Open Text Collections

Sebastian Nordhoff
Corpus Glosés: de la construction à l’exploitation automatique
2023-06-28
Electronic Reference Grammars for Typology: Challenges and Solutions

Sebastian Nordhoff
University of Amsterdam

Electronic publication offers new possibilities for the creation and exploration of grammatical descriptions. This paper lists values influencing the structure of electronic grammatical descriptions. It then investigates challenges and solutions for a grammar authoring software trying to adhere to these values in the domains of data quality, creation of the description, and exploration of the description. The paper closes by discussing possibilities for the standardization of grammatical descriptions on a macroscopic level, complementing the standardization efforts on a more fine-grained level like GOLD or
Boasian Trilogy: publication options

1. **Dictionaries**: many outlets
2. **Grammatical descriptions**: many outlets
3. **Text collections**: no significant outlets
Existing platforms

› **TILA:** [https://www.americanlinguistics.org/?page_id=1830](https://www.americanlinguistics.org/?page_id=1830) (closed access)

› **pangloss:** [https://pangloss.cnrs.fr/](https://pangloss.cnrs.fr/) (eclectic, manual workflow)

› **doreco:** [https://doreco.huma-num.fr](https://doreco.huma-num.fr) (transcribed audio, finished project?)
Open Text Collections

Platform to start in 2023

- open
- prestigious
- interoperable
- start 2023-09-15
Guiding principles

- **texts**, not audio
- **edited**, not naturalistic
- **curated**, not eclectic
- **peer reviewed**, not legacy or opportunistic

- **availability first**: no paywalls, registration, waiting times, click rallyes
- **community**
- **openness**
- **prestige**: seasoned editors
For a project at ZAS Berlin, I analyzed 20,000 ELAN files culled from various endangered language archives.

Take home message: everybody does what they want with ELAN.

Tier names and structures vary wildly.

It is possible to build an API to access glossed material in language archives, but it is painful.
 Backend

- CSV for the web (CSVW)
- text-based
- simple
- extensible
- easy to version
CSVW

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tire la bobinette,</td>
<td>pull.IMP DEF.F.SG bobbin DEF.F.SG</td>
<td>Pull the bobbin and the latch will fall.</td>
<td>stan1290</td>
</tr>
<tr>
<td>la chevillette</td>
<td>latch fall-FUT.3SG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cher-ra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- three obligatory columns
  - vernacular
  - glosses
  - translation
- can be expanded by further columns as necessary
- Leipzig Glossing Rules
- correspondences and constraints between cells in the same row
Output formats

- dump (csv, json-ld, nq, rdf)
- pdf
  - printed pdf = book
- html
- query interface
  - see https://imtvault.org/?q=coconut
Workflow

- every text collection is prepared on GitHub
- DOI via the GitHub-Zenodo bridge
- review via GitHub issues
- collections approved by the regional boards get a new release (1.0), archived on Zenodo, and are accepted in the relevant Zenodo communities.
Quality assurance

› technical review (automated, probably via GitHub actions)
  › number of words match between vernacular and glosses
  › number and type of morphemes match between vernacular and glosses
  › abbreviations are either Leipzig Glossing Rules or listed in separate document

› content review
  › regional boards with area specialists
  › precise setup to be debated
  › can use GitHub Issues
Data model
› **findable**: registered, linked, metadata

› **accessible**: no paywalls

› **interoperable**: many different well-defined and open formats

› **reusable**: open license
3 year grant from DFG, 2023-2026, 1.5 FTE
after that consortial funding via Language Science Press
Mandana Seyfeddinipur, Director ELDP/ELAR at BBAW

Christian Döhler, Papuanist, holder of the Gabelentz Award for the best published grammar (A grammar of Komnzo)

Sebastian Nordhoff, managing director for Language Science Press, extensive experience in Open Access publishing and Linked Data

student assistants
Regional boards

- **Oceania** Christian Döhler, Kilu von Prince (Düsseldorf)
- **Africa** Alena Witzlack-Makarevich (Jerusalem), Jeff Good (Buffalo)
- **Eurasia** Michael Rießler (Joensuu)
- **South America** Matt Coler (Groningen), Nick Emle (Groningen)
- **Caucasus** Diana Forker (Jena)

Further board probably via cooption
Planned text collections

- Komnzo
- Bine
- Daakaka
- Dalkalaen
- Muylaque Aymara
- Iquito
- Two Amazonian languages
- Kawesqar
- Hinuq
- Sanzhi
- Chirag Dargwa
- Tabasaran
- Gawarbati
- Palula
- Saek
Thank you for your attention.

http://opentextcollections.org
Mastodon https://fedihum.org/@otc
https://twitter.com/OpenTextColl