Originally published in: Alfter, David/Volodina, Elena/François, Thomas/Desmet, Piet/Cornillie, Frederik/Jönsson, Arne/Rennes, Evelina (Eds.): Proceedings of the 11th workshop on natural language processing for computer-assisted language learning (NLP4CALL 2022). – Linköping: LiU Electronic Press, 2022. Pp. 108-113. (Linköping Electronic Conference Proceedings 190; NEALT Proceedings Series 47)
DOI: *https://doi.org/10.3384/ecp190*

Metadata Formats for Learner Corpora: Case Study and Discussion

Herbert Lange
IDS Mannheim
lange@ids-mannheim.de

Abstract

Metadata provides important information relevant both to finding and understanding corpus data. Meaningful linguistic data requires both reasonable annotations and documentation of these annotations. This documentation is part of the metadata of a dataset. While corpus documentation has often been provided in the form of accompanying publications, machine-readable metadata, both containing the bibliographic information and documenting the corpus data, has many advantages. Metadata standards allow for the development of common tools and interfaces. In this paper I want to add a new perspective from an archive's point of view and look at the metadata provided for four learner corpora and discuss the suitability of established standards for machine-readable metadata. I am are aware that there is ongoing work towards metadata standards for learner corpora. However, I would like to keep the discussion going and add another point of view: increasing findability and reusability of learner corpora in an archiving context.

1 Introduction

Research data, including linguistic corpus data, usually is not just published as-is, but instead is enriched with so-called metadata. Metadata subsumes a wide range of additional information. Two main functions of metadata are to allow the data to be found and also to be understood by giving additional context.

For researchers the first point might seem more obvious and relevant. If someone publishes data, they typically want other people to be able to find this data. This is accomplished by providing bibliographic or catalog metadata. This kind of metadata can be used in repositories and registries to be able to provide relevant data to a user. Within the CLARIN infrastructure, the Virtual Language Observatory (VLO) (Goosen and Eckart, 2014) provides such a registry harvesting metadata from a wide range of repositories and providing a uniform interface to look for corpus data based on the provided metadata.

But findability is only one of the important aspects. There is also a growing interest in making data reusable. A very vocal initiative promoting this among other values is the FAIR initiative (Wilkinson et al., 2016). FAIR stands for Findable, Accessible, Interoperable, and Reusable and is connected to the Linked Open Data (LOD) movement. Linking various forms of data together enriches its value for future research. Suitable metadata can provide suitable linking.

2 Background: Established Metadata Standards for Corpora

There exist many formats used to provide metadata. They vary in expressively and their use can also depend on the file format used for the corpus data itself. Instead of covering many different formats I will focus on three formats that seem most relevant for learner corpora available in public archives.

2.1 CMDI

The Component Metadata Initiative (CMDI, Broeder et al., 2011) is the metadata standard established within the CLARIN infrastructure. It is used in the CLARIN VLO to find corpus data. Using standardized interfaces such as OAI-PMH[1], it can be automatically harvested from the repository providing the data. As a modular format, researchers

[1] http://www.openarchives.org/OAI/openarchivesprotocol.html

Herbert Lange. Metadata Formats for Learner Corpora: Case Study and Discussion. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022).* Linköping Electronic Conference Proceedings 190: 108–113.

can define profiles matching their data and annotations. It is a very powerful standard which already with a basic profile provides catalog metadata as well as information about the file structure of the corpus.

2.2 Coma

The EXMARaLDA Corpus Manager (Coma) metadata format is often used in combination with EXMARaLDA Partitur-Editor (Schmidt and Wörner, 2014) annotations for audiovisual data. It can contain catalog metadata compatible with Dublin Core. Furthermore, it is designed to provide information about the corpus structure as well as information about the speakers and events. The documentation states: "Coma is [...] used for managing the relation of metadata, transcriptions, recordings, external annotations, and further related files, tying all related data together into a single corpus document." (Schmidt and Wörner, 2014, p. 413) This format can be especially relevant for spoken learner data.

2.3 TEI Header

Another common metadata format is TEI headers. Not a stand-alone format as the other formats, it is a standard for header information to be included in corpus data encoded following the guidelines of the Text Encoding Initiative (TEI, TEI Consortium, 2022). It can contain five main parts:

- a file description containing the bibliographic or catalog information
- an encoding description describing the relationship between an electronic text and its source or sources
- a text profile containing classification and contextual information about the text, such as its subject matter, the situation in which it was produced, the individuals described by or participating in producing it, and so forth
- a container element for other metadata formats allowing easy inclusion of metadata from non-TEI schemes
- a revision history providing a history of changes made during the development of the electronic text

Depending on the application, a TEI header can be quite a simple or a very complex and structured object. Because TEI is more dominant for written data, TEI headers are more relevant for corpora containing written learner data, but it should be noted that the TEI guidelines also cover transcription of spoken language which would make TEI headers also relevant for spoken learner data.

3 Case Study Learner Corpora

To evaluate the current situation of metadata provided for learner data, I selected four corpora out of the large collection of available datasets. Three of these corpora, DISKO, MIKO, and HMAT, are hosted at the IDS, either in the IDS repository[2] or as part of the database of spoken German (DGD, Schmidt, 2017)[3] and thus relevant for all archiving efforts at the IDS. The fourth corpus, SweLL, was selected to include a dataset that is not hosted in-house at the IDS. The selected corpora cover both written and spoken data.

The most relevant aspect of this case study is the metadata formats used. As shown in the overview of metadata formats, the choice of a metadata format is also influenced both by the annotation tools used and the repository hosting the data. Thus, this information is also summarized for each of the datasets.

3.1 SweLL

The Swedish Learner Language corpus (SweLL, Volodina et al., 2019) consists of two sub-corpora, SweLL-pilot and SweLL-gold. Both are collections of written learner essays. The learners are adults learning Swedish. The pilot corpus has been anonynimized and graded according to CEFR levels, the gold corpus has been pseudonymized, normalized and correction annotated. The annotations, sucha as normaliza-tion/correction annotation and pseudonymization, have been created using the SVALA annotation tool (Wirén et al., 2019) and are available in a plain text format and as JSON. Export to XML is possible.

The metadata description is available in human-readable form as Markdown and PDF

[2] https://repos.ids-mannheim.de/
[3] https://dgd.ids-mannheim.de/

following the guidelines by Granger and Paquot (2017b). In addition, learner metadata as well as statistics about pseudonymization and correction labels for the gold corpus are provided as MS Excel spreadsheets. SweLL is hosted at the Swedish Language Bank (Språkbanken Text)[4].

3.2 HaMaTaC

The Hamburg Mapping Task Corpus (HaMaTac, HZSK, 2010) is a spoken learner corpus with elicited speech data using a map task and involves multilingual speakers learning German. The recordings have been transcribed using the EXMARaLDA Partitur-Editor. Manual annotations include disfluency and phonetic phenomena. Part-of-speech tags using a modified STTS tag set (Schiller et al., 1999) as well as lemmatized forms have been added automatically using TreeTagger.

Metadata is provided using the Coma format and additional speaker metadata is present as headers in the transcription files. The Coma file covers catalog metadata following Dublin Core as well as transcription and annotation metadata including annotation structure. The corpus is available both via the Hamburg Center for Language Corpora (HZSK)[5] and as part of the Database of Spoken German (DGD). The HZSK is part of the CLARIN infrastructure, consequently some metadata are also available as CMDI. In addition to machine-readable metadata, corpus documentation is present as PDFs.

3.3 MIKO

The "Mitschreiben in Vorlesungen: Ein multimodales Lehr-Lernkorpus" corpus (MIKO, Spiegel et al., 2022) is a multimodal corpus containing recordings of lectures as well as lecture notes created by students, both L1 and L2 speakers of German. Most of the lectures are transcribed and annotated using EXMARaLDA and stored as machine-readable data. The lecture notes are based on photos of the notes which have been anonymized and stored as PDFs.

Coma metadata is included in the corpus to document speaker information. Additional metadata about both lectures and lecture notes are included as CSV tables. Finally, human-readable corpus documentation as well as description of the metadata variables is included as PDFs. MIKO is also available as part of the DGD. Furthermore, MIKO is present in the IDS repository which is part of the CLARIN infrastructure and thus requires some CMDI metadata.

3.4 DISKO

Finally, the "Deutsch im Studium: Lernerkorpus" corpus (DISKO, Wisniewski et al., 2022) is a written learner corpus consisting of several subcorpora. It was created in the context of the same project as MIKO and shares some similarities. The two main corpora consist of texts created for a writing exercise repeated up to three times with one year intervals by both L1 and L2 speakers of German. Additional corpora are based on language tests for students. Unusual for a written corpus, annotations have been created using an extension of the EXMARaLDA Partitur-Editor. Besides the EXMARaLDA files the data is also available as plain text and ANNIS data as well as the original handwritten documents as PDFs.

For the main parts DISKO L1 and L2 the metadata contain extensive information about the participants including language and socioeconomic background. For the other subcorpora a limited set of metadata is available. Despite the use of EXMARaLDA, no Coma data is present, but the transcription files contain extensive information in the file headers. Also, similar to MIKO, metadata is present as CSV files and documentation of both the corpus itself and the metadata is available as PDFs. DISKO is available in the IDS repository and consequently requires some metadata available as CMDI.

4 Discussion

As one can see from these datasets listed in Section 3, both the metadata formats used and the information included are quite diverse. That shows that we are quite a bit away from an ideal of a single machine-readable metadata standard for learner corpora.

Several good reasons can be listed both in favor of expressive machine-readable metadata

[4] https://spraakbanken.gu.se/en/projects/swell
[5] https://corpora.uni-hamburg.de/hzsk/en/

for (learner) corpora and against it. One reason against the enforcement of metadata standards, e.g., before archiving the created data is the additional overhead. Already the creation of a dataset is time-consuming and sometimes even tedious. Adding the strict requirement for complete, extensive, machine-readable metadata and documentation can be seen as gate-keeping and too high a threshold. Some people might even consider withholding their data instead of releasing it if they have to meet such requirements for publication.

One major point in favor of standardized metadata and corpus documentation is the ability to automatically check for issues in the data set. Especially when archiving corpus data it is necessary to assess the quality of the data to guarantee later reuse. For example within the QUEST project (QUality ESTablished – Testing and Application of Curation Criteria and Quality Standards for Audiovisual Annotated Language Data)[6] it was demonstrated how a semi-automatic quality assurance process can profit from machine-readable corpus information (Arestau, 2022; Wamprechtshammer et al., 2022). For example, as long as the annotation schema is known, either because it follows some standard or if it is documented in a suitable way, it can be checked to be consistent and coherent across the whole data set.

It is also not the case that we have to start completely from scratch. There has been previous work on metadata standards for learner corpora such as (Granger and Paquot, 2017b,a). However, they lack visibility and are currently not generally applied. Another issue is that the draft by Granger and Paquot only specifies the data model, i.e., which fields have to be included and which values are acceptable, but not the representation. Consequently, the standard can be met both by human-readable metadata expressed for example using XML or JSON but also by only human-readable documentation such as MS Word documents or PDFs. Both issues, however, will hopefully be solved soon. Following the 6th International Conference for Learner Corpus Research (LCR 2022), a public call for feedback on a new draft of the core metadata standard has been sent to several relevant mailing lists[7]. Furthemore, at the same conference König et al. (2022) presented their approach to testing the core metadata standard on several corpora and expressing it using CMDI.

The question of representation of metadata is the final issue to be discussed here. As I summarized in the introduction, there is a number of viable and established metadata formats for learner corpora. Most of them are sufficiently expressive or extensible to be used for machine-readable corpus documentation. And there can be good reasons to prefer one over the other, e.g., good integration in the annotation software or in the infrastructure in which the data should be deposited. Sometimes several formats can be "competing" by providing similar functionality: both CMDI and OLAC (Bird and Simons, 2001) formats can be used in metadata harvesting, CMDI within CLARIN infrastructure and OLAC with the Open Language Archives Community. However, each metadata format requires understanding its philosophy to be able to use it in the most suitable way. This can be partially mitigated by using dedicated software for metadata creation and management such as the EXMARaLDA Corpus Manager, LAMETA (Hatton et al., 2021) or various CMDI tools in the CLARIN infrastructure but requires learning how to use the software instead. A minimum viable solution could be based on spreadsheets which are both easy to create and edit and can be automatically read by software. However, spreadsheets lack additional semantics such as a hierarchical structure or controlled vocabulary.

5 Conclusion

There are many good reasons for metadata standards, especially from the perspective of archiving and research data infrastructure. It is easier to deposit data in a repository if a supported set of metadata is provided in a standardized format. Furthermore, having access to suitable metadata, it is possible to auto-

[6] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html

[7] LINGUIST list archive: https://web.archive.org/web/20221124163838/https://list.elra.info/mailman3/hyperkitty/list/corpora@list.elra.info/message/5ITI7JXPYWAADXQ2MWTEXIQITWSVV332/

matically check relevant aspects of the corpus data. These two points would improve both findability and reusability of the deposited data. Especially the increased findability of the created datasets should ideally motivate corpus creators to include a sufficient set of metadata information in addition to their corpus data.

Furthermore, there are established machine-readable metadata formats with infrastructure and ecosystem surrounding them. For example CMDI is already omnipresent for all data published within CLARIN and can be modified to fit the data using profiles. As show by König et al. (2022), it could form a starting point for a unified representation for learner corpora metadata. And because it is a standard format within a large infrastructure, existing tools can be used to create and modify the metadata for learner corpora. Finally, having one metadata format as a pivot for conversion into other formats could be suitable for any additional metadata requirements such as specific formats for a certain archive outside CLARIN as well as for Linked Open Data.

A major challenge is to balance the interests of all parties involved. From an infrastructure point of view it is essential to have machine-readable metadata usable for ingesting the corpus data and providing means for finding relevant data. But when establishing a machine-readable metadata standard we also need to reduce the additional work loaded onto the researcher to document their data. The whole discussion is only relevant when corpus creators are willing to prepare and submit their data. Consequently, we have to collaborate on establishing standards acceptable for all parties.

Acknowledgements

Parts of this work have been funded by the QUEST project by the German Federal Ministry of Education and Research (BMBF) grant 16QK09B (06/2019–05/2022) as well as Text+ funded by the German Research Foundation (DFG) project number 460033370.

I would also like to thank Elena Arestau for the collaboration on quality standards for learner corpora as part of QUEST.

References

Elena Arestau. 2022. Curation of learner corpora. Technical report, University of Hamburg. https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest/ueber-das-projekt/projektergebnisse/arestaulearnercorpora.pdf.

Steven Bird and Gary Simons. 2001. The OLAC Metadata Set and Controlled Vocabularies. In Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In Proceedings of Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011, volume 7 of Balisage Series on Markup Technologies, Montréal, Canada.

Twan Goosen and Thomas Eckart. 2014. Virtual language observatory 3.0: What's new. In CLARIN Annual Conference, Soesterberg, The Netherlands.

Sylviane Granger and Magali Paquot. 2017a. Core metadata for learner corpora. Draft 1.0.

Sylviane Granger and Magali Paquot. 2017b. Towards standardization of metadata for L2 corpora. In Invited talk at the CLARIN workshop on Interoperability of Second Language Resources and Tools, 6-8 December 2017, University of Gothenburg, Sweden.

John Hatton, Gary Holton, Mandana Seyfeddinipur, and Nick Thieberger. 2021. Lameta. https://github.com/onset/laMETA/releases.

HZSK. 2010. HAMATAC - the Hamburg MapTask Corpus. Archived in Hamburger Zentrum für Sprachkorpora. Version 0.3. Publication date 2010-09-16.

Alexander König, Jennifer-Carmen Frey, Egon W. Stemle, Glaznieks Aivars, and Magali Paquot. 2022. Towards standardizing lcr metadata. In 6th International Conference for Learner Corpus Research (LCR 2022), 22.9.2022–24.9.2022, Padova.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset). Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart and Seminar für Sprachwissenschaften, Universität Tübingen. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf, accessed 2022-01-03.

Thomas Schmidt. 2017. DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora

am Institut für Deutsche Sprache (IDS) in Mannheim. 45(3):451 – 463.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. The Oxford handbook of corpus phonology, pages 402 – 419. Oxford Univ. Press, Oxford.

Leonore Spiegel, Maria Parker, Tim Feldmüller, Lisa Lenort, and Katrin Wisniewski. 2022. Mitschreiben in Vorlesungen in der Studieneingangsphase: Das multimodale Lehr-Lernkorpus MIKO. In Katrin Wisniewski, Wolfgang Lenhard, Leonore Spiegel, and Jupp Möhring, editors, Sprache und Studienerfolg bei Bildungsausländerinnen und Bildungsausländern. Waxmann Verlag, Münster.

TEI Consortium. 2022. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. Version 4.4.0. Last updated on 19th April 2022, revision ff9cc28b0.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL Language Learner Corpus : From Design to Annotation. Northern European Journal of Language Technology (NEJLT), 6:67–104. Special Issue of Selected Contributions from the Seventh Swedish Language Technology Conference (SLTC 2018).

Anna Wamprechtshammer, Elena Arestau, Jocelyn Aznar, Hanna Hedeland, Amy Isard, Ilya Khait, Herbert Lange, Nicole Majka, Felix Rau, and Gabriele Schwiertz. 2022. QUEST: Guidelines and specifications for the assessment of audiovisual, annotated language data. Technical report, QUEST project. Forthcoming.

Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1):160018.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018 :, number 159 in Linköping Electronic Conference Proceedings, pages 222–234.

Katrin Wisniewski, Elisabeth Muntschick, and Annette Portmann. 2022. Schreiben in der Studiersprache Deutsch. Das Lernerkorpus DISKO. In Katrin Wisniewski, Wolfgang Lenhard, Leonore Spiegel, and Jupp Möhring, editors, Sprache und Studienerfolg bei Bildungsausländerinnen und Bildungsausländern. Waxmann Verlag, Münster.