BMBF-Verbundprojekt
„**QUEST**: Quality – Established: Erprobung und Anwendung von Kurationskriterien und Qualitätsstandards für audiovisuelle, annotierte Sprachdaten“

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Curation of Learner Corpora

By Elena Arestau

Table of Contents

1. Learner Corpora

This document provides an overview of the curation of learner corpora and presents the main steps that I believe are important for the curation and reuse of such resources. These parts give an overview about relevant aspects and standards and serve as recommendations that should be considered by archives, data centres and persons who need to manage such data.

By learner corpora, we mean an electronic collection of natural or near-natural data produced by foreign or second language learners and assembled based on explicit design criteria (Granger, 2002). Learner Corpora consist of electronic collections of natural or near-natural learners written or/and oral production produced by L2 or L3/Lx learners or users and are built and published according to certain design criteria. Such samples can be used to answer a variety of research questions not only in second language research and pedagogy but also in other disciplines such sociology, psychology and even neurology. The following curation criteria are designed to be representative for a wide range of research objectives.

1.1 Design Criteria

When creating a learner corpus, several recommendations should be considered. Learner corpora need to be specified according to strict, transparent and systematic design criteria. All decisions and specifications of the design criteria should be well documented and made available to the corpus users to increase the reusability of a corpus. If data is gathered without documentation of learner, language- and task variables, the resulting corpus will be not of much use. As Bell/Payant state: “In designing a corpus, careful selection, documentation and justification of all criteria will increase the likelihood that the resulting corpus is methodologically-sound.” (Bell/Payant 2020: 56).

Therefore, the first step in assessing the reusability of learner corpora is to check whether **design considerations for building learner corpora are met**.

For this to be implemented, we recommend taking into account the following decisions and preparatory steps:

- The decision about what will be included in the corpus: spoken or written learner data.
- The decision about the sample of language learners: how much, how often and for how long.
- The decision about the environment in which the data are collected: the difference between target language as a second language and as a foreign language.
- The decision about the target population: e.g. learning environment, age, nationality, mother tongue background of the learner group.

- The decision about how the data to be collected: naturalistic production (e.g. recorded natural speech), elicited production (e.g. role-play) or experimental production.
- The decisions about possible variations in learner corpus design: language-related criteria (e.g. mode, genre, style, topic), task-related criteria (data collection, data elicitation, time limitations, use of references), learner-related criteria (e.g. age, motivation and attitude, learning context, L1 background, L2 proficiency)
- The decision about the educational settings (school / university) or in a natural settings (outside school or university) (cf. Tono 2003 – the best recommendations for the major categories).

1.2 Transcription

For the analysing to be possible, spoken learner corpora should be transcribed and the presence of transcription should be documented. Standardized conventions should be used to ensure consistency in the transcription of data and conversion into other formats.

It is important that the data is transcribed using standardised conventions and (accordingly documented) the transcription is documented.

For instance: CHAT transcription format is frequently used in the community and recommended in a number of articles (MacWhinney 2017a). The CHILDEShandbook offers important information about this transcription format. Transcripts in CHAT format can be automatically converted into the formats required for Praat (praat.org), Phon (phonbank.talkbank.org), ELAN (tla.mpi.nl/tools/elan), CoNLL, ANVIL (anvil-software.org), EXMARaLDA (exmaralda.org), LIPP (ihsys.com), SALT (saltsoftware.com), LENA (lenafoundation.org), Transcriber (trans.sourceforge.net), and ANNIS (corpustools.org/ANNIS) (vgl. HTTPS://TALKBANK.ORG/MANUALS/CHAT.PDF).

Furthermore, using transcript editors like CLAN, Praat or EXMARaLDA can facilitate transcription processes.

The second step to be checked is whether **the data is transcribed and documented using standardised conventions.**

1.3 Annotation

Linguistic annotation makes it possible to sort and compare learner corpora. For learner corpora, the annotation needs to be documented in the file list with function “annotation”. It is possible to retrieve linguistic patterns e.g. errors, grammatical categories. This can support the identification of learner language use. It can be done manually or automatically. One type of manual annotation is error annotation. For the error tagging/annotation a multi-layer corpus standoff architecture is very useful (cf. Lüdeling et al 2005).The most frequent type of automatic

annotation is POS Tagging, Parsing. The annotation should be consistent and accurate and the raw text should always remain recoverable. Important step is the evaluation of annotation: one has a gold standard and evaluates it against this corpus or one uses several annotators to annotate the same sub corpus using the same tag set and guidelines and evaluates how often and where they agree (called inter-annotator agreement). This step assures the consistency of annotation and is very important (cf. Bell / Payant 2020).

This step assures the consistency of annotation and is very important. For this to be measured we recommend the documentation of the annotation conventions (tag set). If the annotation file is documented in the data list “function”, it will allow the automatic check of the consistency of the annotation and then the checker can be used. For the validation of the annotation there is still the manual checking needed.

The third step is to check whether **the annotation tag set is documented and consistent**.

1.4 Metadata

Some efforts have been undertaken to review current (L2) learner corpora metadata to make recommendations about how such domain specific variables can be selected and presented to increase the re-use potential of such resources. The following recommendations are essential for the reuse and findability of learner corpora concerning the metadata (cf. Stemle et al. 2019, Granger and Paquot 2017, Volodina et al. 2016):

- Metadata should follow a specific format used in the community (e.g. XML-like format, minimally in the header of each text making up the corpus), so that results can be compared across different learner corpora;

- Metadata should be collected and published including how they were collected;

- At least a minimal set of core metadata is required, when corpora are shared in the community and should be provided in the file headers or in accompanying databases;

- The variables in the minimal metadata set have to be filtered according to legal aspects. If the data are not available, at least the metadata for the corpus should be provided.

Granger and Paquot (2017) proposed a core metadata set for learner corpora (L2). They design a flexible system that allows, depending on the focus of research, description of the main variables that have an influence on the learner use of L2 with the possibility to add or to expand the elements in the metadata set. This set consists of five main components: administrative metadata, corpus design metadata, corpus annotation metadata, task and learner metadata. Some categories are obligatory and essential for the learner language research such as target language (languages) of the corpus, L1(s) and other L2 language(s)

in the learners' environment, proficiency level, nationality of the participants, place and institution. (vgl. Stemle et al. 2019).

There is no standardised set for learner corpus metadata. Learner corpora metadata are present in a different format and must be manually checked. The automatic evaluation is not possible. The corpus needs to provide metadata in a standardised accessible way including a set for learner characteristics and a set for corpus information. "Without extensive documentation of data with relevant metadata, learner corpora cannot be reused." (Frey et al. 2020)

- Core metadata sets based on the suggestions made by Granger and Paquot (2017) and depending on the focus of research:

- Detailed information about the purpose and circumstances of data collection

- The background of the learners (e.g. L1, age, proficiency, other L2s)

- The writing task itself (target language, genre, writing prompts etc.)

- Information on the provenance of the data (e.g. authors, responsible people for data collection, processing and annotation, time and place of data collections) - licensing information Additional type of metadata

- Questionnaire in which each learner answers a set of questions related to motivation. (triangulating or combining different types of data in such a way, it may be possible to examine motivation as a variable which directly affects language development). Minimal list of required variables and recommendations (core variables in learner languages):

- Administrative metadata - Corpus design metadata: type and character of the metadata

- Corpus annotation metadata: conducted annotation and processing activities

- Research design: cross sectional, pseudo/quasi longitudinal, longitudinal

- Medium: collection of written texts, the transcriptions of spoken interactions, and/or audio-visual data (multimodal corpora)

- Information about the learners: age, languages, L2 learners or L2 users, language learning biography (recorded via a learner profile questionnaire, completed by all learners), stays abroad, information about motivation and attitudes, types of learners (children, teenager, adults), data concerning knowledge of all languages used or studied by the learner (mother tongue(s), home language(s), instructed additional language(s), extensive living abroad experiences), environment (in a second language context or in foreign language context)

- Language proficiency (to increase reliable comparability across corpora and studies): proficiency measure should always be included (how these levels were determined). Learner-centred method: age, institutional status, text-centred method: scores on standardized test

- Task information: style, format, task register, type of communicative task (written/oral, informal/formal, first or final draft, access to external sources), planning time, the amount of time on task, task setting (interactive task/computermediated communication)

- Information about the interviewers: gender, mother tongue, knowledge of the other foreign languages, familiarity with the learner

– Possible influence on the learner production.

Additional variables should also be reported: e. g. socio-economic status, parent's higher level of education, motivation to learn, learning disabilities, attitudes towards one's various languages " inclusion of an 'objective proficiency score' (Gilquin 2015, 30) and 'information […] about incidental learning in everyday life through reading, entertainment, social networking, etc.', which would extend our understanding of the learners' amount of L2 input variable" (Gilquin 2015: 30-31).

References

Bell, P. / Payant, C. (2020): Designing Learner Corpora. Collection, Transcription, and Annotation. In: Paquot, M. / Tracy-Ventura, N. (2020)(eds.): The Routledge Handbook of Second Language Acquisition and Corpora. Routledge.

Frey, J.-C., König, A. Stemle, E. W. (2021): "Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora" Information 12, no. 5: 199. HTTPS://DOI.ORG/10.3390 /info12050199

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 (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press, 9–34.

Granger, S., Gilquin, G., Meunier, F. (2016)(eds.): The Cambridge handbook of learner corpus research. Cambridge: University Press

Granger, S. &Paqout, M. (2017). Towards standardization of metadata for L2 corpora. CLARIN workshop on Interoperability of Second Language Resources and Tools (Gothenburg, Sweden, du 06/12/2017 au 08/12/2017).

Hedeland, H. / Lehmberg, T. / Schmidt, T. / Wörner, K. (2014): Multilingual Corpora at the Hamburg Centre for Language Corpora, in: Best Practices for Speech Corpora in Linguistic Research. Cambridge Scholars Publishing.

Le Bruyn, B. / Paquot, M. (2021)(eds.): Learner Corpus Research Meets Second Language Acquisition. Cambridge: University Press.

Lüdi, G. (2011): Neue Herausforderungen an eine Migrationslinguistik im Zeichen der Globalisierung, in: Stehl, Thomas (eds.): Sprachen in mobilisierten Kulturen; Aspekte der Migrationslinguistik, Universitätsverlag. Potsdam. (=Mobilisierte Kulturen 2).

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C-J., Sundberg, G., Wirén, M., Volodina, E. (2018): Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In Proceedings of the 7th NLP4CALL, SLTC workshop, Stockholm, Sweden.

Paquot, M. / Tracy-Ventura, N. (2020)(eds.): The Routledge Handbook of Second Language Acquisition and Corpora.

Schmidt, T. / Wörner, K. (eds.) (2012): Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: Benjamins.Schmidt, Thomas / Wörner, Kai / Hedeland, Hanna / Lehmberg, Timm (2013): Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache, Mannheim, Institut für Deutsche Sprache.

Schmidt, T. / Schütte, W. / Winterscheid, J. (2015): cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2), Mannheim: Institut für Deutsche Sprache.

Stemle, E., Boyd, A., Janssen, M., Lindström Tiedemann, T., MikelićPreradović, N., Rosen, A., Rosén, D. &Volodina, E. (2019). Working together towards an ideal infrastructure for language learner corpora. In Andrea Abel, AivarsGlaznieks, VerenaLyding& Lionel Nicolas (eds.) Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference. Corpora and Language in Use – Proceedings 5, Louvain: Presses universitaires de Louvain, 427-468.

The LIPPS Group (2000): The LIDES Coding Manual. A document for preparing and analyzing language interaction data. Version 1.1 - July, 1999. In: International Journal of Bilingualism, 4, Nr. 2

Volodina, E., Megyesi, B., Wirén, M., Granstedt, L., Prentice, J., Reichenberg, M., &Sundberg, G. (2016). A friend in need? Research agenda for electronic second language

infrastructure. In Proceedings of the Sixth Swedish Language Technology Conference (SLTC-2016), Umeå, 17-18 November, 2016, 1–4.