details and examples” column provides further information about the metadata fields as well as examples. For example, to account for the variety of registers potentially represented in corpus data, we opted for a double representation by means of an obligatory “situation_register” field with a fixed set of attributes representing broad registers typically found in learner corpora (essay, summary, translation, creative writing, conversation, monologue, other) and an optional and open “situation_register_specific” field where users can use their own register categorization, which we exemplified with terms such as letter of application, MA dissertation, interview, argumentative essay, text messaging, etc. in the “further details and examples” column. Importantly, while we made an effort to select variable names and fixed attributes that are commonly used in the research community, we also refrained from providing closed definitions for all these elements at this stage. This decision stems from the lack of widely agreed-upon definitions for terms such as L1, L2, and proficiency. For example, L1 can be understood as the language(s) that a person learns from birth or infancy or the dominant language(s). As we

perceive it, the metadata schema should be complemented with detailed corpus documentation providing more contextual information about the corpus compilation project. This should include elements such as definitions for the terms used, questionnaires used to collect learner variables, and further details about the assessment methods.

Similarly, it was often challenging for us to determine whether to use fixed attributes for specific metadata fields or keep them open-ended. On one hand, adhering to FAIR principles necessitates the use of a controlled vocabulary, enabling others, both humans and machines, to locate, access, interoperate, and reuse datasets. This underscores the importance of incorporating fixed attributes whenever feasible. On the other hand, for many metadata fields, we felt time was not ripe to set fixed attributes. Particularly for optional learner-related metadata fields concerning the learning context, language exposure and proficiency, we encountered difficulty identifying sets of fixed attributes that could be widely accepted by the research community, given the current state of available learner corpora or L2 research literature. In such cases, we opted to let researchers decide what attributes they wanted to use. While we provide a recommended list of attributes in the "further details and examples" column whenever possible (e.g., for describing listening, speaking, reading, and writing experience), we believe it is crucial for researchers to retain the freedom to select attributes that best suit their requirements. Therefore, for many of the less established variables, we followed Burnard's principle, emphasising the "need for standards which do not limit choice, but rather facilitate an accurate presentation of the choices made" (Burnard, 2017). We anticipate that as these metadata fields gain more usage and preferences emerge, the LCR community will eventually be able to further standardise their utilisation in a future version of the LC-meta, although this process will likely require some time. In the meantime, it is advisable to provide comprehensive descriptions of attributes used in open elements in the corpus documentation.

Future users of the LC-meta may initially find themselves overwhelmed by the number of elements that can potentially be recorded. It should be emphasised however that many of these variables will only require recording once by the corpus compilers (e.g., administrative metadata, corpus design metadata, situational and task-oriented metadata). Additionally, it is important to highlight that only a limited number of variables are obligatory (for example, 15 out of 23 administrative variables are required and only 8 out of the 52 learner variables are strictly obligatory), and many others are optional. The obligatory metadata fields represent what we consider a minimal set of elements that are required to describe a learner corpus. They include, for example, learner variables such as learner age at text production, gender, target language, L1s, type of language context (instructed, immersion or naturalistic), and situational and task characteristics such as mode, medium of communication, situation settings (personal, public, occupational, educational), communicative purpose, general topic domain and broad register.

It is imperative for the field to establish a common set of obligatory elements that are systematically collected in all learner corpus compilation projects. This consensus is crucial for several reasons as discussed above, from adhering to the FAIR principles essential in scientific research to ensuring comparability across studies and facilitating cumulative research efforts. By agreeing upon this 'hard core' of obligatory elements, the field can also provide valuable guidance to newcomers embarking on a learner corpus compilation project. This standardised approach will alleviate the burden of starting from scratch and enhance the value of learner corpora for the research community. Ultimately, a crucial objective of any learner corpus compilation initiative should be to share carefully compiled and enriched data with the community. As briefly mentioned above, even if learner corpora cannot be shared, it should nevertheless remain a priority to share the associated metadata. Optional fields, by contrast, can be used to provide additional information about a range of learner corpus characteristics.

Although they are not mandatory, we have included them for two primary reasons. First, despite their optional nature, many fields could benefit from a standardised defining vocabulary and operationalization. For example, while learners' educational background may not be relevant in all learner corpus compilation projects, it may be beneficial to define the variable uniformly and use consistent terminology and categories across studies that collect it. Second, some of these optional fields align with what researchers in the field including ourselves consider desiderata for LCR (e.g. Tracy-Ventura et al., 2021; Wulff, 2023). Therefore, the metadata schema can also help learner corpus compilers in identifying relevant variables to enhance the value of their learner data for the broader research community. Essentially, it can serve as a guide for learner corpus compilers to understand the expectations and preferences of the field regarding future learner corpora.

The "repeatable" column in the schema identifies metadata fields that can be repeated. This feature proves particularly useful for the treatment of first languages and other known languages. For example, the variable "learner_L1" is repeatable to account for the fact that learners may consider more than one language as their L1s. Similarly, the group of variables that serve to describe the context of use, exposure, and proficiency in other languages can be repeated to describe all known languages in detail, thus providing a more comprehensive account of the potentially multilingual environment of today's learners of second or foreign languages.

4. Future developments

In the short to medium term, there are two main avenues for further development. First, the metadata schema has so far been released exclusively in the form of a spreadsheet. To maximise its utility, the LC-meta should first be made available in more appropriate formats. A first step

in this direction will be to convert the schema into a format that will break free from the inherent limitations of a spreadsheet and make its inherent structure more explicit. This transformation will entail several tasks. First, it will involve identifying the relationships between different components (e.g. recognizing that a text can be linked to various obligatory or optional components). Second, it will necessitate the formal specification of dependencies between variables, most particularly establishing links between conditional variables and their associated variables. For example, if the corpus_text_proficiency_assignment_method variable under the "corpus design" component indicates that proficiency information is available at the text level, then the text_proficiency variable should become obligatory in the "text" component. Finally, the schema should facilitate the straightforward representation of repeatable variables, grouped variables (e.g. all the variables describing L2 languages known should be repeated as a block) and components (e.g. a text could be linked to one or more "learner" components as well as one or more "annotation" components; an "annotation" component may be linked to one or more "annotator" components). While working on data representation, updates to the schema will probably be necessary to better account for more complex types of learner corpora such as longitudinal corpora and multilingual corpora, which are challenging to represent with the current set of variables.

Such a format should be understandable and usable by both humans and machines. For computers, this entails ensuring maximum data consistency and compatibility with other formats such as CLARIN's *Component Metadata Infrastructure* (CMDI, Windhouwer & Goosen, 2022)). We consider it particularly important to convert the schema into CMDI to enable its integration with the CLARIN infrastructure. Implementing a schema that better caters to the needs of L2 researchers within CLARIN should offer several advantages. These include, for example, facilitating updates of the *L2 Learner Corpora Resource Family* by allowing for the incorporation of new data and revisions to existing datasets and enabling more

comprehensive exploration of this resource family (cf. Lindström Tiedemann et al., 2018 on the limited coverage of existing learner corpora in CLARIN and the need for the CLARIN *Virtual Language Observatory* (VLO) to take into account the specific needs of L2 researchers). More generally, as argued by Lehmbert and Wörner (2008), "the extent to which established standards are used is determined by the availability of tools that work with this standardised data and vice versa" (Lehmbert & Wörner, 2008: 484). In line with this, to assist future corpus compilation projects, we also aim to progress towards the release of a survey created using an open-source survey maker tool such as *LimeSurvey*. This survey will be designed for the purpose of collecting core metadata directly from learners. In the long term, however, there may be a need for an online data collection interface that provides tailored access to various types of users, including corpus compilers, teachers, and learners. This interface could be modelled after projects like the *Multilingual Student Translation* (MUST) learner corpus (Granger & Lefer, 2020).

Second, there is a need to develop and integrate additional (optional) modules into the schema to accommodate the complete description of specific types of learner corpora. The LC-meta does not list all the many variables necessary to fully describe specific learner corpora such as multimodal learner corpora, learner translation corpora, learner corpora collected in the classroom, content and language integrated learning (CLIL) data, etc. The development of additional modules will however require specialised expertise from colleagues who have compiled such corpora and are actively using them. Ideally, these researchers would collaborate to develop their own modules and submit them to a team of LCR experts responsible for maintaining the schema for feedback and approval.

Indeed, our aspiration is for the LC-meta to become a shared resource within the learner corpus research community, maintained and further developed under the auspices of internationally recognized associations or institutions. As argued by Higgins (2007),

"[m]etadata 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). Ideally, we would collaborate closely with the newly recognized CLARIN *Knowledge Centre for Learner Corpora*[6] hosted by UCLouvain and the LCA. This collaboration aligns with the missions that the CLARIN K-centre for learner corpora could fulfil, and we also envision a prominent role for the LCA, potentially through the establishment of a working group dedicated to metadata. LCA members with an interest in metadata could join this working group and collaborate on the maintenance and further developments of the schema. One first task for this working group could be to evaluate the current representation of the various languages known within the "learner" component of the LC-meta. In version 2, substantial progress has been made towards a more comprehensive representation of a learner's multilingual profile, while still retaining a widely used representation of known languages in terms of L1(s) and L2(s). We anticipate that future revisions of the schema may necessitate reconsidering this representation, perhaps removing the current distinction between L1(s) and L2(s) altogether.

5. Conclusion

LCR has experienced rapid growth, yet it remains an area in dire need of standardisation. Critical aspects of conducting a learner corpus study are still in need of community-based standards today, spanning from the collection of metadata during corpus compilation to the establishment of consistent data formats and transcription guidelines. Additionally, the methods employed in annotating learner corpora, particularly concerning error annotation, require

[6] https://uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora.html

careful consideration and alignment (e.g. Stemle et al., 2019; Tracy-Ventura et al., 2021). Implementing standardised practices in these areas should not only enhance the rigour and reproducibility of research but also foster collaboration and facilitate cross-study comparisons within the field. As argued by MacWhinney, it is essential to "consider ways of collecting new corpora that can maximise the possibility of making systematic comparisons across learner levels, educational approaches, elicitation methods, and languages" (MacWhinney, 2017: 259).

With the development of the LC-meta, we have sought to address the issue of the lack of community-based standards for metadata. Standardised metadata should contribute to developments in learner corpus research in several ways. First, standardised metadata should contribute to making learner corpus research more FAIR (Findable, Accessible, Interoperable, and Reusable). Second, standardised metadata holds the potential to elevate the overall quality of learner corpus research by providing researchers with the necessary variables to accurately interpret their results. Alongside detailed corpus documentation, consistent and comprehensive metadata can assist researchers in effectively controlling for confounding variables. This, in turn, should enhance the reliability and validity of their findings. Third, standardised metadata can help the field to contribute more to cumulative research. As argued by Stemle et al. (2019),

> [l]ooking 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: 442)

Finally, the development of a standardised metadata schema is also driven by our wish to contribute to the development of best practices for learner corpus compilation and sharing. Corpus design is an essential task in LCR, requiring a high level of expertise. However, despite

its significance, the field currently lacks comprehensive guidelines to inform and support corpus designers. As a result, many learner corpus compilers find themselves navigating new projects without a standardised framework or clear direction. By establishing community-driven guidelines for learner corpus compilation, we can streamline the process, promote consistency, and leverage collective expertise to advance the field more effectively.

Standards, however, are not imposed on a community. Rather, they emerge and evolve through collaborative efforts within the community itself. This principle has been fundamental to our approach in developing the LC-meta. From the outset, our objective has been to foster a collaborative environment where the LCR community actively contributes to shaping the standards to meet the specific needs of the field. We have made deliberate efforts to engage the community and solicit feedback at every stage of the project. This iterative approach has allowed us to incorporate valuable insights and suggestions, refining the schema to better serve the needs of the community. Through these efforts, we have sought to ensure that the metadata schema reflects the diverse perspectives, requirements, and expertise within the LCR community. We have also aimed to develop a metadata schema that not only meets the immediate needs of the LCR community but also lays the foundation for future advancements in learner corpus research.

We are hopeful that the LCR community will start using the LC-meta to describe existing corpora. More importantly, perhaps, we encourage its use in new learner corpus compilation projects. We acknowledge that the uptake of the schema may remain challenging in the absence of appropriate tools (e.g. dedicated questionnaires, data collection interfaces, search interfaces), but we are fully committed to addressing this challenge through our ongoing efforts. In the meantime, our sincere hope is for the community to actively contribute to maintaining and refining the schema. The metadata schema is not intended to remain static. In the future, it will necessarily have to evolve to allow for innovation and change in the field. We

firmly believe that the research community should take an active role in deciding which new metadata elements are included and which are not, which variables are renamed or revised, etc. We advocate for this task to be ideally carried out through the establishment of a dedicated task force under the auspices of the LCA and in collaboration with the CLARIN Knowledge Centre for Learner Corpora.

Finally, we trust that this collaborative effort will also serve as inspiration for other community members to tackle the challenge of standardisation and documentation across various domains of learner corpus research, and corpus research at large.

Open Material badge

This article has been awarded an Open Material badge. The Core Metadata Schema for Learner Corpora is publicly accessible via IRIS at https://www.iris-database.org/details/0Vyia-z4Hgj. It is also available from https://doi.org/10.14428/DVN/AAUEM2. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki

Acknowledgement

We thank the many colleagues who provided feedback on different versions of the *Core Metadata Schema for Learner Corpora*: Ilia Afanasev, Piotr Banski, Alessia Battisti, Alexandru Craevschi, Elisa Di Nuovo, Seda A. Eickhoff, Cristina Grisot, Shin Ishikawa, Ilmari Ivaska, Stephen Jeaco, Mélissa Juillet, Nannan Liu, Surangika Ranathunga, Anita Thomas, Stefania Spina, and Elena Volodina. Special thanks go to Stefanie Wulff who sent us a draft version of a manuscript (Wulff, 2023) that was instrumental in shaping version 2.

References

Barker, F., Salamoura, A. & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 511-533). Cambridge University Press.

Biber, D. & S. Conrad (2019), *Register, genre and style*. Cambridge University Press.

Brown, R. (2021). The importance of data citation. *BioScience*, 71(3), 211. https://doi.org/10.1093/biosci/biab012

Burnard, L. (2004). Developing linguistic corpora: a guide to good practice. Metadata for corpus work. https://users.ox.ac.uk/~martinw/dlc/chapter3.htm .

Carlsen, C. (2012). Proficiency level—A fuzzy variable in computer learner corpora. *Applied Linguistics*, 33(2), 161-183. https://doi.org/10.1093/applin/amr047

Council of Europe (2020), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr.

Frey, J.-C., König, A., Stemle, E. & Paquot, M. (2023). Core Metadata Schema for L2 data [Conference presentation]. 32nd Conference of the European Second Language Association (EUROSLA), 30 August – 2 September 2023, University of Birmingham, UK.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 10-34). Cambridge University Press.

Granger, S. & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation*, 54: 1183-1199. https://rdcu.be/b0MkU

Granger, S. & Paquot, M. (2017). Towards standardization of metadata for L2 corpora. Invited talk at the CLARIN workshop on Interoperability of Second Language Resources and

Tools, 6-8 December 2017, University of Gothenburg, Sweden. https://sweclarin.se/sites/sweclarin.se/files/event_atachements/Granger_Paquot_Metadata_G%C3%B6teborg_final.pdf.

Higgins, S. (2007). What are metadata standards? Digital Curation Centre. Standards Watch Papers. http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards.

Ide, N. (1998). Encoding linguistic corpora. *Sixth Workshop on Very Large Corpora* (pp. 9-17). https://aclanthology.org/W98-1102.

Kerz, E. & Wiechmann, D. (2020). Individual differences. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 394-406). Routledge.

König, A., Frey, J.-C. & Stemle, E. (2021). Exploring reusability and reproducibility for a research infrastructure for L1 and L2 learner corpora. *Information* 12(5): 199, https://doi.org/10.3390/info12050199

Kormos, J. (2020). Specific learning difficulties in second language learning and teaching. *Language Teaching*, 53(2), 129-143. https://doi.org/10.1017/S0261444819000442

Larsson, T., Paquot, M., & Biber, D. (2021). On the importance of register in learner writing: A multi-dimensional approach. In E. Seoane & D. Biber (Eds.), *Corpus based approaches to register variation* (pp. 235-258). Benjamins.

Lehmberg, T. & Wörner, K. (2008). Annotation standards. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics – An international handbook* (volume 1) (pp. 484-501). Walter de Gruyter.

Li, S., Hiver, P., & Papi, M. (2022). Individual differences in second language acquisition: Theory, research, and practice. In S. Li, P. Hiver & M. Papi (Eds.), *The Routledge*

handbook of second language acquisition and individual differences (pp. 3-33). Routledge.

Lindström Tiedemann, T., Lenardič, J., & Fišer, D. (2018). L2 learner corpus survey: Towards improved verifiability, reproducibility and inspiration in learner corpus research. *Proceedings of CLARIN Annual Conference 2018*, Pisa, Italy, 8-10 October 2018, pp. 146-150. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, 67(S1), 254-275. https://doi.org/10.1111/lang.12220

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd edition). Lawrence Erlbaum Associates.

MacWhinney, B. (2024). Tools for analyzing talk. Part 1: The CHAT transcription format. https://doi.org/10.21415/3mhn-0z89

Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal*, 103, 23-38.

Paquot, M. (2023). The Core Metadata Schema for L2 data: Collaborative efforts towards improved data findability, metadata quality and study comparability in L2 research. "Corpus Linguistics and Applied Linguistics Research" series of online talks, Universidad de Murcia, Spain, 30 October 2023. https://www.youtube.com/watch?v=jvnVT40qLcw

Stemle, E. W., Boyd, A., Janssen, M., Tiedemann, T. L., Preradovic, N. M., Rosen, A., Rosén, D., & Volodina, E. (2019). Working together towards an ideal infrastructure for language learner corpora. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.), *Widening the scope of learner corpus research. Selected papers from the fourth Learner Corpus*

Research Conference (pp. 427-468). Corpora and Language in Use – Proceedings 5, Presses Universitaires de Louvain.

Tracy-Ventura, N., Paquot, M. & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds), *The Routledge handbook of second language acquisition and corpora* (pp. 409-424). Routledge.

Volodina, E., Janssen, M., Lindström Tiedemann, T., Mikelic Preradovic, N., Ragnhildstveit, S., Tenfjord, K., & de Smedt, K. (2018). Interoperability of second language resources and tools. *Proceedings of the CLARIN annual conference 2018*, Pisa, Italy, 8-10 October 2018, 90-94. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Windhouwer, M. & Goosen, T. (2022). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *CLARIN: The infrastructure for language resources* (pp. 191-222). De Gruyter. https://doi.org/10.1515/9783110767377-008

Wulff, S. (2023). Corpus research. In J. Cabrelli, A. Chaouch-Orozco, J. González Alonso, S. Pereira Soares, E. Puig-Mayenco, & J. Rothman (Eds.), *The Cambridge handbook of third language acquisition* (pp. 683-695). Cambridge University Press.

Address for correspondence

Magali Paquot

UCLouvain

ILC

Place Cardinal Mercier 31/L3.03.33

1348 Louvain-la-Neuve

Belgium

Magali.paquot@uclouvain.be

https://orcid.org/0000-0001-5687-5074

Co-author information

Egon Stemle

Institute for Applied Linguistics, Eurac Research

Egon.Stemle@eurac.edu

https://orcid.org/0000-0002-7655-5526

Alexander König

CLARIN ERIC

alex@clarin.eu

https://orcid.org/0000-0002-8540-2396

Jennifer-Carmen Frey

Institute for Applied Linguistics, Eurac Research

JenniferCarmen.Frey@eurac.edu

https://orcid.org/0000-0002-7008-6394

Publication history

Date received: 11 April 2024

Date accepted: 12 June 2024