MAX PLANCK INSTITUTE
FOR EVOLUTIONARY ANTHROPOLOGY

Computer-Assisted Approaches to Lexical Typology

Semantic Shifts: From lexicon to grammar. Diachronic and typological perspectives

September 2022

Annika Tjuka and Johann-Mattis List

Department of Linguistic and Cultural Evolution

Agenda

Day 1

1 Introduction

2 Concepticon

3 Norms, Ratings, and Relations

4 Case Study: Cross-linguistic Comparison of Sensory Modality Ratings

Day 2

5 Lexibank

6 Partial Colexifications

7 Case Study: Emotion Colexifications

8 Case Study: Body Colexifications

1

INTRODUCTION

Computer-Assisted Language Comparison (CALC)

Challenge & Goal

- **Lack of standards, agreed-upon methods, and comparable data**
  - Huge variations in naming languages, referencing concepts, or transcribing words.
  - Methods and procedures for inference differ from scholar to scholar.
  - Key aspects of the data have not been unified, as reflected in idiosyncratic elicitation glosses, language names, or transcription systems.

- **Standardizing data that has been published and encouraging scholars to standardize data along with its publication would increase the amount of comparable data out there.**

Computer-Assisted Language Comparison

Core Ideas & Data Integration

- **Data must be human- and machine-readable.**
- **Software is used to preprocess linguistic data and should specifically target linguistic problems rather than build on naive off-the-shelf solutions in machine learning.**
- **Interfaces help linguists to access the data and to post-process and correct machine output.**

- **Assemble data from multiple sources in such a way that we can use aggregated information for various studies.**

Cross-Linguistic Data Formats (CLDF)

Standardization & Curation

(Retro)-Standardization (or *Data Lifting*)

- **Establish and curate *reference catalogs* (large collections of small-scale *constructs* for linguistic research objects, including *languages*, *concepts*, and *sounds*),**
- **Parse digitized data semi-automatically in order to *link* data points to our reference catalogs,**
- **Use test-driven data curation to guarantee the workflow passes our tests.**

Test-Driven Data Curation

- **Versionize the work**
- **Test the basic characteristics of the data automatically with the help of unittests**
- **Write small, targeted web-based applications that enhance the digitization process**

Cross-Linguistic Data Formats

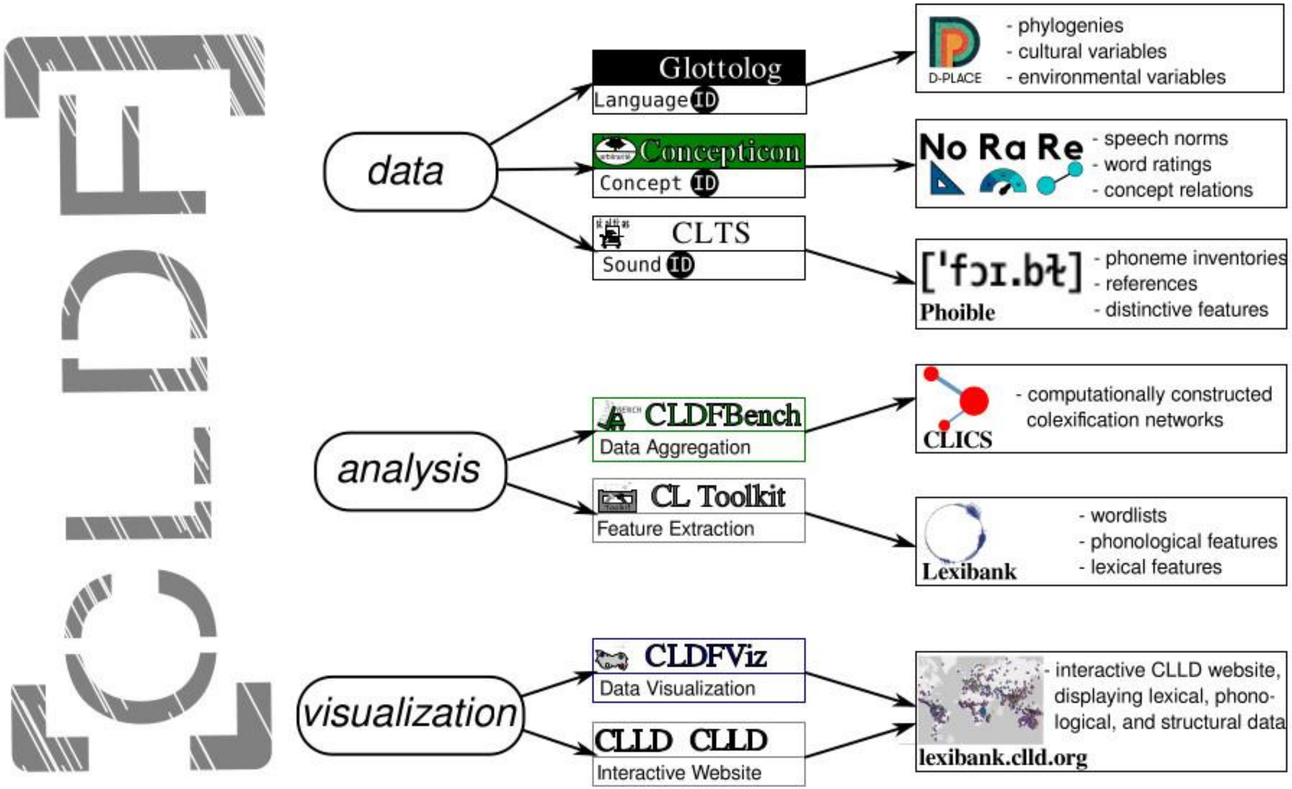
Reference Catalogs

- **Glottolog (https://glottolog.org)**
  - Reference catalogue for language varieties (languages and dialects), providing language identifiers, geolocations, classifications, and references.
- **Concepticon (https://concepticon.clld.org)**
  - Reference catalogue for concepts, which are defined independently of concrete languages, providing concept identifiers, concept metadata, concept relations, and references.
- **Cross-Linguistic Transcription Systems (https://clts.clld.org)**
  - Reference catalogue for speech sounds (across different transcription systems and data sets), offering sound identifiers, feature-based sound descriptions, and references.

Python Package `cldfbench`

- A Python package that does the lifting of data.
- Retro-standardization and conversion of data from other formats can be done with Python code that is testable, modularizable, and transparent.
- Teach more and more people to work with CLDF.
- Examples of how to use the library are published in various forms (e.g., as blog posts at https://calc.hypotheses.org).

Workflow for Data Curation, Analysis, and Visualization



More Information

- CALC: https://digling.org/calc/
- CLDF: https://cldf.clld.org
- Data preparation and examples: https://calc.hypotheses.org/
- CLLD: https://clld.org/

2

CONCEPTICON

Concepticon

- **A resource of concept and word lists that offers standardized concept sets and links to glosses. It serves as a reference catalog for historical and typological language comparison.**

- **Concepticon 2.6.0 (List et al. 2022)**
  - 404 concept lists
  - 3898 concept sets
  - 94,056 glosses mapped to concept sets
  - 40 glossing languages

- **Website: https://concepticon.clld.org/**
- **GitHub: https://github.com/concepticon/concepticon-data**

Concept and Word Lists

Concept Lists

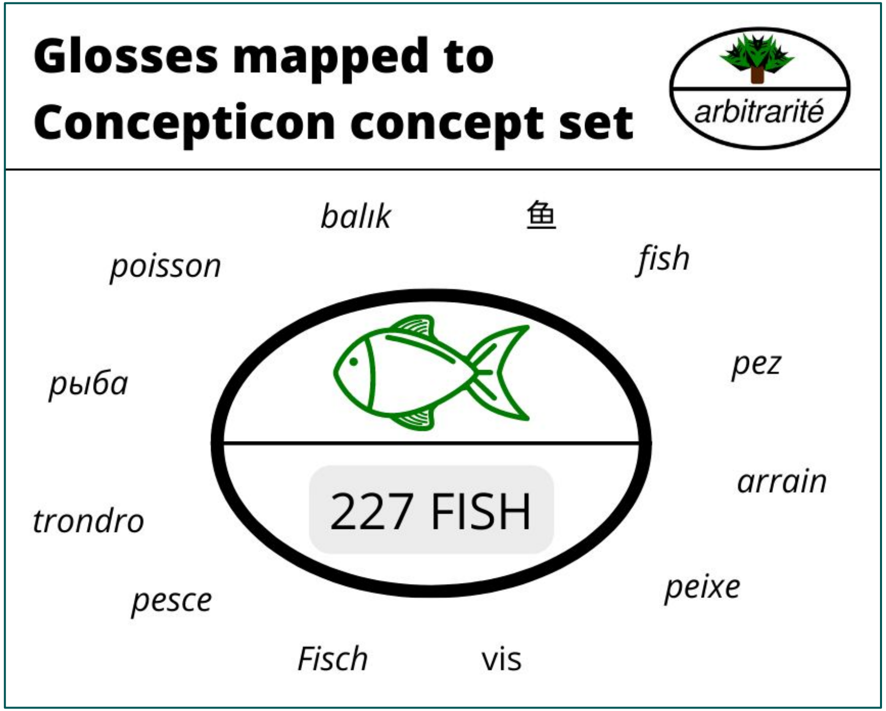
- Include basic vocabulary and cross-linguistically comparable concepts such as HAND, TREE, YOU, or GIVE.
- Used to elicit the glosses for the concepts across languages.
- Compiled by historical linguists and linguistic field workers.
- Often not standardized and adapted.
- Usually small lists of up to 300 concepts.

Word Lists

- Include basic vocabulary and information on word properties.
- Used in psychology to elicit properties of concepts.
- Word properties indicate whether a word is perceived as abstract or concrete, positive or negative, etc.
- Usually include thousands of words.

Concepticon Concept Sets

- Consist of a unique identifier, a label, a definition, a semantic field, and an ontological category.
- Concept identifiers (e.g., “227”) are connected to a unique label (e.g., “FISH”).
- Concepticon concept sets reflect concepts that are deemed interesting for comparison by linguists and occur frequently in concept lists (List et al. 2016).
- Elicitation glosses are established by linguists and are often based on already existing concept lists.



Data Curation

- **Automatic and manual mapping to Concepticon concept sets**
- **Information on data types in metadata.json**
- **Test-driven data curation**
- **Python package: `pyconcepticon` (Forkel, Rzymski & List 2019)**
- **Accessed via command line**
- **Regular releases**

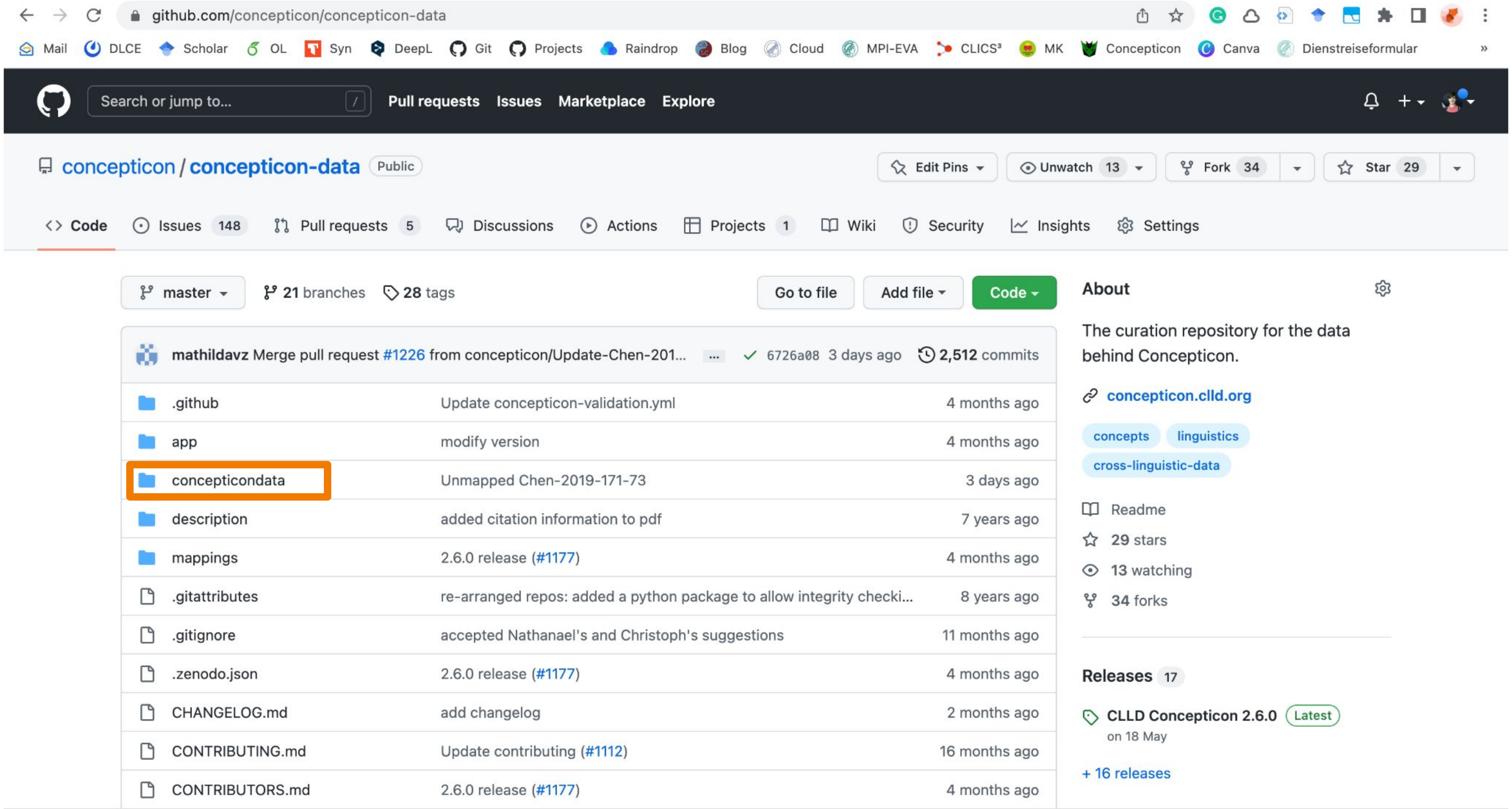
Tutorials

- ○ Tjuka (2020); Tresoldi (2019a; 2019b)

Workflows

- All of our data is curated on GitHub.
- New lists are added via a Pull Request (PR).
- The PRs are reviewed by the editors to discuss mappings, corrections, etc.
- We offer tutorials, examples, and guidelines in form of blog posts.
- So far, we have trained several student assistants to add new data sets.
- Through GitHub people can point us to new lists, improvements, and corrections of our data.

Open: GitHub and Website



concepticon / concepticon-data

mathildavz Merge pull request #1226 from concepticon/Update-Chen-201... 6726a08 3 days ago 2,512 commits

.github	Update concepticon-validation.yml	4 months ago
app	modify version	4 months ago
concepticondata	Unmapped Chen-2019-171-73	3 days ago
description	added citation information to pdf	7 years ago
mappings	2.6.0 release (#1177)	4 months ago
.gitattributes	re-arranged repos: added a python package to allow integrity checki...	8 years ago
.gitignore	accepted Nathanael's and Christoph's suggestions	11 months ago
.zenodo.json	2.6.0 release (#1177)	4 months ago
CHANGELOG.md	add changelog	2 months ago
CONTRIBUTING.md	Update contributing (#1112)	16 months ago
CONTRIBUTORS.md	2.6.0 release (#1177)	4 months ago

About

The curation repository for the data behind Concepticon.

concepticon.clld.org

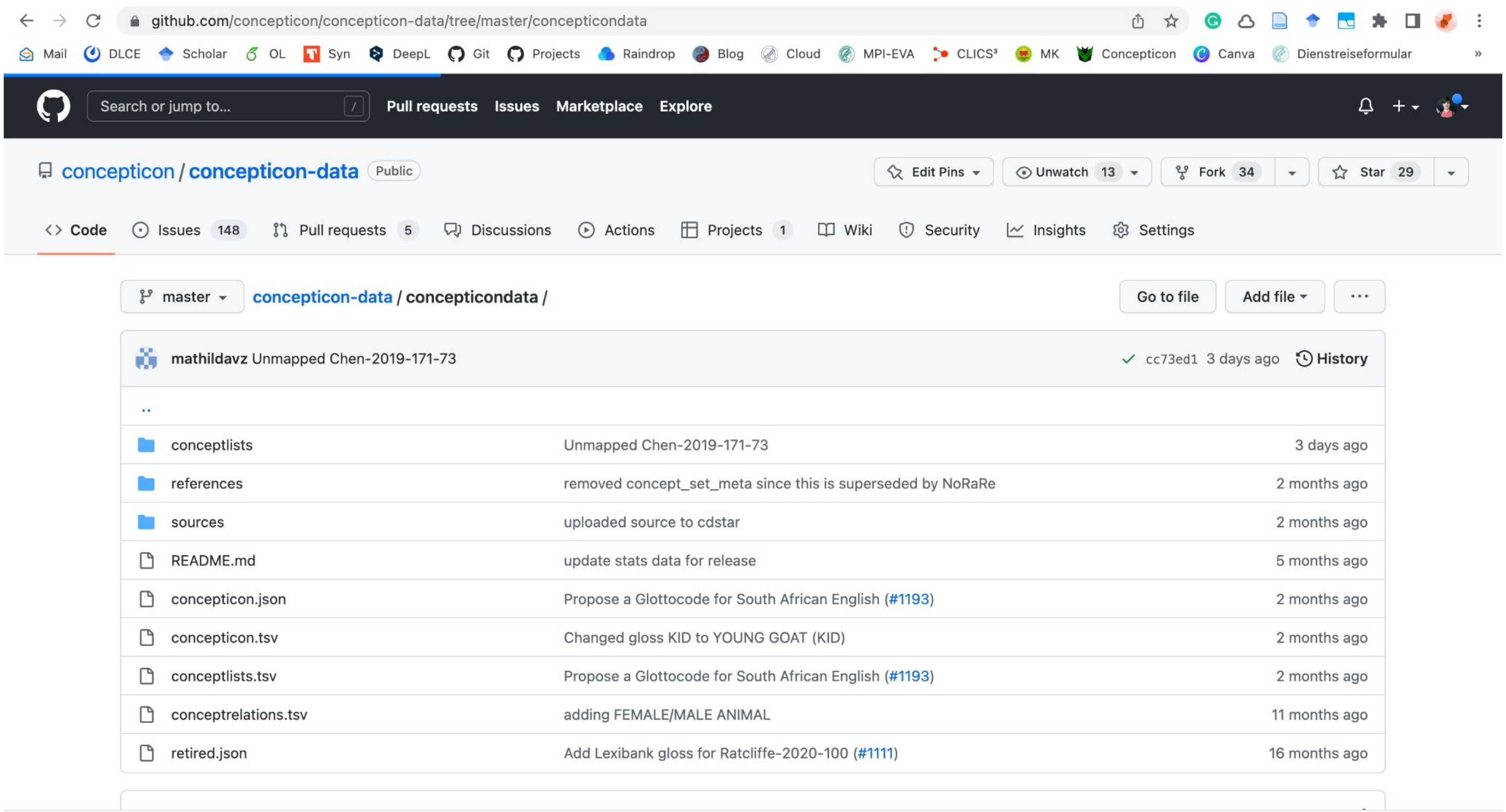
concepts linguistics cross-linguistic-data

Readme
29 stars
13 watching
34 forks

Releases 17

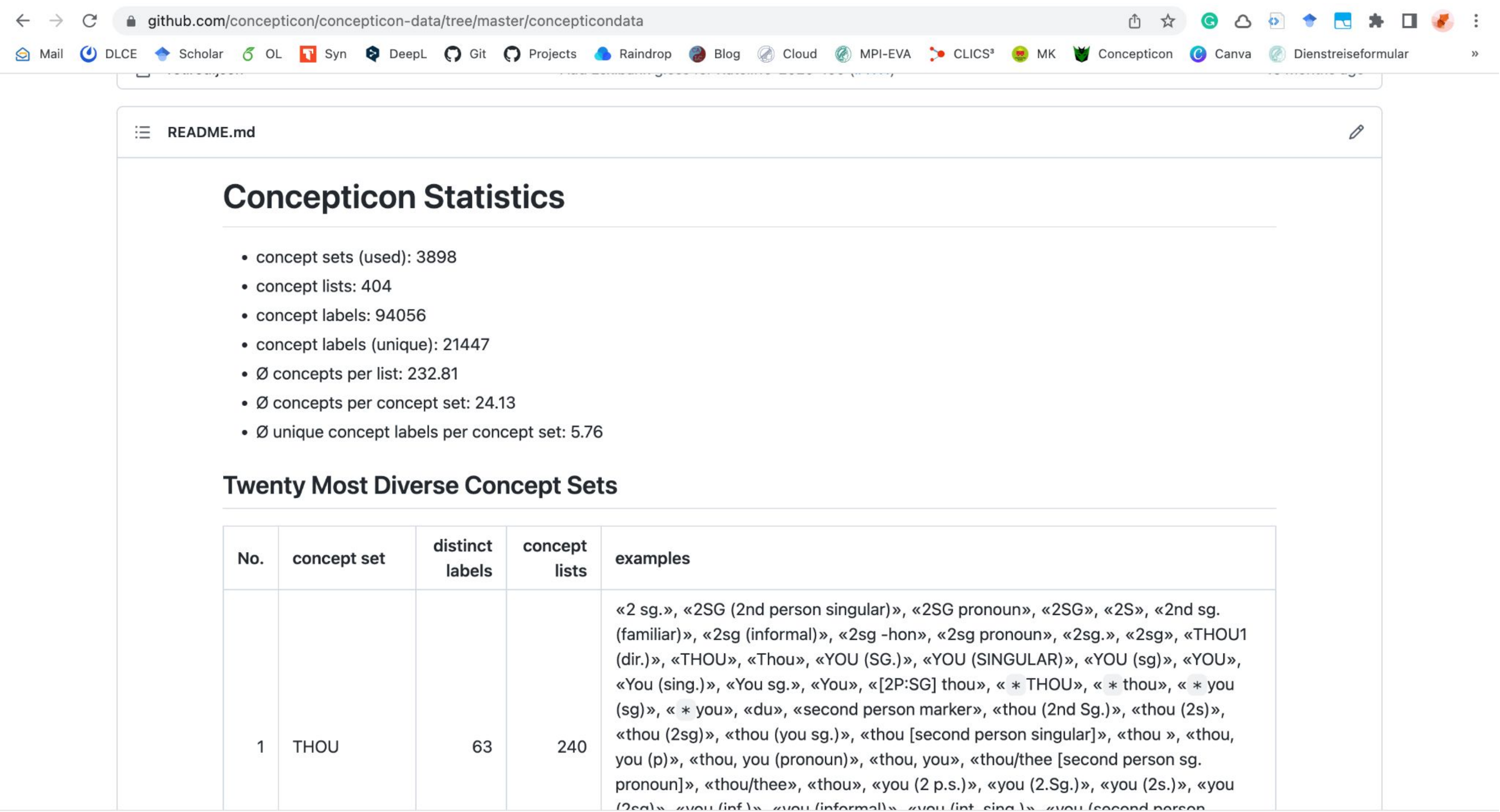
CLLD Concepticon 2.6.0 Latest
on 18 May

+ 16 releases



mathildavz Unmapped Chen-2019-171-73 ✓ cc73ed1 3 days ago

..		
conceptlists	Unmapped Chen-2019-171-73	3 days ago
references	removed concept_set_meta since this is superseded by NoRaRe	2 months ago
sources	uploaded source to cdstar	2 months ago
README.md	update stats data for release	5 months ago
concepticon.json	Propose a Glottocode for South African English (#1193)	2 months ago
concepticon.tsv	Changed gloss KID to YOUNG GOAT (KID)	2 months ago
conceptlists.tsv	Propose a Glottocode for South African English (#1193)	2 months ago
conceptrelations.tsv	adding FEMALE/MALE ANIMAL	11 months ago
retired.json	Add Lexibank gloss for Ratcliffe-2020-100 (#1111)	16 months ago

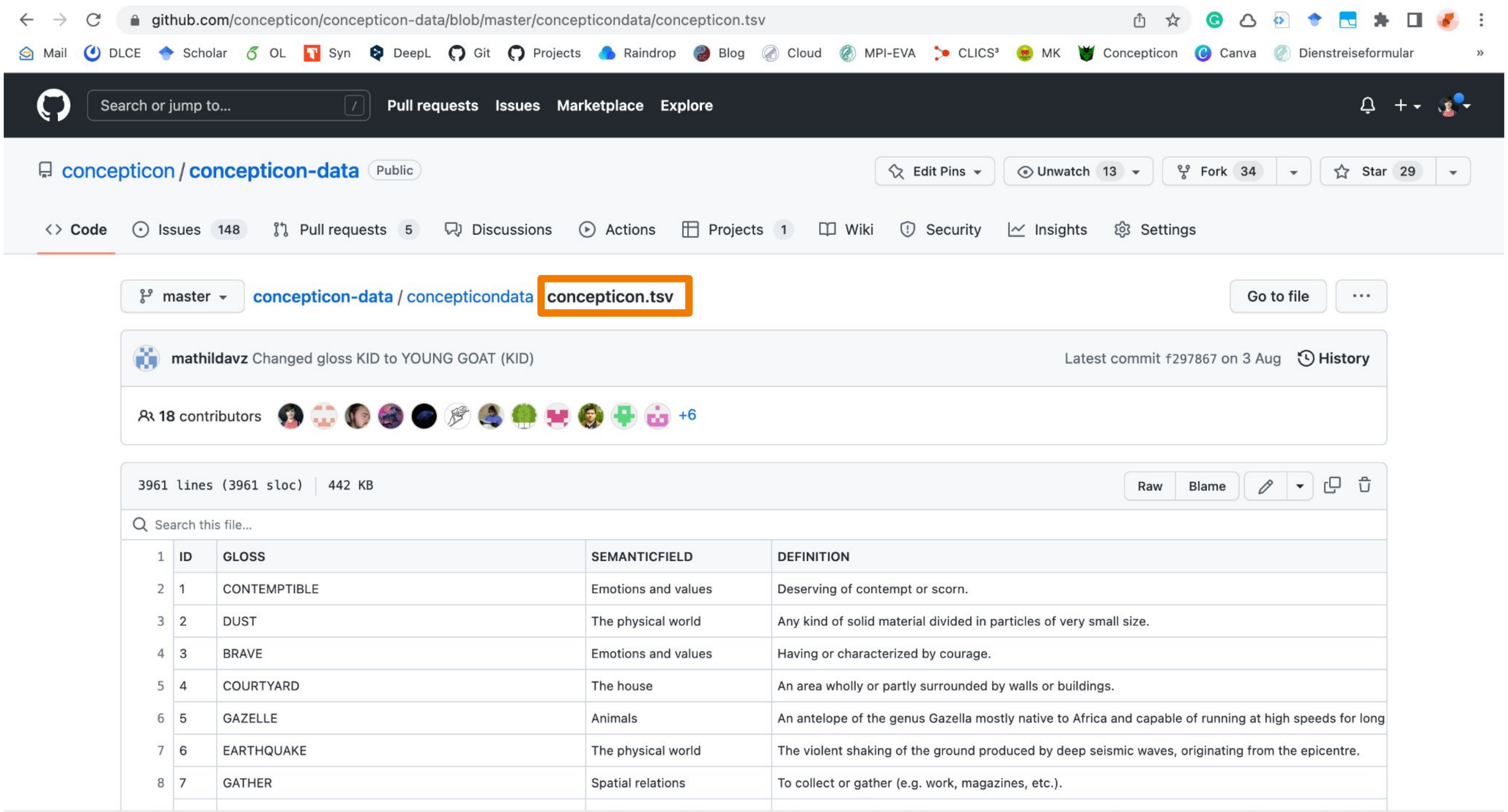


Concepticon Statistics

- concept sets (used): 3898
- concept lists: 404
- concept labels: 94056
- concept labels (unique): 21447
- Ø concepts per list: 232.81
- Ø concepts per concept set: 24.13
- Ø unique concept labels per concept set: 5.76

Twenty Most Diverse Concept Sets

No.	concept set	distinct labels	concept lists	examples
1	THOU	63	240	«2 sg.», «2SG (2nd person singular)», «2SG pronoun», «2SG», «2S», «2nd sg. (familiar)», «2sg (informal)», «2sg -hon», «2sg pronoun», «2sg.», «2sg», «THOU1 (dir.)», «THOU», «Thou», «YOU (SG.)», «YOU (SINGULAR)», «YOU (sg)», «YOU», «You (sing.)», «You sg.», «You», «[2P:SG] thou», «*THOU», «*thou», «*you (sg)», «*you», «du», «second person marker», «thou (2nd Sg.)», «thou (2s)», «thou (2sg)», «thou (you sg.)», «thou [second person singular]», «thou », «thou, you (p)», «thou, you (pronoun)», «thou, you», «thou/thee [second person sg. pronoun]», «thou/thee», «thou», «you (2 p.s.)», «you (2.Sg.)», «you (2s.)», «you (2sg)», «you (inf.)», «you (informal)», «you (int. sing.)», «you (second person

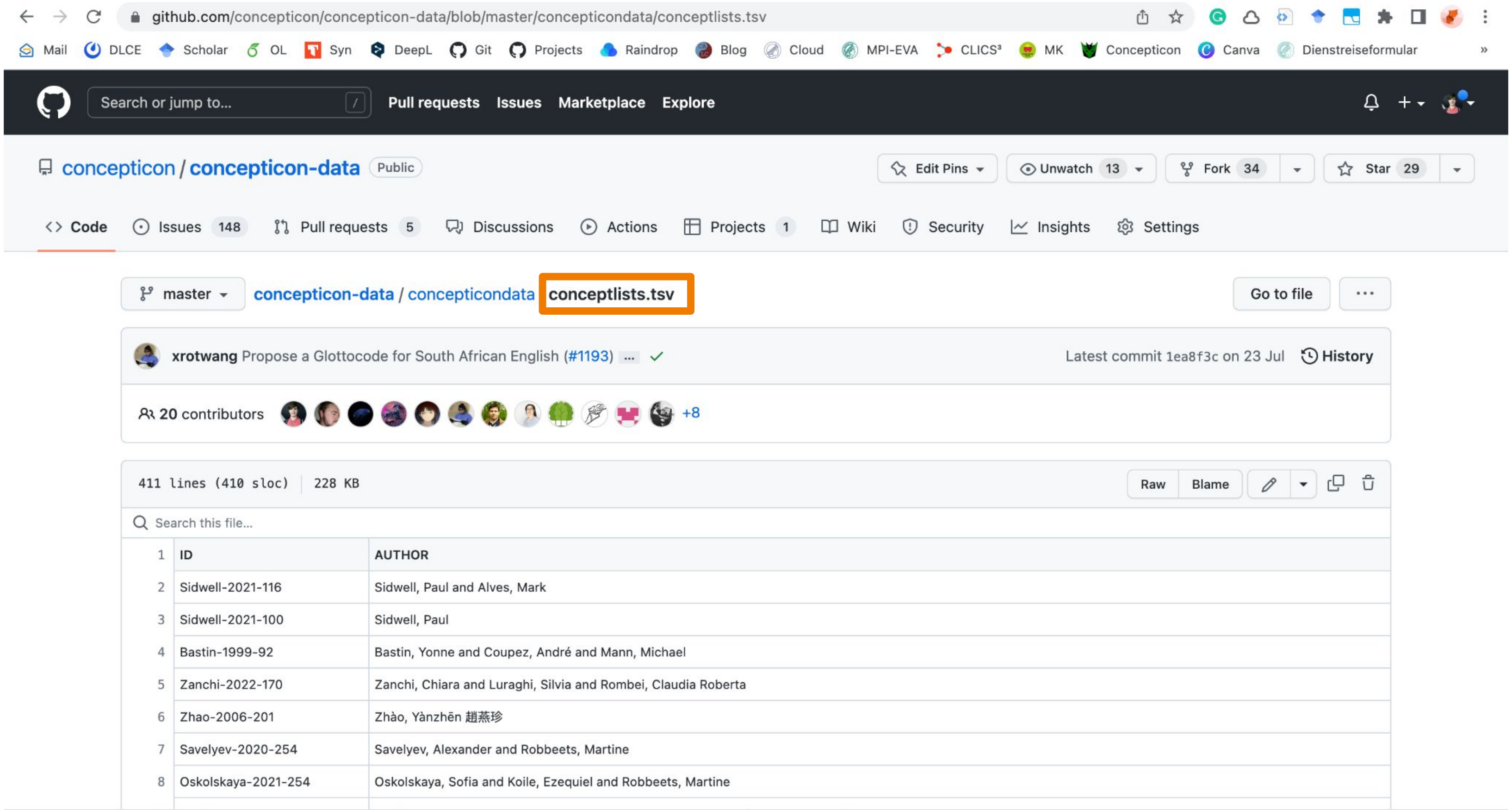


mathildavz Changed gloss KID to YOUNG GOAT (KID)

Latest commit f297867 on 3 Aug

3961 lines (3961 sloc) | 442 KB

ID	GLOSS	SEMANTICFIELD	DEFINITION
1	CONTEMPTIBLE	Emotions and values	Deserving of contempt or scorn.
2	DUST	The physical world	Any kind of solid material divided in particles of very small size.
3	BRAVE	Emotions and values	Having or characterized by courage.
4	COURTYARD	The house	An area wholly or partly surrounded by walls or buildings.
5	GAZELLE	Animals	An antelope of the genus Gazella mostly native to Africa and capable of running at high speeds for long
6	EARTHQUAKE	The physical world	The violent shaking of the ground produced by deep seismic waves, originating from the epicentre.
7	GATHER	Spatial relations	To collect or gather (e.g. work, magazines, etc.).



github.com/concepticon/concepticon-data/blob/master/concepticondata/conceptlists.tsv

concepticon / concepticon-data

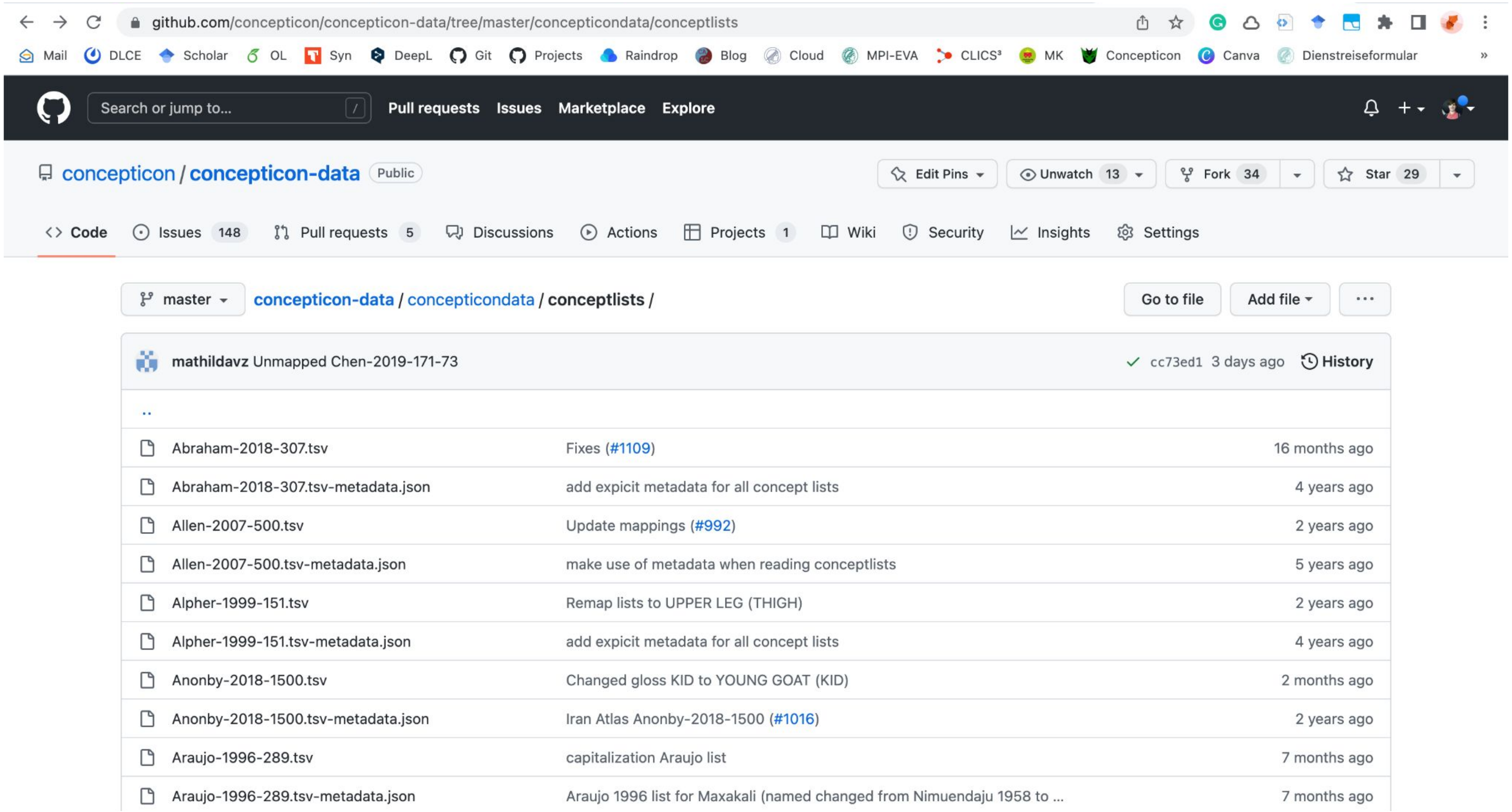
concepticon-data / concepticondata / **conceptlists.tsv**

xrotwang Propose a Glottocode for South African English (#1193) … ✓ Latest commit 1ea8f3c on 23 Jul

20 contributors +8

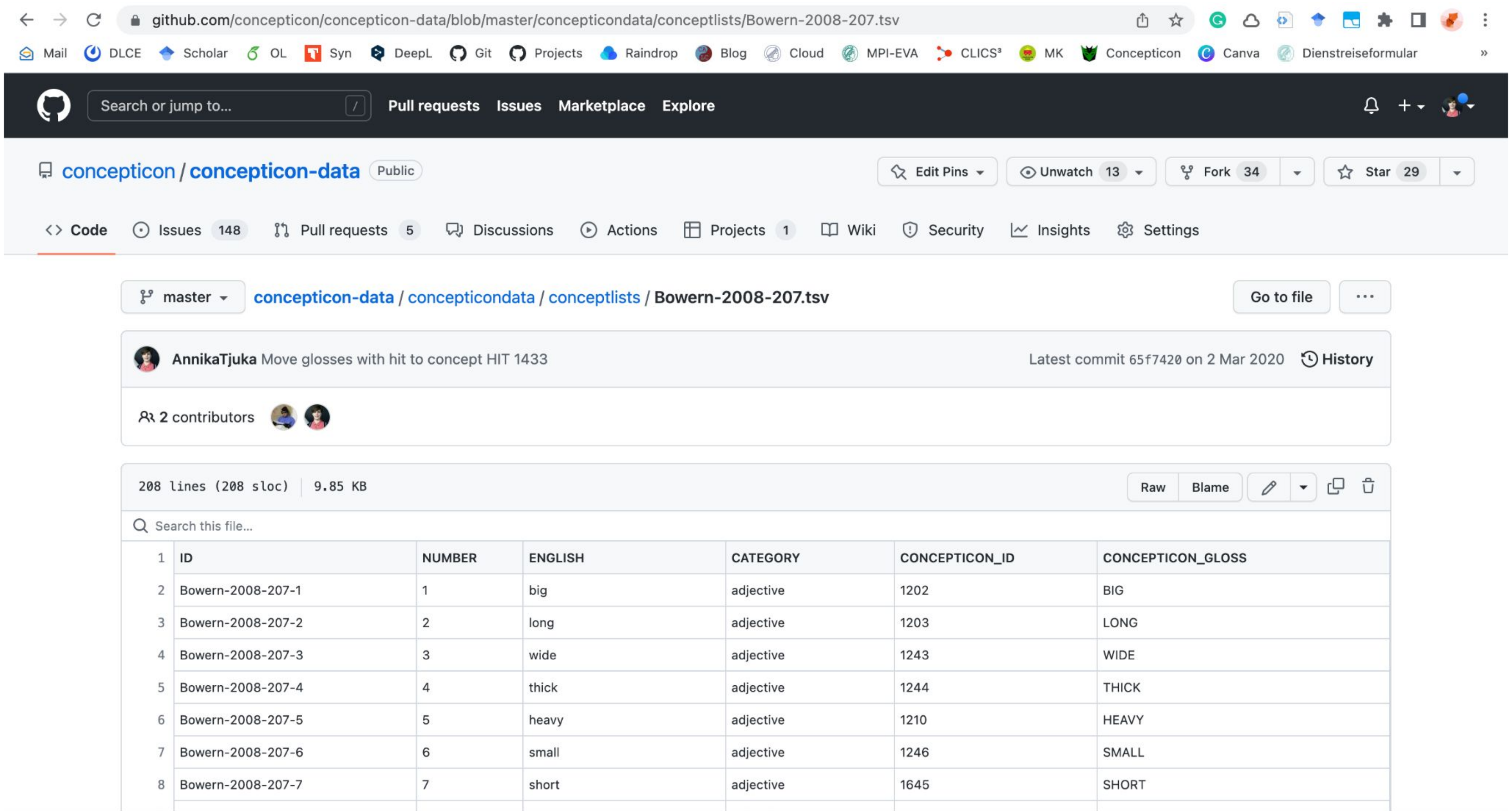
411 lines (410 sloc) | 228 KB

	ID	AUTHOR
2	Sidwell-2021-116	Sidwell, Paul and Alves, Mark
3	Sidwell-2021-100	Sidwell, Paul
4	Bastin-1999-92	Bastin, Yonne and Coupez, André and Mann, Michael
5	Zanchi-2022-170	Zanchi, Chiara and Luraghi, Silvia and Rombei, Claudia Roberta
6	Zhao-2006-201	Zhào, Yànzhēn 趙燕珍
7	Savelyev-2020-254	Savelyev, Alexander and Robbeets, Martine
8	Oskolskaya-2021-254	Oskolskaya, Sofia and Koile, Ezequiel and Robbeets, Martine



mathildavz Unmapped Chen-2019-171-73 ✓ cc73ed1 3 days ago

..		
Abraham-2018-307.tsv	Fixes (#1109)	16 months ago
Abraham-2018-307.tsv-metadata.json	add expicit metadata for all concept lists	4 years ago
Allen-2007-500.tsv	Update mappings (#992)	2 years ago
Allen-2007-500.tsv-metadata.json	make use of metadata when reading conceptlists	5 years ago
Alpher-1999-151.tsv	Remap lists to UPPER LEG (THIGH)	2 years ago
Alpher-1999-151.tsv-metadata.json	add expicit metadata for all concept lists	4 years ago
Anonby-2018-1500.tsv	Changed gloss KID to YOUNG GOAT (KID)	2 months ago
Anonby-2018-1500.tsv-metadata.json	Iran Atlas Anonby-2018-1500 (#1016)	2 years ago
Araujo-1996-289.tsv	capitalization Araujo list	7 months ago
Araujo-1996-289.tsv-metadata.json	Araujo 1996 list for Maxakali (named changed from Nimuendaju 1958 to ...	7 months ago



concepticon-data / concepticondata / conceptlists / Bowern-2008-207.tsv

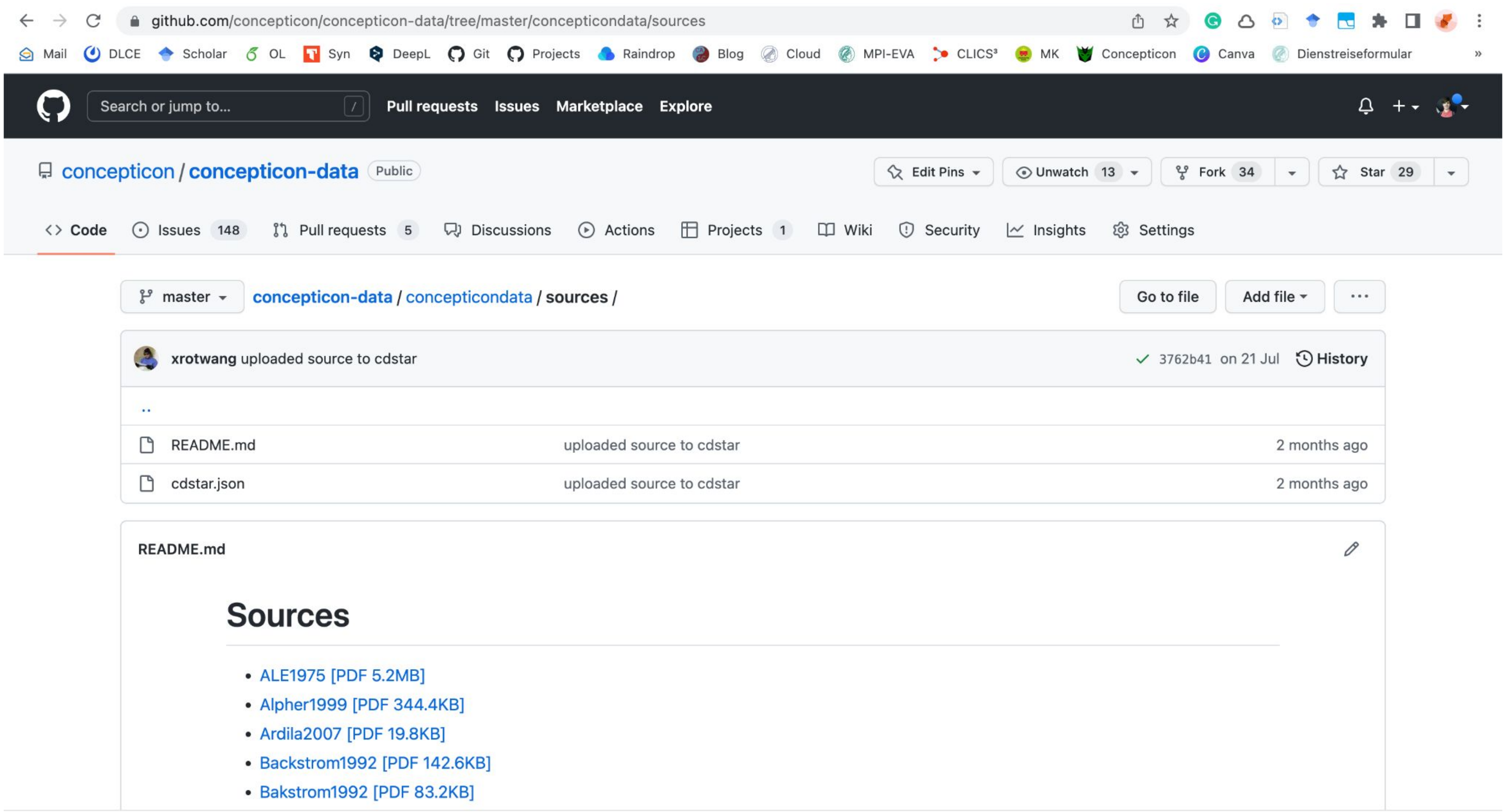
AnnikaTjuka Move glosses with hit to concept HIT 1433

Latest commit 65f7420 on 2 Mar 2020

2 contributors

208 lines (208 sloc) | 9.85 KB

	ID	NUMBER	ENGLISH	CATEGORY	CONCEPTICON_ID	CONCEPTICON_GLOSS
2	Bowern-2008-207-1	1	big	adjective	1202	BIG
3	Bowern-2008-207-2	2	long	adjective	1203	LONG
4	Bowern-2008-207-3	3	wide	adjective	1243	WIDE
5	Bowern-2008-207-4	4	thick	adjective	1244	THICK
6	Bowern-2008-207-5	5	heavy	adjective	1210	HEAVY
7	Bowern-2008-207-6	6	small	adjective	1246	SMALL
8	Bowern-2008-207-7	7	short	adjective	1645	SHORT



concepticon / concepticon-data Public

concepticon-data / concepticondata / sources /

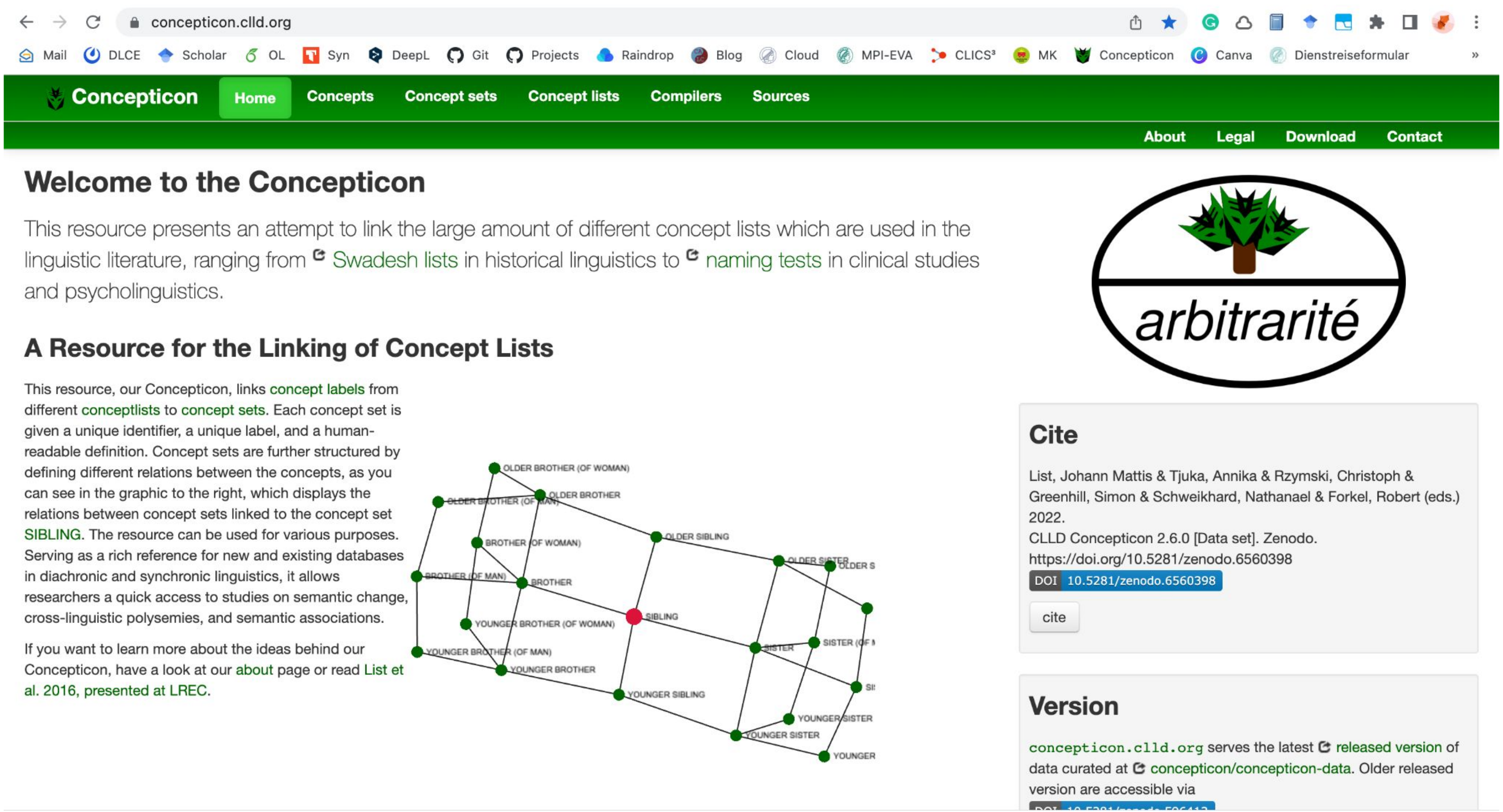
xrotwang uploaded source to cdstar ✓ 3762b41 on 21 Jul

```
..
README.md      uploaded source to cdstar   2 months ago
cdstar.json    uploaded source to cdstar   2 months ago
```

README.md

Sources

- ALE1975 [PDF 5.2MB]
- Alpher1999 [PDF 344.4KB]
- Ardila2007 [PDF 19.8KB]
- Backstrom1992 [PDF 142.6KB]
- Bakstrom1992 [PDF 83.2KB]



Welcome to the Concepticon

This resource presents an attempt to link the large amount of different concept lists which are used in the linguistic literature, ranging from Swadesh lists in historical linguistics to naming tests in clinical studies and psycholinguistics.

A Resource for the Linking of Concept Lists

This resource, our Concepticon, links concept labels from different conceptlists to concept sets. Each concept set is given a unique identifier, a unique label, and a human-readable definition. Concept sets are further structured by defining different relations between the concepts, as you can see in the graphic to the right, which displays the relations between concept sets linked to the concept set SIBLING. The resource can be used for various purposes. Serving as a rich reference for new and existing databases in diachronic and synchronic linguistics, it allows researchers a quick access to studies on semantic change, cross-linguistic polysemies, and semantic associations.

If you want to learn more about the ideas behind our Concepticon, have a look at our about page or read List et al. 2016, presented at LREC.

Cite

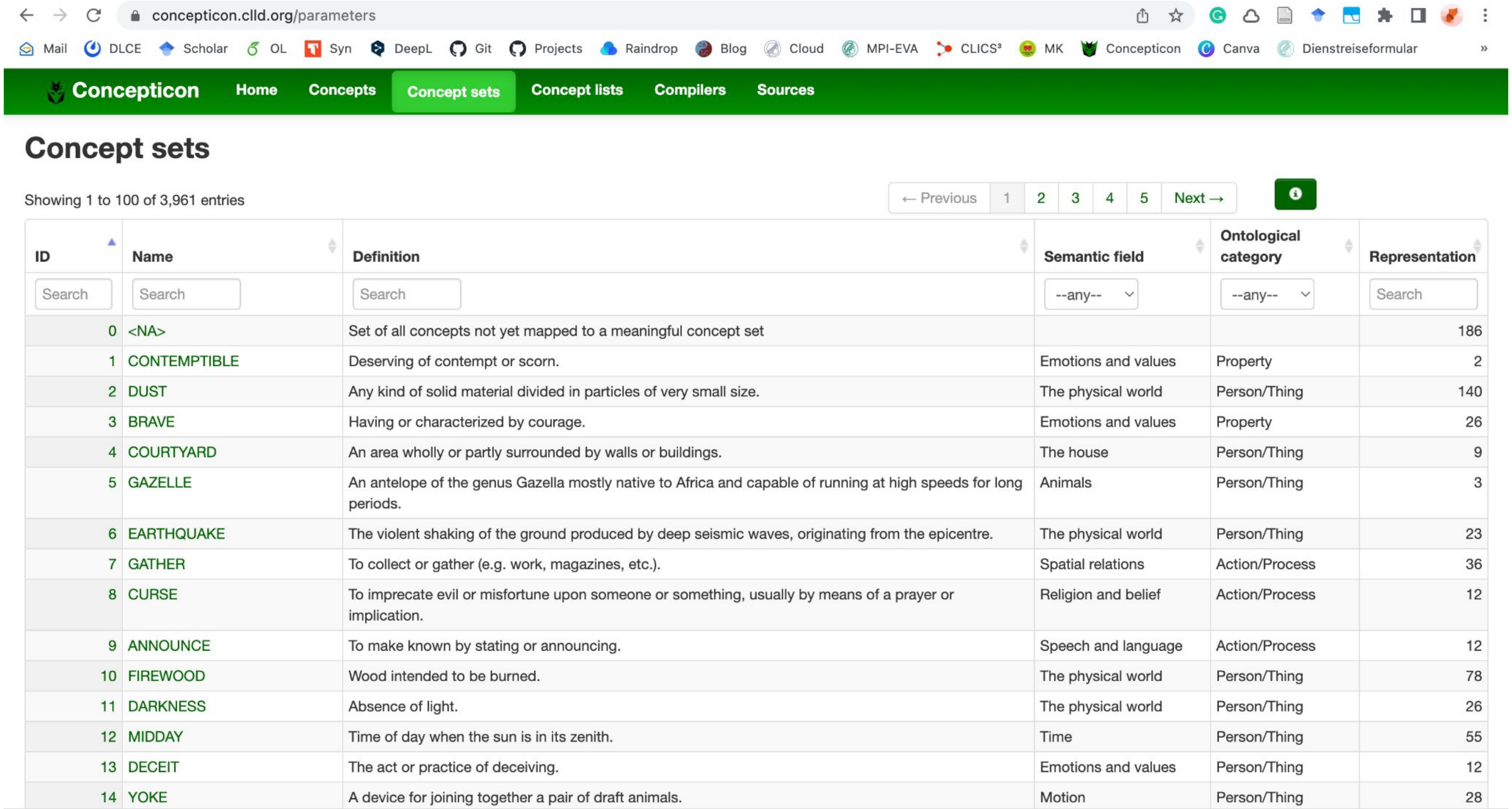
List, Johann Mattis & Tjuka, Annika & Rzymski, Christoph & Greenhill, Simon & Schweikhard, Nathanael & Forkel, Robert (eds.) 2022.
CLLD Concepticon 2.6.0 [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.6560398

DOI 10.5281/zenodo.6560398

cite

Version

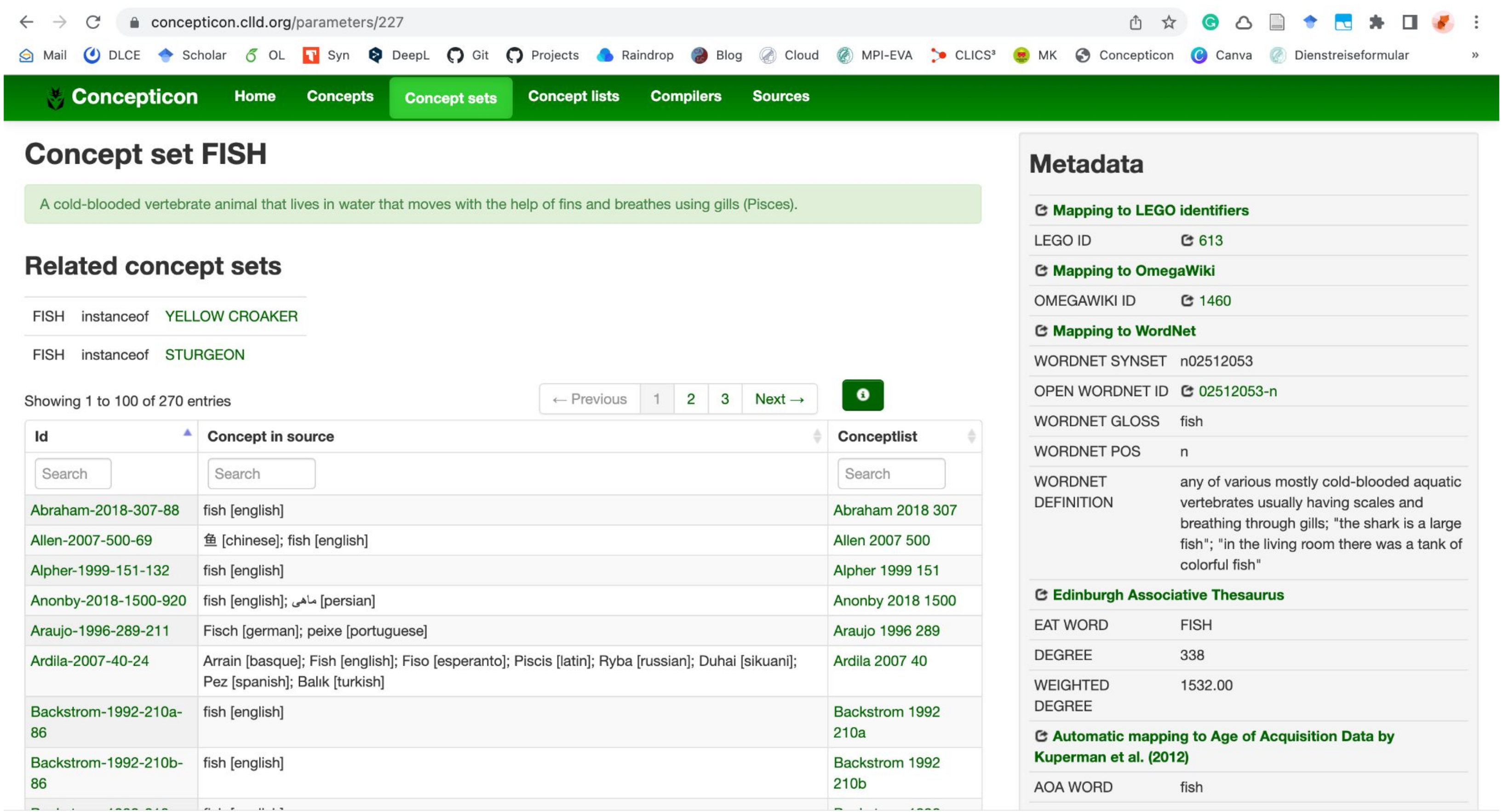
`concepticon.clld.org` serves the latest released version of data curated at concepticon/concepticon-data. Older released version are accessible via



Concept sets

Showing 1 to 100 of 3,961 entries

ID	Name	Definition	Semantic field	Ontological category	Representation
Search	Search	Search	--any--	--any--	Search
0	<NA>	Set of all concepts not yet mapped to a meaningful concept set			186
1	CONTEMPTIBLE	Deserving of contempt or scorn.	Emotions and values	Property	2
2	DUST	Any kind of solid material divided in particles of very small size.	The physical world	Person/Thing	140
3	BRAVE	Having or characterized by courage.	Emotions and values	Property	26
4	COURTYARD	An area wholly or partly surrounded by walls or buildings.	The house	Person/Thing	9
5	GAZELLE	An antelope of the genus Gazella mostly native to Africa and capable of running at high speeds for long periods.	Animals	Person/Thing	3
6	EARTHQUAKE	The violent shaking of the ground produced by deep seismic waves, originating from the epicentre.	The physical world	Person/Thing	23
7	GATHER	To collect or gather (e.g. work, magazines, etc.).	Spatial relations	Action/Process	36
8	CURSE	To imprecate evil or misfortune upon someone or something, usually by means of a prayer or implication.	Religion and belief	Action/Process	12
9	ANNOUNCE	To make known by stating or announcing.	Speech and language	Action/Process	12
10	FIREWOOD	Wood intended to be burned.	The physical world	Person/Thing	78
11	DARKNESS	Absence of light.	The physical world	Person/Thing	26
12	MIDDAY	Time of day when the sun is in its zenith.	Time	Person/Thing	55
13	DECEIT	The act or practice of deceiving.	Emotions and values	Person/Thing	12
14	YOKE	A device for joining together a pair of draft animals.	Motion	Person/Thing	28



Concept set FISH

A cold-blooded vertebrate animal that lives in water that moves with the help of fins and breathes using gills (Pisces).

Related concept sets

FISH	instanceof	YELLOW CROAKER
FISH	instanceof	STURGEON

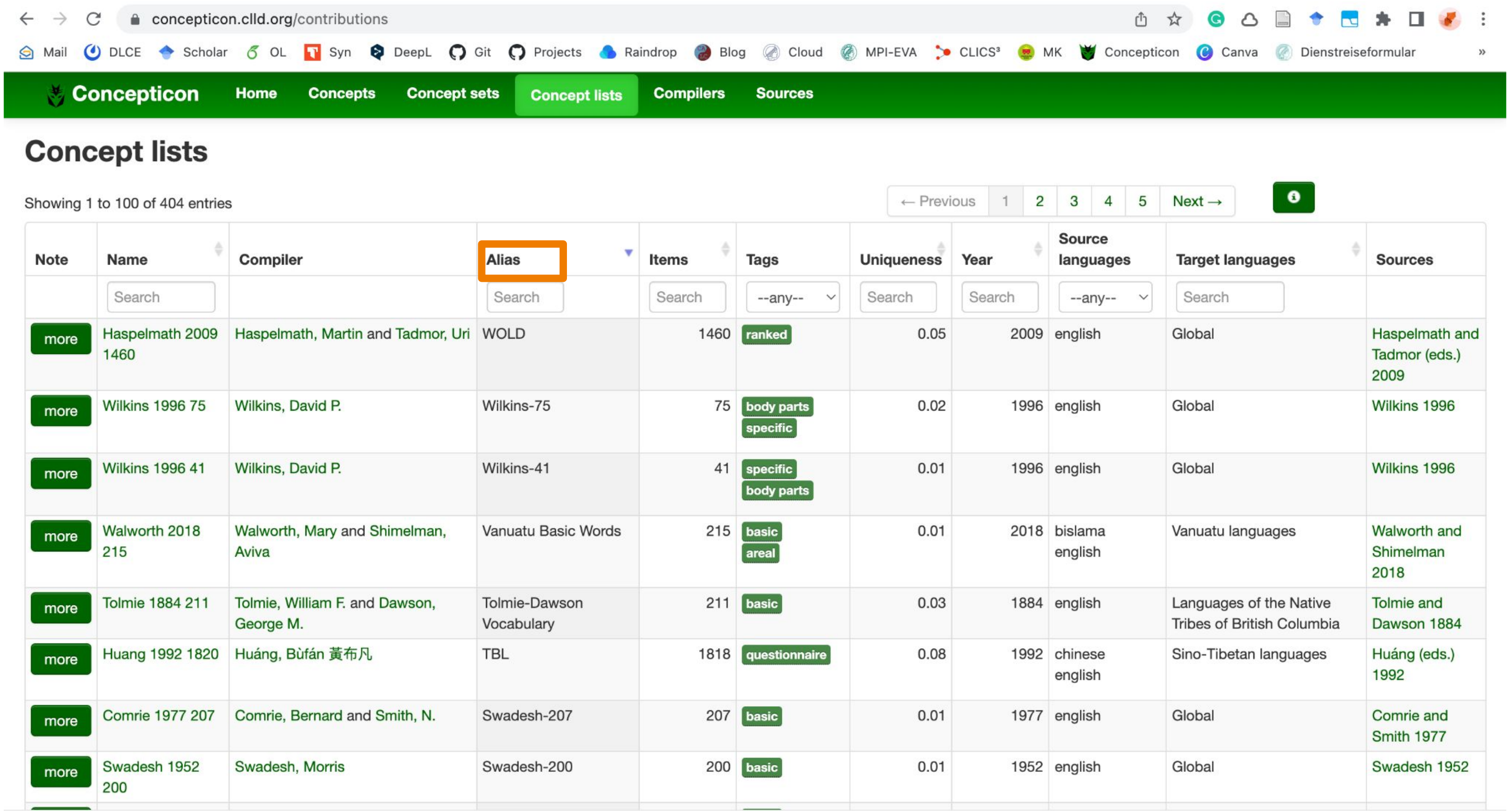
Showing 1 to 100 of 270 entries

← Previous 1 2 3 Next →

Id	Concept in source	Conceptlist
Abraham-2018-307-88	fish [english]	Abraham 2018 307
Allen-2007-500-69	鱼 [chinese]; fish [english]	Allen 2007 500
Alpher-1999-151-132	fish [english]	Alpher 1999 151
Anonby-2018-1500-920	fish [english]; ماهی [persian]	Anonby 2018 1500
Araujo-1996-289-211	Fisch [german]; peixe [portuguese]	Araujo 1996 289
Ardila-2007-40-24	Arrain [basque]; Fish [english]; Fiso [esperanto]; Piscis [latin]; Ryba [russian]; Duhai [sikuani]; Pez [spanish]; Balık [turkish]	Ardila 2007 40
Backstrom-1992-210a-86	fish [english]	Backstrom 1992 210a
Backstrom-1992-210b-86	fish [english]	Backstrom 1992 210b

Metadata

Mapping to LEGO identifiers	
LEGO ID	613
Mapping to OmegaWiki	
OMEGAWIKI ID	1460
Mapping to WordNet	
WORDNET SYNSET	n02512053
OPEN WORDNET ID	02512053-n
WORDNET GLOSS	fish
WORDNET POS	n
WORDNET DEFINITION	any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills; "the shark is a large fish"; "in the living room there was a tank of colorful fish"
Edinburgh Associative Thesaurus	
EAT WORD	FISH
DEGREE	338
WEIGHTED DEGREE	1532.00
Automatic mapping to Age of Acquisition Data by Kuperman et al. (2012)	
AOA WORD	fish

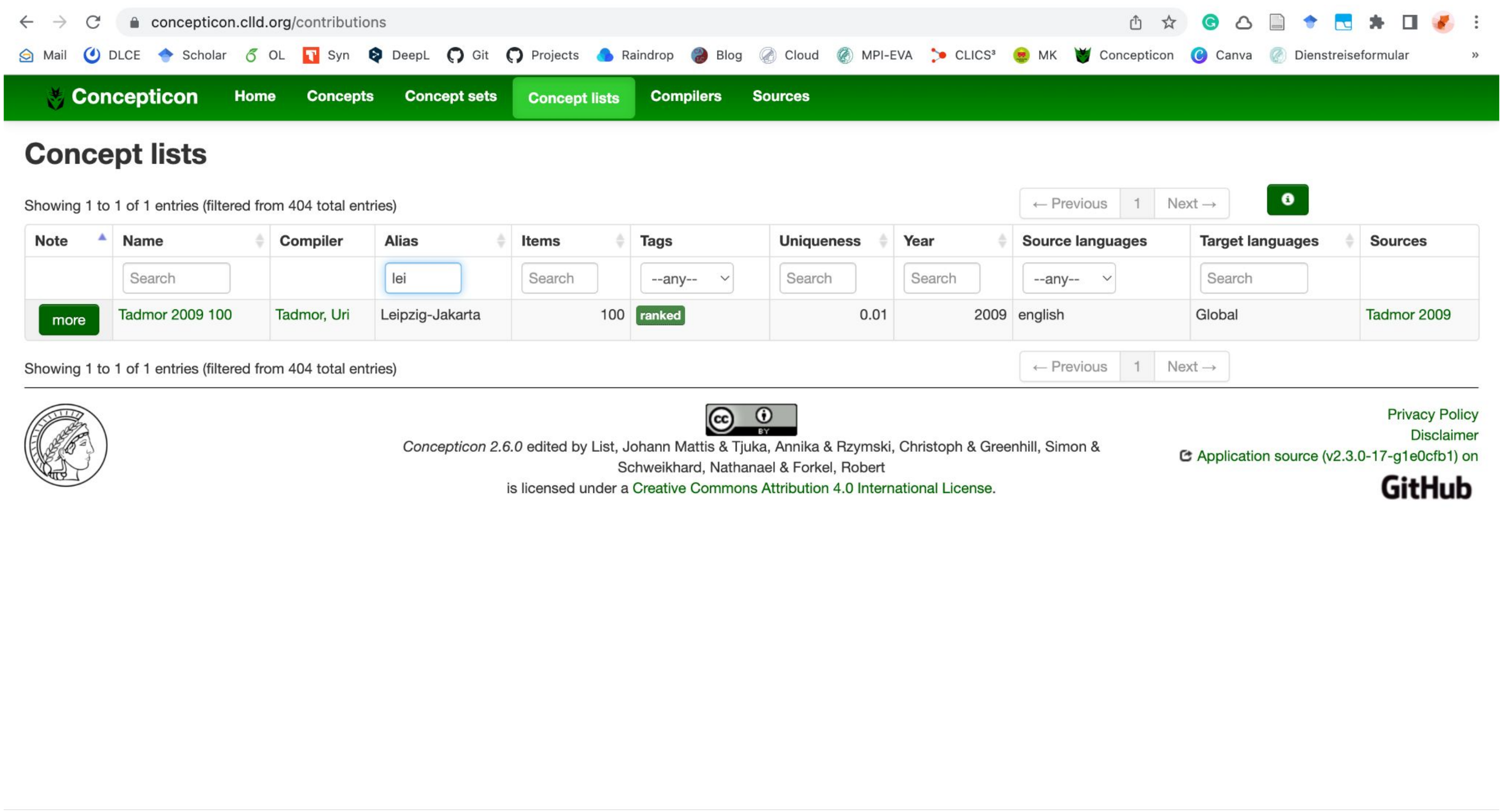


Concept lists

Showing 1 to 100 of 404 entries

← Previous 1 2 3 4 5 Next →

Note	Name	Compiler	Alias	Items	Tags	Uniqueness	Year	Source languages	Target languages	Sources
	Search		Search	Search	--any--	Search	Search	--any--	Search	
more	Haspelmath 2009 1460	Haspelmath, Martin and Tadmor, Uri	WOLD	1460	ranked	0.05	2009	english	Global	Haspelmath and Tadmor (eds.) 2009
more	Wilkins 1996 75	Wilkins, David P.	Wilkins-75	75	body parts specific	0.02	1996	english	Global	Wilkins 1996
more	Wilkins 1996 41	Wilkins, David P.	Wilkins-41	41	specific body parts	0.01	1996	english	Global	Wilkins 1996
more	Walworth 2018 215	Walworth, Mary and Shimelman, Aviva	Vanuatu Basic Words	215	basic areal	0.01	2018	bislama english	Vanuatu languages	Walworth and Shimelman 2018
more	Tolmie 1884 211	Tolmie, William F. and Dawson, George M.	Tolmie-Dawson Vocabulary	211	basic	0.03	1884	english	Languages of the Native Tribes of British Columbia	Tolmie and Dawson 1884
more	Huang 1992 1820	Huáng, Bùfán 黃布凡	TBL	1818	questionnaire	0.08	1992	chinese english	Sino-Tibetan languages	Huáng (eds.) 1992
more	Comrie 1977 207	Comrie, Bernard and Smith, N.	Swadesh-207	207	basic	0.01	1977	english	Global	Comrie and Smith 1977
more	Swadesh 1952 200	Swadesh, Morris	Swadesh-200	200	basic	0.01	1952	english	Global	Swadesh 1952



Concept lists

Showing 1 to 1 of 1 entries (filtered from 404 total entries)

← Previous | 1 | Next →

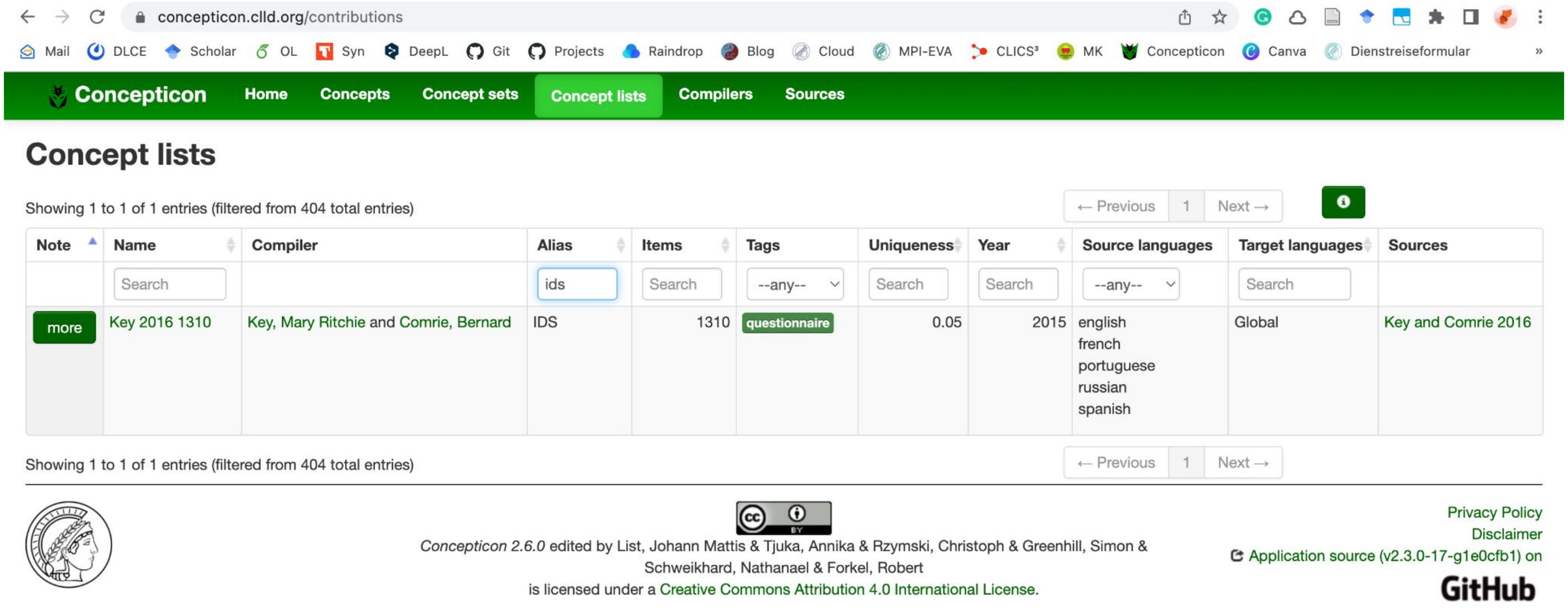
Note	Name	Compiler	Alias	Items	Tags	Uniqueness	Year	Source languages	Target languages	Sources
	Search		lei	Search	--any--	Search	Search	--any--	Search	
more	Tadmor 2009 100	Tadmor, Uri	Leipzig-Jakarta	100	ranked	0.01	2009	english	Global	Tadmor 2009

Showing 1 to 1 of 1 entries (filtered from 404 total entries)

← Previous | 1 | Next →

Concepticon 2.6.0 edited by List, Johann Mattis & Tjuka, Annika & Rzymski, Christoph & Greenhill, Simon & Schweikhard, Nathanael & Forkel, Robert
is licensed under a Creative Commons Attribution 4.0 International License.

Privacy Policy
Disclaimer
Application source (v2.3.0-17-g1e0cfb1) on
GitHub



Concept lists

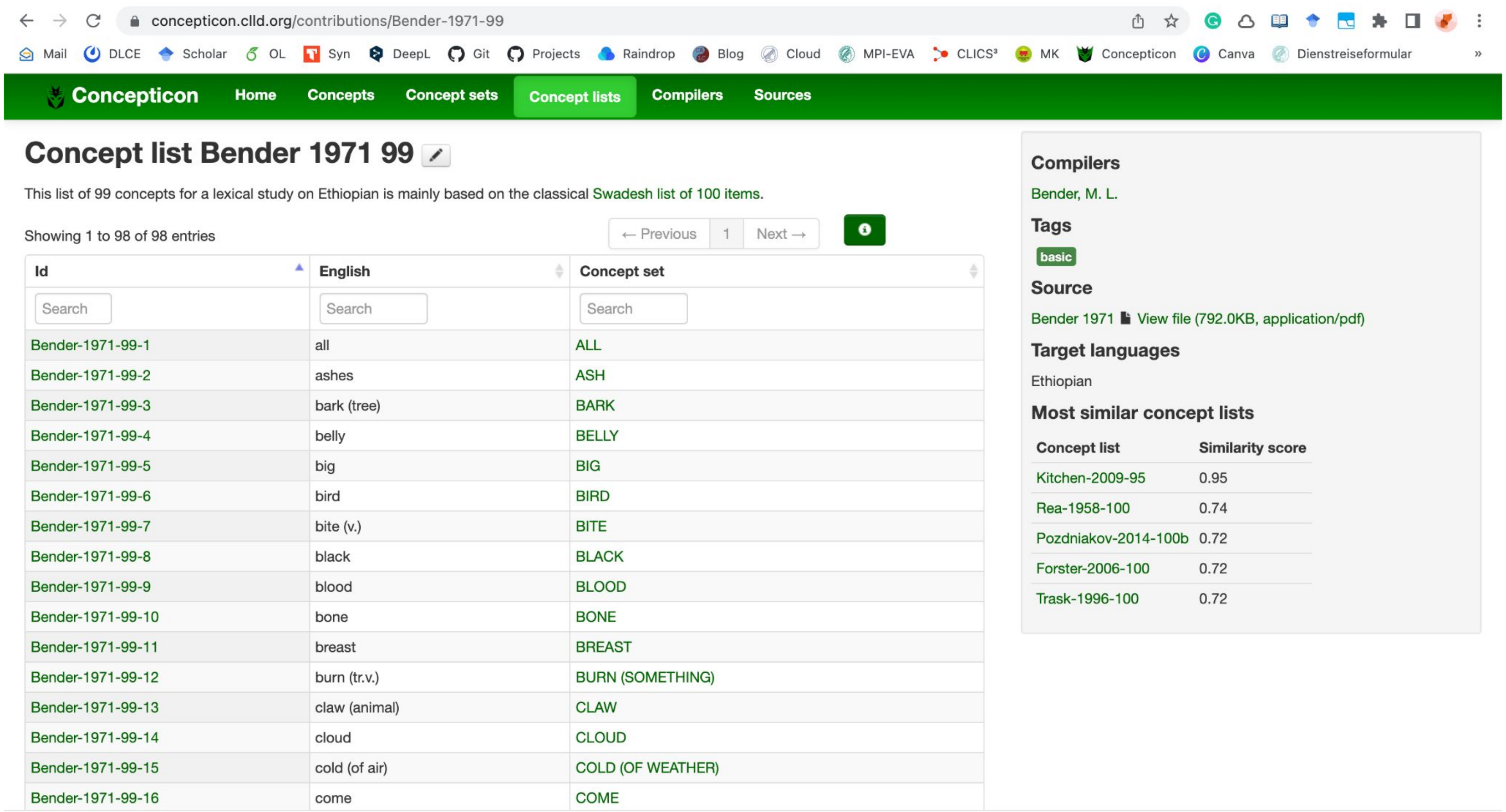
Showing 1 to 1 of 1 entries (filtered from 404 total entries)

Note	Name	Compiler	Alias	Items	Tags	Uniqueness	Year	Source languages	Target languages	Sources
	Search		ids	Search	--any--	Search	Search	--any--	Search	
more	Key 2016 1310	Key, Mary Ritchie and Comrie, Bernard	IDS	1310	questionnaire	0.05	2015	english french portuguese russian spanish	Global	Key and Comrie 2016

Showing 1 to 1 of 1 entries (filtered from 404 total entries)

Concepticon 2.6.0 edited by List, Johann Mattis & Tjuka, Annika & Rzymski, Christoph & Greenhill, Simon & Schweikhard, Nathanael & Forkel, Robert is licensed under a Creative Commons Attribution 4.0 International License.

Privacy Policy
Disclaimer
Application source (v2.3.0-17-g1e0cfb1) on GitHub



Concept list Bender 1971 99

This list of 99 concepts for a lexical study on Ethiopian is mainly based on the classical Swadesh list of 100 items.

Showing 1 to 98 of 98 entries

Id	English	Concept set
Search	Search	Search
Bender-1971-99-1	all	ALL
Bender-1971-99-2	ashes	ASH
Bender-1971-99-3	bark (tree)	BARK
Bender-1971-99-4	belly	BELLY
Bender-1971-99-5	big	BIG
Bender-1971-99-6	bird	BIRD
Bender-1971-99-7	bite (v.)	BITE
Bender-1971-99-8	black	BLACK
Bender-1971-99-9	blood	BLOOD
Bender-1971-99-10	bone	BONE
Bender-1971-99-11	breast	BREAST
Bender-1971-99-12	burn (tr.v.)	BURN (SOMETHING)
Bender-1971-99-13	claw (animal)	CLAW
Bender-1971-99-14	cloud	CLOUD
Bender-1971-99-15	cold (of air)	COLD (OF WEATHER)
Bender-1971-99-16	come	COME

Compilers

Bender, M. L.

Tags

basic

Source

Bender 1971 View file (792.0KB, application/pdf)

Target languages

Ethiopian

Most similar concept lists

Concept list	Similarity score
Kitchen-2009-95	0.95
Rea-1958-100	0.74
Pozdniakov-2014-100b	0.72
Forster-2006-100	0.72
Trask-1996-100	0.72

Editors and Contributors

- **Current Editorial Team**
  - Johann-Mattis List, Annika Tjuka, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, and Robert Forkel

- **Contributors**
  - 67 researchers have contributed data.
  - They point us to missing lists, provide scans, translations, and corrections.
  - Full list: https://github.com/concepticon/concepticon-data/blob/master/CONTRIBUTORS.md

3

NORMS, RATINGS, AND RELATIONS

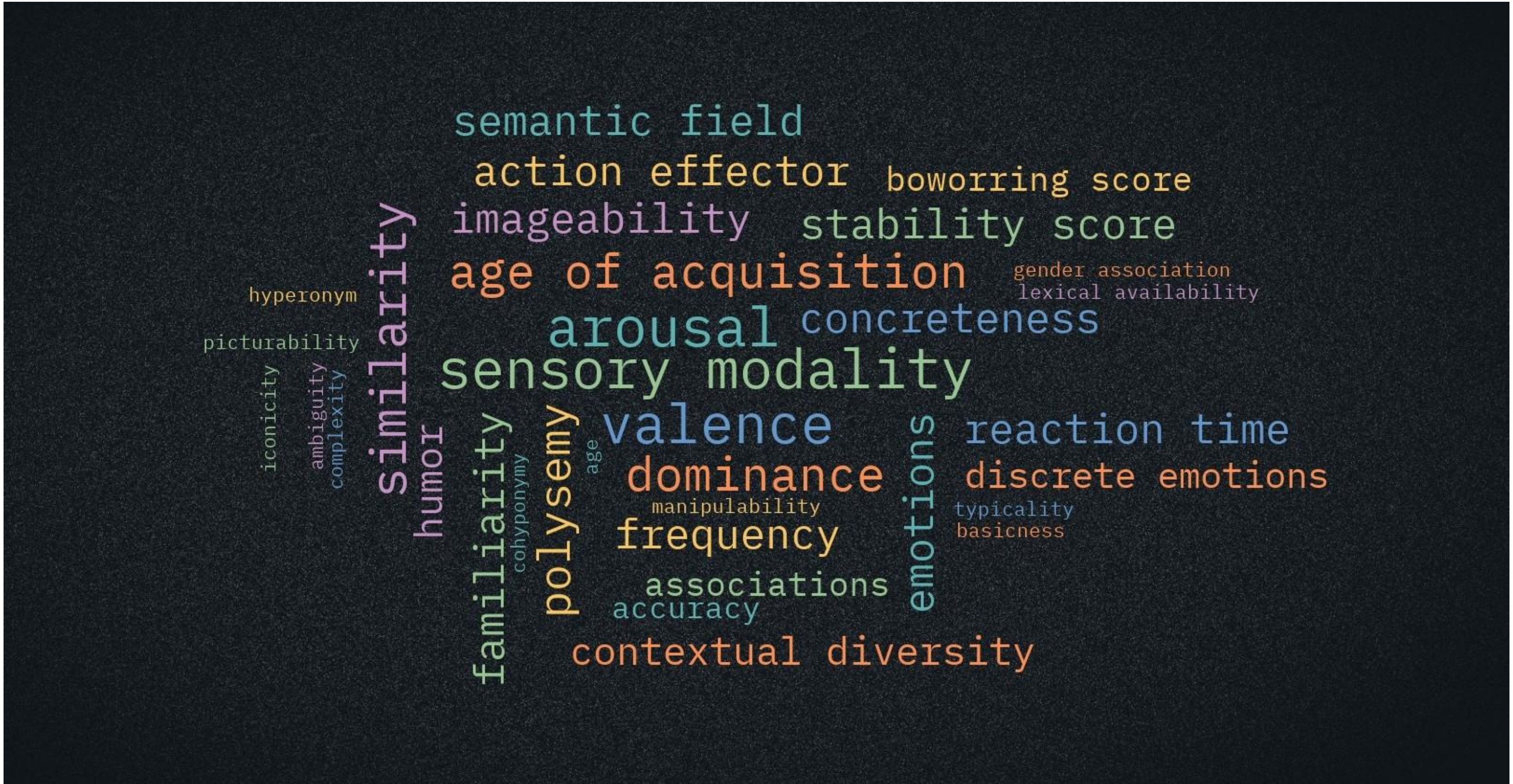
No Ra Re

- **A cross-linguistic database of norms, ratings, and relations for words and concepts. Building on Concepticon, it integrates data from psychology and linguistics.**

- **NoRaRe 0.2 (Tjuka et al. 2021)**
  - 98 data sets
  - 65 unique word properties
  - 40 languages

- **Website: https://digling.org/norare/**
- **GitHub: https://github.com/concepticon/norare-data**
- **Article: https://doi.org/10.3758/s13428-021-01650-1**

Word Properties



Norms, Ratings, and Relations

Norms

- Include data that are collected by taking samples from a total quantity.
- Collected and applied predominantly in the field of psychology.
- Examples:
  - word frequency
  - lexical decision

Ratings

- Based on participant judgments of a given word in a particular language either on a scale or on other measures.
- Examples:
  - age-of-acquisition
  - emotional states
  - sensory modality

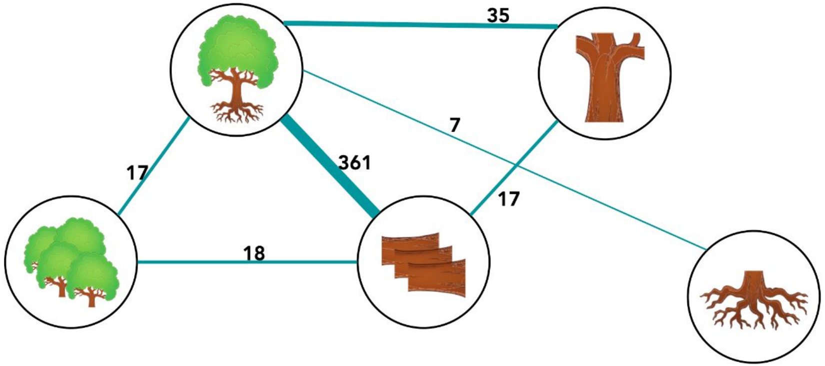
Relations

- Information on the relation between two words or concepts.
- Collected in the field of comparative linguistics and Natural Language Processing (NLP).
- Examples:
  - colexifications
  - stability rankings
  - associations

Data Structure

		402	906	1803	344	670
		FOREST	TREE	WOOD	TREE TRUNK	ROOT
🇬🇧	Frequency (log10)	3.06	3.52	2.78	3.00	2.72
🇩🇪	Frequency (log10)	3.09	3.01	2.74	2.42	0.47
🇫🇷	Arousal (mean)	4.00	2.97	3.51	3.18	NA
🇪🇸	Arousal (mean)	4.00	3.63	3.95	4.45	4.4

Cross-linguistic colexifications



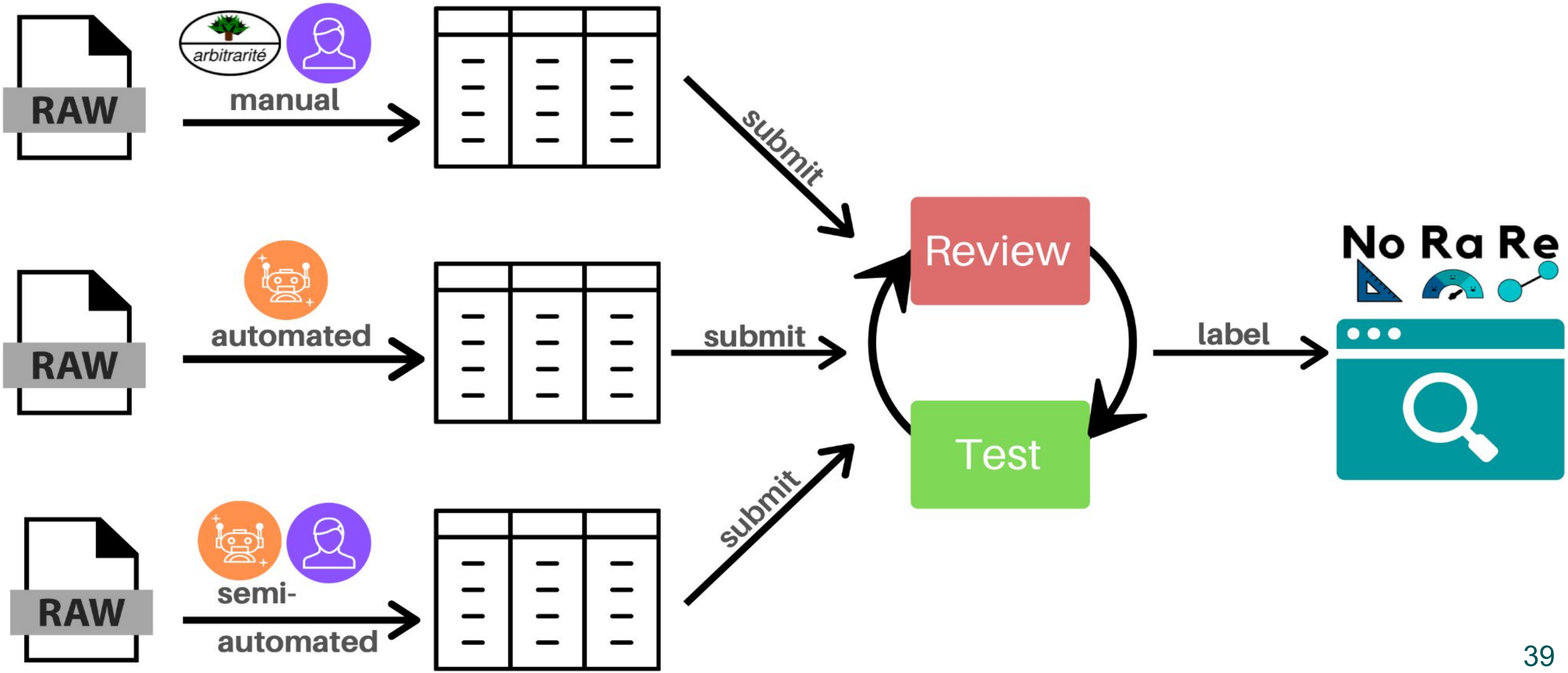
Data Curation

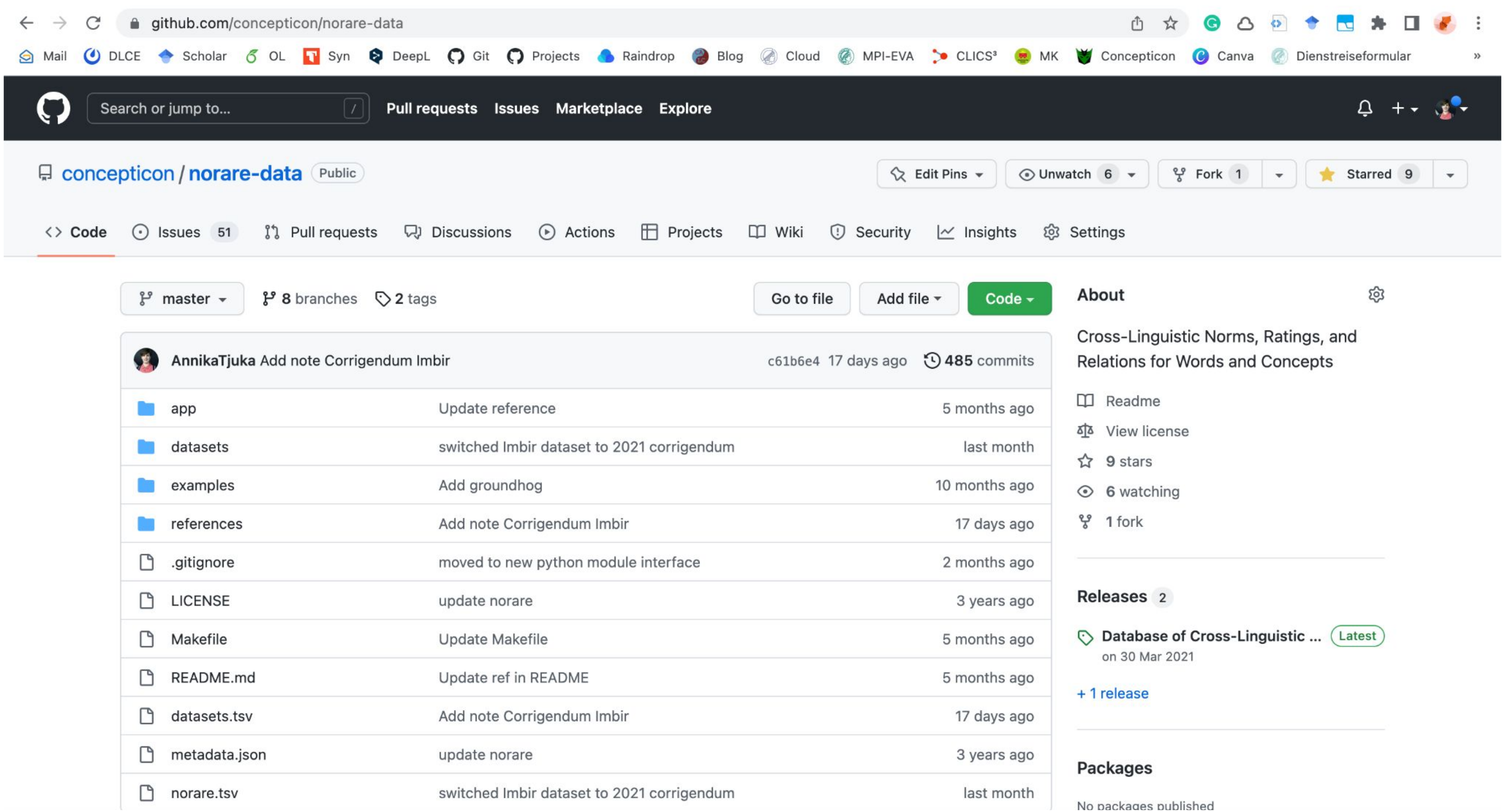
- **Manual, automated, and semi-automated mapping to Concepticon concept sets**
- **Information on data types in metadata.json**
- **Test-driven data curation**
- **Python package: `pynorare` (List & Forkel 2020)**
- **Accessed via command line**
- **Regular releases**

Tutorials

  - Tjuka (2021a; 2021b)

Overview of Data Curation Workflows





concepticon / norare-data Public

8 branches 2 tags

AnnikaTjuka Add note Corrigendum Imbir c61b6e4 17 days ago 485 commits

app	Update reference	5 months ago
datasets	switched Imbir dataset to 2021 corrigendum	last month
examples	Add groundhog	10 months ago
references	Add note Corrigendum Imbir	17 days ago
.gitignore	moved to new python module interface	2 months ago
LICENSE	update norare	3 years ago
Makefile	Update Makefile	5 months ago
README.md	Update ref in README	5 months ago
datasets.tsv	Add note Corrigendum Imbir	17 days ago
metadata.json	update norare	3 years ago
norare.tsv	switched Imbir dataset to 2021 corrigendum	last month

About

Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts

Readme

View license

9 stars

6 watching

1 fork

Releases 2

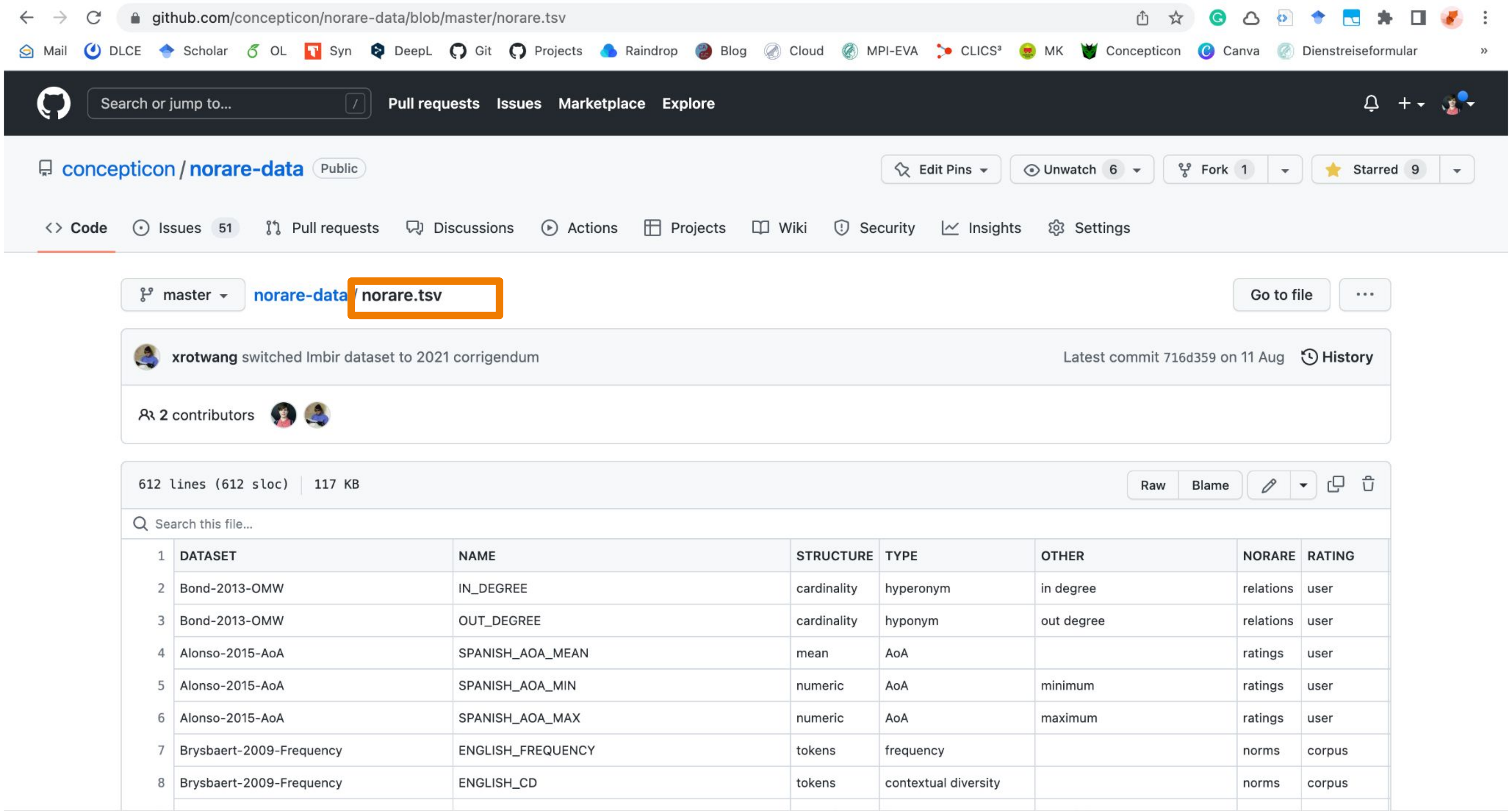
Database of Cross-Linguistic ... Latest
on 30 Mar 2021

+ 1 release

Packages

No packages published

Open: GitHub and Website

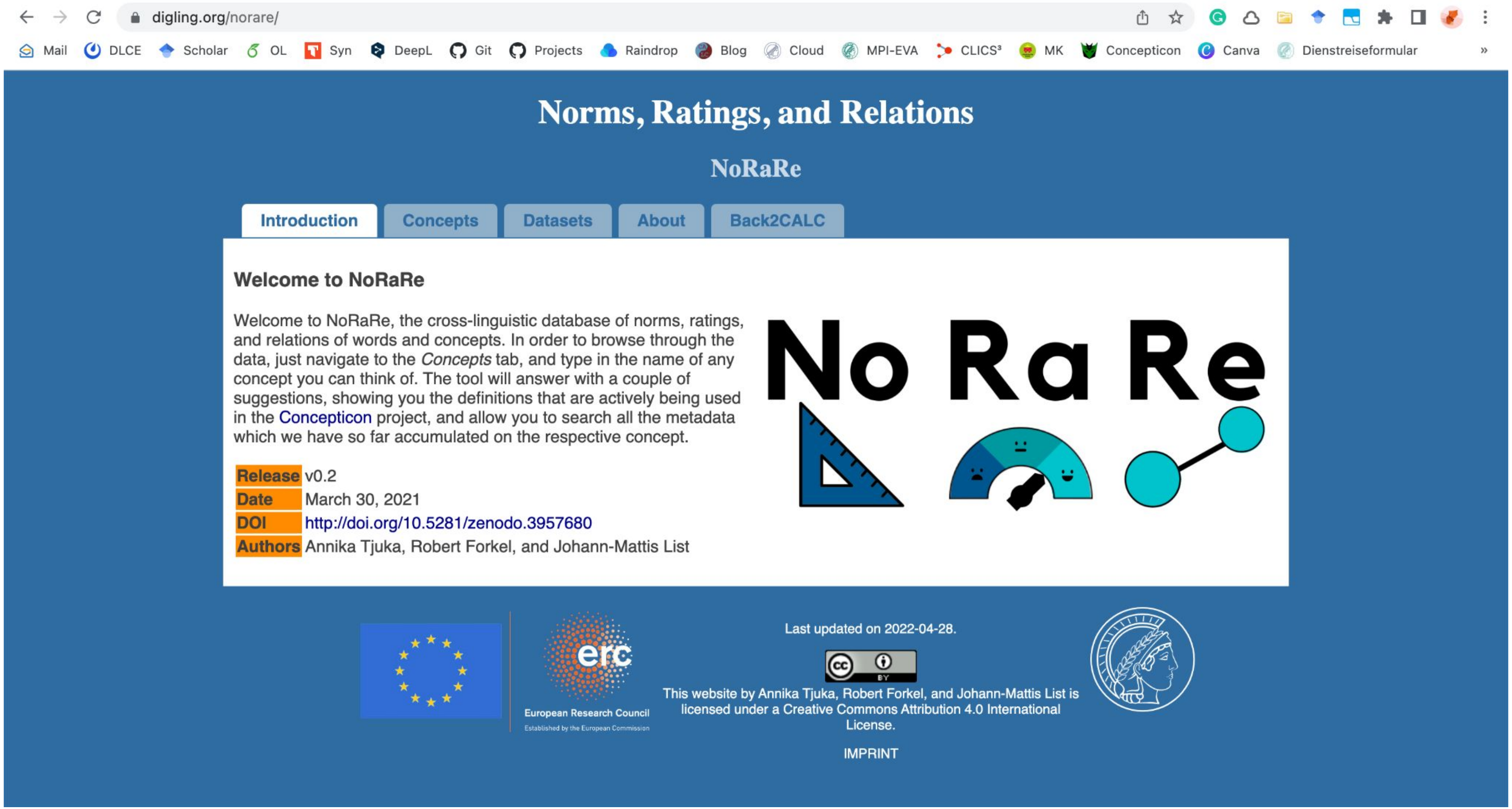


xrotwang switched lmbir dataset to 2021 corrigendum

Latest commit 716d359 on 11 Aug

612 lines (612 sloc) | 117 KB

DATASET	NAME	STRUCTURE	TYPE	OTHER	NORARE	RATING
Bond-2013-OMW	IN_DEGREE	cardinality	hyperonym	in degree	relations	user
Bond-2013-OMW	OUT_DEGREE	cardinality	hyponym	out degree	relations	user
Alonso-2015-AoA	SPANISH_AOA_MEAN	mean	AoA		ratings	user
Alonso-2015-AoA	SPANISH_AOA_MIN	numeric	AoA	minimum	ratings	user
Alonso-2015-AoA	SPANISH_AOA_MAX	numeric	AoA	maximum	ratings	user
Brysbaert-2009-Frequency	ENGLISH_FREQUENCY	tokens	frequency		norms	corpus
Brysbaert-2009-Frequency	ENGLISH_CD	tokens	contextual diversity		norms	corpus



Norms, Ratings, and Relations

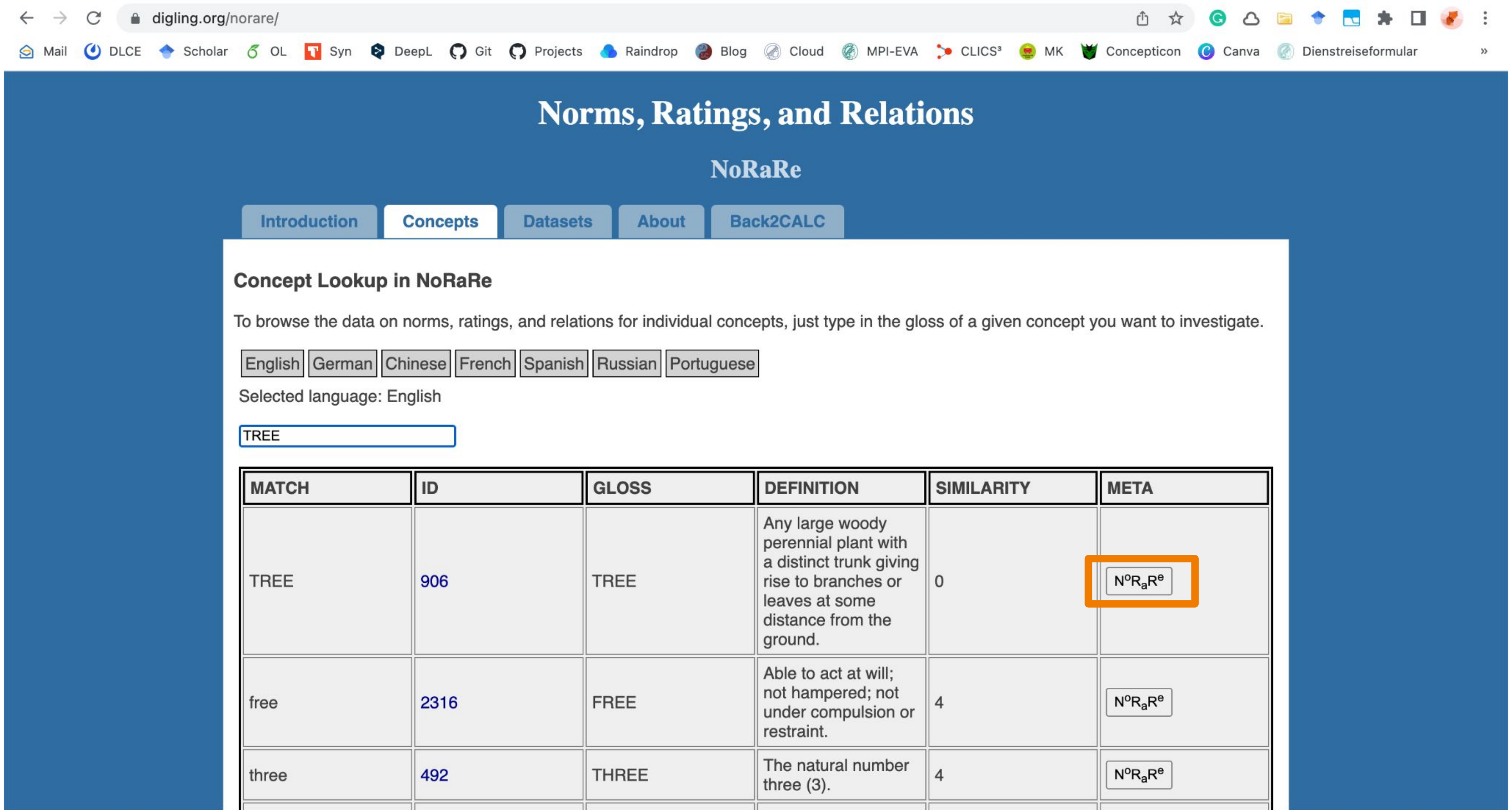
NoRaRe

Welcome to NoRaRe

Welcome to NoRaRe, the cross-linguistic database of norms, ratings, and relations of words and concepts. In order to browse through the data, just navigate to the *Concepts* tab, and type in the name of any concept you can think of. The tool will answer with a couple of suggestions, showing you the definitions that are actively being used in the Concepticon project, and allow you to search all the metadata which we have so far accumulated on the respective concept.

Release	v0.2
Date	March 30, 2021
DOI	http://doi.org/10.5281/zenodo.3957680
Authors	Annika Tjuka, Robert Forkel, and Johann-Mattis List

Last updated on 2022-04-28.



Norms, Ratings, and Relations

NoRaRe

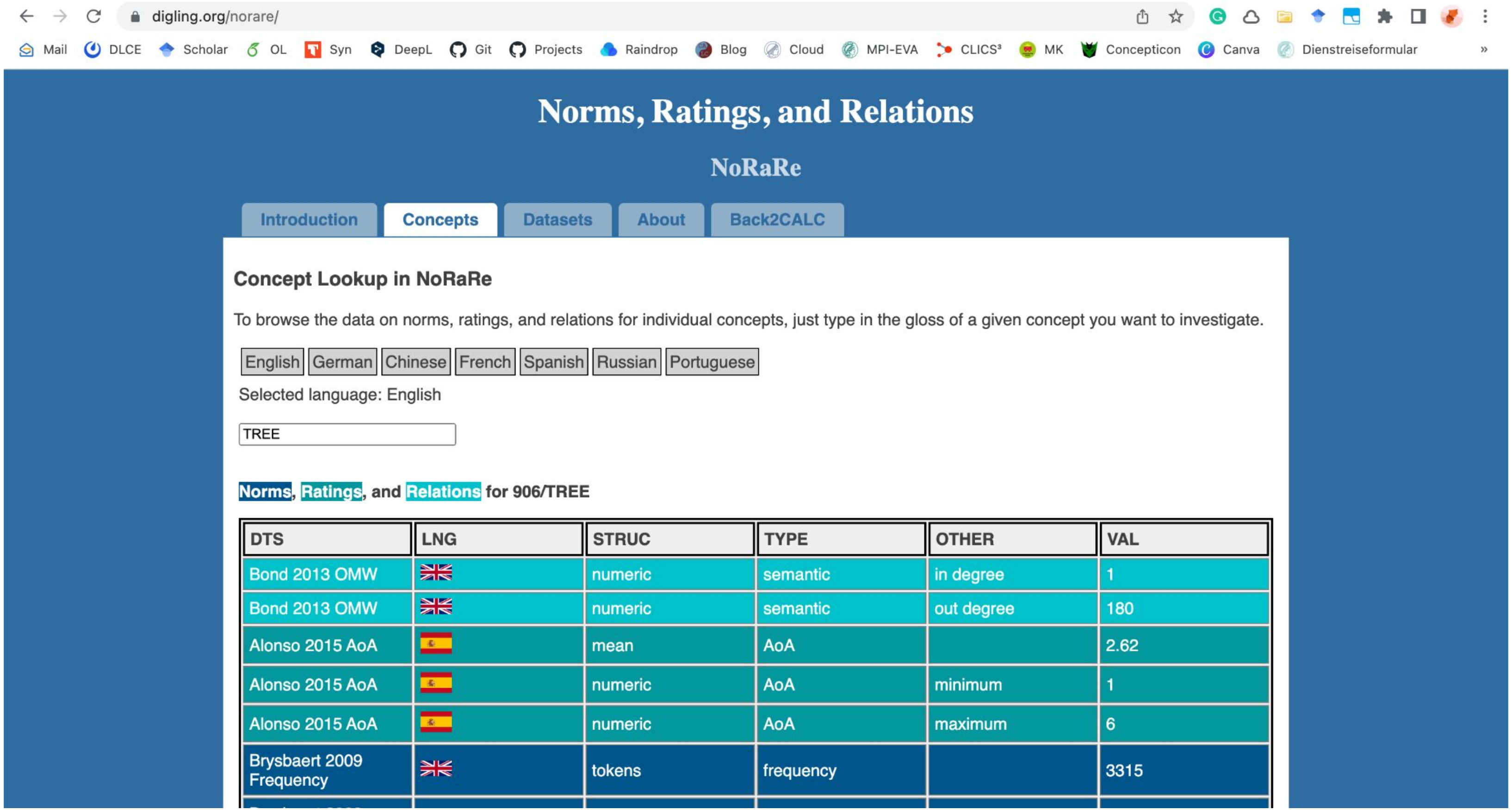
Concept Lookup in NoRaRe

To browse the data on norms, ratings, and relations for individual concepts, just type in the gloss of a given concept you want to investigate.

Selected language: English

TREE

MATCH	ID	GLOSS	DEFINITION	SIMILARITY	META
TREE	906	TREE	Any large woody perennial plant with a distinct trunk giving rise to branches or leaves at some distance from the ground.	0	$N^oR_aR^e$
free	2316	FREE	Able to act at will; not hampered; not under compulsion or restraint.	4	$N^oR_aR^e$
three	492	THREE	The natural number three (3).	4	$N^oR_aR^e$



Norms, Ratings, and Relations

NoRaRe

Concept Lookup in NoRaRe

To browse the data on norms, ratings, and relations for individual concepts, just type in the gloss of a given concept you want to investigate.

Selected language: English

TREE

Norms, Ratings, and Relations for 906/TREE

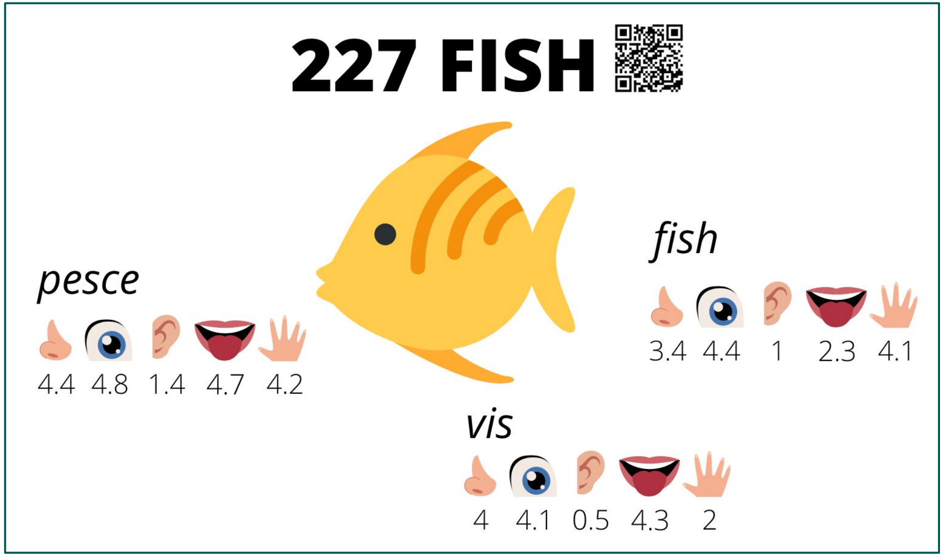
DTS	LNG	STRUC	TYPE	OTHER	VAL
Bond 2013 OMW	🇬🇧	numeric	semantic	in degree	1
Bond 2013 OMW	🇬🇧	numeric	semantic	out degree	180
Alonso 2015 AoA	🇪🇸	mean	AoA		2.62
Alonso 2015 AoA	🇪🇸	numeric	AoA	minimum	1
Alonso 2015 AoA	🇪🇸	numeric	AoA	maximum	6
Brysbaert 2009 Frequency	🇬🇧	tokens	frequency		3315

4

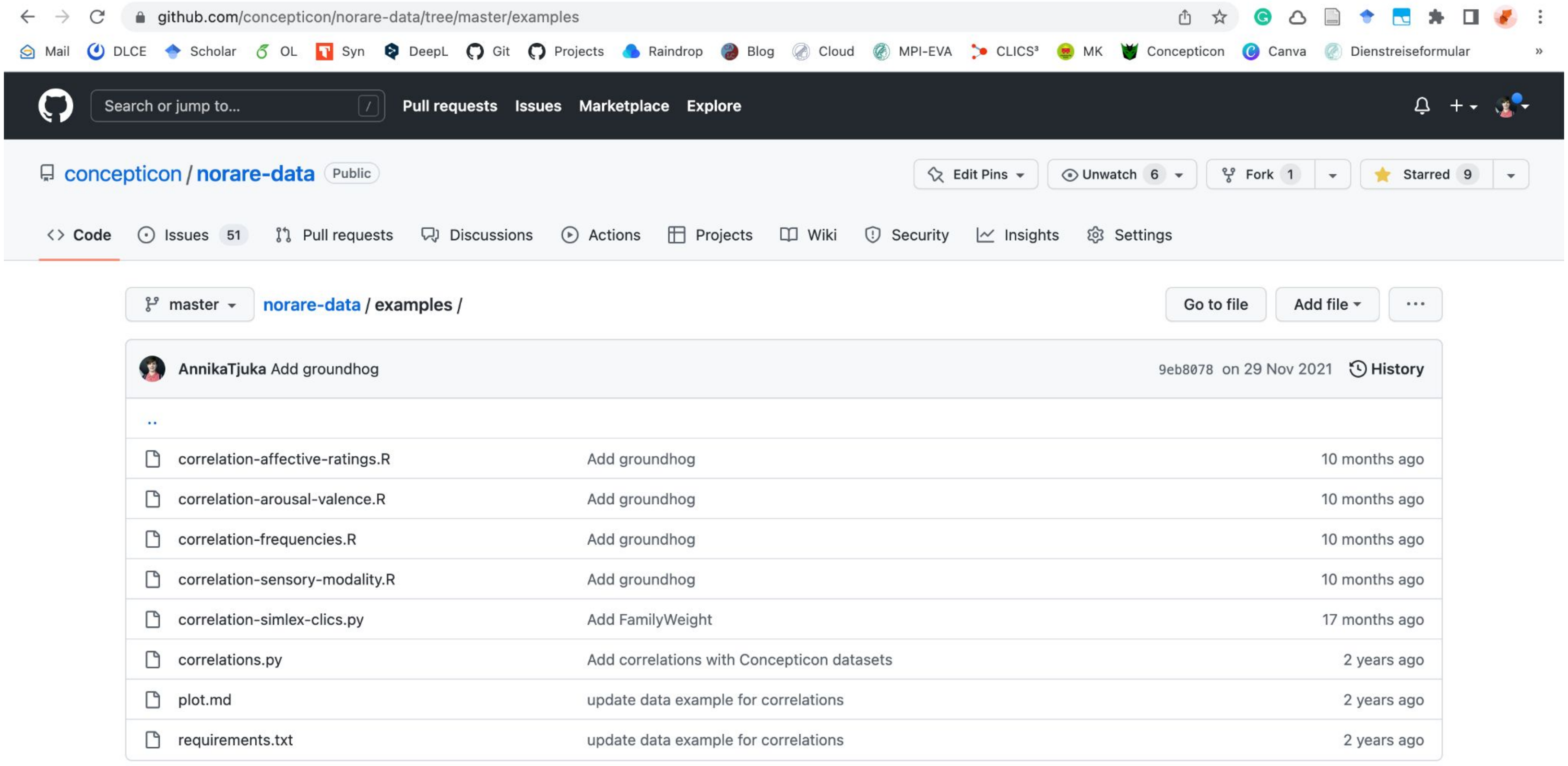
CASE STUDY: CROSS-LINGUISTIC COMPARISON OF SENSORY MODALITY RATINGS

Cross-Linguistic Comparison of Word Properties

- Variation in word meaning is an inevitable phenomenon that needs further investigation to broaden our understanding of human minds.
- Main questions: Are words in different languages expressing the same concept represented similarly?
- Study: Comparing sensory modality ratings of five dimensions (haptic, visual, olfactory, gustatory, and auditory) aross English, Italian, and Dutch.
- Data: English (Lynott et al., 2020), Dutch (Speed & Brysbaert, 2022), and Italian (Vergallito et al., 2020)



Open: GitHub

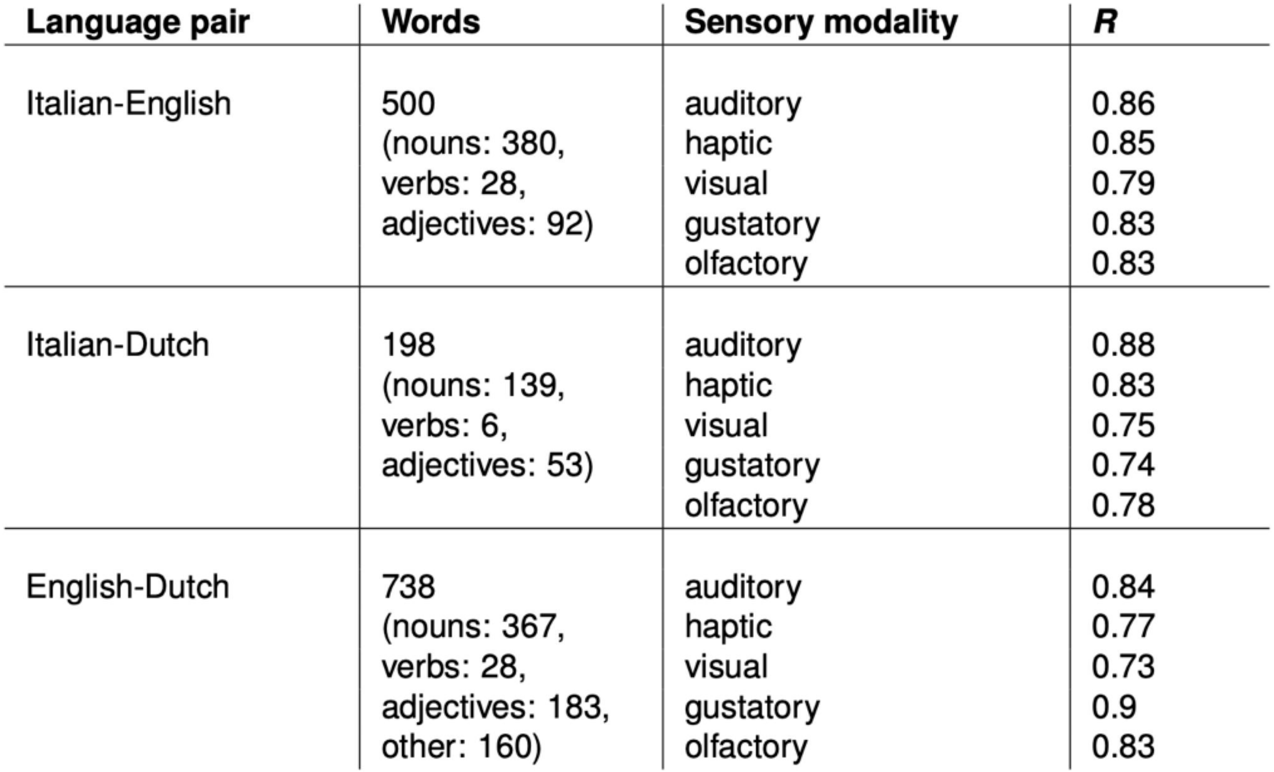


AnnikaTjuka Add groundhog · 9eb8078 on 29 Nov 2021

..		
correlation-affective-ratings.R	Add groundhog	10 months ago
correlation-arousal-valence.R	Add groundhog	10 months ago
correlation-frequencies.R	Add groundhog	10 months ago
correlation-sensory-modality.R	Add groundhog	10 months ago
correlation-simlex-clics.py	Add FamilyWeight	17 months ago
correlations.py	Add correlations with Concepticon datasets	2 years ago
plot.md	update data example for correlations	2 years ago
requirements.txt	update data example for correlations	2 years ago

Cross-Linguistic Comparison of Word Properties

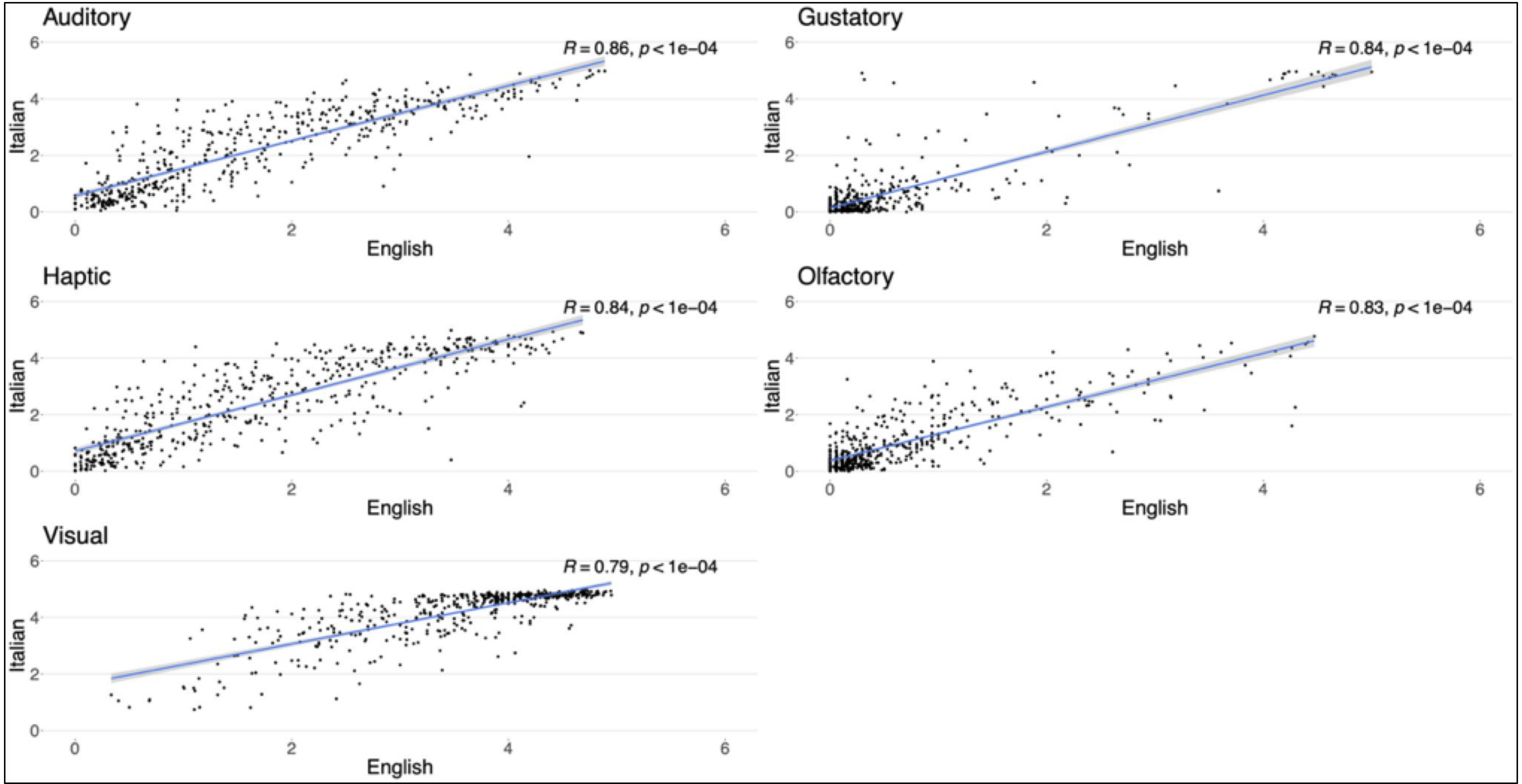
Results



Language pair	Words	Sensory modality	*R*
Italian-English	500 (nouns: 380, verbs: 28, adjectives: 92)	auditory haptic visual gustatory olfactory	0.86 0.85 0.79 0.83 0.83
Italian-Dutch	198 (nouns: 139, verbs: 6, adjectives: 53)	auditory haptic visual gustatory olfactory	0.88 0.83 0.75 0.74 0.78
English-Dutch	738 (nouns: 367, verbs: 28, adjectives: 183, other: 160)	auditory haptic visual gustatory olfactory	0.84 0.77 0.73 0.9 0.83

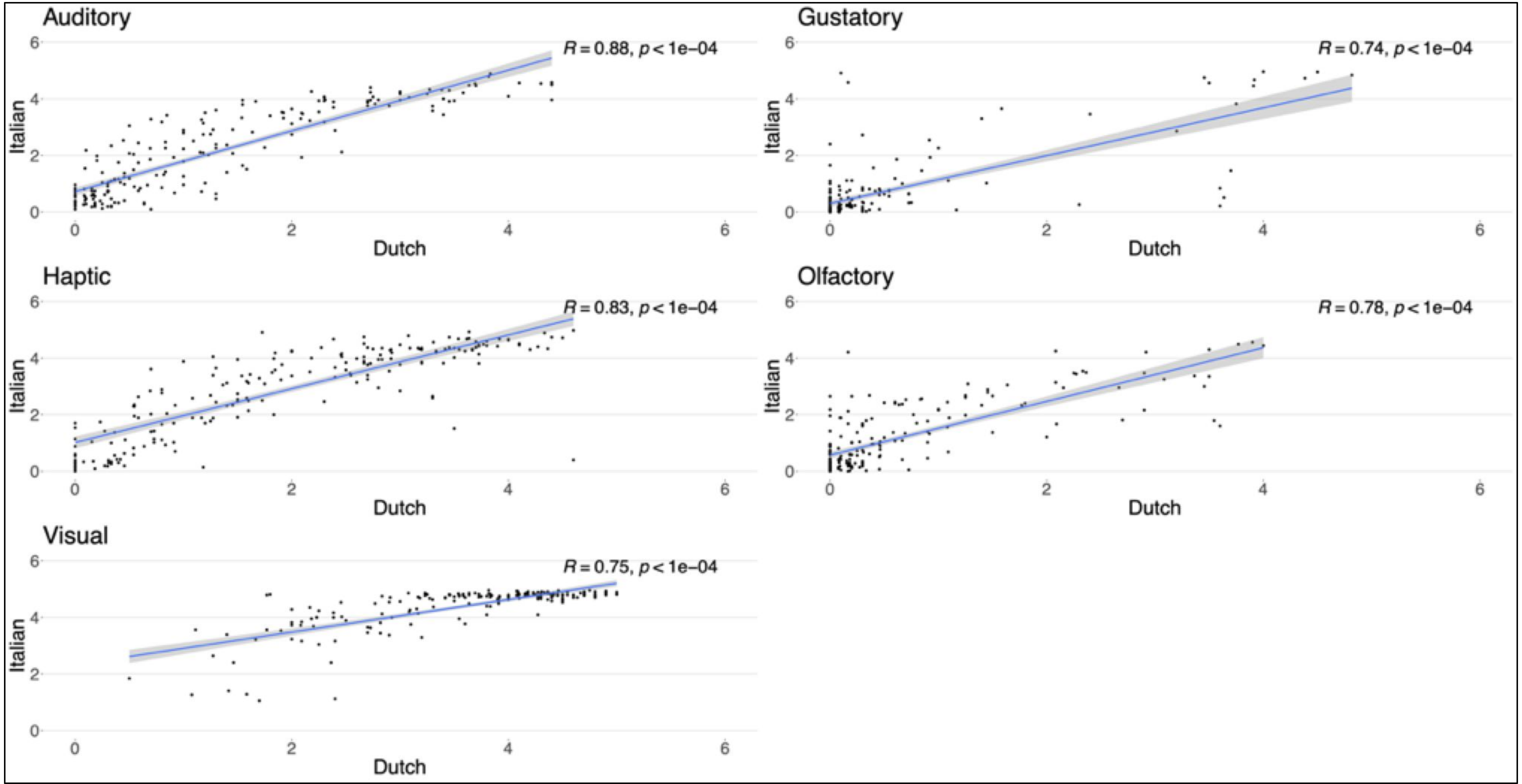
Cross-Linguistic Comparison of Word Properties

Italian-English Comparison



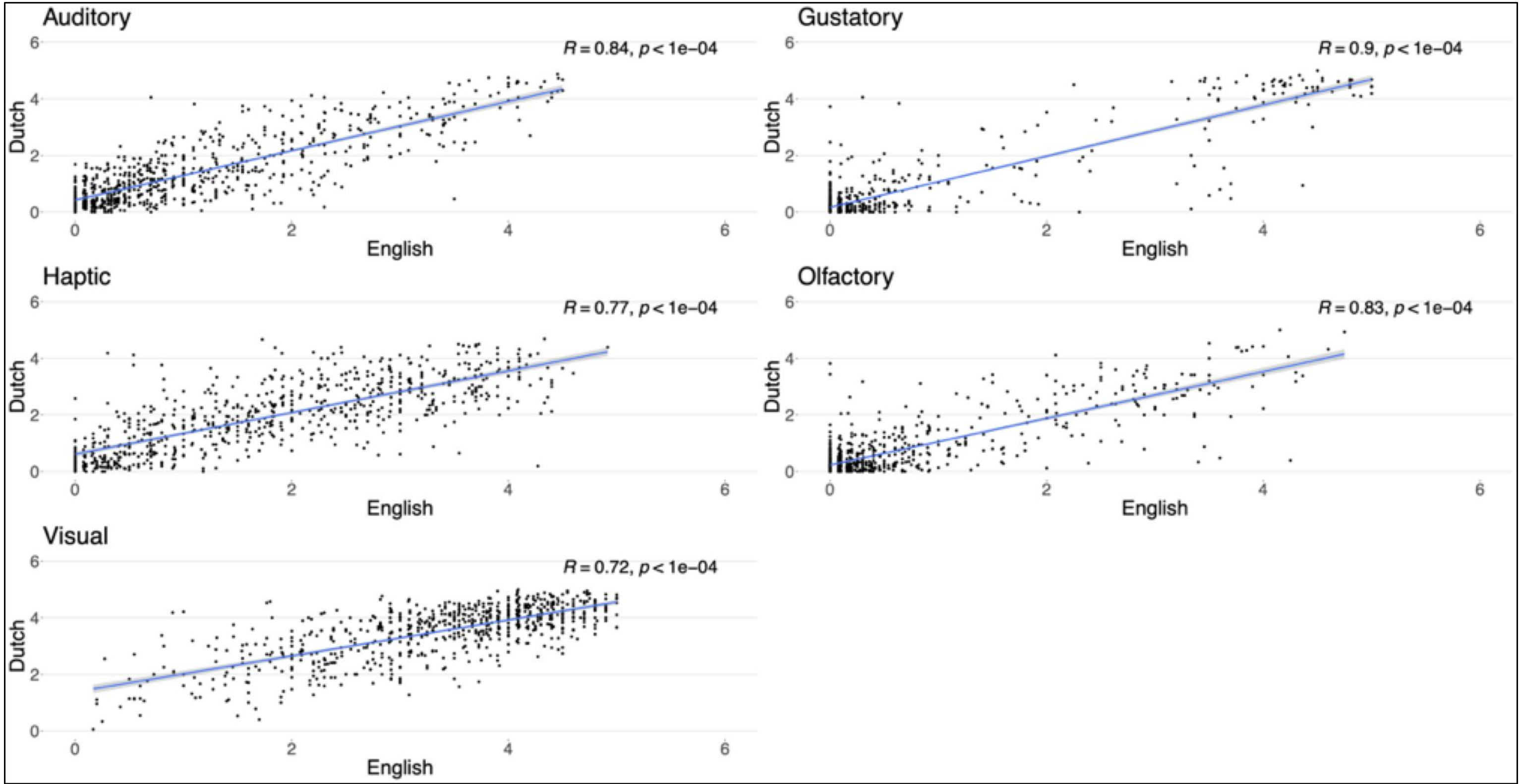
Cross-Linguistic Comparison of Word Properties

Italian-Dutch Comparison



Cross-Linguistic Comparison of Word Properties

Dutch-English Comparison



Cross-Linguistic Comparison of Word Properties

Conclusions & Implications

- **Sensory properties of words are perceived similarly across English, Dutch, and Italian speakers.**

- **Only slight differences across individual sensory modalities and language pairs occur.**

- **Additional data for various languages with the same rating scale need to be collected before a general claim can be made about the perception of sensory properties of words across cultures.**

Interim Summary

- Data must be standardized, human- and machine-readable.
- We use a test-driven data curation approach.
- Regular data releases ensure transparency, improvements, and extensions.
- Interoperability of the data allows correlations studies.

MANY THANKS
FOR YOUR ATTENTION

If you have any questions, please contact:

Annika Tjuka

annika_tjuka@eva.mpg.de

@AnnikaTjuka

Johann-Mattis List

mattis_list@eva.mpg.de

@LinguList

Blog Posts

List, Johann-Mattis. 2018. Towards a history of concept list compilation in historical linguistics. Blog. History and Philosophy of the Language Sciences. https://hiphilangsci.net/2018/10/31/concept-list-compilation/. (29 December, 2020).

Tjuka, Annika. 2020. Adding concept lists to Concepticon: A guide for beginners. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2225. (28 December, 2020).

Tjuka, Annika. 2021a. How to review concept lists in collaboration (How to do X in linguistics 6). Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2680. (25 March, 2021).

Tjuka, Annika. 2021b. Adding data sets to NoRaRe: A guide for beginners. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2890. (31 December, 2021).

Tjuka, Annika. 2021c. Comparing NoRaRe data sets: Calculation of correlations and creation of plots in R. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/3109. (31 December, 2021).

Tresoldi, Tiago. 2019a. Using pyconcepticon to map concept lists. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/1820. (24 March, 2021).

Tresoldi, Tiago. 2019b. Using pyconcepticon to map concept lists (II). Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/1844. (24 March, 2021).

References

Forkel, Robert. 2022. CLDFViz. A Python library providing tools to visualize data from CLDF datasets (Version 0.8.0). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5162667.

Gast, Volker & Maria Koptjevskaja-Tamm. 2019. The areal factor in lexical typology. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), Aspects of Linguistic Variation, 43–82. Berlin/New York: Walter de Gruyter. https://doi.org/10.1515/9783110607963-003.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. Language 86(3). 663–687. https://doi.org/10.1353/lan.2010.0021.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray & Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. Science 366. 1517–1522. https://doi.org/10.1126/science.aaw8160.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1379/.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9(1). 316. https://doi.org/10.1038/s41597-022-01432-0.

References

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel. 2018. CLICS²: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. Linguistic Typology 22(2). 277–306. https://doi.org/10.1515/lingty-2018-0010.

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle & Matthias Urban. 2014. CLICS: Database of cross-linguistic colexifications. Marburg: Forschungszentrum Deutscher Sprachatlas. http://clics.lingpy.org.

List, Johann-Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt & Robert Forkel. 2021. Concepticon. A resource for the linking of concept lists (Version 2.5.0). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.596412.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. Scientific Data 7(1). 1–12. https://doi.org/10.1038/s41597-019-0341-x.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2021. NoRaRe. A database of cross-linguistic norms, ratings, and relations for words and concepts (Version 0.2). Jena, Germany: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4647878.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking norms, ratings, and relations of words and concepts across multiple language varieties. Behavior Research Methods 54. 864–884. https://doi.org/10.3758/s13428-021-01650-1.

MAX PLANCK INSTITUTE
FOR EVOLUTIONARY ANTHROPOLOGY

Computer-Assisted Approaches to Lexical Typology

Semantic Shifts: From lexicon to grammar. Diachronic and typological perspectives
September 2022

Annika Tjuka and Johann-Mattis List
Department of Linguistic and Cultural Evolution

Agenda

Day 1

1 Introduction

2 Concepticon

3 Norms, Ratings, and Relations

4 Case Study: Cross-linguistic Comparison of Sensory Modality Ratings

Day 2

5 Lexibank

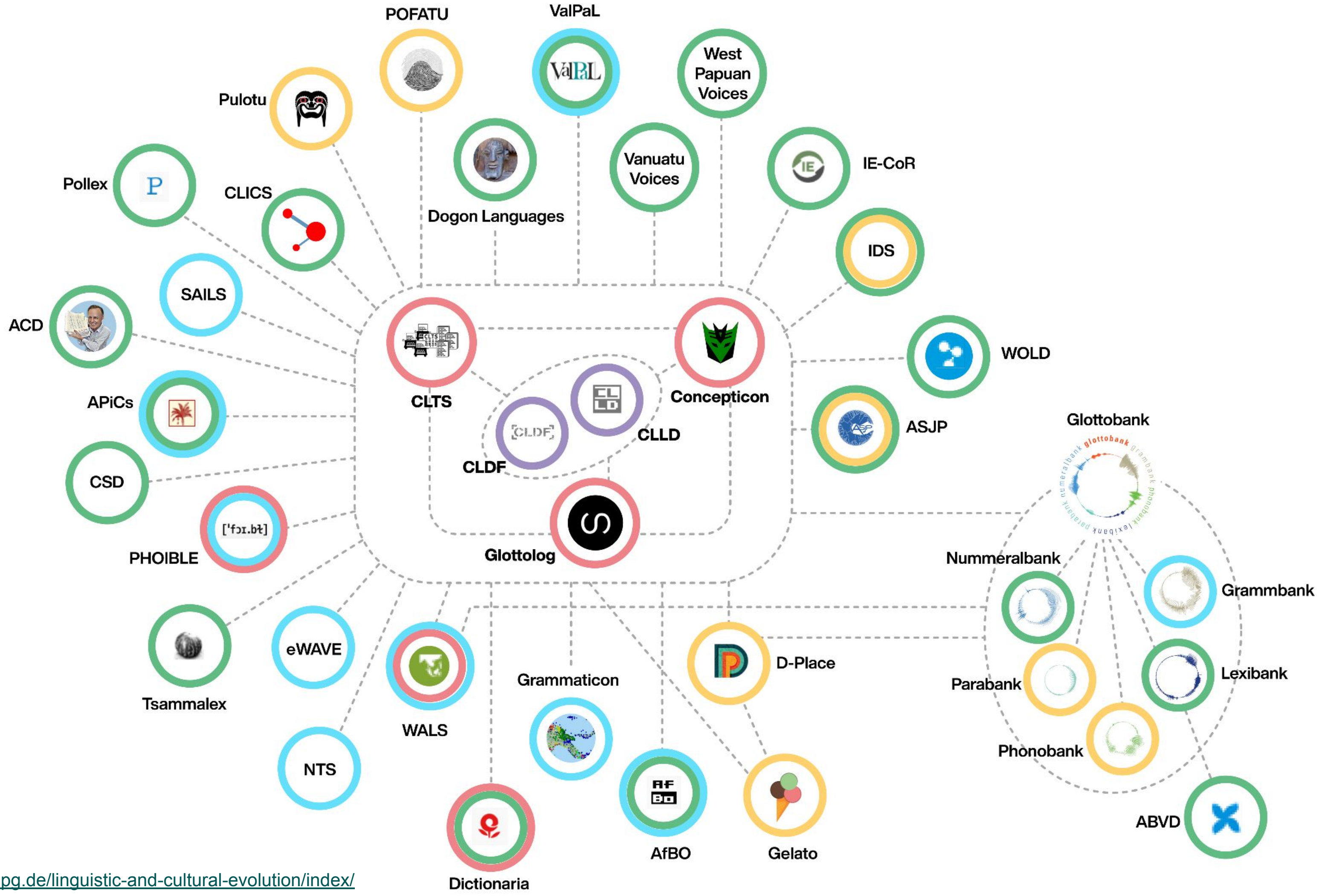
6 Partial Colexifications

7 Case Study: Emotion Colexifications

8 Case Study: Body Colexifications

5

LEXIBANK



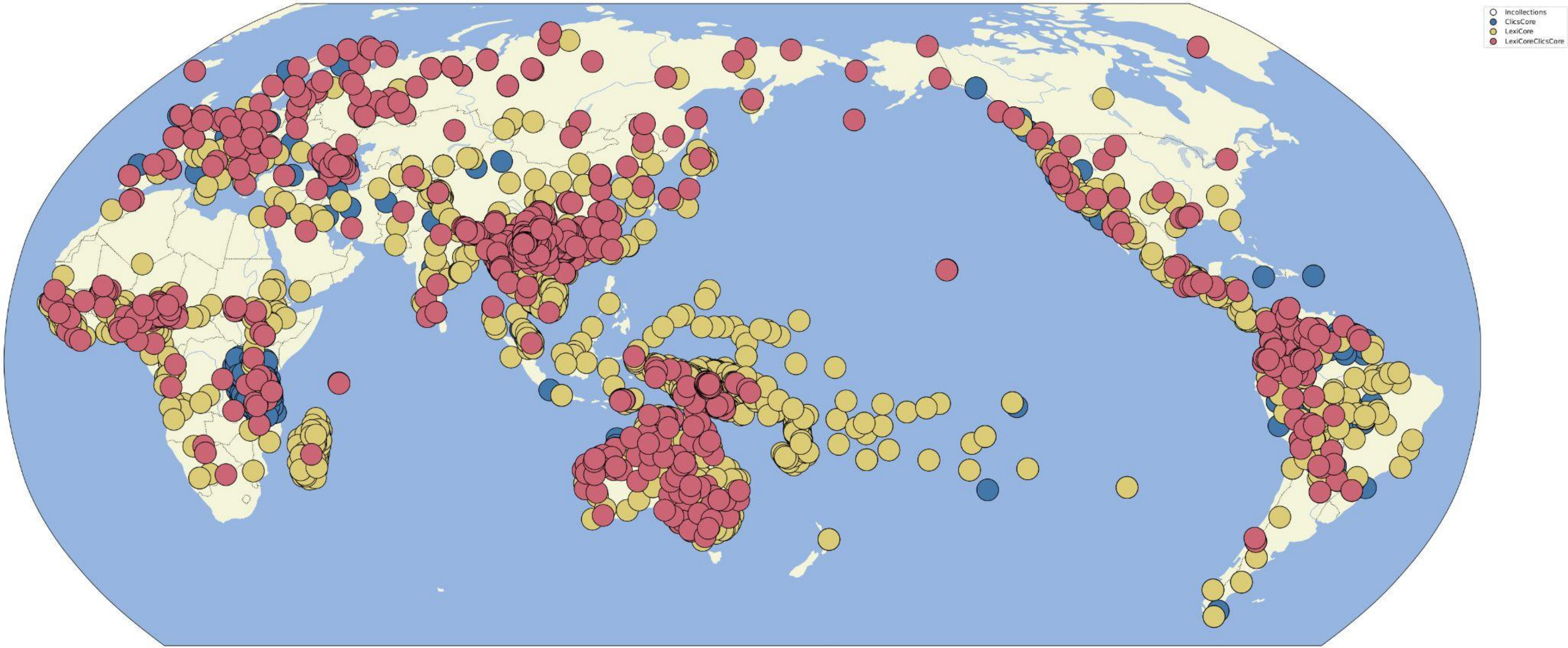
https://www.eva.mpg.de/linguistic-and-cultural-evolution/index/

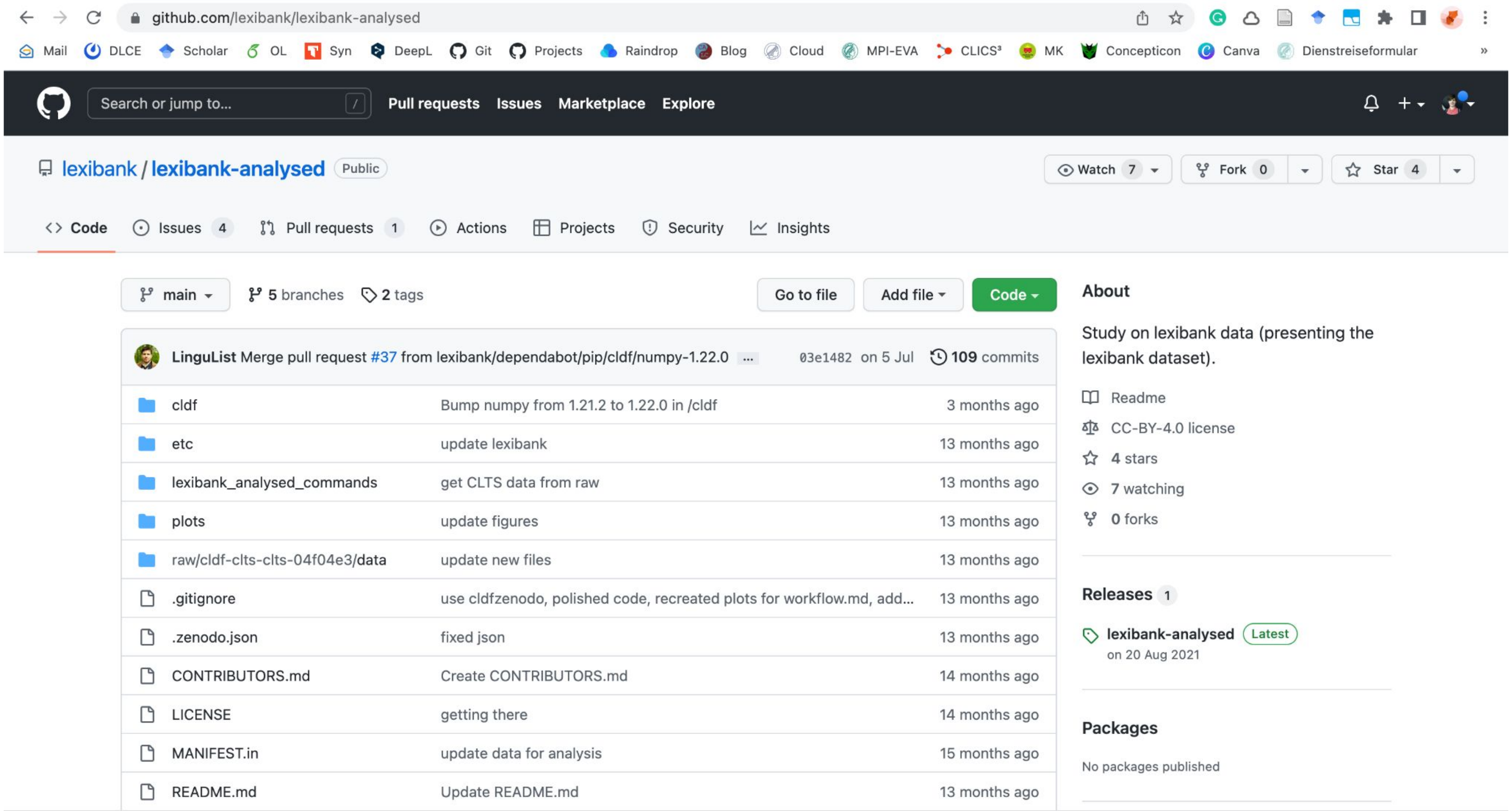
Lexibank

- **A public repository of standardized concept lists with computed phonological and lexical features. The data include unified phonetic transcriptions covering more than 2000 language varieties. It is the largest cross-linguistic lexical data collection to date.**

- **Lexibank 0.1 (List et al. 2022)**
  - 100 CLDF data sets based on 4069 concept lists from 2456 language varieties
    - LexiCore: 3320 concept lists from 94 data sets with fully standardized phonetic transcriptions for at least 80 word forms; 1806 concept lists from 52 data sets for at least 250 word forms
    - CogCore: 1441 concept lists from 49 data sets with information on etymologically related words
    - ClicsCore: 50 data sets across 1784 different language varieties with at least 250 concepts
- **GitHub: https://github.com/lexibank/lexibank-analysed**
- **Article: List et al. (2022)**
  - FAZ article (only in German): https://tinyurl.com/mwa73xn8

Data Distribution





lexibank / lexibank-analysed

LinguList Merge pull request #37 from lexibank/dependabot/pip/cldf/numpy-1.22.0 … 03e1482 on 5 Jul 109 commits

cldf	Bump numpy from 1.21.2 to 1.22.0 in /cldf	3 months ago
etc	update lexibank	13 months ago
lexibank_analysed_commands	get CLTS data from raw	13 months ago
plots	update figures	13 months ago
raw/cldf-clts-clts-04f04e3/data	update new files	13 months ago
.gitignore	use cldfzenodo, polished code, recreated plots for workflow.md, add…	13 months ago
.zenodo.json	fixed json	13 months ago
CONTRIBUTORS.md	Create CONTRIBUTORS.md	14 months ago
LICENSE	getting there	14 months ago
MANIFEST.in	update data for analysis	15 months ago
README.md	Update README.md	13 months ago

About

Study on lexibank data (presenting the lexibank dataset).

- Readme
- CC-BY-4.0 license
- 4 stars
- 7 watching
- 0 forks

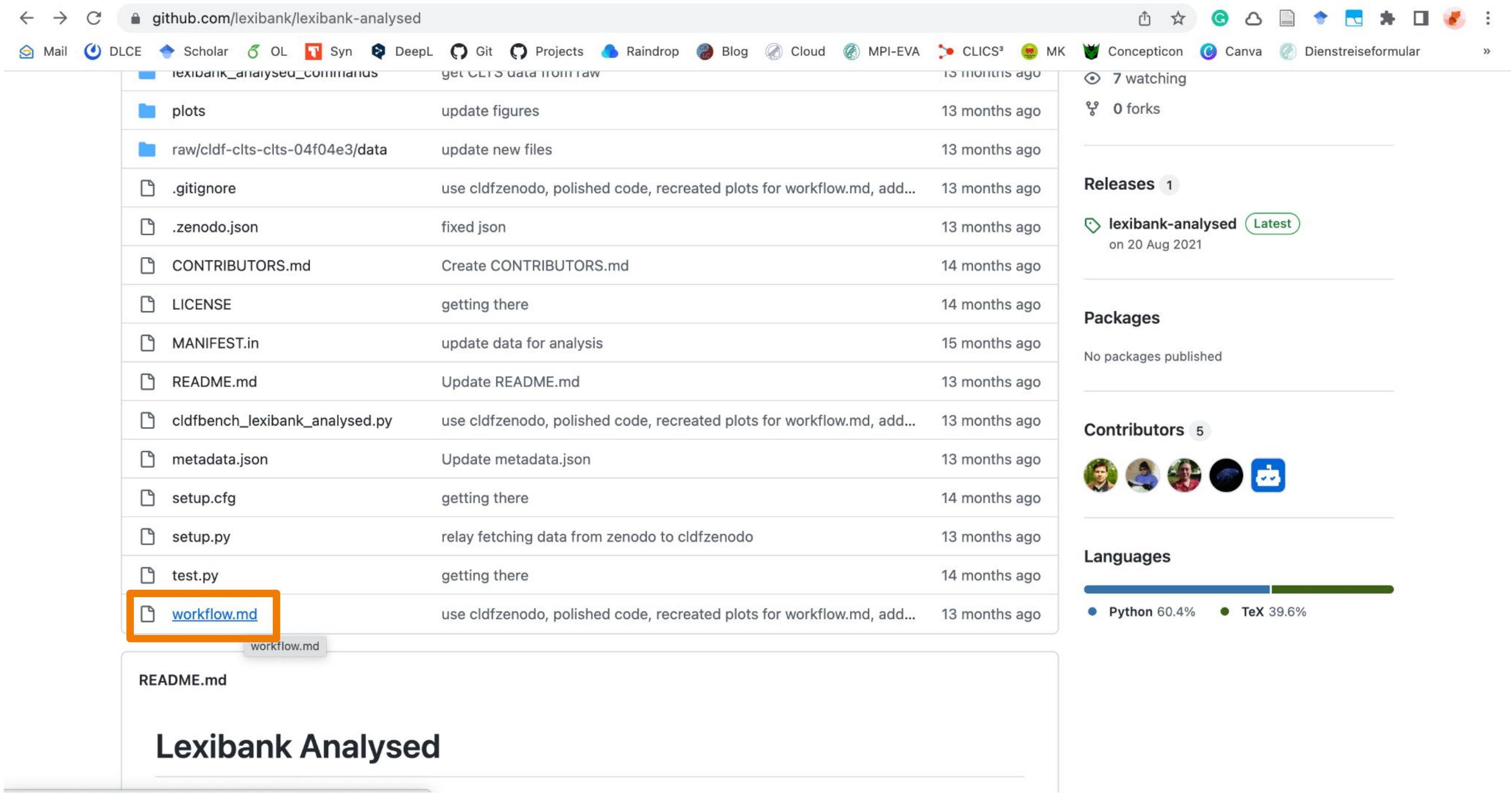
Releases 1

lexibank-analysed Latest
on 20 Aug 2021

Packages

No packages published

Open: GitHub



lexibank_analysed_commands	get CLTS data from raw	13 months ago
plots	update figures	13 months ago
raw/cldf-clts-clts-04f04e3/data	update new files	13 months ago
.gitignore	use cldfzenodo, polished code, recreated plots for workflow.md, add...	13 months ago
.zenodo.json	fixed json	13 months ago
CONTRIBUTORS.md	Create CONTRIBUTORS.md	14 months ago
LICENSE	getting there	14 months ago
MANIFEST.in	update data for analysis	15 months ago
README.md	Update README.md	13 months ago
cldfbench_lexibank_analysed.py	use cldfzenodo, polished code, recreated plots for workflow.md, add...	13 months ago
metadata.json	Update metadata.json	13 months ago
setup.cfg	getting there	14 months ago
setup.py	relay fetching data from zenodo to cldfzenodo	13 months ago
test.py	getting there	14 months ago
workflow.md	use cldfzenodo, polished code, recreated plots for workflow.md, add...	13 months ago

7 watching

0 forks

Releases 1

lexibank-analysed Latest
on 20 Aug 2021

Packages

No packages published

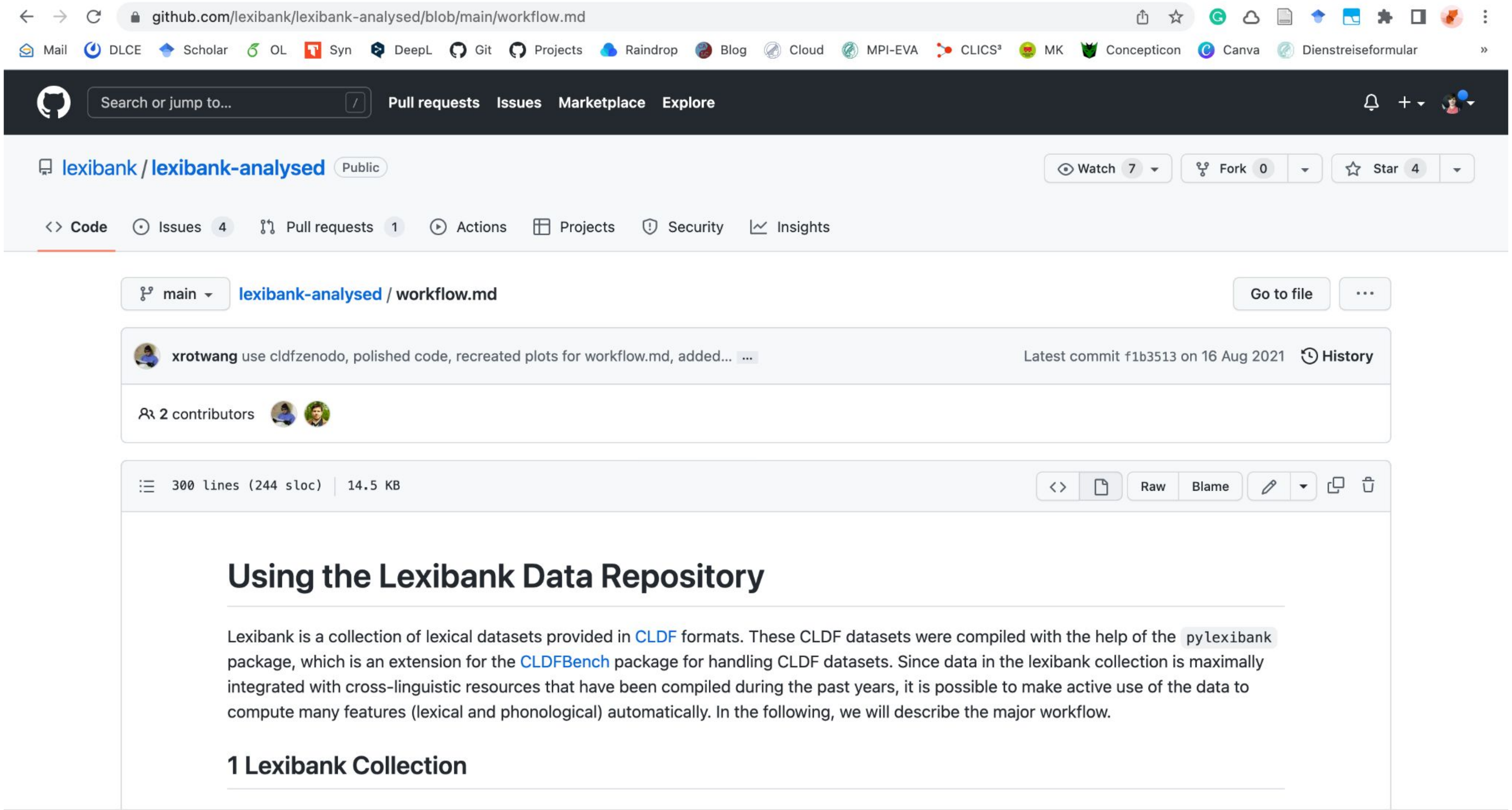
Contributors 5

Languages

● Python 60.4% ● TeX 39.6%

README.md

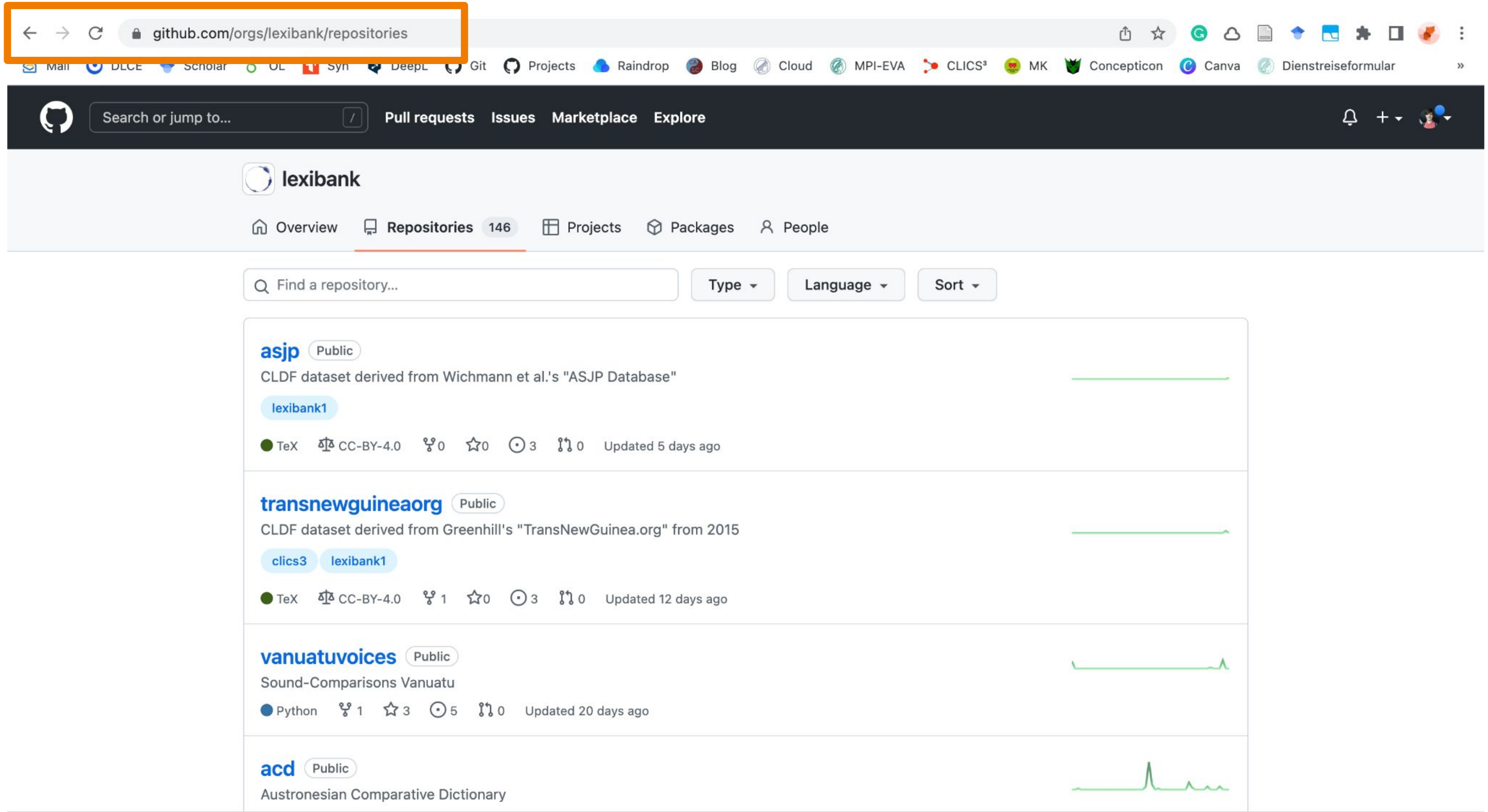
Lexibank Analysed



Using the Lexibank Data Repository

Lexibank is a collection of lexical datasets provided in CLDF formats. These CLDF datasets were compiled with the help of the `pylexibank` package, which is an extension for the CLDFBench package for handling CLDF datasets. Since data in the lexibank collection is maximally integrated with cross-linguistic resources that have been compiled during the past years, it is possible to make active use of the data to compute many features (lexical and phonological) automatically. In the following, we will describe the major workflow.

1 Lexibank Collection



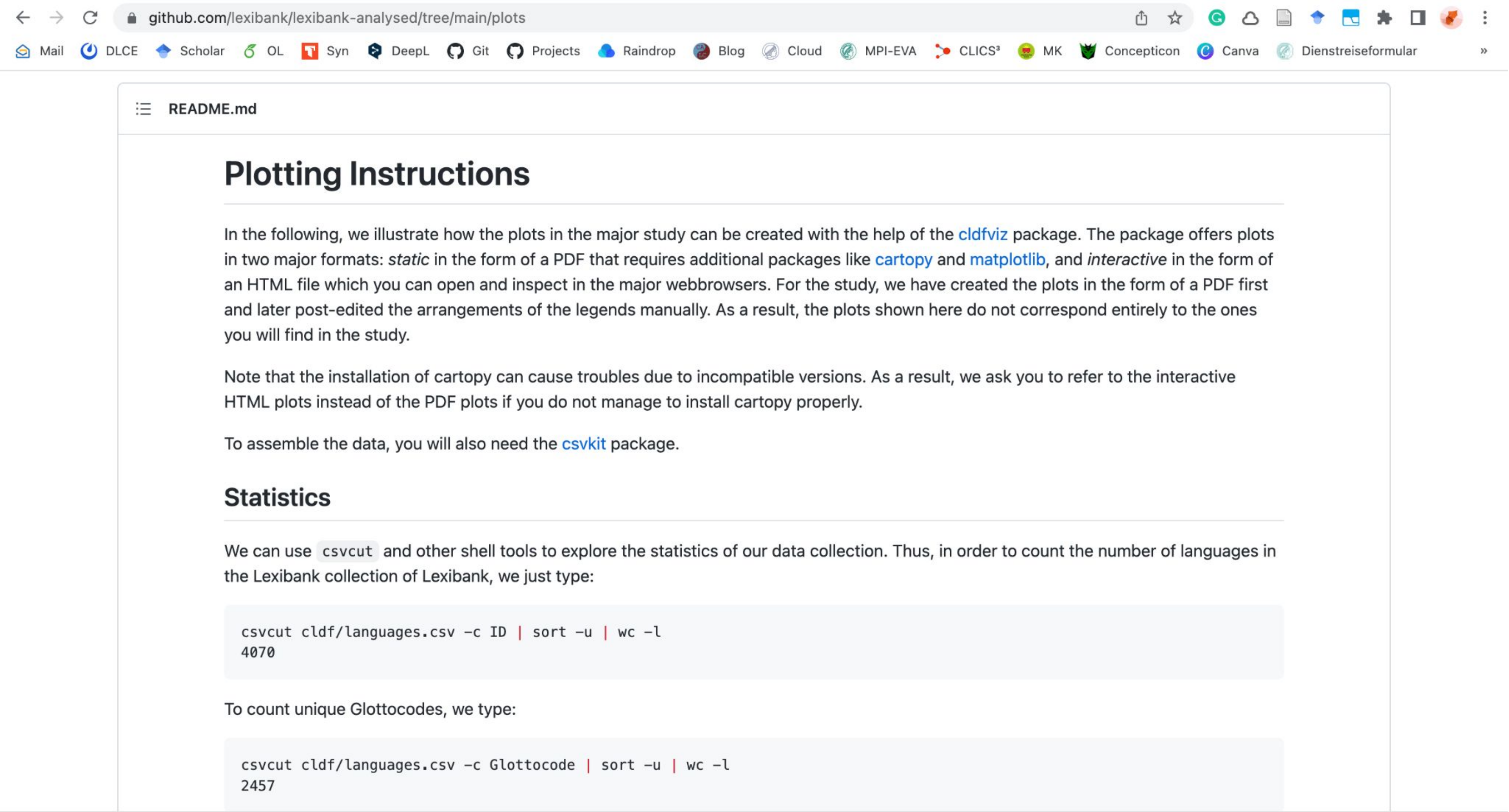
lexibank

asjp Public
CLDF dataset derived from Wichmann et al.'s "ASJP Database"
lexibank1
TeX CC-BY-4.0 0 ☆0 3 0 Updated 5 days ago

transnewguineaorg Public
CLDF dataset derived from Greenhill's "TransNewGuinea.org" from 2015
clics3 lexibank1
TeX CC-BY-4.0 1 ☆0 3 0 Updated 12 days ago

vanuatuvoices Public
Sound-Comparisons Vanuatu
Python 1 ☆ 3 5 0 Updated 20 days ago

acd Public
Austronesian Comparative Dictionary



Plotting Instructions

In the following, we illustrate how the plots in the major study can be created with the help of the cldfviz package. The package offers plots in two major formats: *static* in the form of a PDF that requires additional packages like cartopy and matplotlib, and *interactive* in the form of an HTML file which you can open and inspect in the major webbrowsers. For the study, we have created the plots in the form of a PDF first and later post-edited the arrangements of the legends manually. As a result, the plots shown here do not correspond entirely to the ones you will find in the study.

Note that the installation of cartopy can cause troubles due to incompatible versions. As a result, we ask you to refer to the interactive HTML plots instead of the PDF plots if you do not manage to install cartopy properly.

To assemble the data, you will also need the csvkit package.

Statistics

We can use `csvcut` and other shell tools to explore the statistics of our data collection. Thus, in order to count the number of languages in the Lexibank collection of Lexibank, we just type:

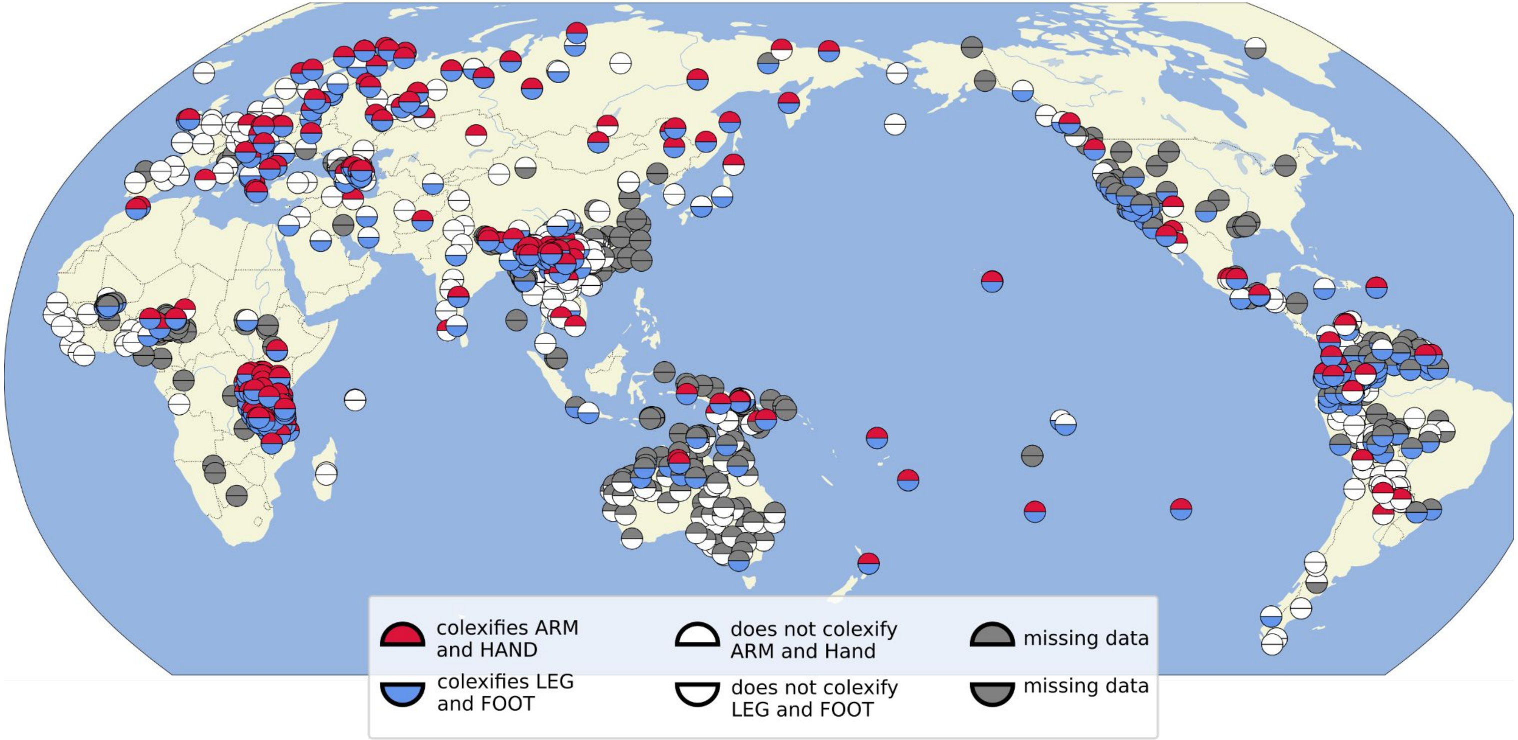
```
csvcut cldf/languages.csv -c ID | sort -u | wc -l
4070
```

To count unique Glottocodes, we type:

```
csvcut cldf/languages.csv -c Glottocode | sort -u | wc -l
2457
```

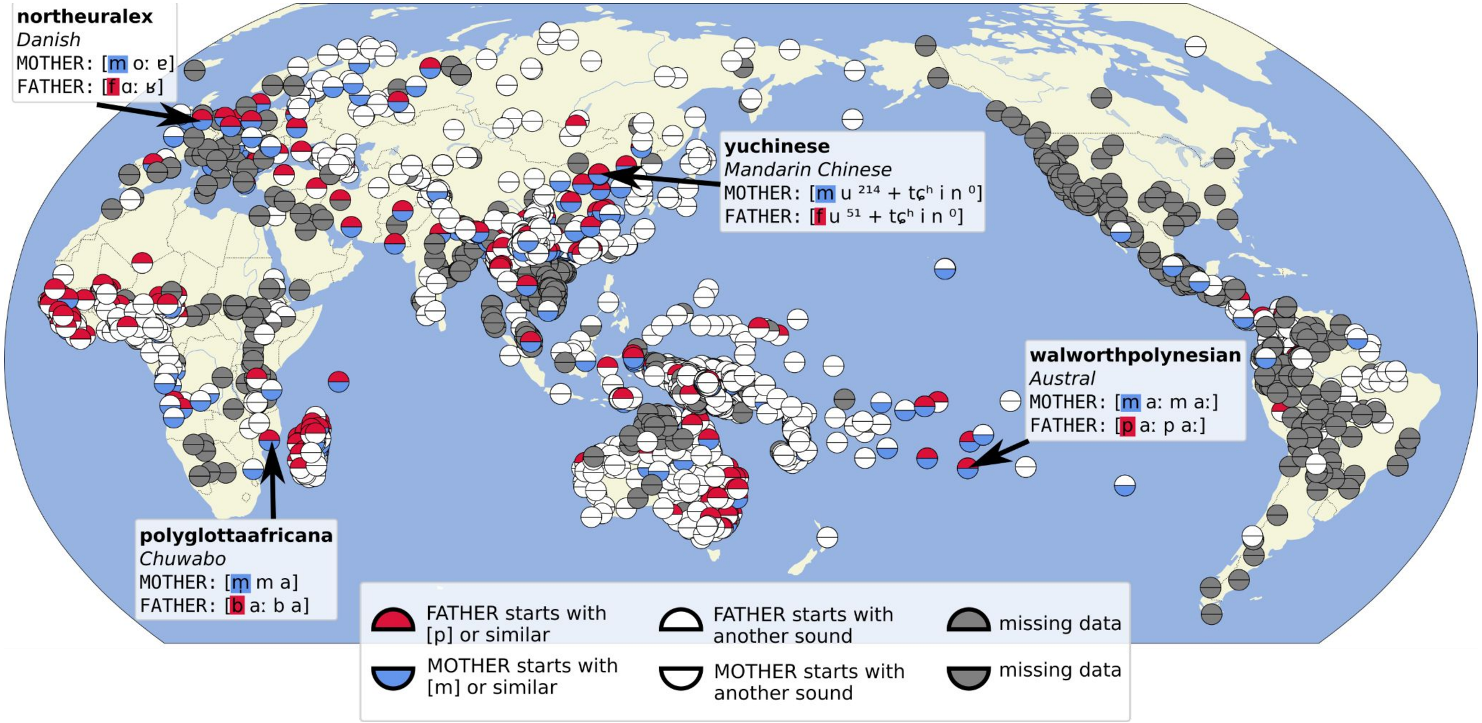
Lexical Features

Colexification of HAND-ARM (`HandAndArm`) and FOOT-LEG (`FootAndLeg`)



Sound-Symbolic Features

Start Sounds of FATHER (`FatherWithP`) and MOTHER (`MotherWithM`)



Lexibank

Conclusions & Implications

- **Lexibank can be seen as our current endpoint in a longer quest to assemble standardized lexical data for cross-linguistic large-scale approaches to lexical typology.**
- **The different versions of the CLICS database can be seen as the predecessors of Lexibank.**

- **Lexibank is a repository of standardized individual data sets, it is not a data set itself and should not be treated as such, this means specifically:**
  - In order to contribute, one has to standardize a data set in CLDF and we'll test it.
  - Errors must be changed on the level of individual data sets, not on the level of Lexibank itself.
  - Using Lexibank for one's own analyses requires a careful selection of useful data sets in order to arrive at a balanced sample of the data.

6

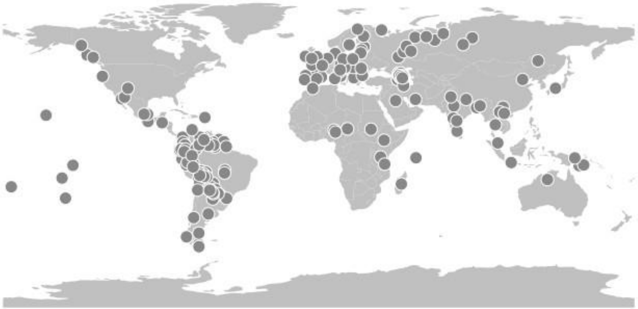
PARTIAL COLEXIFICATIONS

Database of Cross-Linguistic Colexifications (CLICS)

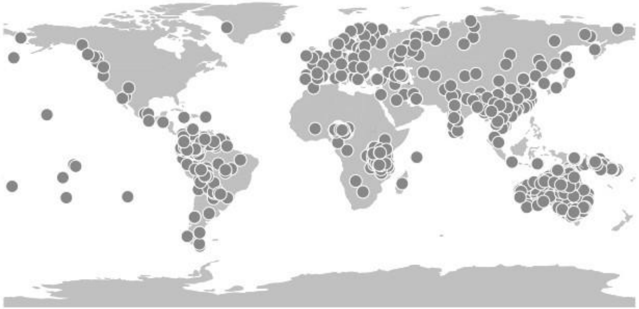
- Assemble lexical data sets coded in CLDF to extract colexification patterns automatically
- Use this to replace the not-so-easy-maintainable CLICS-1.0 database (List et al. 2014)
- Restrict the curation of the data to the selection of a couple of base data sets
- Add minimal Python code to infer colexifications from the data
- Create networks with Infomap algorithm for community detection
- Represent the data as a CLLD app

- GitHub: https://github.com/clics/clics3
- Article: Rzymski et al. (2020)

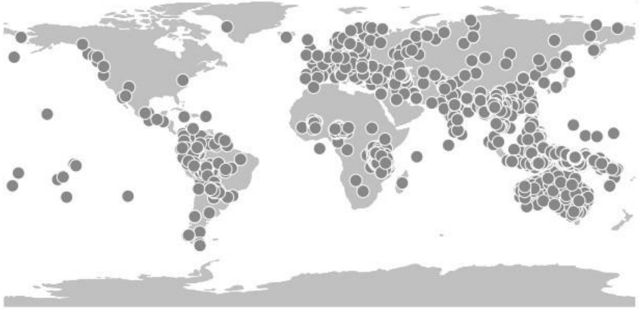
Database of Cross-Linguistic Colexifications



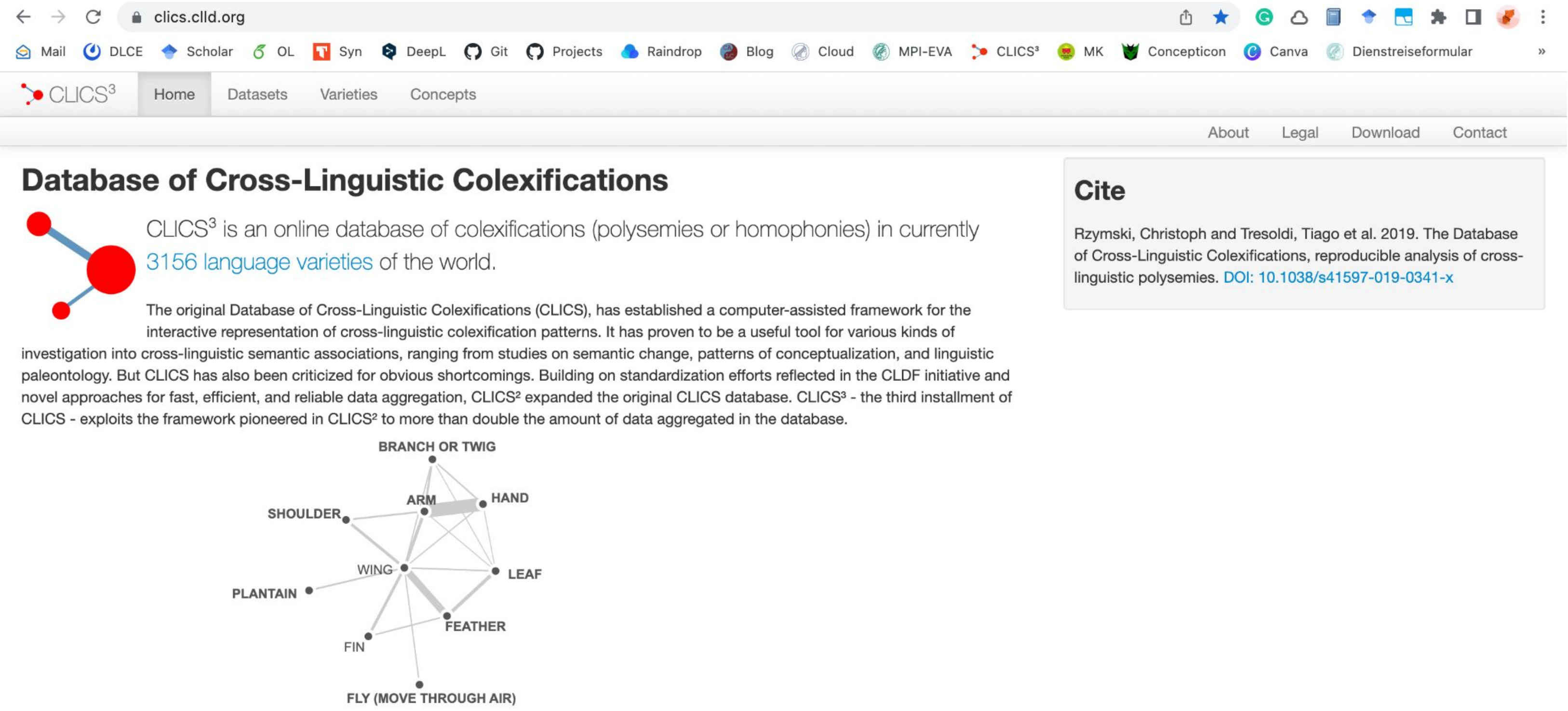
CLICS¹ (2014)



CLICS² (2018)



CLICS³ (2020)



Database of Cross-Linguistic Colexifications

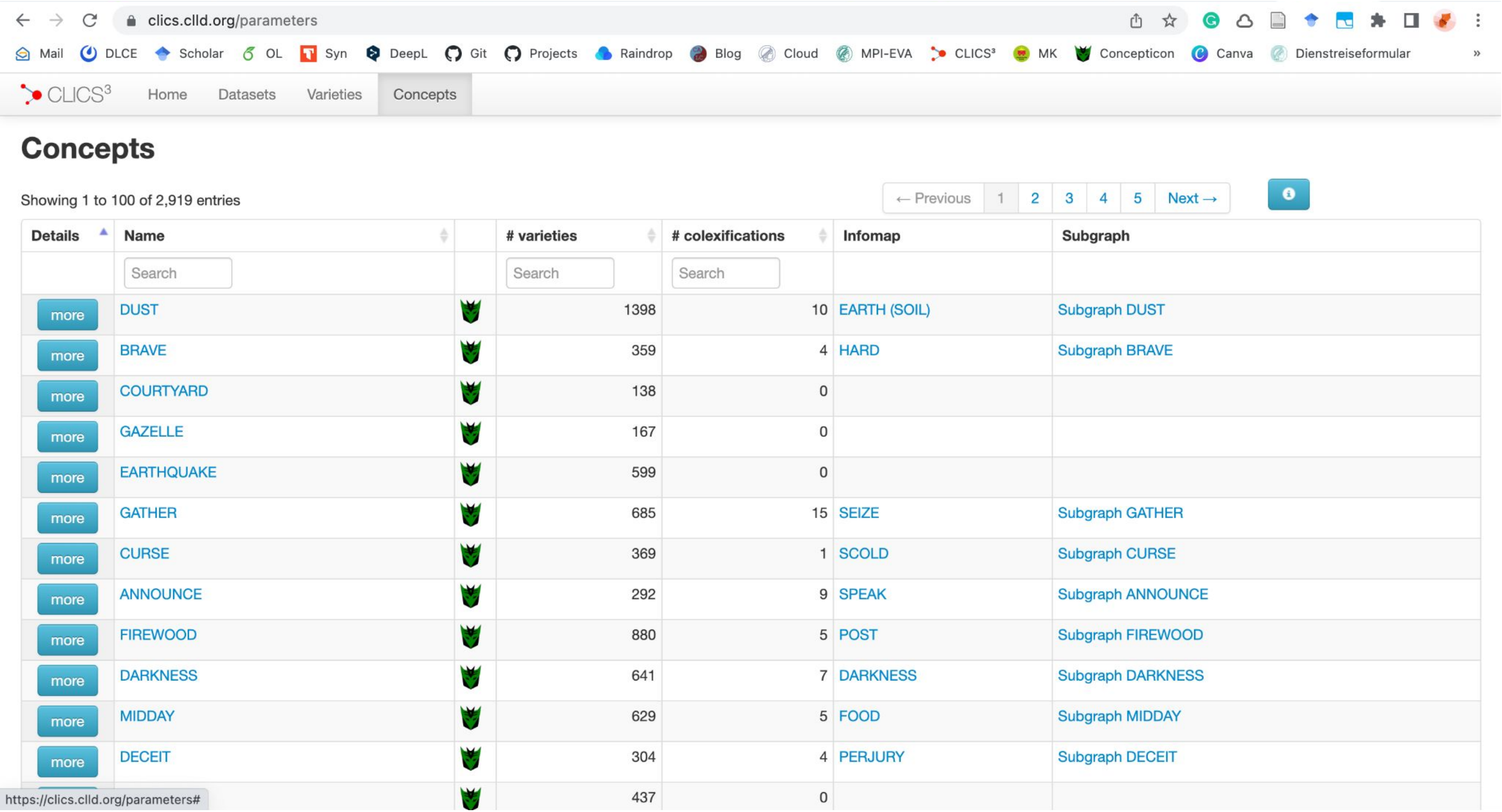
CLICS³ is an online database of colexifications (polysemies or homophonies) in currently 3156 language varieties of the world.

The original Database of Cross-Linguistic Colexifications (CLICS), has established a computer-assisted framework for the interactive representation of cross-linguistic colexification patterns. It has proven to be a useful tool for various kinds of investigation into cross-linguistic semantic associations, ranging from studies on semantic change, patterns of conceptualization, and linguistic paleontology. But CLICS has also been criticized for obvious shortcomings. Building on standardization efforts reflected in the CLDF initiative and novel approaches for fast, efficient, and reliable data aggregation, CLICS² expanded the original CLICS database. CLICS³ - the third installment of CLICS - exploits the framework pioneered in CLICS² to more than double the amount of data aggregated in the database.

Cite

Rzymski, Christoph and Tresoldi, Tiago et al. 2019. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. DOI: 10.1038/s41597-019-0341-x

Open: CLICS



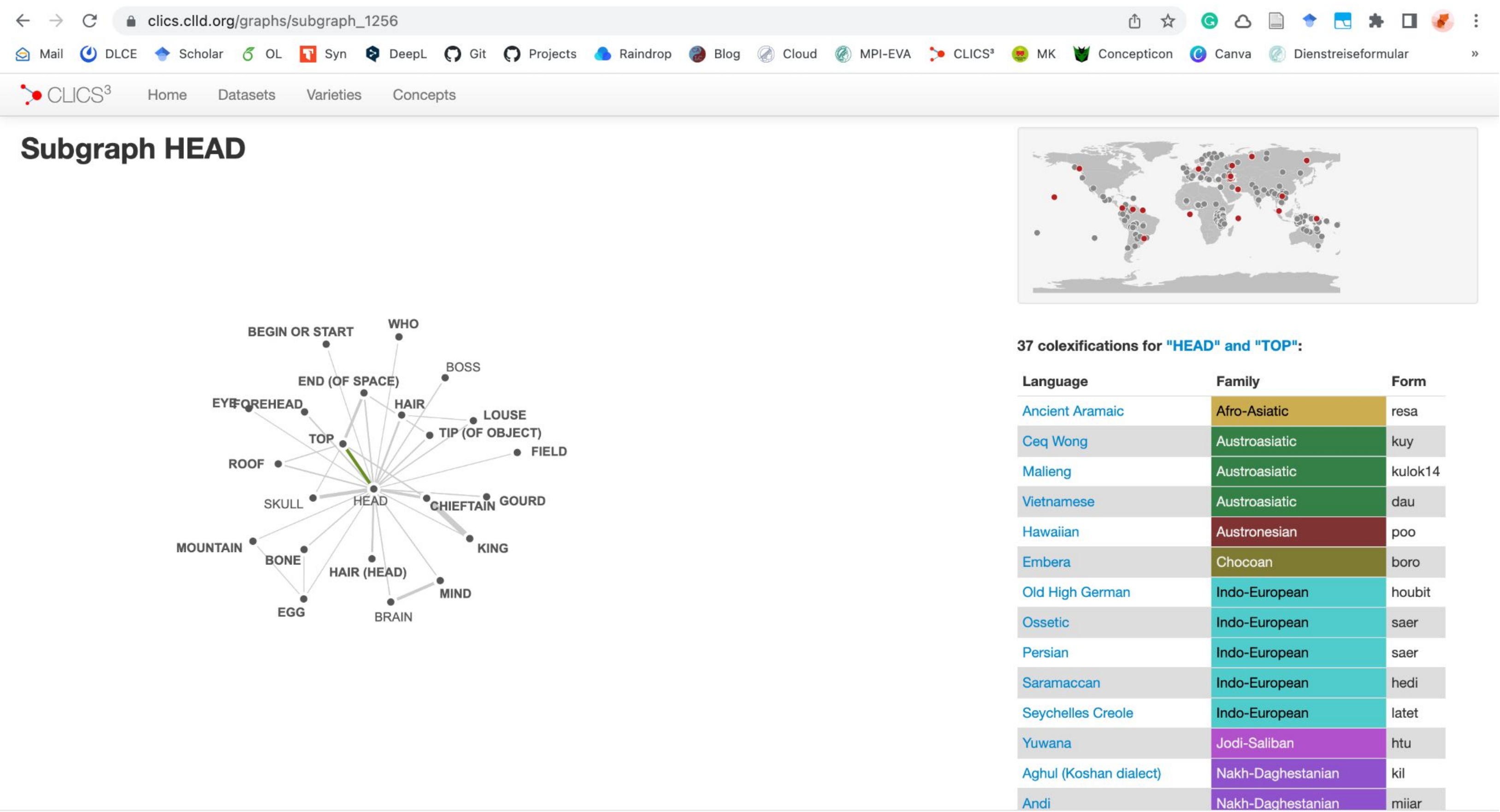
Concepts

Showing 1 to 100 of 2,919 entries

← Previous 1 2 3 4 5 Next →

Details	Name		# varieties	# colexifications	Infomap	Subgraph
	Search		Search	Search		
more	DUST		1398	10	EARTH (SOIL)	Subgraph DUST
more	BRAVE		359	4	HARD	Subgraph BRAVE
more	COURTYARD		138	0		
more	GAZELLE		167	0		
more	EARTHQUAKE		599	0		
more	GATHER		685	15	SEIZE	Subgraph GATHER
more	CURSE		369	1	SCOLD	Subgraph CURSE
more	ANNOUNCE		292	9	SPEAK	Subgraph ANNOUNCE
more	FIREWOOD		880	5	POST	Subgraph FIREWOOD
more	DARKNESS		641	7	DARKNESS	Subgraph DARKNESS
more	MIDDAY		629	5	FOOD	Subgraph MIDDAY
more	DECEIT		304	4	PERJURY	Subgraph DECEIT
			437	0		

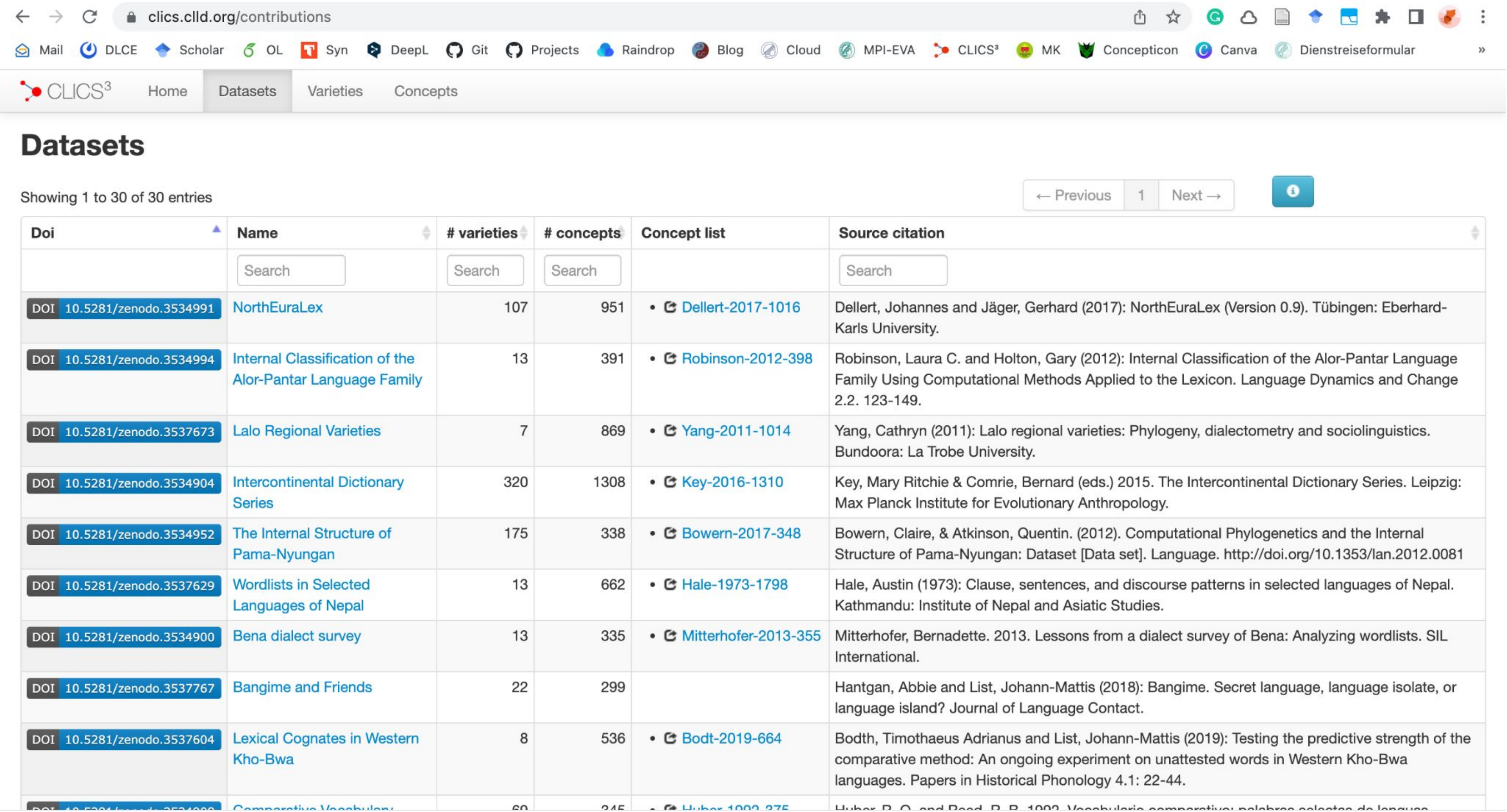
https://clics.clld.org/parameters#



Subgraph HEAD

37 colexifications for "HEAD" and "TOP":

Language	Family	Form
Ancient Aramaic	Afro-Asiatic	resa
Ceq Wong	Austroasiatic	kuy
Malieng	Austroasiatic	kulok14
Vietnamese	Austroasiatic	dau
Hawaiian	Austronesian	poo
Embera	Chocoan	boro
Old High German	Indo-European	houbit
Ossetic	Indo-European	saer
Persian	Indo-European	saer
Saramaccan	Indo-European	hedi
Seychelles Creole	Indo-European	latet
Yuwana	Jodi-Saliban	htu
Aghul (Koshan dialect)	Nakh-Daghestanian	kil
Andi	Nakh-Daghestanian	miiar



Datasets

Showing 1 to 30 of 30 entries

Doi	Name	# varieties	# concepts	Concept list	Source citation
	Search	Search	Search		Search
DOI 10.5281/zenodo.3534991	NorthEuraLex	107	951	• Dellert-2017-1016	Dellert, Johannes and Jäger, Gerhard (2017): NorthEuraLex (Version 0.9). Tübingen: Eberhard-Karls University.
DOI 10.5281/zenodo.3534994	Internal Classification of the Alor-Pantar Language Family	13	391	• Robinson-2012-398	Robinson, Laura C. and Holton, Gary (2012): Internal Classification of the Alor-Pantar Language Family Using Computational Methods Applied to the Lexicon. Language Dynamics and Change 2.2. 123-149.
DOI 10.5281/zenodo.3537673	Lalo Regional Varieties	7	869	• Yang-2011-1014	Yang, Cathryn (2011): Lalo regional varieties: Phylogeny, dialectometry and sociolinguistics. Bundoora: La Trobe University.
DOI 10.5281/zenodo.3534904	Intercontinental Dictionary Series	320	1308	• Key-2016-1310	Key, Mary Ritchie & Comrie, Bernard (eds.) 2015. The Intercontinental Dictionary Series. Leipzig: Max Planck Institute for Evolutionary Anthropology.
DOI 10.5281/zenodo.3534952	The Internal Structure of Pama-Nyungan	175	338	• Bowern-2017-348	Bowern, Claire, & Atkinson, Quentin. (2012). Computational Phylogenetics and the Internal Structure of Pama-Nyungan: Dataset [Data set]. Language. http://doi.org/10.1353/lan.2012.0081
DOI 10.5281/zenodo.3537629	Wordlists in Selected Languages of Nepal	13	662	• Hale-1973-1798	Hale, Austin (1973): Clause, sentences, and discourse patterns in selected languages of Nepal. Kathmandu: Institute of Nepal and Asiatic Studies.
DOI 10.5281/zenodo.3534900	Bena dialect survey	13	335	• Mitterhofer-2013-355	Mitterhofer, Bernadette. 2013. Lessons from a dialect survey of Bena: Analyzing wordlists. SIL International.
DOI 10.5281/zenodo.3537767	Bangime and Friends	22	299		Hantgan, Abbie and List, Johann-Mattis (2018): Bangime. Secret language, language isolate, or language island? Journal of Language Contact.
DOI 10.5281/zenodo.3537604	Lexical Cognates in Western Kho-Bwa	8	536	• Bodt-2019-664	Bodth, Timothaeus Adrianus and List, Johann-Mattis (2019): Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. Papers in Historical Phonology 4.1: 22-44.
DOI [illegible]	Comparative Vocabulary	69	345	• Huber-1992-375	Huber, R. Q. and Reed, R. B. 1992. Vocabulario comparativo: palabras selectas de lenguas

Unpublished work! Please refrain from taking pictures.

Partial Colexifications

Background

- **CLICS measures colexifications only if the same complete word form is used to express two different concepts.**
- **In many cases, however, we can note that words share certain morphemes without being completely identical.**
- **These *partial colexifications* can point to interesting patterns of *lexical motivation* in the sense of Koch (2001), reflecting the semantic and pragmatic processes underlying the formation of new words.**

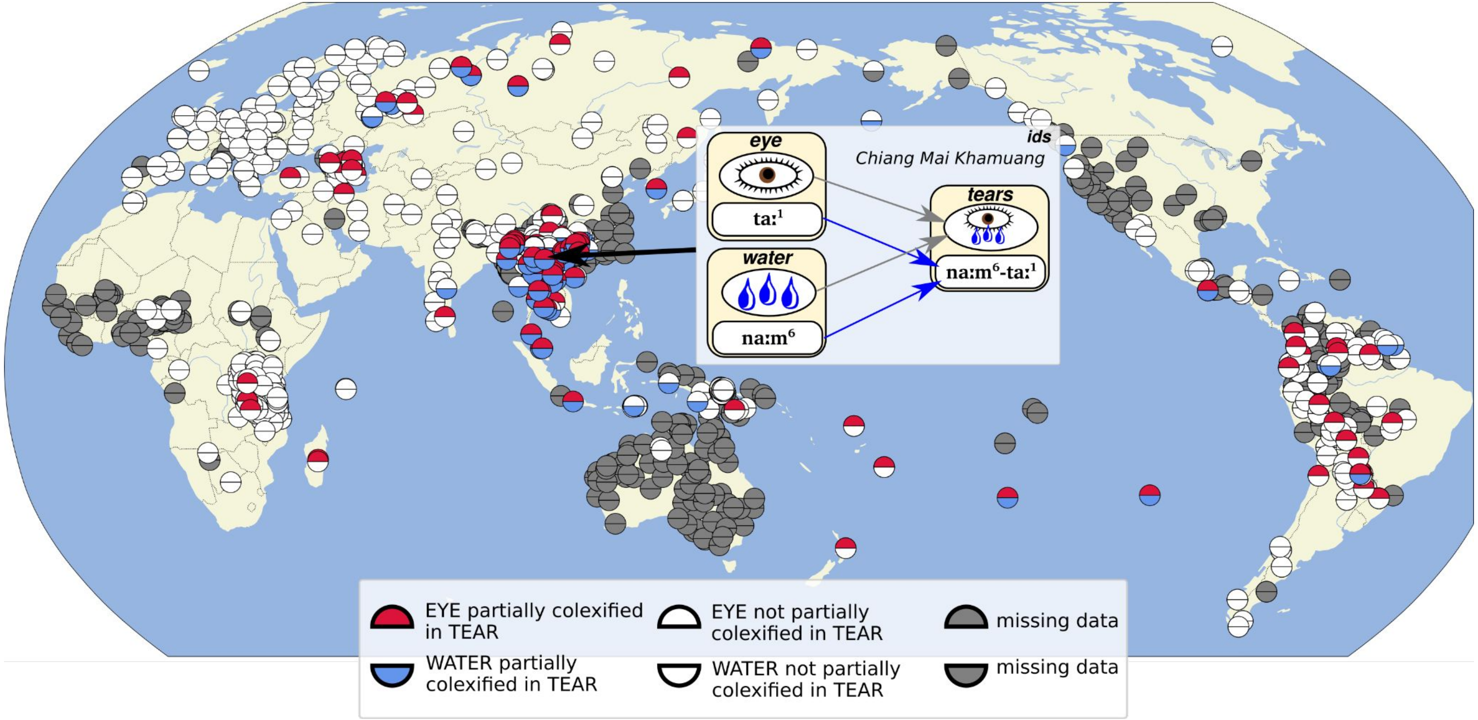
- **Article: List (in preparation)**

Partial Colexifications

- **Partial colexifications can be defined in the same way as lexical features in Lexibank.**
- **Affix colexification:**
  - If a word X in a language A recurs in the word Y in language A.
  - Modeled in the form of a directed, weighted graph (see Urban 2011, List in preparation).
  - A link is drawn from the concept expressed by the “affix” (in strict formal terms) to the concept expressing the “full” word.
- **Substring colexification:**
  - If a word X and a word Y in language A share a common substring
  - Modeled in the form of a weighted undirected graph, similar to the CLICS networks.
- **Together, affix and substring colexifications (which both represent instances of partial colexifications) can be used to extend CLICS networks by offering different perspectives on lexical data.**

Affix Colexifications

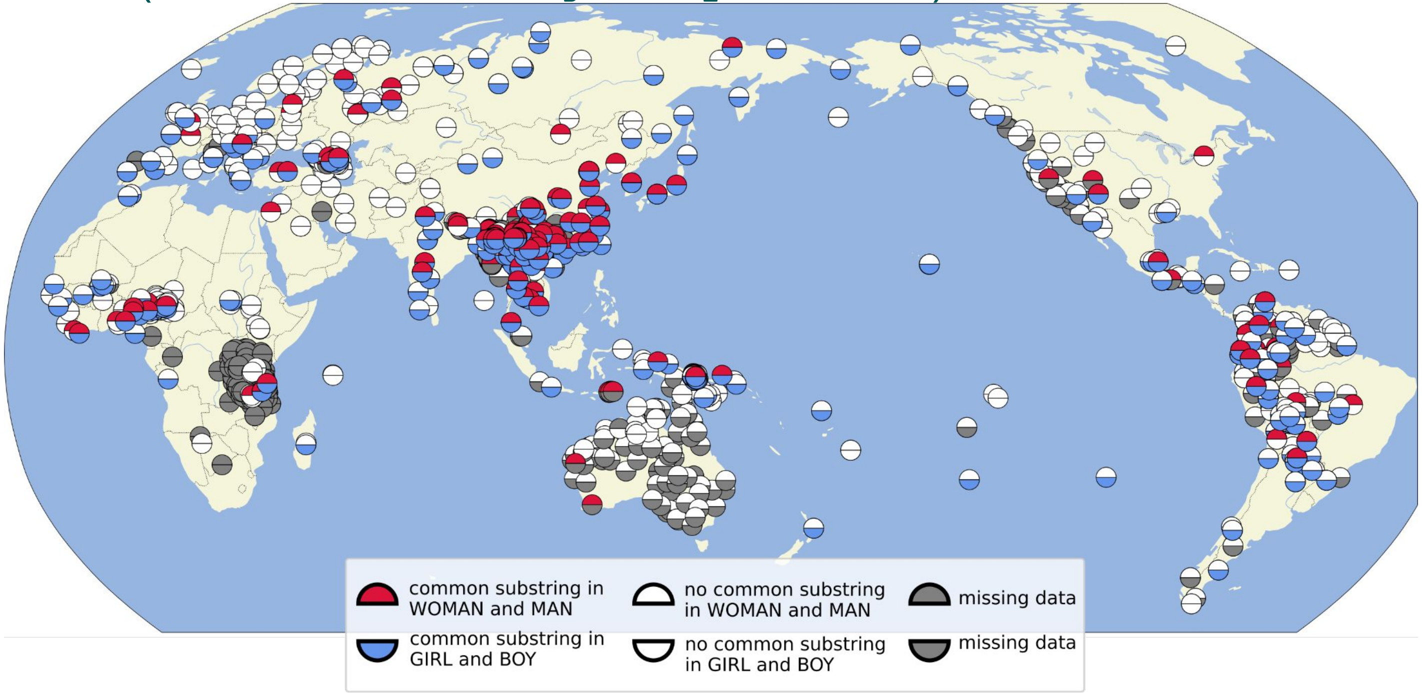
EYE-TEAR (`EyeInTear`) and WATER-TEAR (`WaterInTear`) → *eye water*


eye
ta:^{1}
water
na:m^{6}
ids
Chiang Mai Khamuang
tears
na:m^{6}-ta:^{1}
EYE partially colexified in TEAR
EYE not partially colexified in TEAR
missing data
WATER partially colexified in TEAR
WATER not partially colexified in TEAR
missing data

Substring Colexifications

WOMAN-MAN (`CommonSubstringInManAndWoman`) and

BOY-GIRL (`CommonSubstringInBoyAndGirl`)


common substring in WOMAN and MAN
no common substring in WOMAN and MAN
missing data
common substring in GIRL and BOY
no common substring in GIRL and BOY
missing data

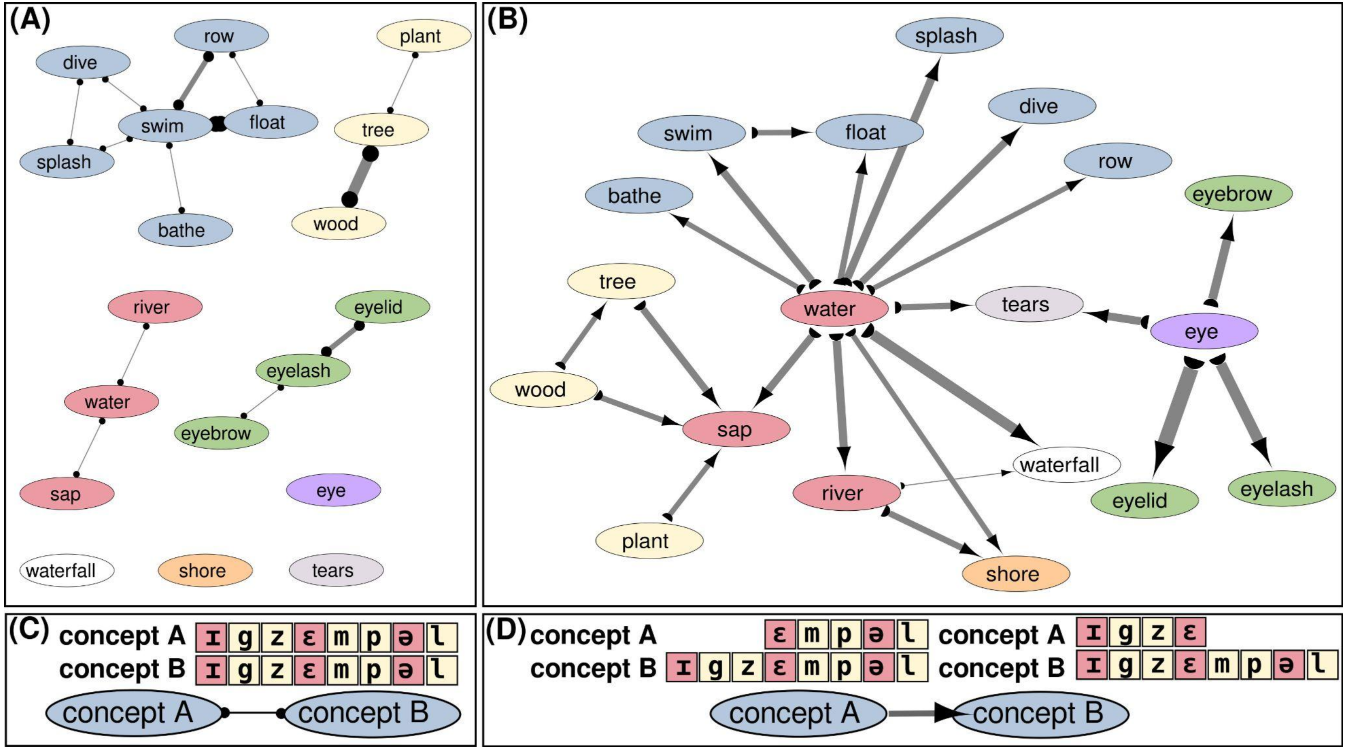
Directed Partial Colexification Networks

List (in preparation)

- **Initial experiments and pilot studies show that the visualization of affix colexifications in the form of directed weighted networks is promising.**
- **Our current experiments draw on smaller data sets and use an extended workflow by which partial colexifications in the form of affix colexifications can be inferred from standardized Lexibank data and later visualized in the form of a directed weighted network.**
- **These networks will be investigated in more detail in the upcoming research project *Productive Signs* (ERC Consolidator Grant awarded to J.-M. List), to start in January 2023.**

Directed Partial Colexification Networks

List (in preparation)



Directed Partial Colexification Networks

Summary

- **Partial colexifications bear a lot of potential that has so far not been readily explored.**
- **Generating networks of partial colexification is more difficult due to the increase of erroneous or coincidental matches in the data (increase of noise).**
- **Targeted experiments with Lexibank as well as initial pilot studies with slightly adjusted workflows yield promising results.**

CASE STUDY: EMOTION COLEXIFICATIONS

Scientific Debates about Emotions

Background

- **Early theories on emotion go back to Darwin and assumed that there is a discrete number of universal emotions which are – similar to primary colors – shared by all humans.**
- **Recent investigations argued, however, that emotions vary systematically across cultures and languages and that seemingly universal emotions like ANGER and GRIEF cannot be derived from concrete structures in the human brain (Lindquist et al. 2012).**
- **Investigating emotion semantics across different language families could give us some interesting insights into the question of whether and to what degree emotion concepts vary cross-linguistically.**

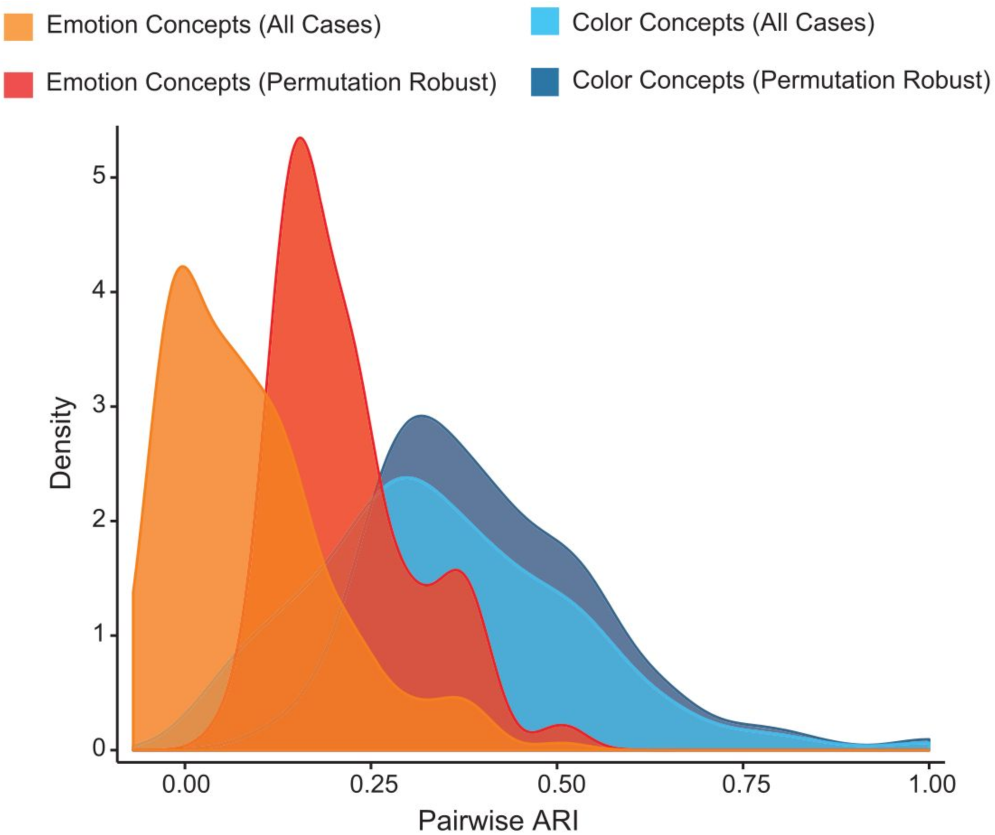
- **Article: Jackson et al. (2019)**

Studying Emotions with CLICS Data

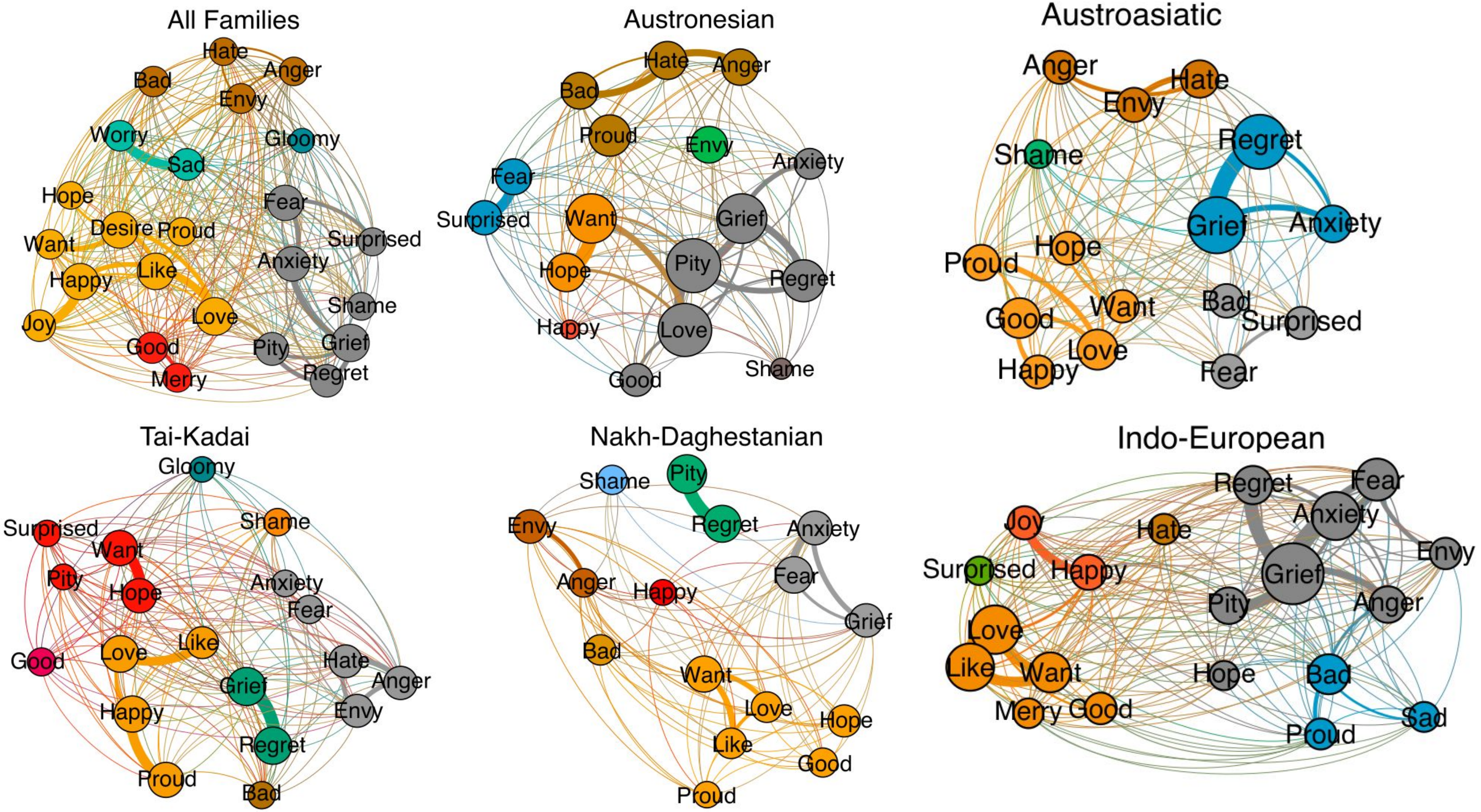
Approach

- Using CLICS³ (Rzymski et al. 2020) offers a large amount of data with more than 2400 language varieties.
- 24 emotion concepts were selected, showing a good coverage in the aggregated database.
- Community networks of emotion concepts were inferred for the 20 largest language families.
- The networks were systematically compared with respect to their structure (using Adjusted Rand Index) to examine the extent to which they would differ from one family to another.

Variation in Network Structure across Language Families

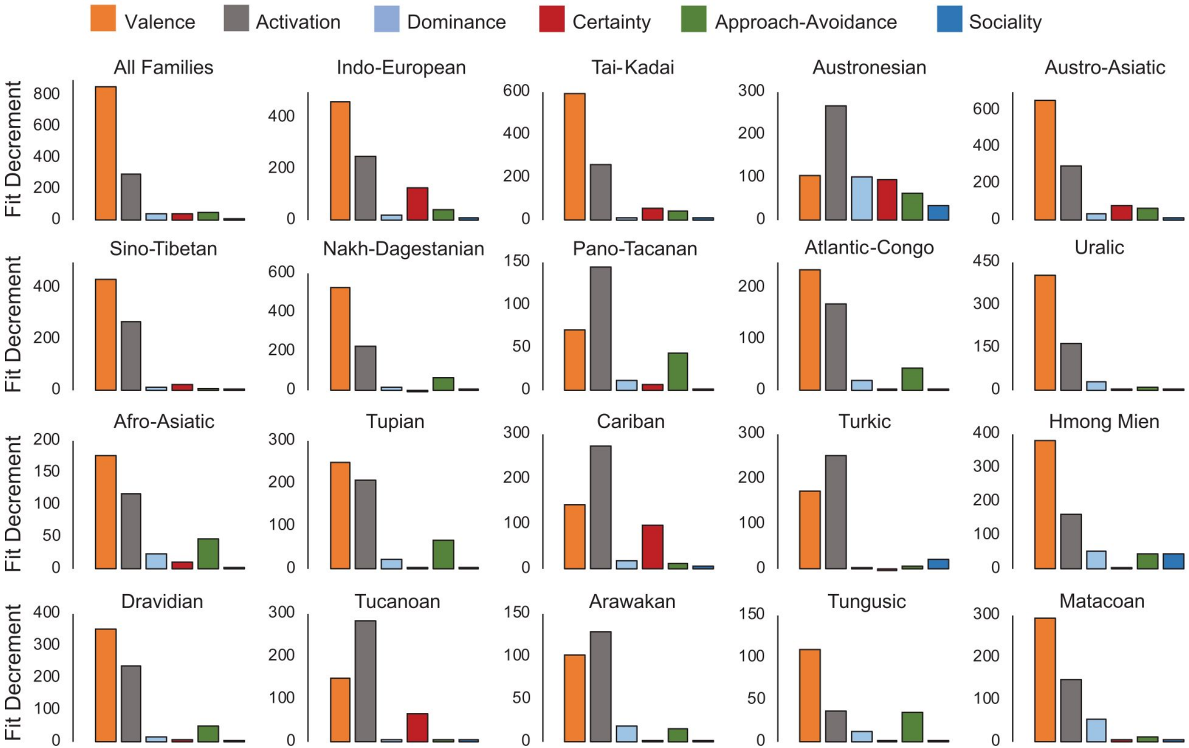


Emotion Colexifications across Language Families



Emotion Colexifications

Cross-Linguistic Tendencies in Ratings



Emotion Colexifications across Language Families

Conclusions

- The results showed strong variation in emotion semantics across language families.
- Not all seems to be “constructed” individually, however. There seems to be some universal core reflected in the importance of the categories of valence and activation which predict the emotion networks in individual language families.

8

CASE STUDY: BODY COLEXIFICATIONS

Unpublished work! Please refrain from taking pictures.

Scientific Debates about Body Parts

Background

- **All humans have a body with almost the same parts.**
- **Why do languages have a single word while other languages have seperate words for the same body parts?**
- **Claim 1:**
  - There are universal categorization principles in how languages refer to body parts (Brown 1976; Andersen 1978).
- **Claim 2:**
  - Not all languages follow these categorization principles (Majid et al. 2006).
- **Article: Tjuka, Blasi, and List (in preparation)**

Studying Body Colexifications with Lexibank Data

Approach

- **Using the Lexibank framework (List et al. 2022), specifically the ClicsCore data sets, to analyse data from 15 language families.**
- **Applying network analysis to investigate the connections between body concepts.**
- **Accounting for shared ancestry of languages by implementing cognate detection.**
- **Coding the data for adjacency, shape, and function.**

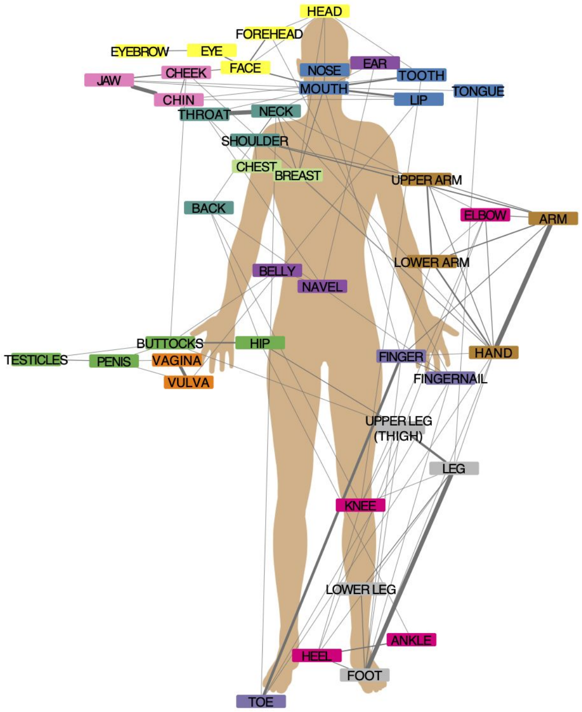
Body Colexifications across Language Families

Preliminary Results

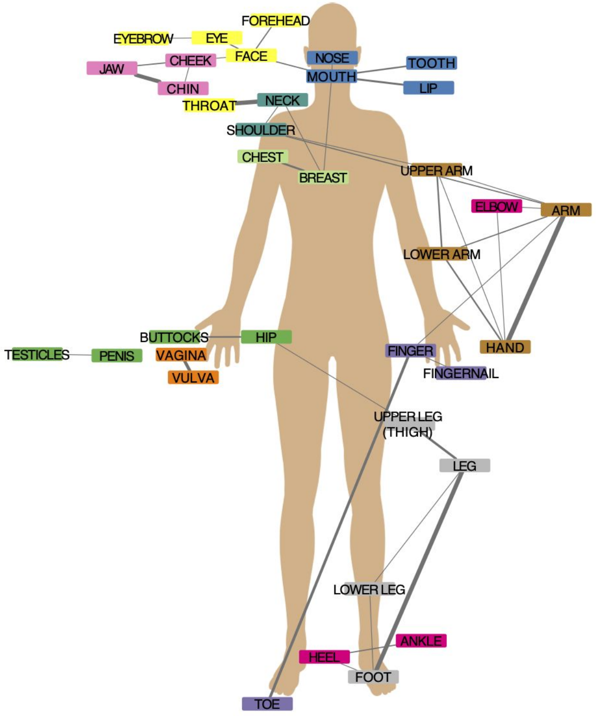
- **93 body colexifications in a set of 49 body concepts across 15 language families were found.**
- **60 out of 93 were coded for *adjacency*, 15 for *shape*, 18 for *function*.**
- **8 colexifications were coded for two features and 5 for all three (e.g., LEG-LOWER LEG).**
- **From the 35 non-adjacent colexifications, FINGER-TOE stands out because it occurs across 54 languages (7 families). All other non-adjacent colexifications occur in 1-5 languages (1-2 families).**

Body Colexifications across Language Families

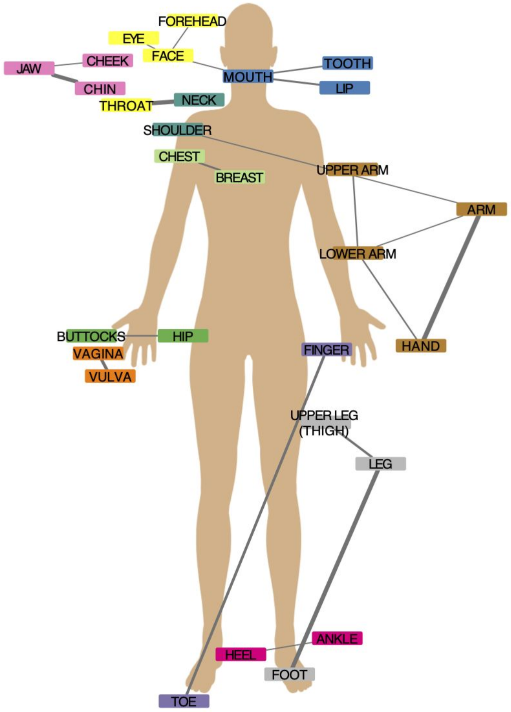
Global networks with different thresholds



no threshold



> 1 lang. fam.



> 2 lang. fam.

Body Colexifications across Language Families

Preliminary Conclusions

- **A strong tendency exists for a few universal colexifications as opposed to many language family-specific ones.**
- **Adjacency is the most frequent which indicates that languages don’t acknowledge a discontinuity (i.e., a joint) as a separating factor.**
- **Shape is less frequent, but leads to interesting connections such as LIP-NAVEL that seem to be specific to a particular language family.**
- **Function seems to occur in only a few colexifications such as ANKLE-WRIST.**

Summary

- Lexibank offers standardized lexical data for cross-linguistic large-scale approaches.
- Automatic computation of phonological and lexical features with Lexibank.
- Partial colexifications point to patterns of lexical motivation.
- Emotion and body colexifications provide insights into linguistic diversity and universality.

MANY THANKS
FOR YOUR ATTENTION

If you have any questions, please contact:

Annika Tjuka

annika_tjuka@eva.mpg.de

@AnnikaTjuka

Johann-Mattis List

mattis_list@eva.mpg.de

@LinguList

Blog Posts

List, Johann-Mattis. 2018. Towards a history of concept list compilation in historical linguistics. Blog. History and Philosophy of the Language Sciences. https://hiphilangsci.net/2018/10/31/concept-list-compilation/. (29 December, 2020).

Tjuka, Annika. 2020. Adding concept lists to Concepticon: A guide for beginners. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2225. (28 December, 2020).

Tjuka, Annika. 2021a. How to review concept lists in collaboration (How to do X in linguistics 6). Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2680. (25 March, 2021).

Tjuka, Annika. 2021b. Adding data sets to NoRaRe: A guide for beginners. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2890. (31 December, 2021).

Tjuka, Annika. 2021c. Comparing NoRaRe data sets: Calculation of correlations and creation of plots in R. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/3109. (31 December, 2021).

Tresoldi, Tiago. 2019a. Using pyconcepticon to map concept lists. Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/1820. (24 March, 2021).

Tresoldi, Tiago. 2019b. Using pyconcepticon to map concept lists (II). Blog. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/1844. (24 March, 2021).

References

Forkel, Robert. 2022. CLDFViz. A Python library providing tools to visualize data from CLDF datasets (Version 0.8.0). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5162667.

Gast, Volker & Maria Koptjevskaja-Tamm. 2019. The areal factor in lexical typology. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), Aspects of Linguistic Variation, 43–82. Berlin/New York: Walter de Gruyter. https://doi.org/10.1515/9783110607963-003.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. Language 86(3). 663–687. https://doi.org/10.1353/lan.2010.0021.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray & Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. Science 366. 1517–1522. https://doi.org/10.1126/science.aaw8160.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1379/.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9(1). 316. https://doi.org/10.1038/s41597-022-01432-0.

References

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel. 2018. CLICS^2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. Linguistic Typology 22(2). 277–306. https://doi.org/10.1515/lingty-2018-0010.

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle & Matthias Urban. 2014. CLICS: Database of cross-linguistic colexifications. Marburg: Forschungszentrum Deutscher Sprachatlas. http://clics.lingpy.org.

List, Johann-Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt & Robert Forkel. 2021. Concepticon. A resource for the linking of concept lists (Version 2.5.0). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.596412.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. Scientific Data 7(1). 1–12. https://doi.org/10.1038/s41597-019-0341-x.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2021. NoRaRe. A database of cross-linguistic norms, ratings, and relations for words and concepts (Version 0.2). Jena, Germany: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4647878.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking norms, ratings, and relations of words and concepts across multiple language varieties. Behavior Research Methods 54. 864–884. https://doi.org/10.3758/s13428-021-01650-1.