Digital grammaticography

towards data-rich corpus-based interactive hypertext grammars

Florian Matter

University of Oregon

SLE
Societas
Linguistica
Europaea

Digital grammaticography

towards data-rich corpus-based interactive hypertext grammars

Florian Matter

University of Oregon

SLE Societas Linguistica Europaea

Description and documentation

- grammatical descriptions should be based on naturalistic data (corpora)
- language description based on language documentation (Himmelmann 1998; McDonnell et al. 2018)
- usual scenario:
  1. `.wav` and ELAN files in archive (imported FLEx annotations?)
  2. description written in word processor → PDF/book
- separate (but digital!) products

A new approach

- text is written in markdown
- data is stored in a database
- different output formats can represent entities differently

grammars are prose interspersed with data[1]

Why would one do this?

- database (analytical) updates are directly reflected in final output
- non-linear and data-rich output formats (backwards compatibility with paper)
- increased "reproducibility" of grammatical descriptions
  - writers are forced to explicitly reference database entries
  - skeptical readers can search the underlying corpus
- data is easily reusable by other researchers (no copypasting data out of a PDF)

Implementation

- text is written in markdown (what else?)
- choice of database: CLDF (Forkel et al. 2017) and CLLD (Forkel et al. 2019)

Prose

- the text module of the cldfviz library[2] allows embedding CLDF database entries in markdown output
- pylingdocs (Matter 2023b): a general-purpose application for data-rich linguistic documents
  - different output formats (LaTeX, HTML, CLLD)
  - cross-references and example references
  - tables
  - multi-file documents
- database pointers: `[f](form-1)`, or more complex: `[ex](ex-04,ex-15?example_id=plur&title=Different functions of the pluractional suffix&with_primaryText=True)`

Data

- CLDF components for linguistic corpora (Matter 2023a)
  - available as a python library
  - integrated into *box and FLEx conversion libraries (Matter 2022a, 2023c)
  - lightweight, shareable datasets
- corresponding CLLD database models available as plugins

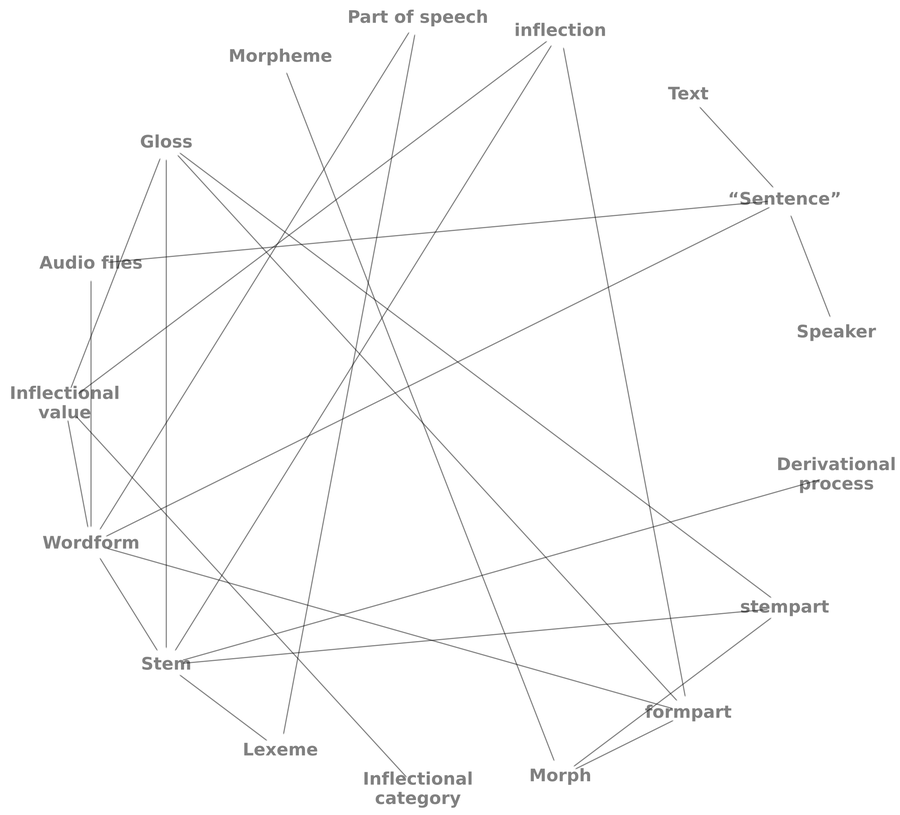
Interactive output

- InDiCoGram: interactive, digital, corpus-based grammar (Matter 2022b)
- lightweight CLLD template using several plugins
- everything is a link (almost)
- audio!
- example: yawarana-sketch.herokuapp.com

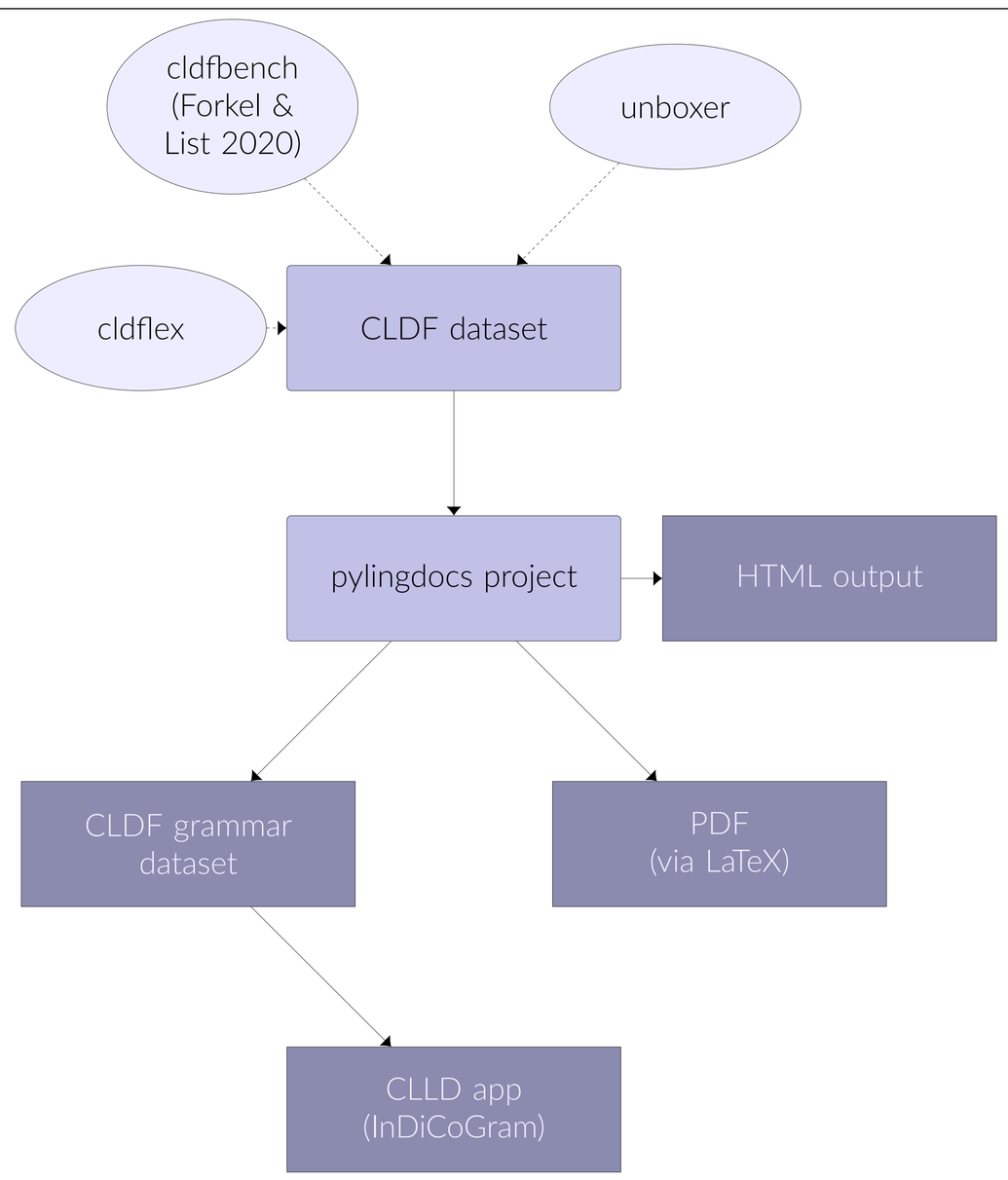
[1] See Nordhoff (2012)

[2] github.com/cldf/cldfviz/

Current database tables



Workflow



References

Forkel, Robert & Johann-Mattis List (May 2020). "CLDF-Bench: Give Your Cross-Linguistic Data a Lift". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association: 6995–7002. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.864.

Forkel, Robert et al. (2017). *CLDF 1.0*. DOI: 10.5281/zenodo.1117644. URL: https://doi.org/10.5281/zenodo.1117644.

Forkel, Robert et al. (2019). "clld: a toolkit for cross-linguistic databases". DOI: 10.5281/zenodo.3239095. Online: https://doi.org/10.5281/zenodo.3239095.

Himmelmann, Nikolaus P. (1998). "Documentary and descriptive linguistics". In: *Linguistics* 36 (1): 161–195.

Matter, Florian (Nov. 2022a). *cldflex*. Version 0.1.0. URL: https://github.com/fmatter/cldflex.

— (Oct. 2022b). *InDiCoGram*. Version 0.0.1.dev. URL: https://github.com/fmatter/indicogram.

— (June 2023a). *cldf-ldd*. Version 0.0.6. URL: https://github.com/fmatter/cldf-ldd.

— (July 2023b). *pylingdocs*. Version 0.0.12. URL: https://github.com/fmatter/pylingdocs/releases/tag/0.0.12.

— (July 2023c). *unboxer*. Version 0.0.2. URL: https://github.com/fmatter/unboxer.

McDonnell, Bradley, Andrea L. Berez-Kroeker & Gary Holton, eds. (2018). *Reflections on Language Documentation: 20 Years after Himmelmann 1998*. Honolulu: University of Hawai'i Press.

Nordhoff, Sebastian (2012). "The grammatical description as a collection of form-meaning-pairs". In: *Electronic Grammaticography*. Sebastian Nordhoff (ed.). Manoa: University of Hawai'i Press: 33–62.