A first glimpse at a workflow for writing digital grammars

Florian Matter

February 21, 2023

Data-based grammaticography

- grammatical descriptions should be based on data
- claims need to be supported with naturalistic data (corpora)
- language description based on language documentation (Himmelmann 1998; McDonnell, Berez-Kroeker, and Holton 2018)
- usual scenario:
  1. `.wav` and ELAN files in archive (imported FLEx annotations?)
  2. description written in word processor → PDF/book
- disparate “products”
- unused possibilities: digital grammars

Digital grammaticography

1. how to structure a digital grammar?
    a. what kind of information is stored?
    b. in what format is it stored?
2. how to write a digital grammar?
3. how to consume/explore/read a digital grammar?

How should a digital grammar be structured?

- ideally modelled according to a standard ontology for language description
- RDF[1] triples could be used to encode statements about linguistic entities (Good 2012)
  - [Language X] hasPhoneme [/t/]
  - machines can evaluate data
  - ..and visualize them for humans?
- no such ontology
- **grammars are prose interspersed with data** (Nordhoff 2012)

[1] Resource Description Framework

Combining prose and data

- usual approach: **copying** data (from somewhere) into a document (in some format)
    - potential for analytical discrepancies between data and text
    - manual formatting
    - not straightforward to do
    - data in PDF is hard to access
- my approach: prose containing only **references** to data
    - every datapoint is an entity in the database
    - representation depending on output format
    - structure of grammar: text + database
    - ontology-independent (!)

Combining prose and data: implementation

- obvious candidate: Markdown
    - widely used
    - lightweight and easy to use
    - adaptable
- established for data-rich text
    - for R: `rmarkdown`
    - for python: `jupyter`

Combining prose and data: implementation

- what kind of database? should be...
  - open
  - flexible
  - shareable & accessible
- my choice: CLDF (cross-linguistic data format)
  - born out of the CLLD (cross-linguistic linked data) project (known for WALS, glottolog, dictionaria, DoReCo, *bank)
  - CSV[2] data, JSON[3] metadata
  - easily convertible to CLLD database → powerful web app

[2] comma-separated-values

[3] JavaScript Object Notation

Combining prose and data: implementation

- R. Forkel introduced text module to `cldfviz`
- link notation is "hijacked"
  - `[label](http://www.target.com)`
  - `[label](FormTable#cldf:form-1)`
  - rendered with `Jinja2` templates
- added functionality with `pylingdocs`:
  - simpler data references (`[f](form-1)`)
  - tables (as CSV files)
  - multi-file documents
  - cross and example references
  - different output formats
  - general-purpose application for data-rich linguistic documents

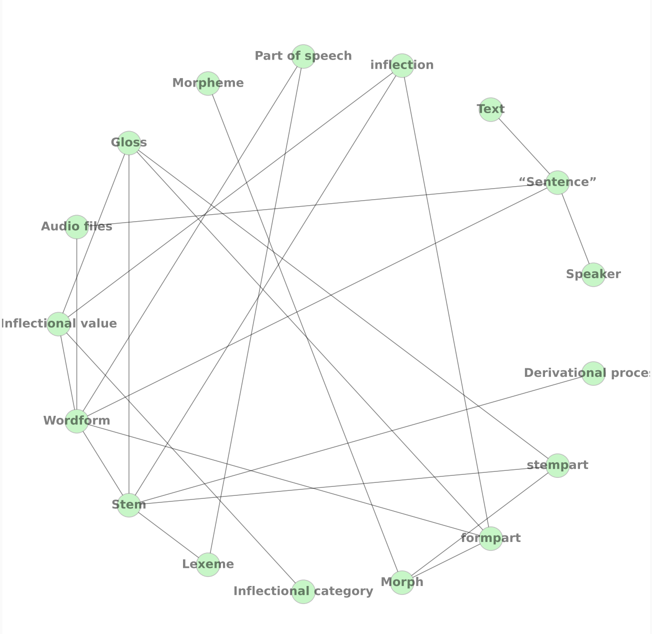
Writing a digital grammar: prose

- markdown is plaintext; can be written in application of your choice
- pylingdocs is not for editing, only rendering
- first option: Sublime Text with a plugin
- second option: browser-based pylingdocs-gui

Writing a digital grammar: database

- CLDF ontology rather limited
    1. typological parameters
    2. comparative wordlists
    3. simple dictionaries
    4. parallel texts
- implemented additional components, based on structure of fieldwork corpora: cldf-ldd

Writing a digital grammar: current ontology



Writing a digital grammar: creating a CLDF dataset

- brew your own with `cldfbench`
- FLEx: `cldflex`
- *box: `unboxer`

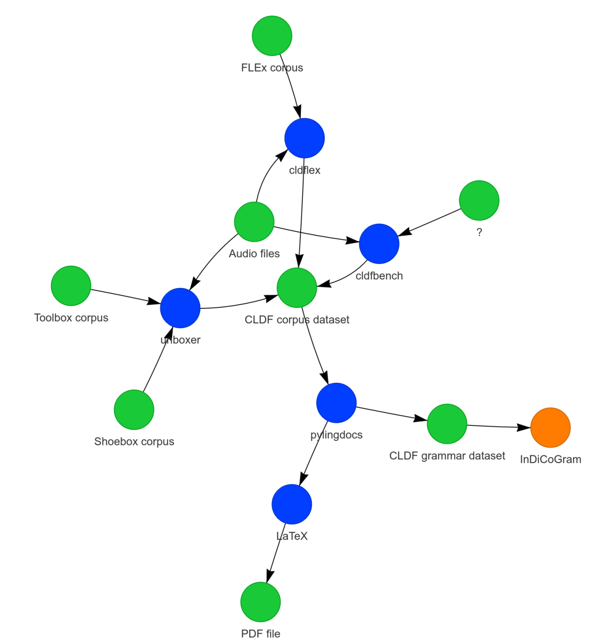
Consuming a digital grammar

- two target formats for `pylingdocs`:
    - producing PDF output via LaTeX (→ print product)
    - adding `chapters.csv` to an existing CLDF dataset → CLLD web app
- CLLD plugins:
    - `clld-markdown-plugin` (w/ R. Forkel)
    - `clld-document-plugin` (chapters, example references, tables...)
    - `clld-morphology-plugin`
    - `clld-corpus-plugin`
- bundled in InDiCoGram template

Digital grammaticography

1. how to structure a digital grammar?
   a. what kind of information is stored? **prose and database**
   b. in what format is it stored? **markdown and CLDF**
2. how to write a digital grammar? **pylingdocs, cldflex**
3. how to consume/explore/read a digital grammar? **PDF or CLLD app**

Pipeline


FLEx corpus
cldflex
?
Audio files
cldfbench
Toolbox corpus
CLDF corpus dataset
unboxer
Shoebox corpus
pylingdocs
CLDF grammar dataset
InDiCoGram
LaTeX
PDF file

Advantages

- data accessible for and easily shareable with other researchers (CLDF dataset)
- "reproducibility"; all reference to data is explicit
- nonlinear consumption
- audio
- different writing process

Issues

- practical:
    - grammar is for humans, not computers
    - publishing?
    - onomasiology?
    - not enough buttons
- ontology:
    - meaning?
    - non-concatenative processes?
    - kinds of allomorphy?
    - syntactic structures?
    - …

DIY

- FLEx database to CLLD tutorial

Comparison: other approaches

- Abesabesi grammar (Lau 2022, 2021)
  - structure: XML description + FLEx converted to better XML
  - writing: manually coding XML
  - consumption: web app
- online grammars of Eastern Cree (Junker 2000--2014) and Nunggubuyu (Thieberger, Musgrave, and Baker n.d.; Musgrave and Thieberger 2012)
  - structure, writing, consumption: HTML

References

Good, Jeff. 2012. "Deconstructing Descriptive Grammars." In *Electronic Grammaticography*, edited by Sebastian Nordhoff, 2–32. Manoa: University of Hawai'i Press.

Himmelmann, Nikolaus P. 1998. "Documentary and Descriptive Linguistics." *Linguistics* 36: 161–95.

Junker, Marie-Odile, ed. 2000--2014. "The Interactive East Cree Reference Grammar." 2000--2014. https://www.eastcree.org/cree/en/grammar/.

Lau, Jonas. 2021. "A Digital Reference Grammar of Abesabesi: Towards a Data Format for Digital Reference Grammars." PhD thesis, University of Cologne.

———. 2022. "Abesabesi Grammar." 2022. https://abesabesi.cceh.uni-koeln.de.

References

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

Musgrave, Simon, and Nick Thieberger. 2012. "Language Description and Hypertext: Nunggubuyu as a Case Study." In *Electronic Grammaticography*, edited by Sebastian Nordhoff, 63–77. Manoa: University of Hawai'i Press.

Nordhoff, Sebastian. 2012. "The Grammatical Description as a Collection of Form-Meaning-Pairs." In *Electronic Grammaticography*, edited by Sebastian Nordhoff, 33–62. Manoa: University of Hawai'i Press.

References

Thieberger, Nick, Simon Musgrave, and Brett Baker, eds. n.d. "Nunggubuyu Grammar Online." Accessed May 20, 2020. http://users.monash.edu.au/~smusgrav/Nunggubuyu/.