Combining description and documentation: a digital Boasian trilogy

Florian Matter

Institute of Linguistics, University of Bern

👇 Go to page 2 to skip boring background 👇

Language description

- grammatical descriptions / "grammars": structural description of a language as a system
- long-standing tradition in linguistics, with beginning of structuralism
- expanding knowledge of what is possible cross-linguistically, but also capturing aspects of a single language
- urgent, given extinction rate of (undescribed) languages

Natural speech in language description

- elicited material easier to get and process
- but it is much less naturalistic
- alternative approach: collect more natural utterances, such as conversations or narratives
- collection serves as text corpus, provides illustrative examples
- so-called "Boasian trilogy": **grammar, dictionary, texts** (Evans & Dench 2006: 10–16)

Recorded language and language description

- both technological possibilities and mutual accessibility of researchers and speech communities have increased
- audio and later video recording is logical next step
- results in lasting record of described language
- transcription for text corpus is possible at later stage :)

Language documentation (Himmelmann 1998)

- recording, transcription and translation as separate process from grammatical analysis
- annotated speech corpus as separate product from illustrations in grammar
- different methodological issues than grammatical analysis, but intimately connected

Usual interplay of linguistic description and documentation

- corpora are gathered as part of fieldwork and are transcribed, translated, and annotated
- they serve as naturalistic source of information for grammatical analysis and enable replicability
- proper annotation and translation rely on linguistic analysis
- often done by same person(s), but entirely distinct final products:

Boasian trilogy

grammar, dictionary, texts

published as monograph(s)

vs

"Himmelmannian trilogy" (Good 2018)

recordings, metadata, grammatical annotation

ideally deposited in archive (e.g. ELAR)

Try it yourself: linking products of language description and documentation

- high-quality typologically informed grammar of a Cariban language: Cáceres (2011)
- based on extensive speech corpus, contains six glossed texts in appendix
- good ELAR corpus (Cáceres 2014)

"Homework":

1. find random example sentence from corpus in PDF (available here)
   - will have sources like `ConvChur.016:Anl`
2. download corresponding text from ELAR corpus (link)
3. locate audio snippet corresponding to example sentence

- how easily did you find the raw data used as an example?

CLLD: a linked data approach to linguistics

- cross-linguistic linked data (Forkel et al. 2019)
- was created for WALS Online (Dryer & Haspelmath 2013)
- linguistic databases
  - relatively easy to set up and curate
  - citable
  - consistent
- see `https://clld.org/datasets.html` for examples

CLDF

- crosslinguistic linked data format (Forkel et al. 2017), externalized model from CLLD
- standardized data format for (cross-)linguistic data
  - wordlists
  - structural information
  - simple dictionaries
  - cognate set collection
  - …or anything else!
- human-readable formats:
  - data in `.csv` files (comma-separated values)
  - metadata as JSON (JavaScript Object Notation)
- collaborative data curation via Git

CLDF and CLLD combined

- data curation and dissemination in CLDF
  - lightweight
  - no proprietary format
  - easily parseable
  - publish digitally on e.g. Zenodo
  - example: PHOIBLE CLDF Dataset
- CLLD database built from CLDF dataset
  - for humans (users / readers)
  - visual representation of data
  - interactive and explorative
  - example: PHOIBLE

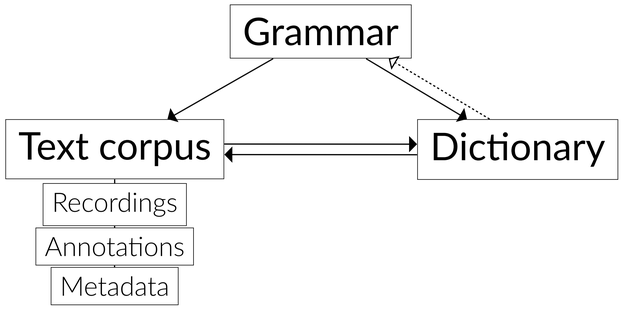
Combining description and documentation: a digital Boasian trilogy

Florian Matter

Institute of Linguistics, University of Bern

Introducing GRAMR

- digital grammar, with text corpus and (basic) dictionary
- currently a prototype!
- application of CLDF and CLLD frameworks to close gap between language documentation and description
- text corpus has audio and metadata
- everything is linked (almost):



Check out the online demo!

`https://florianmatter.gitlab.io/gramr`

Some impressions

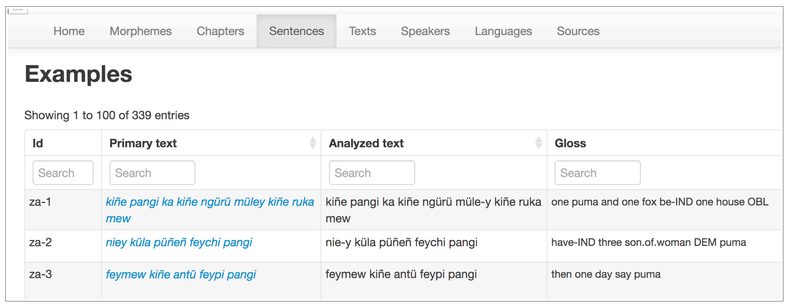


Figure 1. Browsing sentences

Some impressions

Figure 2. Sentence with dictionary links and audio

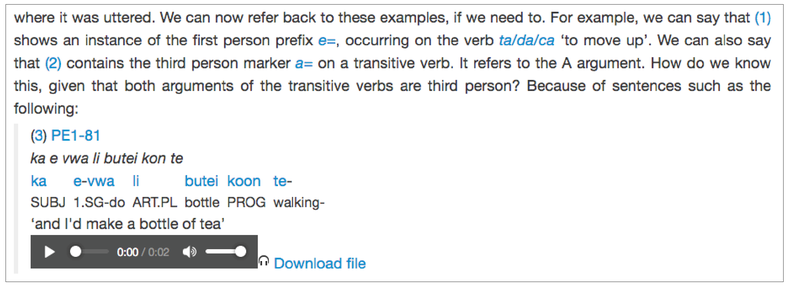


Figure 3. Dictionary links and sentence with audio in running text

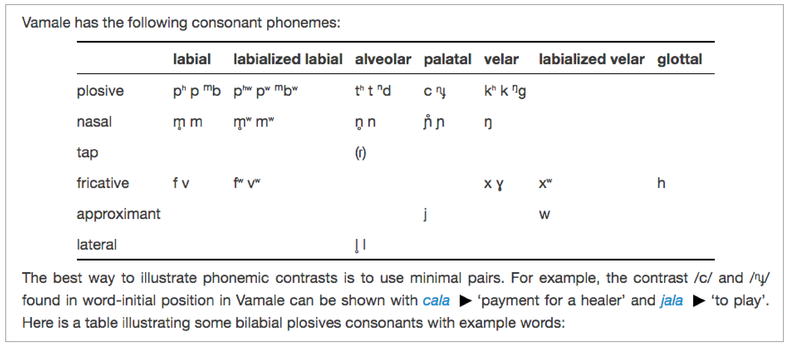


Figure 4. A table

Behind the scenes

Currently employed workflow:

1. recording
2. transcription & translation in ELAN
3. export as `.flextext`
4. import to FLEx
5. annotate in FLEx
6. export corpus as `.flextext`
7. export dictionary as LIFT
8. convert both to CLDF
9. provide grammatical description as `.txt`
10. provide metadata as `.csv`
11. feed CLLD app

- converting to CLDF from other formats should be feasible
- needed information:
  - example ID
  - surface line
  - object line
  - glossing line
  - translation line
  - speaker ID
  - text ID + internal number
  - time stamps
  - dictionary morpheme IDs

ELAN to FLEx to CLDF

- existing workflow for exporting from ELAN with translations, time stamps, speaker IDs… (Gaved & Salffner 2014; Visser 2019):
  - ELAN metadata are already contained in `.flextext`
  - are kept when re-exported from FLEx, but not visible in app
- I have created CLDFLex
  - set of scripts for converting to and from CLDF and FLEx
  - for both dictionaries and texts

Combining description and documentation: a digital Boasian trilogy

Florian Matter

Institute of Linguistics, University of Bern

Author input

- grammatical description is written as simple `.txt` files
- combination of conventional Markdown and custom directives for grammar writing
- custom markdown directives:

Markdown	Result
`obj:<string>`	*form*
`'<string>'`	'meaning'
`morph:<id>`	*morpheme*
`morph_a:<id>`	*morpheme* ▶
`crossref:<id>`	Section title
`src:<bibkey>[<page>]`	Author 2004: 34
`psrc:<bibkey>[<page>]`	(Author 2004: 34)
`ex:<id>`	(example from corpus)
`exref:<id>`	(1)

- providing metadata:
  - simple `.csv` files for...
    - texts
    - speakers
    - chapter/section titles & structure
    - additional info for single dictionary entries & back references to grammar
- `.txt` and `.csv` are lightweight, non-proprietary, easy to create edit, and convert to from other formats

ID	Title	Description	Filename	Type
za	The clever fox	An epew, taken from src:zuniga2006habla[268--282].	clever_fox.wav	Myth
fc	Fake conversation	A fake conversation without audio.		Conversation
djp	Kalamang pear story	A pear story	pear_djusman.wav	Story

Figure 5. The text table

Advantages

- easily navigable due to omnipresent hyperlinks
- accountability and reproducibility of of presented analysis, testable against corpus
- comparative ease of using non-elicited examples
- examples containing specific forms or meanings easily findable
- increased ease of navigability of corpus during documentation and grammar writing
  - stronger engagement with primary data and more adequate analysis
- more accessible for laypeople
- enforced consistency of analysis (segmentation and glossing in grammar and corpus is identical)
- automatic availability of lexical data in CLDF format for quantitatively oriented comparative linguists

Technical issues

- FLEx export is really messy
- no other input formats supported at the moment
- custom markdown is rather ad-hoc, proper ontology needed
- no user-friendly interface for authors
- tables are the worst

Practical issues

- public visibility of texts (& recordings) is problematic, levels of accessibility should be definable
- who wants to transcribe and gloss **all** their material?!
- digital literacy: collaborative efforts via Git, not wiki-like
- who wants to fund research infrastructure?

References

Cáceres, Natalia (2011). "Grammaire fonctionnelle-typologique du ye'kwana, langue caribe du Venezuela". Ph.D. dissertation. Université Lumière Lyon 2.

– (2014). "Documentation of Ye'kwana in the Erebato river in Venezuela". Online: https://elar.soas.ac.uk/Collection/MPI154072 (visited on 10/14/2020).

Dryer, Matthew S. & Martin Haspelmath, eds. (2013). "WALS Online". Online: https://wals.info/ (visited on 05/20/2020).

Evans, Nicholas & Alan Dench (2006). "Introduction: Catching language". In: *Catching language: The standing challenge of grammar writing*. Felix K. Ameka, Alan Charles Dench & Nicholas Evans (eds.). Berlin: de Gruyter: 1–40.

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). "cldf: a toolkit for cross-linguistic databases". DOI: 10.5281/zenodo.3239095. Online: https://doi.org/10.5281/zenodo.3239095.

Gaved, Tim & Sophie Salffner (2014). *Working with ELAN and FLEx together: an ELAN-FLEx-ELAN teaching set*. URL: https://www.soas.ac.uk/elar/helpsheets/file122785.pdf.

Good, Jeff (2018). "Reflections on the scope of language documentation". In: *Reflections on Language Documentation: 20 Years after Himmelmann 1998*. Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.). Honolulu: University of Hawai'i Press: 13–21.

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

Matter, Florian (2020a). "gramr (source code)". Online: https://gitlab.com/florianmatter/gramr/.

– (Aug. 2020b). "Integrating grammatical description, text collection and dictionary: Language documentation and description for the digital age". SLE 2020, Online.

Visser, Eline (Oct. 2019). "All field data in one place: how to make ELAN and FLEx compatible and truly search your whole corpus". UPPLADOC Workshop on language documentation: multilingual settings and technological advances, University of Uppsala.