# The Core Metadata Schema for Learner Corpora: developing user-friendly resources and addressing key challenges

**Magali Paquot,[1] Jennifer-Carmen Frey,[2] Alexander König,[3] Herbert Lange,[4] Hubert Naets,[5] Egon Stemle,[6] Julien Zakhia Doueihi[7]**

[1] CKL2CORPORA, CECL, FNRS – UCLouvain, Belgium, magali.paquot@uclouvain.be

[2] Eurac Research, Italy, jennifercarmen.frey@eurac.edu

[3] CLARIN ERIC, The Netherlands, alex@clarin.eu

[4] University of Gothenburg, Sweden, herbert.lange@svenska.gu.se

[5] CKL2CORPORA, CENTAL – UCLouvain, Belgium, hubert.naets@uclouvain.be

[6] Eurac Research, Italy, egon.stemle@eurac.edu

[7] CKL2CORPORA, CENTAL – UCLouvain, Belgium, julien.zakhia@uclouvain.be

**Keywords: metadata, FAIR, resources**

In 2024, Paquot et al. introduced the *Core Metadata Schema for Learner Corpora* (LC-meta) in response to repeated calls for increased metadata standardization in learner corpus research (LCR). LC-meta comprises a set of obligatory and optional variables designed to describe any type of learner corpus. These variables are structured into eight interrelated components: administrative details, corpus design, learner-related variables, text-related variables, situational and task-related metadata, and where applicable, annotation, annotator, and transcriber metadata.

Since its release, LC-meta has attracted considerable attention, with several teams adopting it in learner corpus compilation projects. Examples include the *Young German Learner English* (YGLE) corpus (https://www.ygle.de/home; Werner et al., 2024; Pahor de Maiti Tekavčič et al., 2025), DAKODA corpora (https://project.dakoda.org/) and the *STAR* corpus (De Cock et al., 2025). Building on this growing uptake and sustained community engagement, a dedicated working group has been established to maintain and further develop the schema. Operating under the auspices of the Learner Corpus Association and the CLARIN K-Centre for Learner Corpora, the group is responsible for integrating community feedback, aligning LC-meta with evolving best practices in L2 research, and coordinating dissemination and training activities.

The main objective of this presentation is to report on the work undertaken by a subset of members of the LC-meta Working Group, focusing specifically on the development of supporting resources. A key challenge for ensuring wider community uptake lies in the need for user-friendly tools that assist learner corpus researchers in both collecting new data and describing learner corpora in a standardized way. To address this challenge, we are currently engaged in two priority initiatives.

The first focuses on the development of an online questionnaire intended to support the systematic and standardized collection of learner-related variables at the data collection stage. This involves translating metadata variables into questions addressed to learners, raising issues related to terminology

and the operationalization of learner demographics—for example, whether a single formulation can be recommended or whether alternative formulations should be proposed depending on learners' age and proficiency—as well as technical challenges stemming from the complexity of relationships between individual questions and groups of questions.

The second initiative concerns the definition of metadata format representations, together with the tools and converters needed to support them. The aim is to ensure that LC-meta is used to describe and represent learner corpora in a standardized manner aligned with the FAIR principles, thereby enabling the (meta)data to be easily shared, integrated into existing research infrastructures, and reused across projects, supporting interoperability and long-term sustainability. At present, we are translating the LC-meta schema into CMDI with a view to facilitating its integration into CLARIN-based infrastructures. In parallel, we are working on an additional representation based on linked data, which offers greater flexibility, particularly for modelling complex or evolving metadata.

Both initiatives raise additional challenges that we hope to discuss with the LCR community, in particular the absence of dedicated funding and personnel to support this work. As the work is currently carried out on a voluntary basis, the time available for technical development and long-term maintenance is inevitably limited. Nonetheless, we consider these developments essential for enabling the community to fully adopt and use LC-meta and, more broadly, for fostering an interoperable ecosystem that supports the collection, documentation, sharing, and long-term reuse of learner corpus data across projects and research infrastructures.

**References**


De Cock, S., Egbert, J., Gilquin, G., Granger, S., Grixoni, F., Holmberg, A.J., Jadoulle, P., Larsson, T. & Paquot, M. (2025, July 7-8). *The STAR corpus: Student speech and writing across registers*.[Conference presentation]. Register and task variation in Learner Corpus Research (VAR4LCR) conference, UCLouvain, Belgium.

Pahor de Maiti Tekavčič, K, Paquot, M. & Fuchs, R. (2025, March 26). An interview with Magali Paquot and Robert Fuchs. *Tour de CLARIN*. https://www.clarin.eu/blog/tour-de-clarin-interview-magali-paquot-and-robert-fuchs

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024a). The Core Metadata Schema for Learner Corpora: Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research* 10(2), 280–300. https://doi.org/10.1075/ijlcr.24010.paq

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024b). "Core Metadata Schema for Learner Corpora (version 2)", https://doi.org/10.14428/DVN/AAUEM2, Open Data @ UCLouvain, UNF:6:D46/69S0DuhuxwMnT7rn9A== [fileUNF]

Werner, V., Fuchs, R., Rosen, A., Bracke, L., & Stoddard, B. (2024, September 26-28). *Introducing the Corpus of Young German Learner English* [Conference Presentation]. Learner Corpus Research (LCR) Conference, Tartu, Estonia. https://drive.google.com/file/d/17CUTTZ6BxzbNzR9aPNqjZPloOOEobMIM/view