Topic: Curate existing language learners databases & pathological databases and make them searchable

ID: T004, T005

Author: Henk van den Heuvel, Maarten van Gompel

What exists:

- Basilex and Basiscript corpora, both delivered in FoLiA
- Jasmin-CGN and Diglin databases
- FoLiA

What must be adapted / extended / created anew:

- Basilex corpus currently has FoLiA validation errors that need to be fixed (0.3PM)
- Jasmin-CGN and Diglin must be must be:
  - curated into FoLiA format
  - Metadated
  - Student annotations : 7500 EUR (300 uur)
  - Inclusion into search interface (e.g. OpenSoNaR+)
- Datacuratie management over 5 years
  - The exact list of corpora/datasets to be curated must be determined with CLARIAH's EB.

Why important for CLARIAH (scientific impact):

Access and preservation, FoLiA is a de-facto CLARIN standard.

Targeted/Actual users: Linguists, Language Learners

Actual use (quantify!): Many

Social Impact (concrete examples): curated resources can be easily be integrated in other FoLiA-based resources

Lead: Henk van den Heuvel

Proposed PM estimation (try to justify): 9.2 PM + students assistants 7500 euro in total

Allocated PMs: 10 PM

Proposed participants + PMs:

- Henk van den Heuvel datacuratie over 4 years 0,1 fte 4 * 12 * 0.1 = 4.8 PM
- Maarten van Gompel, Curation of corpora into FoLiA: 2 PM
- Programmer for inclusion into Search interface: 3.2 PM

Deliverables

1. Curated Basilex corpus (in FoLiA)
2. Curated Jasmin-CGN (in FoLiA)
3. Curated Diglin (in FoLiA)
4. Search interface (e.g. OpenSoNaR+) with all corpora

Time plan

- Curation databases: Sep 2019 - Sep 2022
- Inclusion into search interface: April 2020 - Dec 2023