

QUEST

Quality–Established

Anna Wamprechtshammer, Elena Arestau & Amy Isard

„Generic and discipline-specific approaches to the quality of audiovisual, annotated language data in the BMBF project QUEST“



Z A S



Outline

1. About the project (Basic Information, Project Goals)

2. Evaluation Criteria

2.1 Quality Standards & Curation Criteria

2.2 Curation Criteria Use Case: Learner Corpora

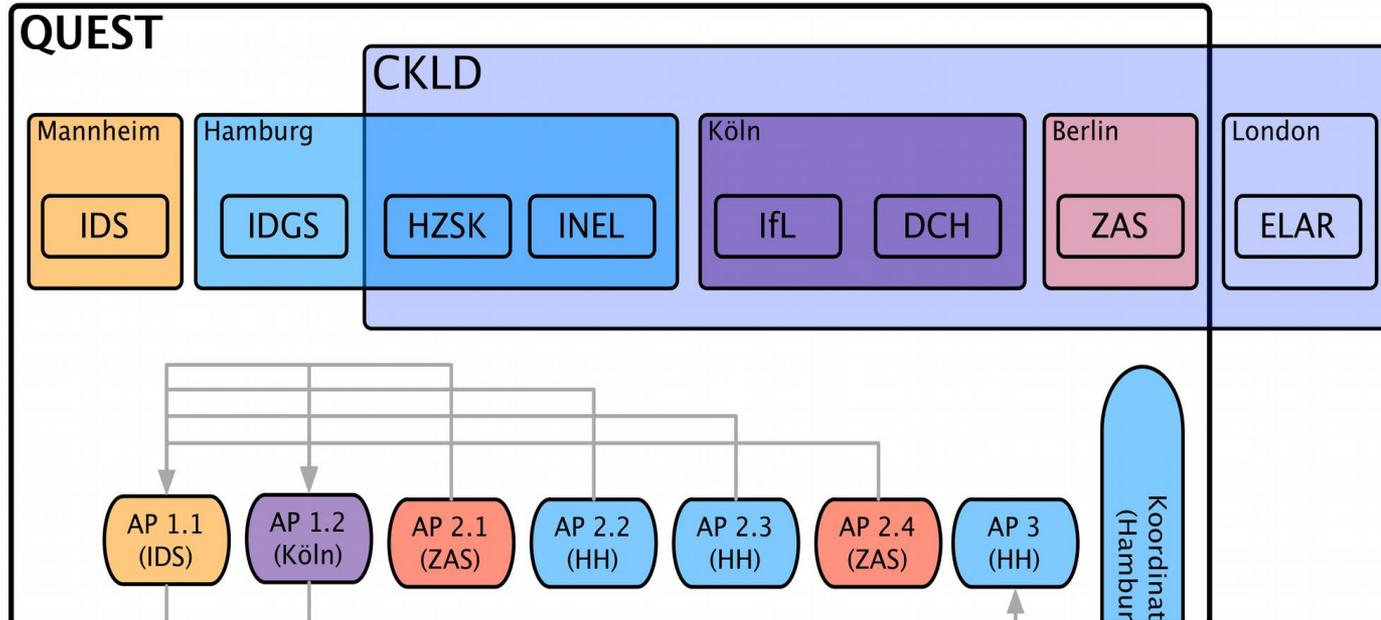
3. Quality Assurance Measures

3.1 Gloss Checking Tool

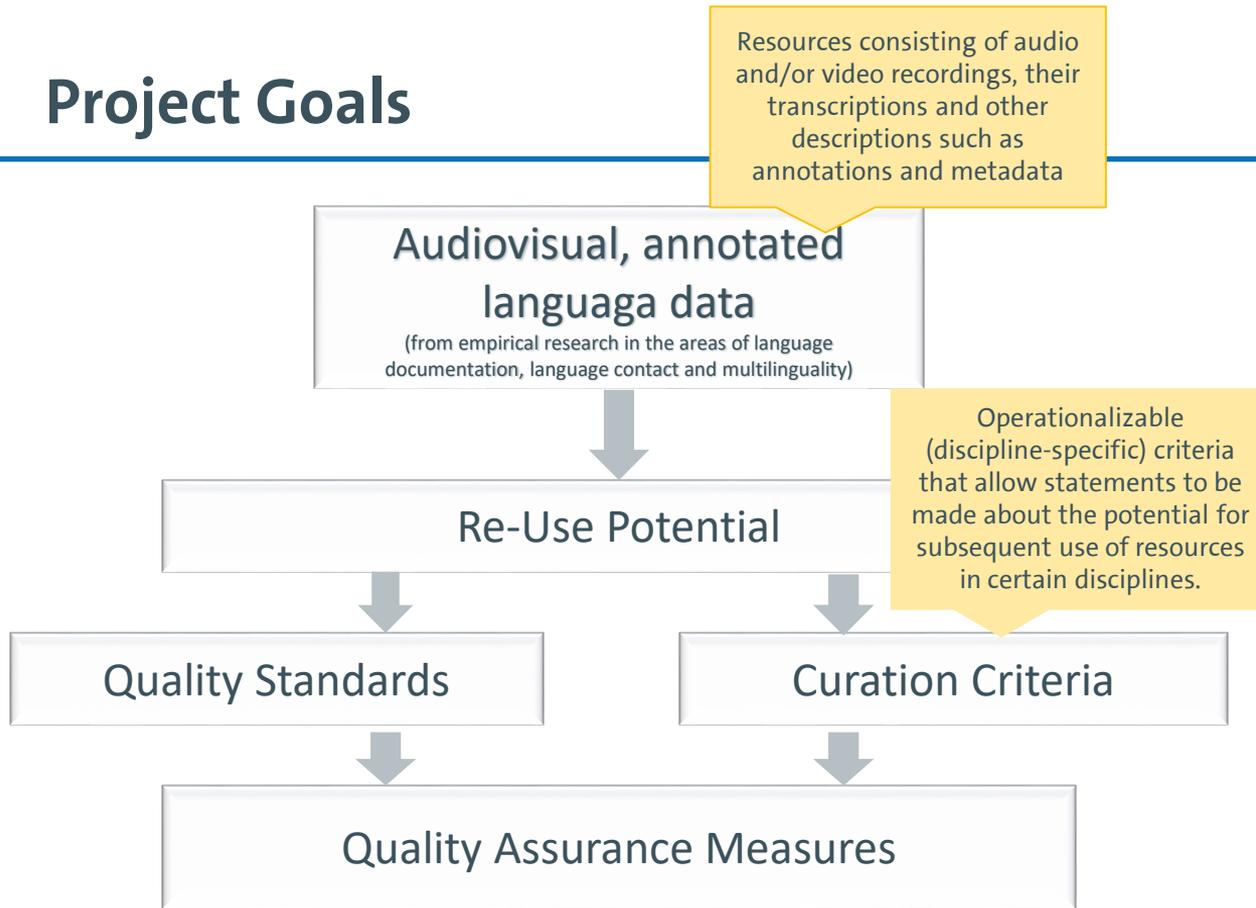
3.2 Evaluation System

4. Contact

„QUEST: Quality – Established: Erprobung und Anwendung von Qualitätsstandards und Kurationskriterien für audiovisuelle, annotierte Sprachdaten“ (2019 – 2022)



Project Goals



Evaluation Criteria – Generic & Discipline-specific Approach

Quality Standards

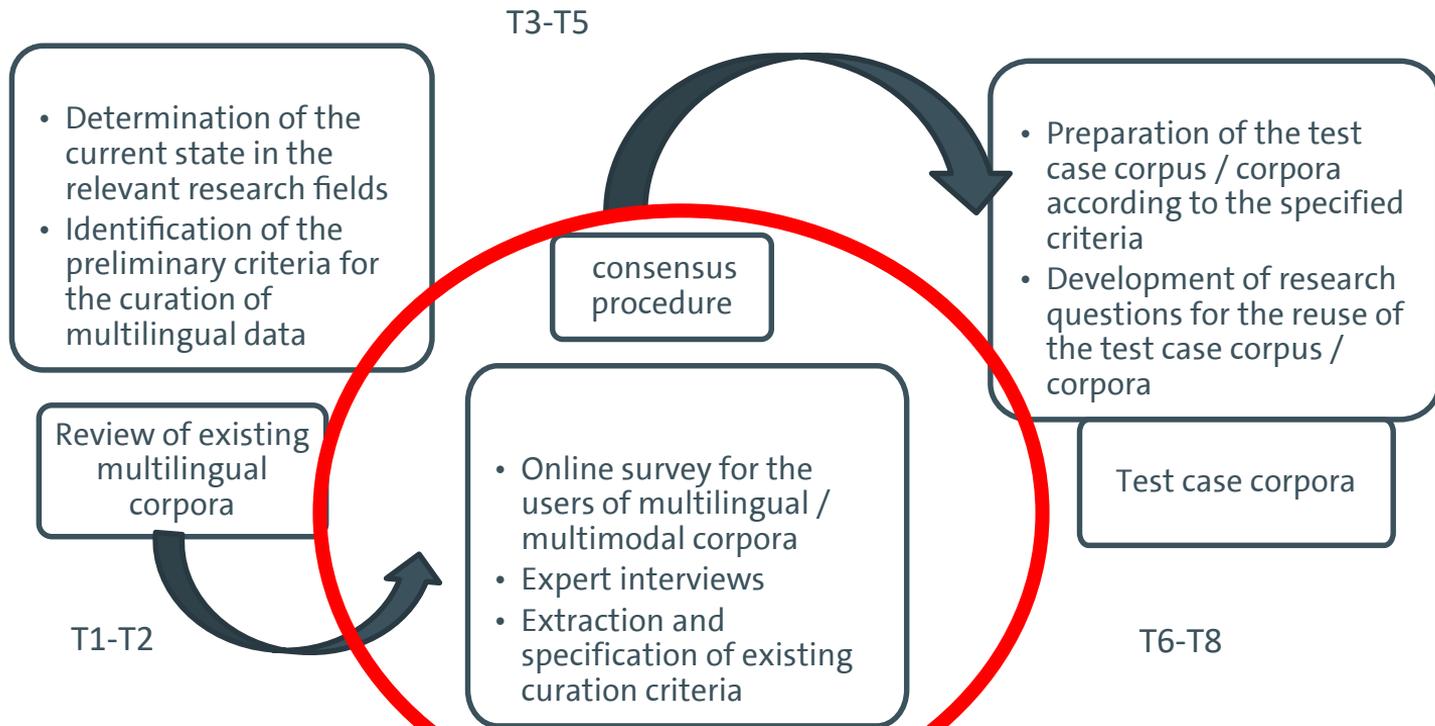
- Generic quality criteria regardless of intended usage scenario
- Quality standards subproject to develop basic generic standards for relevant resource types and their metadata
- Formulated for compliance with technical standards, structural, formal correctness of the data the methodological aspects of data creation
- Examples: All files need to be in recommended file formats, data are to be described with a minimal set of catalogue metadata...

Curation Criteria

- With regard to concrete re-use scenarios for research data from the fields of language documentation, multilingualism research, sign language and oral history, requirements for data, their structure and content with regard to interdisciplinary reuse as well as use of the data within the framework of a third mission are developed.

2.2 Curation Criteria

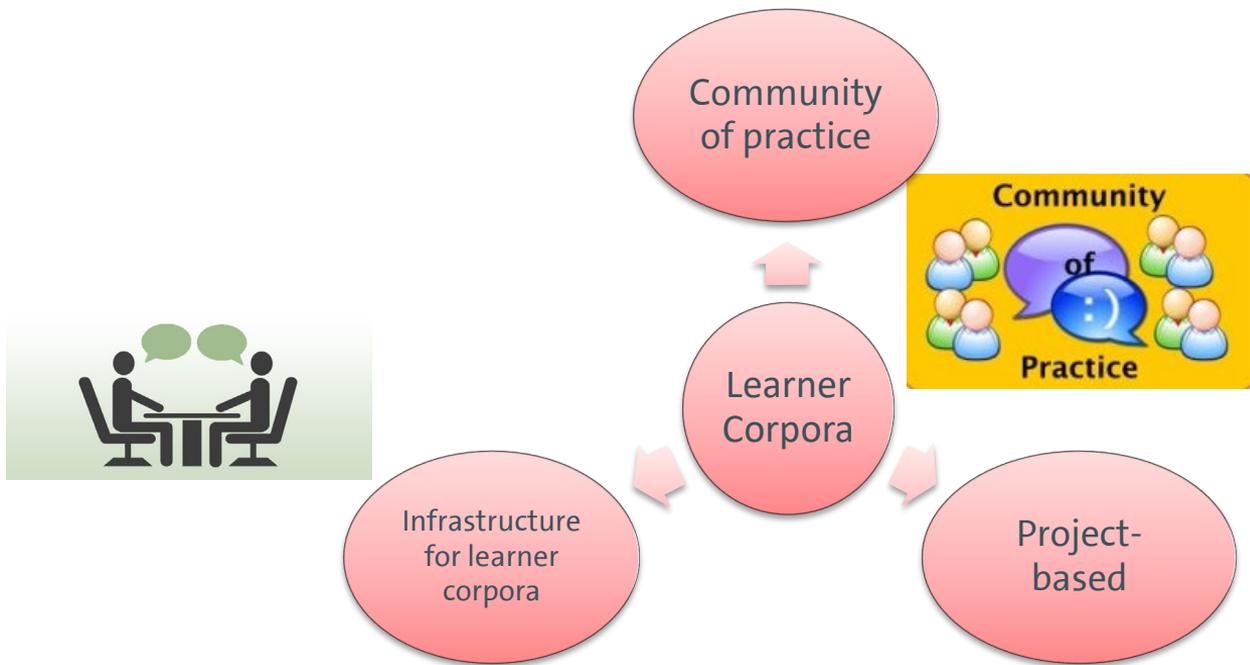
Use Case: Learner Corpora



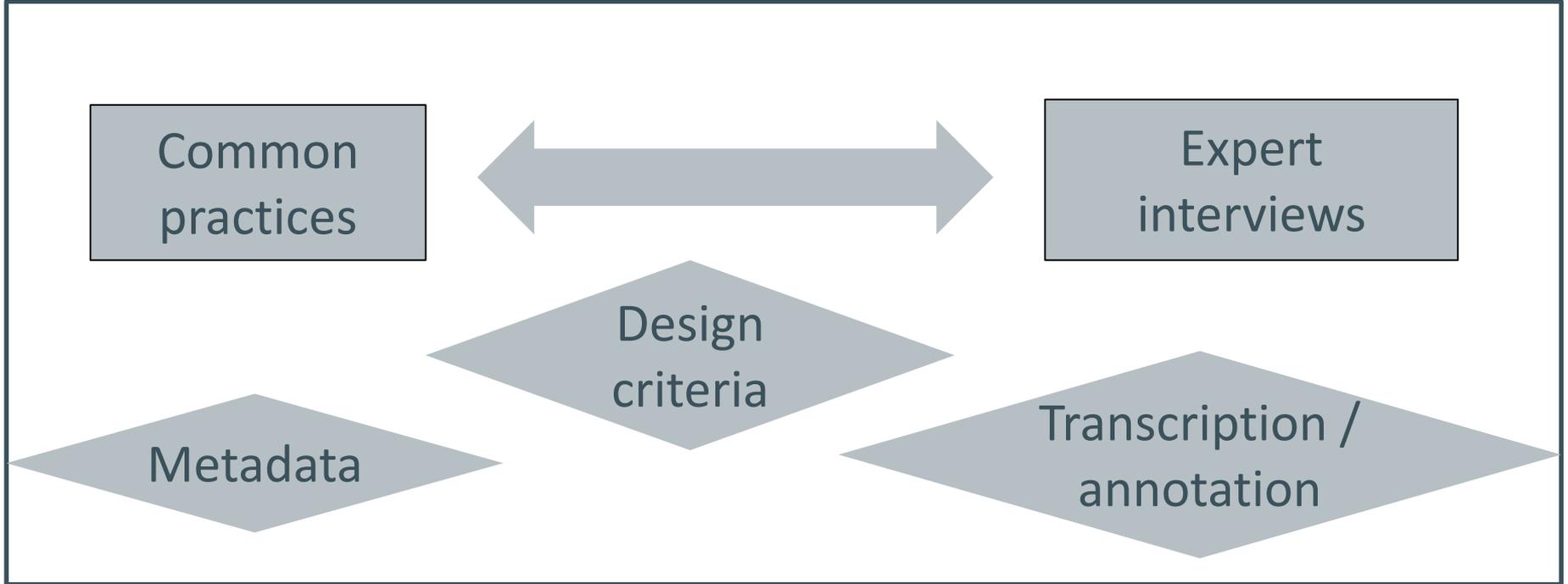
Learner Corpus: Definition

- electronic collection of natural or near-natural learner's written and/or oral production
- produced by L2 or L3/Lx learners or users
- built and published according to certain design criteria

Why learner corpora?

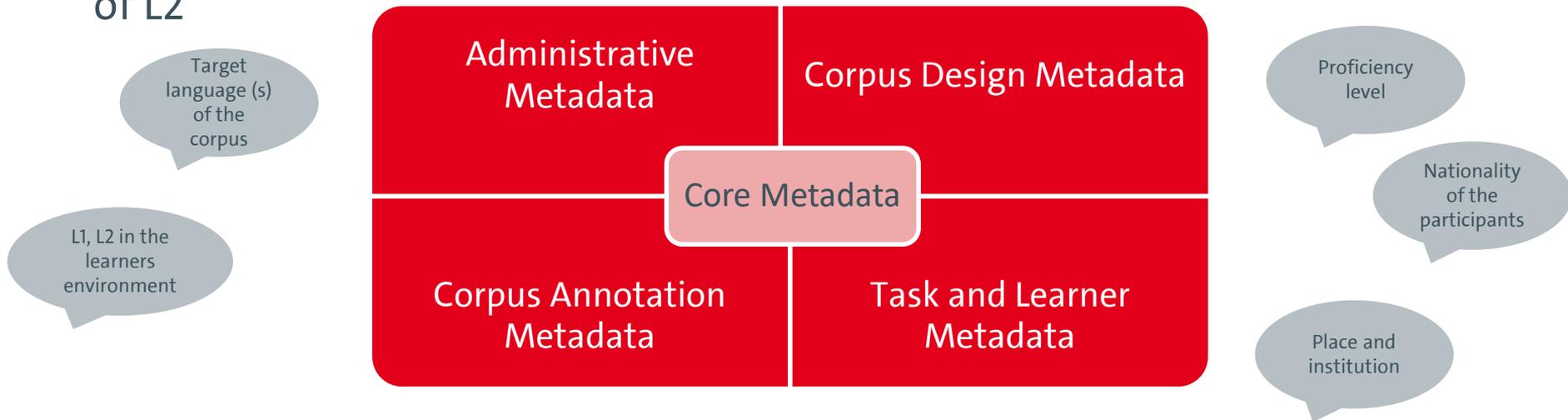


Developing criteria for the curation



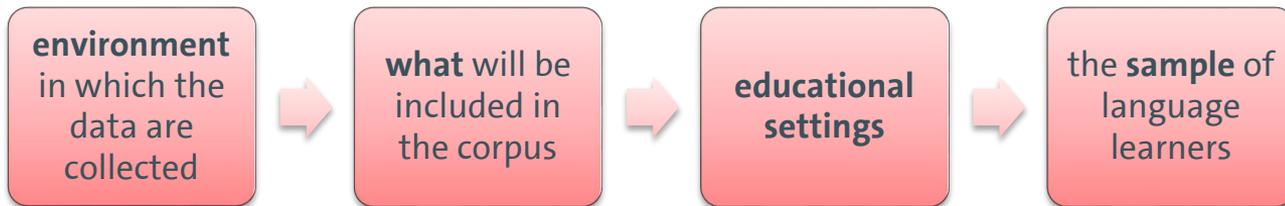
Curation criteria for the evaluation: Metadata

- Granger and Paquot (2017)
- Core metadata set for learner corpora
- Description of the main variables that have influence on the learner use of L2



Curation criteria for the evaluation: Design Criteria

“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: 54).



Criteria for annotation (transcription, POS-tagging)

Standardised conventions
(e.g. CHAT transcription
format MacWhinney 2017 a,b)

resources available and the
research purposes

transcript editors like
CLAN, Praat or
EXMARaLDA

Summary

Standardisation

Availability of documentation

Accessibility

Reliable

Replicable

Reusable

3. Quality Assurance Measures

The third sub-goal:

- implementation of concrete methods of quality assurance to guide and support the construction and processing of audiovisual language data according to the quality standards and curation criteria which have been developed

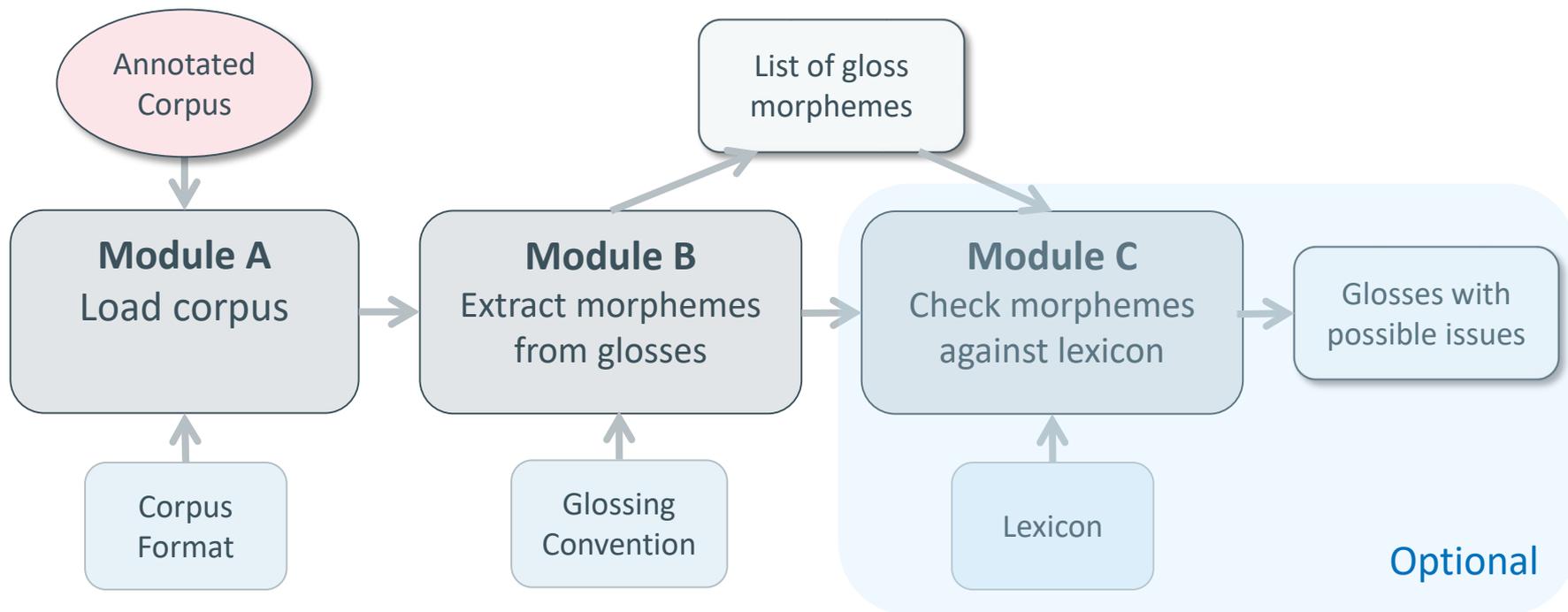
3.1 Gloss Checking Tool

- Automatic checking of corpus gloss annotations
- Gloss in this context: morpheme-by-morpheme translation/definition
- Many different glossing conventions e.g.:
 - **LEIPZIG GLOSSING RULES**
 - **SIGNBANK ID GLOSSES (CORMIER, CRASBORN AND BANK, 2018)**

Gloss checking tool

- Currently under development
- Will be available for download from Quest portal
- Designed for offline use
- Requires:
 - **LOADER FOR CORPUS FORMAT (E.G. ELAN, EXMARALDA, ILEX)**
 - **MODULE INPUT FOR GLOSSING CONVENTION (E.G. LEIPZIG GLOSSING RULES)**
 - **(OPTIONAL) LEXICON OF GLOSS LANGUAGE (E.G. ENGLISH, GERMAN)**
- Input: Annotated Corpus with gloss tier
- Output:
 - **LIST OF GLOSS MORPHEMES**
 - **(OPTIONAL) LIST OF GLOSSES WITH POTENTIAL ISSUES**

Tool Modules



Gloss checking tool

- Corpus Formats:
 - **ILEX (COMPLETED)**
 - **ELAN (IN PROGRESS)**
 - **SIGNBANK (COMING SOON)**
- Glossing Conventions
 - **DGS KORPUS GLOSS RULES (COMPLETED)**
 - **LEIPZIG GLOSSING RULES (IN PROGRESS)**
 - **SIGNBANK ID GLOSSES (COMING SOON)**
- Lexicon of Written Languages
 - **ENGLISH (COMPLETED)**
 - **GERMAN (COMPLETED)**
 - **DUTCH (COMING SOON)**

3.2 Evaluation System



quest.multilingua.uni-hamburg.de

Anna Wamprechtshammer, M.A.

Projektkoordination

Institut für Finnougristik/Uralistik

Überseering 35, Postfach #29

22297 Hamburg

Tel.: +49 40 42838 2578

E-Mail: anna.wamprechtshammer@uni-hamburg.de

References

- Bell, Philippa / Payant, Caroline (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.
- Granger, Sylviane, Gilquin, Gaetanelle, Meunier, Fanny (2016) (eds.): *The Cambridge handbook of learner corpus research*. Cambridge: University Press.
- Granger, Sylviane, Paquot, Magali (2017): Core metadata for learner corpora. Draft. CLARIN Workshop on Interoperability of Second Language Resources and Tools, <https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools>.
- MacWhinney, Brian. 2017a. *Tools for analyzing talk*. Part 1: The CHAT program. Available at: <http://talkbank.org/manuals/CHAT.pdf>. Accessed: 19 August 2017.
- MacWhinney, Brian. 2017b. *Tools for analyzing talk*. Part 2: The CLAN program. Available at: <http://talkbank.org/manuals/CLAN.pdf>. Accessed 19 August 2017.
- Tono, Yukio (2003): “Learner corpora: Design, development and applications”, in: Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.): *Proceedings of the Corpus 27 Linguistics 2003 Conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 800–809. < <http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf> > [29.07.2021].