

Combining language description and documentation

a digital Boasian trilogy

Florian Matter
Department of Linguistics
University of Bern

2020-10-01

Webinar
General Linguistics
University of Lund

Table of Contents

- 1 Language description & documentation
- 2 CLLD and CLDF: a linked data approach to linguistics
- 3 GRAMR

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
- capturing aspects of a single language
- urgent, given extinction rate of (undescribed) languages

A grammar of...

1. The Ye'kwana and their language
2. Phonetics and phonology
3. Morphology – lexical categories
4. Morphology – person markers and associated prefixes
5. Morphology – tense, aspect, mood
6. Lexical strategies of spatial meanings
7. Syntax of simple phrases
8. Syntax of complex phrases

“Grammaire fonctionnelle-typologique du ye'kwana, langue caribe du Venezuela” (Cáceres 2011)

Language description and natural speech

- elicited material easier to get and process
- ...but less naturalistic
- instead: collect (more) natural utterances (conversations, narratives...)
- serves as text corpus, provides illustrative examples
- “Boasian trilogy” (Evans & Dench 2006: 10–16):
grammar, dictionary, texts

Language description and recording

- increasing technological advances & ease of access to speech communities
- audio and later video recording logical next step
- lasting record of described language
- transcription possible at later stage :)

Language documentation (Himmelman 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**

Language description and documentation

- corpora gathered as part of fieldwork
- transcribed, translated, and annotated
- serve as naturalistic source of information for grammatical analysis
- enables replicability thereof
- proper annotation & translation relies on linguistic analysis
- often done by same person
- **but entirely distinct final products**

Products of language description and documentation

1. Boasian trilogy: grammar, dictionary, texts
 - published as one or multiple books (or PDF...)
2. “Himmelmanian trilogy” (Good 2018): recordings, metadata, grammatical annotation
 - (ideally) deposited in archive (e.g. ELAR)

Products of language description and documentation: an example

- “Grammaire fonctionnelle-typologique du ye'kwana, langue caribe du Venezuela” (Cáceres 2011)
- distributed as [PDF file](#)
- contains six glossed texts (!)
- rest of corpus in [ELAR archive](#)

Table of Contents

- 1 Language description & documentation
- 2 **CLLD and CLDF: a linked data approach to linguistics**
- 3 GRAMR

CLLD

- cross-linguistic linked data (Forkel et al. 2019)
- was created in the course of publication of WALS Online (Dryer & Haspelmath 2013)
- linguistic databases
 - “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
 - ...
- human-readable formats:
 - data in CSV files (comma-separated values)
 - metadata as JSON (JavaScript Object Notation)
- collaborative data curation via [Git](#)

CLDF and CLLD

- 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](#)

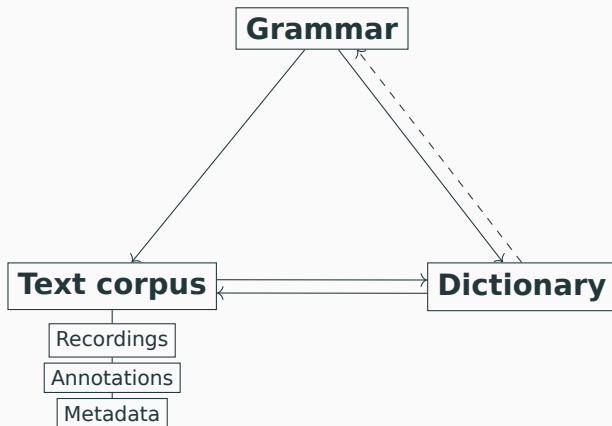
Table of Contents

- 1 Language description & documentation
- 2 CLLD and CLDF: a linked data approach to linguistics
- 3 GRAMR**

GRAMR

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

GRAMR: combination of components



GRAMR: **demo**

online demo

Behind the scenes

current 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

walwal - FieldWorks Language Explorer

File Send/Receive Edit View Data Insert Format Tools Parser Window Help

Texts & Words Texts

Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics

Title Kiriyime pen waytopon

Text

Title wal Kiriyime pen waytoponho
Eng The story of Kiriyime's death

Info | Baseline | Gloss | Analyze | Tagging | Print View | Text Chart

wal

3 Word

Morphemes	nejamronhiri	mokjatkepe	hara
Lex. Entries	nejamro	moku	hara
Lex. Gloss	3PRO.PL	moku	hara
Lex. Gram. Info.	<Not Sure>	-jatkepe	again
Word Gloss	3PRO.PL	-jatkepe	again
Word Cat.	<Not Sure>	REM.PL	***

Free They who went came back again.

Note Eng 221

wal

4 Word

Morphemes	kir'wanhe	so	rma	mokjatkepe
Lex. Entries	kir'wanhe	so	rma	moku
Lex. Gloss	good	PL	still	come
Lex. Gram. Info.	<Not Sure>	<Not Sure>	<Not Sure>	<Not Sure>
Word Gloss	good	PL	self	come
Word Cat.	***	***	***	***

Free They were well when they came.

Note Eng 221

wal

5 Word

Morphemes	jihŋ'iri	me	ka	mokjakpe
Lex. Entries	jihŋ'i	-ri	me	kapu
Lex. Gloss	jihŋ'i	-ri	me	kapu
Lex. Gram. Info.	begin	NMLZ	ADVZ	for.now
Word Gloss	begin	NMLZ	ADVZ	for.now
Word Cat.	begin	NMLZ	ADVZ	for.now

25/Sep/2020 Queue: (-/-) No Parser Loaded 2/2

ELAN and FLEx

- workflow for exporting from ELAN with translations, time stamps, speaker IDs... (Gaved & Salfner 2014; Visser 2019)
- metadata already in `.flectext`
- is kept when re-exported from FLEx, but not visible in app

FLEx and CLDF

- **CLDFLex**, set of scripts for converting to and from CLDF and FLEx
- for both dictionaries and texts

GRAMR: **grammatical description**

- written as simple `.txt` files
- combination of conventional **Markdown** and custom directives for grammar writing

Custom markdown directives

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

Providing metadata

- simple CSV files for...
 - texts
 - speakers
 - chapter/section titles & structure
 - additional info for single dictionary entries & back references to grammar

Advantages

- easily navigable due to omnipresent hyperlinks
- accountability and reproducibility 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

Advantages

- enforced consistency of analysis (segmentation and glossing in grammar and corpus is identical)
- automatic availability of lexical data in CLDF format for (quantitative) historical linguists






Open technical issues

- FLEx export is really messy
- no other input formats supported at the moment (but should not be too hard)
- custom markdown is messy
- 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
- funding

References

-  Cáceres, Natalia (2011). “Grammaire fonctionnelle-typologique du ye’kwana, langue caribe du Venezuela”. Ph.D. dissertation. Université Lumière Lyon 2.
-  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](https://doi.org/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](https://doi.org/10.5281/zenodo.3239095). Online: <https://doi.org/10.5281/zenodo.3239095>.

References

-  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.
-  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.

Thank you!