lexibank

progress report

Robert Forkel[1], Johann-Mattis List[2] and Simon Greenhill[1]

Jena, 5th Glottobank workshop

[1] Max Planck Institute for the Science of Human History
[2] CRLAO/EHESS and Équipe AIRE/UPMC, Paris

Outline

1. `glottobank/lexibank-data`
2. The `lexibank` web app
3. Open Questions

glottobank/lexibank-data

https://github.com/glottobank/lexibank-data

The public GitHub repository `glottobank/lexibank-data` provides

- a workbench to curate data
- an API to access the data
- a python package `pylexibank` wrapping the API.

`lexibank` Datasets

Data in `glottobank/lexibank-data` is organized in **datasets** –

- self-contained,
- citeable,
- internally homogeneous

units – comprising

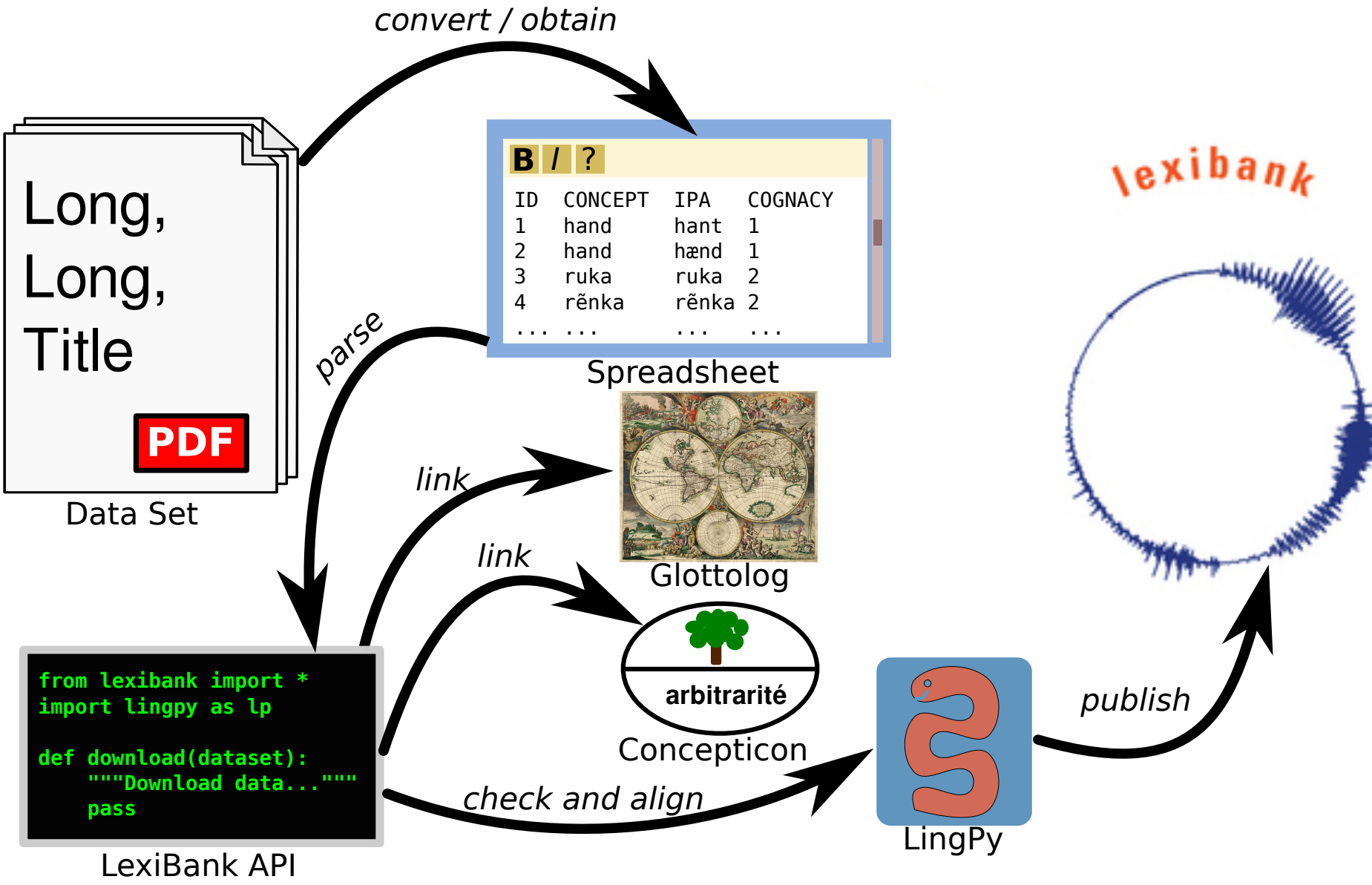
- metadata
- raw data (or a method to retrieve the raw data)
- code to convert the raw data to lexibank's **CLDF** format.

`lexibank` Standards

`lexibank` uses

CLDF	as common data format across datasets
Glottolog	as reference catalog for language/variety identification
Concepticon	as reference catalog for semantic concepts across datasets
CLPA	as reference for transcriptions
LingPy	as a less strict reference for transcriptions and to automatize cognate sets and alignments where they are missing

lexibank Workflow



`lexibank` Data Model

- The core data item in `lexibank` are wordlist items, i.e. triples `(Language,Concept,Form)`.
- These items can be extended with additional attributes, e.g. segmentation, alternative orthographies, etc.
- Cognate sets can be encoded as lists of cognacy judgements, relating a wordlist item to a cognate set.
- Cognacy judgements can include an alignment.

So what?

This infrastructure allows us to

- implement methods on one dataset
- ...and effortlessly apply them to others
- implement quality metrics taking edge cases into account

Good Practices in Scientific Computing

Good Enough Practices in Scientific Computing

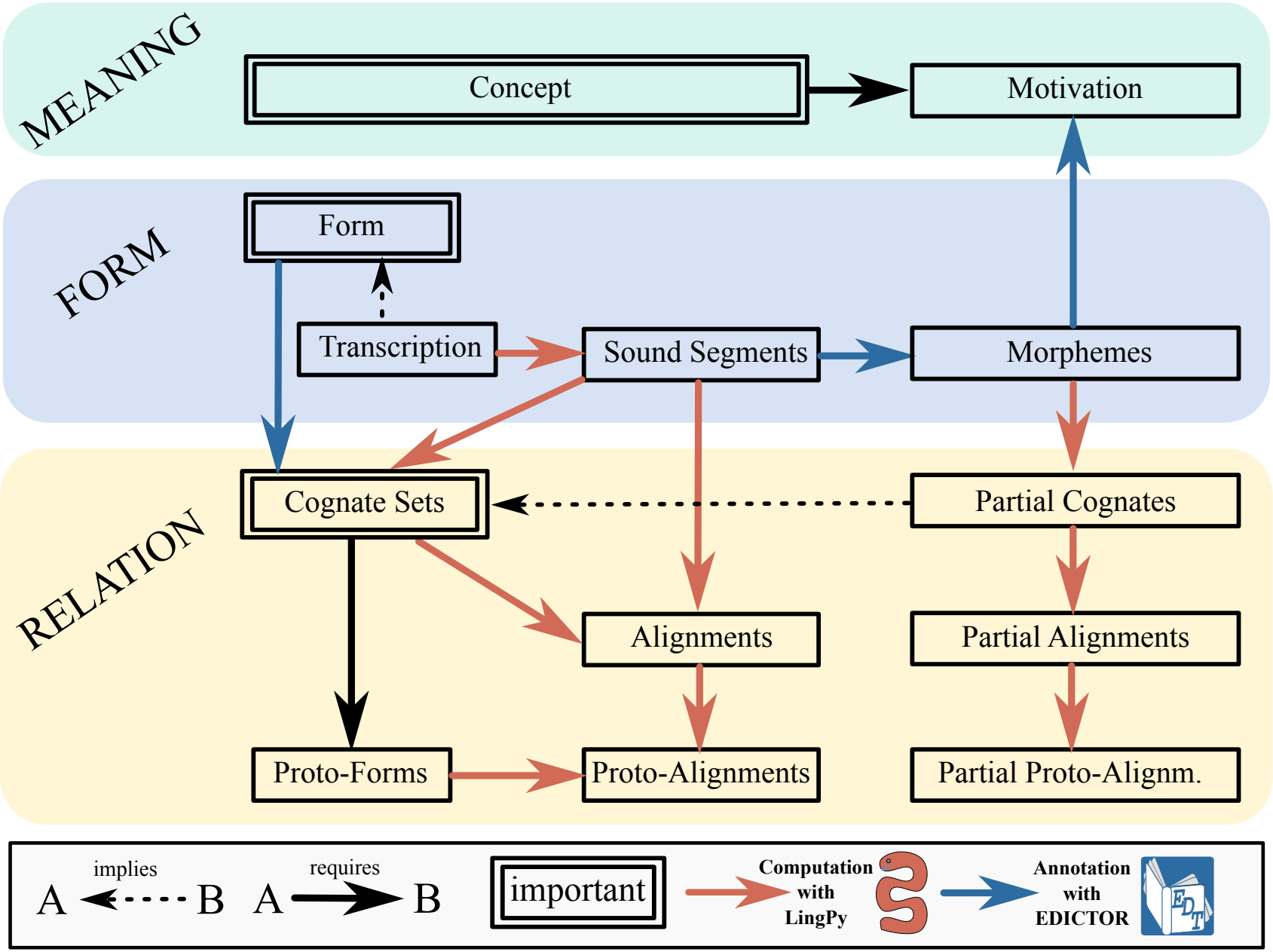
Greg Wilson[1,‡*], Jennifer Bryan[2,‡], Karen Cranston[3,‡], Justin Kitzes[4,‡],
Lex Nederbragt[5,‡], Tracy K. Teal[6,‡]

`lexibank-data` implements all recommendations from a recent paper on best practices in scientific computing:

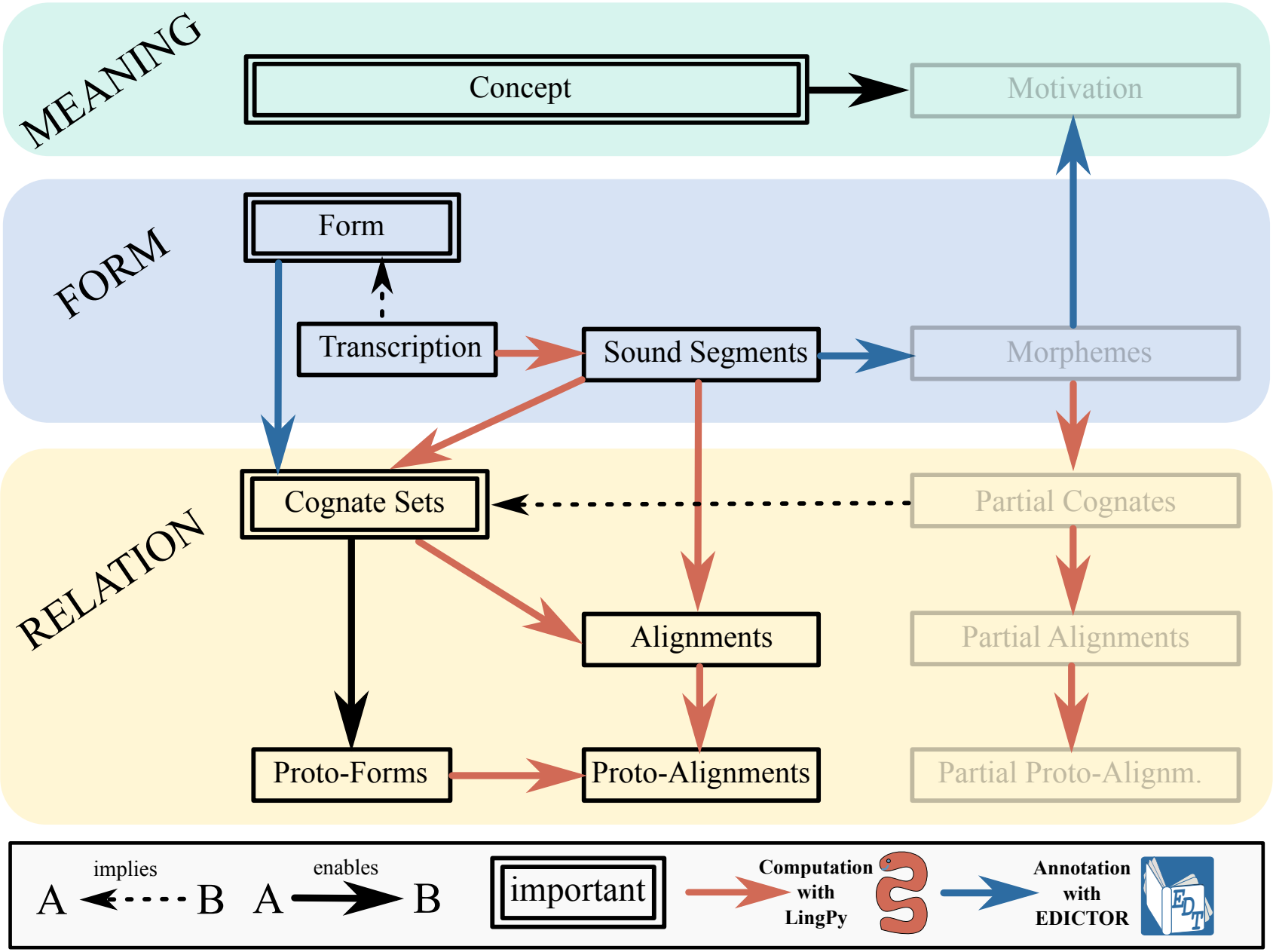
- ✓ Save the raw data: a dataset's `download` function.
- ✓ Create the data you wish to see in the world: a dataset's `cldf` function
- ✓ Create analysis-friendly data
- ✓ Record all the steps used to process data: the `lexibank` workflow and a dataset's python module
- ✓ Anticipate the need to use multiple tables: Baked into CLDF.
- ✓ Submit data to a reputable DOI-issuing repository so that others can access and cite it: See "Open Questions" below.

Thus providing a significant service to the field.

Examples: CLDF Annotation Hierarchy



Examples: CLDF Annotation Hierarchy



Examples: CLDF/**lexibank** Quality Metrics

Bai Dialect Survey

Cite the source dataset as

> Allen, Bryan. 2007. Bai Dialect Survey. SIL International.

Available online at http://www.sil.org/resources/publications/entry/9121

Statistics

Glottolog 100% | Concepticon 100% | Source 100% | LingPy 96% | CLPA 92%

- **Varieties:** 9
- **Concepts:** 499
- **Lexemes:** 4,493
- **Synonymy:** 1.00
- **Cognacy:** 0 cognates in 0 cognate sets
- **Invalid lexemes:** 0
- **Tokens:** 20,092
- **Segments:** 97 (4 LingPy errors, 8 CLPA errors, 4 CLPA modified)
- **Inventory size (avg):** 54.33

Examples: CLDF/**lexibank** Quality Metrics

Cognates in the Bai Dialect Survey

Cite the source dataset as

> List, Johann-Mattis. 2016. Cognates in Bryan Allen's Bai Dialect Survey.

Available online at +++pending+++

Statistics

Glottolog 100% | Concepticon 100% | Source 100% | LingPy 96% | CLPA 92%

- **Varieties:** 9
- **Concepts:** 499
- **Lexemes:** 4,493
- **Synonymy:** 1.00
- **Cognacy:** 3,846 cognates in 671 cognate sets
- **Invalid lexemes:** 0
- **Tokens:** 20,092
- **Segments:** 97 (4 LingPy errors, 8 CLPA errors, 4 CLPA modified)
- **Inventory size (avg):** 54.33

Examples: CLDF/**lexibank** Quality Metrics

Austronesian Basic Vocabulary Database

Cite the source dataset as

> Greenhill, S.J., Blust. R, & Gray, R.D. (2008). The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. Evolutionary Bioinformatics, 4:271-283.

This dataset is licensed under a https://creativecommons.org/licenses/by/4.0/ license

Available online at http://language.psy.auckland.ac.nz/austronesian/

Statistics

Glottolog 95% | Concepticon 100% | Source 84% | LingPy 78% | CLPA 28%

- **Varieties:** 1,367
- **Concepts:** 210
- **Lexemes:** 266,128
- **Synonymy:** 1.14
- **Cognacy:** 191,597 cognates in 14,854 cognate sets
- **Invalid lexemes:** 10
- **Tokens:** 1,346,295
- **Segments:** 1,577 (344 LingPy errors, 1141 CLPA errors, 184 CLPA modified)
- **Inventory size (avg):** 36.53

Examples: CLDF/**lexibank** Quality Metrics

Austroasiatic dataset for phylogenetic analysis

Cite the source dataset as

> Sidwell, Paul 2015, Austroasiatic dataset for phylogenetic analysis: 2015 version, Mon-Khmer Studies: a journal of Southeast Asian Languages and cultures, vol. 44, pp. lxviii-ccclvii.

This dataset is licensed under a https://creativecommons.org/licenses/by-nc-sa/4.0/ license

Available online at http://dx.doi.org/10.5281/zenodo.34092

Statistics

Glottolog 99% | Concepticon 100% | Source 100% | LingPy 92% | CLPA 41%

- **Varieties:** 122
- **Concepts:** 200
- **Lexemes:** 47,992
- **Synonymy:** 1.07
- **Cognacy:** 19,364 cognates in 2,944 cognate sets
- **Invalid lexemes:** 1
- **Tokens:** 198,340
- **Segments:** 881 (67 LingPy errors, 520 CLPA errors, 118 CLPA modified)
- **Inventory size (avg):** 54.89

Examples: Detailed Transcription Reports

Lexicostatistic Wordlist of Semitic Languages

Cite the source dataset as

> Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Andrew Kitchen, Christopher Ehret, Shiferaw Assefa, Connie J. Mulligan. Proc. R. Soc. B 2009 -; DOI: 10.1098/rspb.2009.0408. Published 29 April 2009

See also http://rspb.royalsocietypublishing.org/content/early/2009/04/27/rspb.2009.0408

Statistics

Glottolog 89% | Concepticon 100% | Source 0% | LingPy 91% | CLPA 57%

- **Varieties:** 25
- **Concepts:** 95
- **Lexemes:** 4,484
- **Synonymy:** 1.00
- **Cognacy:** 1,731 cognates in 322 cognate sets
- **Invalid lexemes:** 0
- **Tokens:** 19,888
- **Segments:** 190 (18 LingPy errors, 82 CLPA errors, 37 CLPA modified)
- **Inventory size (avg):** 43.16

Examples: Detailed Transcription Reports

Detailed transcription record

Segments

No	Segment	Occurrence	LingPy	CLPA
1	a	1097	✓	✓
2	ɛ	730	✓	✓
3	r	529	✓	✓
4	s	525	✓	✓
5	n	522	✓	✓
6	m	457	✓	✓
7	t	428	✓	✓
8	ɨ	412	✓	✓
9	k	387	✓	✓
10	b	320	✓	✓

Examples: Detailed Transcription Reports

184	sʸ	1	✓	✓
185	ç	1	✓	✓
186	ʟ	1	✓	✓
187	oa	1	✓	?
188	zh	1	✓	?
189	sh	1	✓	✓
190	[illegible]	1	?	?
191	aˡ	1	✓	?
192	oɐi	1	✓	?

Examples: Detailed Transcription Reports

Words

No	ID	LANGUAGE	CONCEPT	VALUE	SEGMENTS
1	Kitchen2012-1	Ge'ez	All	kʷɨllu	kʷ ɨ ll u
2	Kitchen2012-2	Tigre	All	kɨllu	k ɨ ll u
3	Kitchen2012-3	Tigrinya	All	kullu	k u ll u
4	Kitchen2012-4	Amharic	All	hullu	h u ll u
5	Kitchen2012-5	Argobba	All	diyyu	d iyyu
6	Kitchen2012-6	Harari	All	kulluzo:m	k u ll u z o: m
7	Kitchen2012-7	Zway	All	hullɨn	h u ll ɨ n
8	Kitchen2012-8	Walani	All	ulllmka	u ll l m k a

The **lexibank** web app

http://lexibank.clld.org

The `lexibank` web app provides

- a browsable catalog of (releases of) the data in `glottobank/lexibank-data`
- a showcase of what can be built on top of standardized data
- a platform for extended visualizations of the data, e.g. to explore colexifications, or to inspect multiple alignments

lexibank contains more than **1,200,000** lexical items from **4,871** Glottolog languages from **211** families.

http://lexibank.clld.org: Data Sets

Id	Name	Cite	# languages	# concepts	# lexemes
Search	Search	Search	Search	Search	Search
ids	Intercontinental Dictionary Series	Key, Mary Ritchie & Comrie, Bernard (eds.) 2015. The Intercontinental Dictionary Series. Leipzig: Max Planck Institute for Evolutionary Anthropology.	279	1305	442,555
abvd	Austronesian Basic Vocabulary Database	Greenhill, S.J., Blust. R, & Gray, R.D. (2008). The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. Evolutionary Bioinformatics, 4:271-283.	784	210	253,812
numerals	Numeral Systems of the World's Languages	Eugene Chan 2016. Numeral Systems of the World's Languages. Jena: Max Planck Institute for the Science of Human History.	3730	32	120,702
transnewguinea	TransNewGuinea.org	Simon Greenhill (2015) TransNewGuinea.org: An Online Database of New Guinea Languages. PLoS ONE 10(10): e0141563. doi:10.1371/journal.pone.0141563	758	238	100,197
wold	The World Loanword Database	Haspelmath, Martin & Tadmor, Uri (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wold.clld.org	41	1457	64,408
huntergatherer	Hunter - Gatherer Language Database	Bowern, Claire, Patience Epps, Jane Hill, and Patrick McConvell. Hunter - Gatherer Language Database. https://huntergatherer.la.utexas.edu/ Accessed[date].	226	184	38,778
grollemundbantu	Grollemund Bantu Database	Grollemund et al. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals	330	100	37,722
mayanbvd	Mayan Basic Vocabulary Database	Simon Greenhill 2016. Mayan Basic Vocabulary Database	31	729	33,554
huberandreed	Dataset of Huber and Reed's "Comparative Vocabulary"	Cysouw, M. and Prokić, J. and Bouda, P. and Moran, S. (2012): Dataset of Huber and Reed's "Comparative Vocabulary". Marburg: Philipps-University Marburg.	67	348	26,492
sidwellaustroasiatic	Austroasiatic dataset for phylogenetic analysis	Sidwell, Paul 2015, Austroasiatic dataset for phylogenetic analysis: 2015 version, Mon-Khmer Studies: a journal of Southeast Asian Languages and cultures, vol. 44, pp. lxviii-ccclvii.	100	200	23,789
cogdetbench	Cognate Detection Benchmark	List, J.-M. Sequence comparison in historical linguistics. Düsseldorf: Düsseldorf University Press. 2014.	72	747	22,887

Figure 1: Some of the datasets in lexibank

http://lexibank.clld.org: Cognate Sets

lexibank - - Mozilla Firefox

lexibank -

localhost:6543/cognatesets/sidwellaustroasiatic-fire-1428

lexibank Home Datasets Concepts Languages Cognatesets Sources

Cognate set sidwellaustroasiatic-fire-1428

From dataset Austroasiatic dataset for phylogenetic analysis cite as

Sidwell, Paul 2015, Austroasiatic dataset for phylogenetic analysis: 2015 version, Mon-Khmer Studies: a journal of Southeast Asian Languages and cultures, vol. 44, pp. lxviii-ccclvii.

34 Counterparts

Search:

Counterpart	Language	Concept	certain	alignment	alignment method	loan
ʔɔs	Chewong [Cheq-Wong]	FIRE	yes	- - - - ʔ ɔ - s -	SCA-library	no
ʔɔs	Jah Hut [Jah-Hut]	FIRE	yes	- - - - ʔ ɔ - s -	SCA-library	no
ʔɔs	Jehai [Jahai]	FIRE	yes	- - - - ʔ ɔ - s -	SCA-library	no
met ʔɔs	Kensiu [Kensiw]	FIRE	yes	m e t ʔ ɔ - s -	SCA-library	no
ʔus	Besisi [Mah-Meri]	FIRE	yes	- - - - ʔ u - s -	SCA-library	no
ʔɔs	Minriq [Maniq]	FIRE	yes	- - - - ʔ ɔ - s -	SCA-library	no
ʔɔ:s	Semai [Semai]	FIRE	yes	- - - - ʔ ɔ: - s -	SCA-library	no
ʔus	Semelai [Semelai]	FIRE	yes	- - - - ʔ u - s -	SCA-library	no
ʔo:s	Semnam [Semnam]	FIRE	yes	- - - - ʔ o: - s -	SCA-library	no
ʔo:s	Temiar [Temiar]	FIRE	yes	- - - - ʔ o: - s -	SCA-library	no
ʔolh	Cua [Cua]	FIRE	yes	- - - - ʔ o - lh -	SCA-library	no
ʔous	Koho [Sre]	FIRE	yes	- - - - ʔ o u s -	SCA-library	no
u̯:ih	Western Bru [Bru-QuangTri]	FIRE	yes	- - - - - u̯: i h -	SCA-library	no
ʔujh	Ong [Bru-Talan]	FIRE	yes	- - - - ʔ u - jh -	SCA-library	no
ʔu:s	Eastern Bru [Bru-Ubon]	FIRE	yes	- - - - ʔ u: - s -	SCA-library	no
ʔujh	Ir [In]	FIRE	yes	- - - - ʔ u - jh -	SCA-library	no
ʔujh	Kataang [Katang]	FIRE	yes	- - - - ʔ u - jh -	SCA-library	no

Figure 2: Cognate sets can be displayed including alignments.

http://lexibank.clld.org: Colexifications

lexibank - - Mozilla Firefox

lexibank.clld.org/parameters/1673.colexifications.html

lexibank Home Datasets Concepts Languages Cognatesets Sources

Colexifications for concept ARM

Concepts which are colexified with *ARM* in languages of more than one family are highlighted

Search:

lexeme	concept	lexeme	concept	language	family	dataset
pɛ̃	ARM	pɛ̃	MANY	Kaingang	Nuclear-Macro-Je	Intercontinental Dictionary Series
dob	ARM	dob	ONE	Gobasi	East Strickland	TransNewGuinea.org
dob	ARM	dob	ONE	Gobasi	East Strickland	TransNewGuinea.org
cap	ARM	cap	TWO	Jakalteko, Western	Bookkeeping	Mayan Basic Vocabulary Database
cab	ARM	cab	TWO	Mam	Mayan	Mayan Basic Vocabulary Database
cab	ARM	cab	TWO	Mam	Mayan	Mayan Basic Vocabulary Database
di	ARM	di	FISH	Odoodee	East Strickland	TransNewGuinea.org
mès	ARM	mès	DOG	Meyah	East Bird's Head	TransNewGuinea.org
mès	ARM	mès	DOG	Meyah	East Bird's Head	TransNewGuinea.org
mari	ARM	mari	LEAF	Dedua	Nuclear Trans New Guinea	TransNewGuinea.org
more	ARM	more	LEAF	Migabac	Nuclear Trans New Guinea	TransNewGuinea.org
met	ARM	met	LEAF	Sialum	Nuclear Trans New Guinea	TransNewGuinea.org
me	ARM	me	LEAF	Momare	Nuclear Trans New Guinea	TransNewGuinea.org
met	ARM	met	LEAF	Ono	Nuclear Trans New Guinea	TransNewGuinea.org
lai	ARM	lai	LEAF	Molof	isolate	TransNewGuinea.org
kedʸ	ARM	kedʸ	SKIN	Erzya	Uralic	Intercontinental Dictionary Series
хил	ARM	хил	FEATHER	Tabassaran	Nakh-Daghestanian	Intercontinental Dictionary Series
i-ma-y	ARM	i-ma-y	HAND	Highland Puebla Nahuatl	Uto-Aztecan	Intercontinental Dictionary Series
maai-tl	ARM	maai-tl	HAND	Classical Nahuatl	Uto-Aztecan	Uto-Aztecan Basic Vocabulary Database
nob	ARM	nob	HAND	Southeastern Tepehuan	Uto-Aztecan	Uto-Aztecan Basic Vocabulary Database

Figure 3: Colexifications can be computed for each concept.

Open Questions

Open Questions: General Questions

- Which datasets to include for the launch? Numerals, or should we restrict to a minimal number of concepts and languages?
- Require explicit licenses (CC-BY)? If so, we should insist of derivative works being allowed.
- Should lexibank be a publication platform – e.g. assign DOIs to otherwise unpublished datasets? We would have quite a few usecases (Tukano data by Thiago, Sino-Tibetan data by Mattis)
- How to choose snapshots of evolving databases like ABVD?

Open Questions: Practical Problems

- optimizing workflow and workload (students for concept set / language mapping, data extraction, bibliographic management, scans, preparation of sources)
- information and communication with contributors (enhance the way we can invite people to participate)
- documentation of CLDF/CLPA (think of a publication to introduce the standard and the quality metrics, or make it part of a LexiBank publication)
- make the appearance of LexiBank more official (editorial board, LexiBank email address etc.) to make it easier to approach scholars for their data

Open Questions: Issues and Milestones

- all current issues at `https://github.com/glottbank/lexibank-data`
- major issues for first release
  - concept mappings
  - representation of dialect varieties
  - bibliography
  - expansion of CLPA
  - licensing
  - policy for unpublished datasets (MPI DOIs)
- minor issues
  - representation of partial cognates
  - representation of proto-forms

Possible Uses/Publications?

- A lexibank paper (incl. CLDF/CLPA?)
- Do words evolve at similar rates across families?
- Stability of basic words across cultures and times?
- Does the neogrammarian hypothesis hold? How many exceptions in sound correspondences are “normal”?
- Can we accurately recover phoneme inventories from lexicon?
- What drives sound change: language-specific settings or language-independent preference laws?
- How well do automatic approaches (phylogenetic reconstruction, cognate detection, phonetic alignment) correspond with experts?
- Can we find deeper, better phylogenies with these data?