L2 Learner Corpus Survey – Towards improved verifiability, reproducibility and inspiration in LCR

Therese Lindström Tiedemann[1], Jakob Lenardič [2], Darja Fišer[2,3]

[1] University of Helsinki, [2] University of Ljubljana, [3] Jožef Stefan Institute
therese.lindstromtiedemann@helsinki.fi, jakob.lenardic@ff.uni-lj.si, darja.fiser@ff.uni-lj.si

CLARIN ERIC
Common Language Resources and Technology Infrastructure

FIN-CLARIN

Introduction and survey set up

- Learner corpus research (LCR) has increased a lot since the early 1990s (McEnery 2018).
- L2 Learner corpora are still mainly for English L2 or with English as L1.
- The L2 research community would profit from more extensive availability of learner corpora and information about those that have already been compiled.
- **The aim of the survey**: identify valuable L2 learner corpora and their integration in the CLARIN infrastructure.

L2 corpora in the CLARIN infrastructure

- The following table gives an overview of availability and key metadata of the identified L2-corpora.

	Written	Spoken	Multimodal	Total
English L2	5	4	1	10 (29%)
Other L2	13	7	4	24 (71%)
In VLO	15	11	5	31 (91%)
Download/concordancer	12	10	1	23 (68%)
Through CLARIN rep.	8	7	/	15 (44%)
L1 Language	7	6	2	15 (44%)
Metadata on size	16	6	2	24 (71%)
Metadata on annotation	4	5	/	9 (26%)
Info. on licence	16	9	3	28 (82%)
TOTAL IN CLARIN	18	11	5	34 (19%)
TOTAL NON-CLARIN				146 (81%)
TOTAL				180

- **Identification: 34** CLARIN corpora (in VLO or repositories); difficult to find L2 learner corpora in the VLO.
- **Metadata**: great inconsistency w.r.t availability and precision of information was noted in the survey.
  - **Example:** often difficult to distinguish between target (L2) language and L1 backgrounds in a corpus. Information on latter often missing (cf. *Merlin Written Learner Corpus for Czech, German, and Italian 1.1).*
- **VLO recommendation:** special facet for learner corpora & clearer mark-up of L1 vs. L2.

Documentation guidelines

- A set of clear guidelines for including relevant metadata should be adopted.
- Tentative suggestions for the required metadata (often missing in existing corpora):
  1. The L1 background(s) of the participants
  2. The number of participants
  3. The participants' age
  4. Clear distinction between L1 and L2 languages and their distribution in the corpus
  5. The types of learning tasks in the corpus
  6. Detailed information about proficiency levels
- Consensus on certain metadata is needed: e.g. does L1 mean what the informant sees as their L1 or does it mean parents' L1(s).

Conclusions

- The corpora listed make a good contribution to the observed need for more non-English corpora (L1 / L2).
- Many valuable corpora could still be added **(cf. 146 non-CLARIN vs. 34 CLARIN corpora)**.
- Tailoring the VLO search functionalities to L2 learner corpora is needed.
- More outreach to the L2 researchers about what CLARIN has to offer is needed.
- A consensus on L2 corpus metadata is crucial; comprehensive implementation of the consensus is needed.

A European Research Infrastructure

CLARIN ERIC was established in 2012; it is a landmark in the 2016 ESFRI roadmap.

www.clarin.eu

clarin@clarin.eu
@CLARINERIC
facebook.com/CLARINERIC
github.com/clarin-eric

Acknowledgements

We would like to thank everyone who contributed to the survey through the CLARIN workshop in Gothenburg in Dec 2017, and after that through the online spreadsheet.
We would also like to thank the anonymous reviewers for their helpful comments, and the CLARIN ERIC User Involvement representatives and National Coordinators.
The work described has been supported by *Riksbankens jubileumsfond P17-0716:1* (the first author) and the H2020 *PARTHENOS: Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies* (second and third author) projects.