

The clld toolkit

Robert Forkel and Sebastian Bank

Max Planck Institute for Evolutionary Anthropology, Leipzig

October 7, 2014

Outline

The CLLD project

The `cld` toolkit

- The data model

- ROA, REST and ...

- ... Linked Data

- Versioning, updating, preservation

Towards a domain specific API

- Decoupling database and visualization

- Semantic interoperability

The CLLD project: Overview

Funded by the Max Planck Society for 4 years.

Creates infrastructure for publishing cross-linguistic datasets, including

- ▶ organization: a publication platform <http://clld.org> supporting two publication models:
 - ▶ Standalone databases following an "edited series" model, like WALS, WOLD, ...
 - ▶ Two journals for cross-linguistic datasets
- ▶ infrastructure: Glottolog, a language catalog and comprehensive bibliography
- ▶ technology: the `clld` toolkit powering our applications

The CLLD project: Datasets

Typological:

- ▶ **WALS** - the World Atlas of Language Structures - a database of structural properties of more than 2600 languages
- ▶ **APiCS** - the Atlas of Pidgin and Creole Language Structures
- ▶ **SAILS** - the South American Indigenous Language Structures
- ▶ **PHOIBLE** - a repository of cross-linguistic phonological inventory data

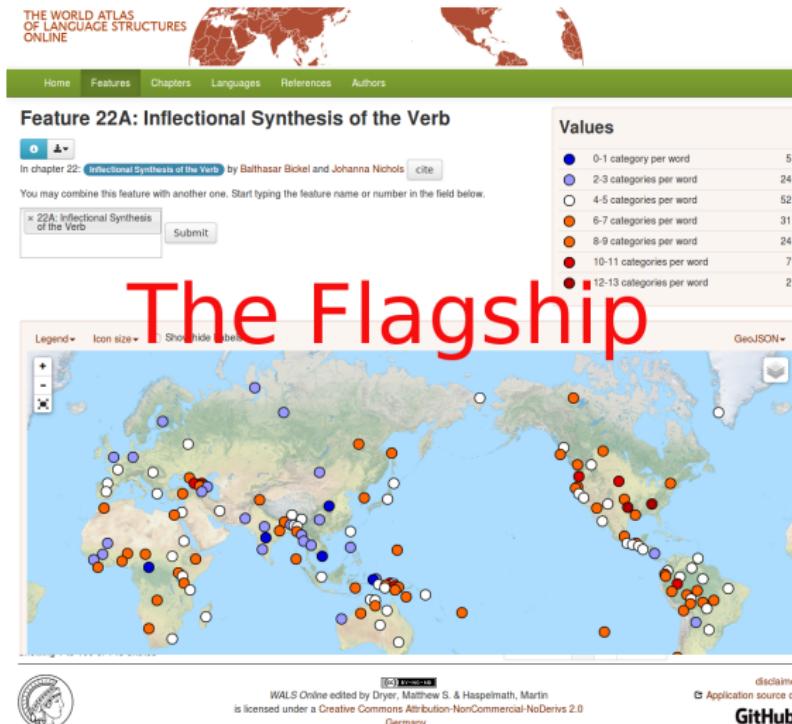
Lexical:

- ▶ **WOLD** - the World Loanword Database contains vocabularies of 41 languages from around the world annotated for loanword status
- ▶ **Tsammalex** - a multilingual lexical database on plants and animals
- ▶ **IDS** - the Intercontinental Dictionary Series (to be published in CLLD in 2014)
- ▶ **ASJP** - the Automated Similarity Judgement Project (to be published in 2014)

Encyclopedic:

- ▶ **Glottolog** - a language catalog and comprehensive bibliography

The CLLD project: WALS



The CLLD project: WOLD

The screenshot shows the WOLD interface. At the top, there's a navigation bar with tabs: Home, Vocabularies (which is selected), Meanings, Languages, and Authors. Below the navigation bar is a decorative graphic featuring several overlapping circles in shades of green, blue, and yellow, each containing a small icon related to language or linguistics.

Vocabulary English

by Anthony Grant cite

The vocabulary contains 1515 meaning-word pairs ("entries") corresponding to core LWT meanings from the recipient language English. The corresponding text chapter was published in the book Loanwords in the World's Languages. The language page English contains a list of all loanwords arranged by donor language.

Meaning-word pairs Description

Showing 1 to 7 of 7 entries (filtered from 1,516 total entries)

Word form	LWT code	Meaning	Core list	Borrowed status	Source words
land	1.21	the land	True	5. no evidence for borrowing	
soil	1.212	the soil	True	1. clearly borrowed	seull 'ground' French solum 'ground, floor, sole' Latin
dust	1.213	the dust	True	5. no evidence for borrowing	
mud	1.214	the mud	True	3. perhaps borrowed	modde 'bog' Dutch
sand	1.215	the sand	True	5. no evidence for borrowing	
give	11.21	to give	True	1. clearly borrowed	giva 'to give' Old Norse
plaintiff	21.21	the plaintiff	True	1. clearly borrowed	plaintif 'plaintiff' French (Anglo-Norman) planctus 'lamentation' Latin

Showing 1 to 7 of 7 entries (filtered from 1,516 total entries)

– Previous 1 Next –

WOLD edited by Haspelmath, Marin & Tadmor, Uri
is licensed under a Creative Commons Attribution 3.0 Germany License.

disclaimer
Application source on GitHub

The CLLD project: APiCS

THE ATLAS OF PIDGIN AND CREOLE LANGUAGE STRUCTURES ONLINE

Home Languages Features WALS-APiCS Examples Sources Authors

1 Order of subject, object, and verb

Description

This feature (based on [WALS feature 81](#), by Matthew S. Dryer) concerns the ordering of subject, object and verb in non-contrastive, non-focussed transitive clauses without special topicalization, more specifically declarative clauses with both the subject and object realized as full nouns (not as pronouns).

We use subject and object in a semantic sense, to refer to the agent-like and patient-like constituents in a monotransitive clause, as in e.g. French [*Les souris mangent le fromage*] 'The mice eat the cheese'. As can be seen from this example, French has SVO order (Subject-Verb-Object), because the subject *les souris* 'the mice' precedes the verb and the object *le fromage* 'the cheese' follows it. Since we only consider non-contrastive, non-focussed, non-topicalized clauses, cases like English *It is the cheese that the mice eat* (=OSV) are disregarded here.

There are six logically possible orders of subject, object and verb, as shown in the list of feature values. Languages can have several word orders (e.g. German is SVO and VSO in main clauses and SOV in subordinate clauses), so several values can be true for this feature.

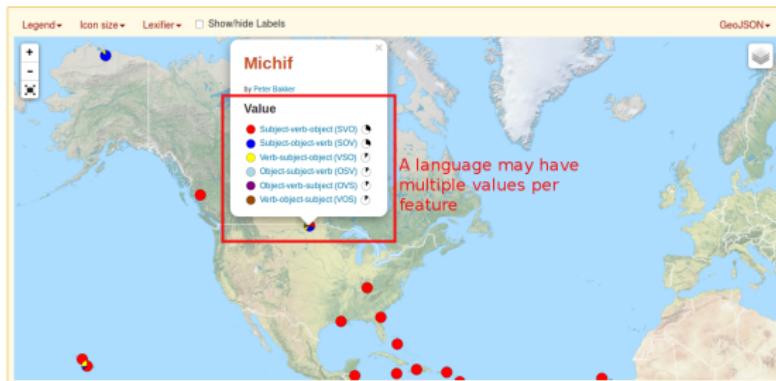
Author

Magnus Huber and the APiCS Consortium [cite](#)

Values

	excl	shrd	all
Subject-verb-object (SVO)	61	10	71
Subject-object-verb (SOV)	1	11	12
Verb-subject-object (VSO)	0	7	7
Verb-object-subject (VOS)	0	3	3
Object-subject-verb (OSV)	0	3	3
Object-verb-subject (OVS)	0	2	2

Representation: 76



The CLLD project: Glottolog

Home Languoids Langdoc News Languages Families Search Languoids information
Glottocode: [east2387](#) ISO 639-3: [limb](#)

East Limba 

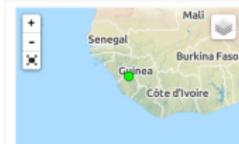
Classification 

- ▼ Atlantic-Congo (1461)
 - ▶ Limba (2)
 - ▶ East Limba 
 - ▶ Northern Limba
 - ▶ Southern Limba
 - ▶ West-Central Limba
 - ▶ Mansanka-Fore-Mbotozi (5)
 - ▶ Mel (13)
 - ▶ Nalu
 - ▶ North-Central Atlantic (46)
 - ▶ Volta-Congo (1394)

[Open East Limba](#) [Expand all](#) [Collapse all](#)

Glottolog:
Language catalog and comprehensive bibliography

Map



[show big map](#)

Countries

Links

Alternative names

References

Showing 1 to 5 of 5 entries

Details*	Name	Title	ca	Year	Pages	Doctype	ca	Provider	da
more	Clarke, Mary Lane 1929	A Limba-English (English-Limba) dictionary / Tempen ta ka talun ta ka hulimba ha in hunkilisi ha	▲	1929	150	dictionary	▲	ebal, webal	▼
more	Clarke, Mary Lane 1922	A Limba-English dictionary / Tempen ta ka talun ta ka hulimba in hunkilisi ha		1922	150	dictionary	hh, ebal, webal	▼	▼
more	Thomas, Northcote Wheridge 1916	Specimens of languages from Sierra Leone		1916	62	overview, wordlist	hh, ebal, webal	▼	▼
more	Clarke, Mary Lane 2005	A Limba-English dictionary; or Tempen ta ka talun ta ka hulimba ha in hunkilisi ha		2005	150	dictionary	▲	mpieva	▼
more	Clarke, Mary Lane 1971	A Limba-English Dictionary or Tempen Ta Ka Talun Ta Ka Hulimba Ha In Huinkilisi Ha		1971		dictionary	▲	aspip2010	▼

Showing 1 to 5 of 5 entries

[- Previous](#) [1](#) [Next -](#)

 [disclaimer](#)  [Application source on GitHub](#)

Glottolog 2.3 edited by Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Nordhoff, Sebastian
is licensed under a [Creative Commons Attribution-ShareAlike 3.0 Unported License](#).

The CLLD project: AfBo

AfBo: A world-wide survey of affix borrowing

Home About Map Languages Affix functions References About Legal Download Contact

Latin affixes in Basque

Subjects are recipient/donor pairs rather than single languages

Summary

Affix function	number of borrowed affixes
Search	Search
adjective	2
augmentative	2
diminutive	2
gender (human)	1
nominal derivation (miscellaneous)	4
nominalizer: abstract	17
nominalizer: agent	4
nominalizer: social group	2
privative	1

Description

Information and examples are from Segura Munguia and Etxebarria Ayesta (1996) and Huilde and Urbina (2003). See also Haase (1992: 48–51), who focuses on French and Gascon influence on the Basque variety of lower Navarra, and Mujika (1982).

2 diminutive suffixes

-ita, -ito 'diminutive', e.g. neskaita 'little girl' (from neska 'girl'), lehatita 'little window' (from leho 'window'), andrakita 'doff' (from andra 'woman'), astokito 'little donkey' (from asto 'donkey'), gizonito 'little man' (from gizon 'man') (Huilde 2003a: 331) (see also Haase 1992: 49; Segura Munguia and Etxebarria Ayesta 1996: 84, 89)

-itx, -itz, -itza, -itzta 'diminutive', e.g. emekitzia 'very softly' (from emeki 'softly'), batitz 'a little one' (from bat 'one') (Huilde 2003a: 331; Segura Munguia and Etxebarria Ayesta 1996: 89)

Recipient language: Basque
Donor language: Latin

Reliability of borrowed status/affiliation: high
Borrowed affixes: 35
Interrelated affixes: 35

References

Haase, Martin 1992 Sprachkontakt und Sprachwandel im Baskenland : die Einflüsse des Gasconischen und Französischen auf das Baskische
Huilde, José Ignacio 2003 Segmental phonology
Huilde, José Ignacio 2003 Derivation
Huilde, José Ignacio and Urbina, Jon Ortiz de 2003 A grammar of Basque
Mujika, Luis Mat 1982 La fina eta erromantikaren eragina euskanan. Euskal lexicaren azterketa bideetan
Segura Munguia, Santiago and Etxebarria Ayesta, Juan Manuel 1996 Del latín al euskara = Latinetik euskarara
Trask, R.L. 2003 The Noun Phrase: nouns, determiners and modifiers; pronouns and names



AfBo: A world-wide survey of affix borrowing by Seifert, Frank
is licensed under a Creative Commons Attribution 3.0 Unported License.

disclaimer
Application source on
[GitHub](#)

The CLLD project: eWAVE

FRIAS
TEACHING INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

**THE ELECTRONIC WORLD ATLAS
OF VARIETIES OF ENGLISH**

Home Varieties Features Informants Examples Sources Map List

2 He/him used for inanimate referents

● A - feature is pervasive or obligatory	0
● B - feature is neither pervasive nor extremely rare	12
● C - feature exists, but is extremely rare	8
● D - attested absence of feature	42
● X - feature is not applicable (given the structural make-up of the variety/P/C)	11
● ? - no information on feature is available	3

Feature area:
Pronouns, pronoun exchange, nominal gender
Typical example:
I bet thee canst climb he [= a tree]
Example source:
Southwest (Wagner 2008: 425)

Legend • Icon size • Type • Show/hide Labels

Non-ISO languages

GeoJSON •

disclaimer
eWAVE edited by Kortmann, Bernd & Lunkenheimer, Kerstin
is licensed under a Creative Commons Attribution 3.0 Unported License .

Application source on GitHub

The CLLD project: SAILS

SAILS Home Features Languages Sources Designers

Languages

Icon size ▾ Family ▾ Show/hide Labels GeoJSON ▾

SAILS:
Macro-area specific

Showing 1 to 1 of 1 entries (filtered from 167 total entries)

Name	Iso-639-3	Family	Features
Cholón	cht	Hibito-Cholón	148

Showing 1 to 1 of 1 entries (filtered from 167 total entries)

– Previous 1 Next –

SAILS Online edited by Hammarstrom, Harald
is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 2.0 Germany](#).

disclaimer
Application source on
[GitHub](#)

The CLLD project: PHOIBLE

[fɔɪ.bɪ] Home Inventories Languages Segments Sources

Inventory Standard English (SPA)

Source name: English

Segment list IPA chart

Consonants (Pulmonic)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glossal
Positive	p b	t d	k g	l ɬ	ʈ ɖ	c ɺ	ts ɻ	ʃ ʂ	χ χ	q ʡ	ʔ
Nasal	m n	ɳ	ɳ	ɳ		ɳ	ɳ	ɳ	ɳ	ɳ	
Trill	r										r
Tap or Flap	v̚										
Fricative	f v	θ ð	s z	ʃ ʂ	tʃ ʈʂ	x ʐ	χ ʐ	x ʐ	x ʐ	χ ʐ	h ʐ
Lateral											
Inclusive			t̪ ʈ̪								
Approximant	w					ɹ					
Lateral approximant											

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Consonants (Non-Pulmonic)

Clicks	Voiceless implosives
○ Bilabial	ɓ Bilabial
Dental	ɗ Dental/velvocal
(Post)alveolar	ʄ Palatal
+ Palatoalveolar	ɠ Velar
Alveolar lateral	ʗ Uvular

Vowels

Front Central Back

	Close	Close-mid	Open-mid	Open
Front	i: ɪ ɿ ɿ	e: ə ɻ ɻ	a: ɻ ɻ ɻ	ɑ: ɻ ɻ ɻ
Central	ʊ: ʊ ʊ	ɔ: ɔ ɔ	ɛ: ɛ ɛ	ə: ə ə
Back	u: ʊ ʊ	o: ɔ ɔ	ɛ: ɛ ɛ	ɑ: ɑ ɑ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

Contributor

Stanford Phonology Archive [cite](#)

Sources

Trnka, Bohumil 1968
O'Conor, J. D. 1973
Gimson, A. C. 1962
Halle, Morris 1973
Fudge, Erik 1975

Standard English

Map

Coordinates WGS84 53°N, 1°W
53.00, -1.00

Links

PHOIBLE:
more than maps ...

Other Segments

- ʍ LATIN SMALL LETTER TURNED W
ȝ LATIN SMALL LETTER O - COMBINING DOWN TACK BELOW - MODIFIER LETTER TRIANGULAR COLON
ȝ̥ LATIN SMALL LETTER K - MODIFIER LETTER SMALL H
dʒ LATIN SMALL LETTER D - COMBINING MINUS SIGN BELOW - LATIN SMALL LETTER EZH
p̥h LATIN SMALL LETTER P - MODIFIER LETTER SMALL H
t̥h LATIN SMALL LETTER T - MODIFIER LETTER SMALL H

The CLLD project: Tsammalex

Tsammalex Home Languages Species Ecoregions

Asian lion

Map Pictures Names

Biological classification: order: Carnivora
- family: Felidae
-- genus: Panthera
-- species: Panthera leo

Countries: Botswana (BW)
Mozambique (MZ)
Namibia (NA)
South Africa (ZA)
Zimbabwe (ZW)

Ecoregions: AT1309 Kalahari xeric savanna

Links: edit wikipedia

Lineage Icon size Show hide Labels GeoJSON

Supplemental files

Date: 2008-10-05 Place: Zoo Leipzig, Germany Author: Christfried Naumann Permission: public domain Comments: "Angola lion", P. angolensis = P. l. bleyenbergh?	Date: 2008-10-05 Place: Zoo Leipzig, Germany Author: Christfried Naumann Permission: public domain Comments: "Angola lion", P. angolensis = P. l. bleyenbergh?	Date: 2009-03-29 Place: Auob river, Kgalagadi Transfrontier Park, South Africa Author: Taa Dobé team (Boden/Güldemann /Naumann) Permission: public domain
--	--	--

Showing 1 to 41 of 41 entries

Language	Lineage	Word form	Generic term	IPA	Grammatical notes	Exact meaning	Categories	General notes	References
Afrikaans	Germanic	leeu	leeu			diere	soogdiere	Search	wikipedia
ǂAmkoe	K'a	qa_beé_qó							Berhold, Fakö and Linds Gerlich 2011
ǂAmkoe	K'a	njħali							Berhold, Fakö and Linds Gerlich 2011

The CLLD project: Where's my dataset?

Have a dataset in need of publication and presentation on the web?

- ▶ Submit to Harald's **Journal of Cross-Linguistic Databases** or
- ▶ submit to Martin's edited series of cross-linguistic databases **clld.org** or
- ▶ get a seasoned python programmer for a month to build your own app on top of the `clld` toolkit!

```
robert@astroman:/tmp/phoible$ cloc --exclude-dir=tests,data phoible/  
    38 text files.  
    36 unique files.  
    28 files ignored.
```

Language	files	blank	comment	code
Python	17	230	173	954
CSS	1	25	49	159
Javascript	1	1	0	0
SUM:	19	256	222	1113

The cld toolkit: Motivation

Survey databases are all alike.

Can we extract functionality needed to build WALS, WOLD, and APiCS into a reusable piece of software?

Design goals:

- ▶ There must be a core database model, which allows for as much shared functionality as possible.
- ▶ User interfaces of applications must be fully customizable.
- ▶ It must be easy to re-implement legacy applications using the framework.
- ▶ Optimize for maintainability, i.e. minimize lines-of-code for apps built with the framework.
- ▶ Find the right level of abstraction!

`cld`: A CMS for cross-linguistic data

The `cld` toolkit is an open source Python package hosted on GitHub providing

- ▶ an extensible core data model
- ▶ a web application framework
 - ▶ powering all CLLD databases
 - ▶ providing a basic API built on Linked Data principles
 - ▶ "reference implementation" of a dataset browser
 - ▶ `cld` apps are web applications built as small layer of code on top of the `cld` framework.
 - ▶ `cld` works with python 2.7 and 3.4 and has a test suite with 100% coverage.

Intermezzo: Disambiguation

- ▶ **CLLD**: The project.
- ▶ **clld.org**: The publisher/brand.
- ▶ **clld**: The software, aka toolkit, aka framework.
- ▶ **clld app**: A web application built using the `clld` framework.

In the remainder of this presentation we will talk about the latter two.

clld data model: Design

The design of the data model was guided by three principles:

- ▶ All the target datasets have to “fit in” without loss.
- ▶ The data model must be as abstract as necessary, as concrete as possible.
- ▶ The data model must be extensible.

clld data model: Entities

- ▶ **Dataset** holds metadata about a dataset like license and publisher information.
- ▶ **Language** may be a languoid (Glottolog) or doculect (ASJP).
- ▶ **Parameter** a feature that can be determined and coded for a language – e.g. a word meaning, or a typological feature.
- ▶ **ValueSet** set of values measured/observed/recorded for one language and one parameter, i.e. the points in the Language-Parameter-matrix.
- ▶ **Value** a single measurement (different types of scales can be modeled using custom attributes).
- ▶ **Unit** parts of a language system that are annotated, such as sounds, words or constructions.
- ▶ **UnitParameter** a feature that can be determined for a unit.
- ▶ **UnitValue** measurement for one unit and one unitparameter.
- ▶ **Contribution** ValueSets can be partitioned into separate contributions sharing provenance.

cldd data model: Relationships

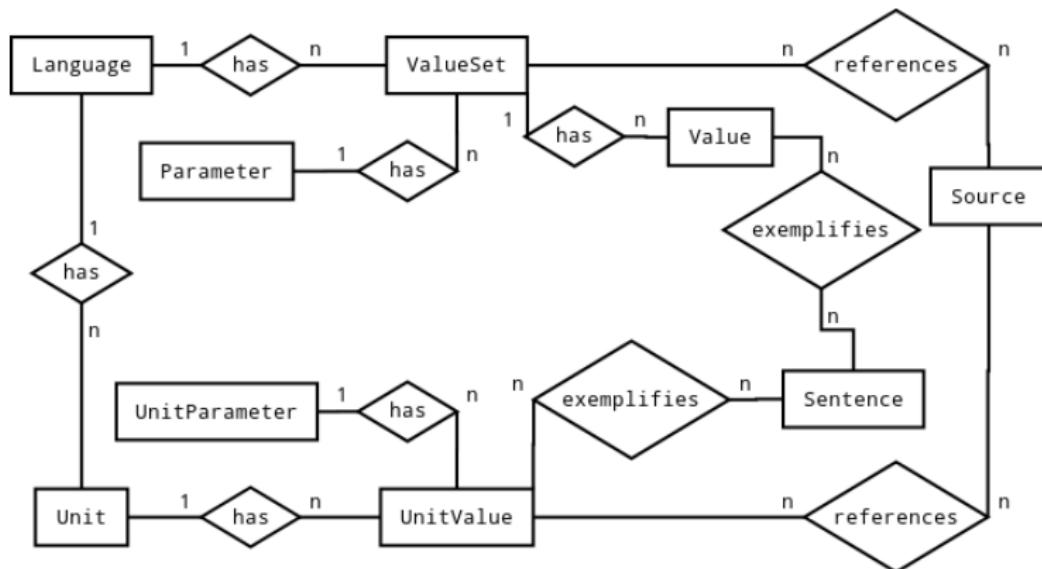


Figure 1: The default cldd data model. Note: Modelling constructions as Units and features as UnitParameters the case mentioned by Harald fits in.

cld data model: Extensibility

cld uses *joined table inheritance* as implemented in SQLAlchemy to provide extensibility of the core data model:

- ▶ Each core model can be specialised/customized in a cld app, adding columns or relationships.

```
@implementer(ILanguage)
class Languoid(Language, CustomModelMixin):
    ...
```

- ▶ The ORM (Object Relational Mapper) transparently joins the two corresponding tables when querying, retrieving the specialized object, i.e. the full set of columns.
- ▶ Additional models can be added freely, reusing cld functionality to enable functionality like versioning, etc.

clld data model: Lexical data

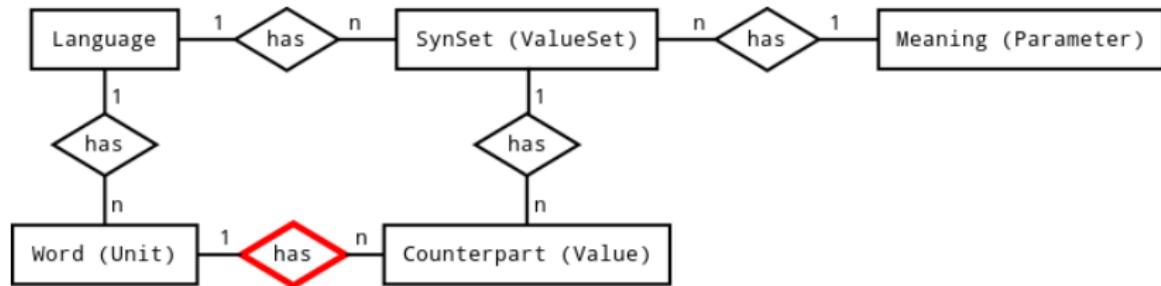


Figure 2: The WOLD instantiation of the data model.

```
@implementer(interfaces.IValue)
class Counterpart(Value, CustomModelMixin):
    ...
    word_pk = Column(Integer, ForeignKey('word.pk'))
    word = relationship(Word, backref='counterparts')
    ...
```

clld data model: Lexical data

Meaning 4.33: the hand

Description:	The body
Typical context:	
Semantic field:	The body
Semantic category:	Noun
Borrowed score	0.15
Age score	0.87
Simplicity score	0.99

Counterpart words in the World Loanword Database

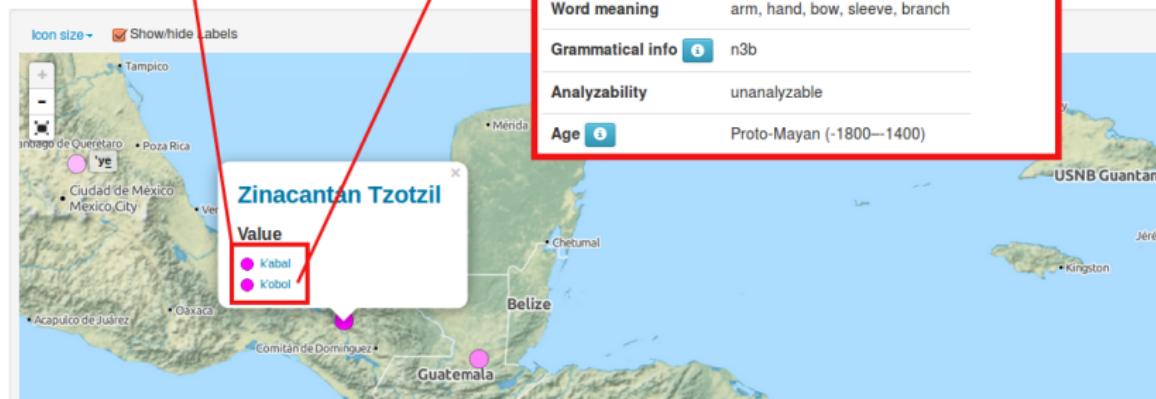


Figure 3: Many-to-many relation between words and meanings in WOLD.

clld data model: Glottolog

Classification ?

open Bunak

- ▼ Timor-Alor-Pantar (22)
 - Alor-Pantar (18)
 - ▼ East Timor-Bunaq (4)
 - Bunak (green circle)
 - East_Timor (3)

Subclassification references

- Schapper, Antoinette and Huber, Juliette and van Engelenhoven, Aone 2012

Figure 4: In Glottolog genealogy is implemented via a self-referential father relation on Language.

```
@implementer(ILanguage)
class Languoid(Language, CustomModelMixin):
    ...
    father_pk = Column(Integer, ForeignKey('languoid.pk'))
    children = relationship(
        'Languoid',
        foreign_keys=[father_pk],
        backref=backref('father', remote_side=[pk]))
    ...
    ...
```

cld resources: Overview

Data done the Web way.

cld implements a Resource Oriented Architecture.

- ▶ Data model is good basis to support shared behaviour across apps.
- ▶ Resource concept makes model entities actionable.
 - ▶ Resources are the things we describe and publish.
 - ▶ Resources define the level of granularity that is of interest.
- ▶ cld knows how to display filtered lists of resources of the same type
- ▶ and detail views of single resources.

`cldd` resources: Adaption

- ▶ ZCA (Zope Component Architecture) provides machinery to register behaviour tied to interfaces, e.g. to resources.
- ▶ Resources can be adapted to representations:
 - ▶ Glottolog: Language represented as family tree in newick format.
 - ▶ ASJP: Contribution serialized in ASJP wordlist format.
 - ▶ All lists can be represented as feeds.
- ▶ The web pages created by a `cldd` app are just resources adapted to HTML.
- ▶ These registry entries can be overridden by `cldd` apps, e.g. providing custom DataTables, custom map markers, custom maps.
- ▶ Again it's about the right level of abstraction: Writing a `cldd` app as declarative as possible, just implement adapters.

clld resources: Adaption

Language Khoekhoeogowab

Compiled by

Showing 1 to 78 of 78 entries

No.	Meaning	Word	Loan
Search	Search	Search	--any--
1	I	ta	False
1	I	KHOEKHOEGOWAB(Kho.CENTRAL_KHOISAN Khoisan,SouthernAfrica,Central,Nama @Khoe-Kwadi,Khoe,Khoekhoe) 1 -25.50, 18.00 251100 kho naq	
1	I	ta, tita, tir //	
2	you	2 you c, sac, s, sas // 1+2. masc, 3+4. fem	
3	we	3 we da, sida, sada //	
11	one	11 one !gui //	
12	two	12 two !gam, !gama //	
18	person	18 person kh-oei //	
2	you	19 fish !aub, !ganub, sui //	
2	you	sas	False
3	we	da	False
3	we	sida	False
3	we	sada	False
11	one	!gui	False
12	two	!gam	False
12	two	!gama	False
18	person	kh-oei	False

GlottoCode: nama1264 ISO 639-3: naq



Classification

WALS
Kho.CENTRAL_KHOISAN
Glotolog
Khoe-Kwadi > Khoe > Khoekhoe
Ethnologue
Khoisan > SouthernAfrica > Central > Nama

Figure 5: Two adaptions of a Language object in ASJP.

clld resources: Extensibility



Showing 1 to 15 of 15 entries (filtered from 114 total entries)

Code	Name	Category	Status	
Search	sava	-any-	-any-	Search
AT0705	East Sudanian savanna	Tropical & Subtropical Grasslands, Savannas & Shrublands	critical or endangered	WWF
AT0707	Guinean forest-savanna mosaic	Tropical & Subtropical Grasslands, Savannas & Shrublands	critical or endangered	WWF

Figure 6: Tsammalex defines a new resource type EcoRegion. EcoRegions behave just like other resources, i.e. they can be listed, bookmarked and associated with maps.

cld and Linked Data

- ▶ We regard Linked Data principles as rules of best practice for publishing data on the web.
- ▶ How do cld apps fare with respect to the five-star rating for Linked Data?
 - * Make your stuff available on the web (whatever format).
 - ** Make it available as structured data (e.g. excel instead of image scan of a table).
 - *** Non-proprietary format (e.g. csv instead of excel).
 - **** Use URLs to identify things, so that people can point at your stuff.
 - ***** Link your data to other people's data to provide context.

cld and Linked Data: three stars

Make your stuff available on the web, as structured data in non-proprietary formats.

- ▶ cld apps do just that.
- ▶ Most CLLD datasets are published under CC-BY, i.e. open, licenses.

cld and Linked Data: three stars

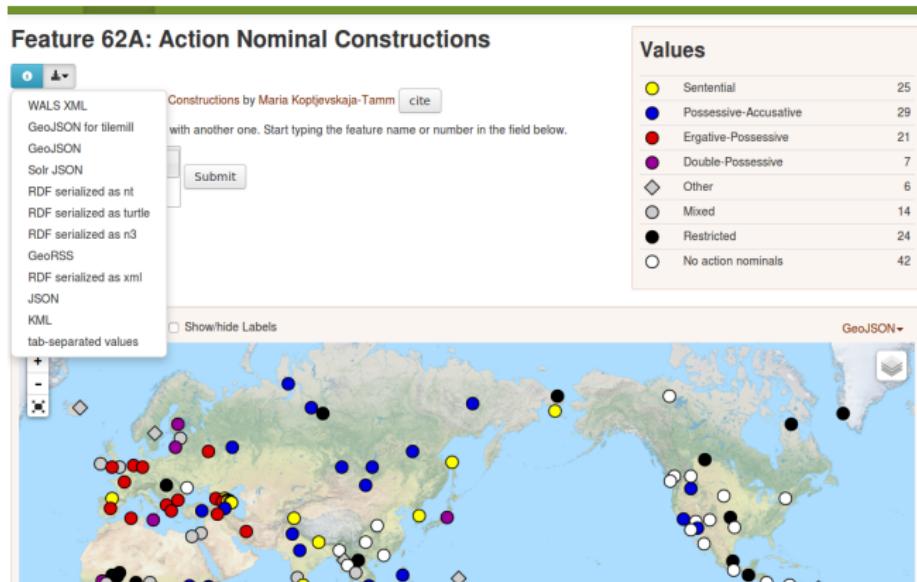


Figure 7: The data of a WALS feature is available in various formats. Note that the map on the page is created by calling the WALS API to retrieve the GeoJSON representation.

cld and Linked Data: four stars

Use URLs to identify things, so that people can point at your stuff.

- ▶ “People” includes yourself
- ▶ forces you to think about the things you want to describe and at which level of granularity
- ▶ enables distributed development of data and the basis for merging via globally unique identifiers
- ▶ puts coarse provenance information in each identifier

clld and Linked Data: four stars

<http://wals.info/valuesets/138A-lat>

Datapoint Latvian / Tea

Language: Latvian

Feature: Tea by Östen Dahl

Value: Words derived from Min Nan Chinese te

cite

comment

Examples

Sentence 4343:

tēja

References

- Malherbe and Rosenberg 1996

Comments

By: Peter Arkadiev

Mon, 09 Dec 2013 02:03:42 -0800

The Latvian word for 'tea' tēja is clearly derived from Min Nan Chinese te, and not from Sinitic cha.

History

2014-07-03 Words derived from Min Nan Chinese te

2008-04-21 Words derived from Sinitic cha

Figure 8: The level of granularity of the WALS data allows to link comments, history and examples to datapoints.

clld and Linked Data: 4-out-of-5 stars

Generally, the usefulness of “4-out-of-5 stars” Linked Data has to be stressed:

- ▶ Linked Data as uniform data access API (following the “crawler” paradigm)
- ▶ enables distributed databases,
- ▶ allows follow-your-nose API discovery (cf. REST),
- ▶ plays well with the web at large (Internet archive, bookmarking, google, etc.),
- ▶ allows easy hosting (thus helps with sustainability, and is attractive for developers/administrators as well) – which cannot be said about SPARQL endpoints.

cld and Linked Data: API and storage format

Publishing Linked Data can be as easy as putting a bunch of files on a web server.

- ▶ cld apps will be able to fall back to that, i.e. dumping the resources they serve as static files by enumerating their URL space.
- ▶ This allows for a graceful degradation of service:
 - ▶ When served from the app, resources will point to a canonical URI using the appropriate HTTP Link header.
 - ▶ These URLs will still resolve in the static-files-on-webserver scenario.
 - ▶ So when served as static files from a plain HTTP server, most things will still work

clld and Linked Data: the 5th star

Link your data to other people's data to provide context.

While HTML provides the prime example of embedding links to provide context, for structured data and common domains RDF models are more useful.

- ▶ Again “other people” includes yourself.
- ▶ VOID is used to convey basic provenance and license information.
- ▶ Typically all statements of linguistic interest (i.e. value assignments) are linked to sources.

clld and Linked Data: the 5th star

- ▶ Our publication platform does spit out RDF.
- ▶ The RDF model for a particular clld app can be completely customized.
- ▶ But should it?
- ▶ Balance between
 - ▶ uniform access across CLLD apps and
 - ▶ semantic interoperability with existing infrastructure.
 - ▶ Is it more useful to model resources as having multiple types or provide mappings?
- ▶ Example: Model lexical data using lemon?
- ▶ Generally, in terms of user-friendliness, the problem is not a choice of RDF models but consumable formats (csv, Newick, ...)

clld and Linked Data: the 5th star

- ▶ Glottolog as hub in the CLLD Linked Data cloud:
 - ▶ language catalog (linking in turn to lexvo, dbpedia, etc.), iso639-3 is often not sufficient.
 - ▶ shared bibliography
- ▶ WOLD as catalog for comparison meanings
(cf. Leipzig-Jakarta list) – a *concepticon*, or an *ontology*.
- ▶ PHOIBLE may play such a role for phonological segments,
e.g. as reference for transcriptions.
- ▶ filling in blanks: Identify phonological descriptions for
languages missing in PHOIBLE by inspecting Glottolog.
- ▶ fill in missing values in WALS for phonological features by
looking up PHOIBLE.

CLLD and Linked Data: A workflow for research based on CLLD data

1. Identify suitable datasets.
 2. Aggregate the data in a triple store (crawling/importing dumps).
 3. Filter data in the triple store (using provenance information, etc.).
 4. Export data to suitable format for analysis.
- CLLD and Linked Data will mainly play a role during aggregation of raw data.

cldd utilities: Versioning/updating/preservation

Several models are possible:

- ▶ versioned data in database
- ▶ only current data in database, archived older versions (ZENODO)
- ▶ updates via database migration scripts (versioned together with the software)

cld utilities: SAILS archived with ZENODO

The screenshot shows the Zenodo website interface. At the top, the Zenodo logo and the tagline "Research. Shared." are visible. Below the header, there are navigation links: Search, Communities, Browse ▾, Upload, and Get started ▾. On the right, there are "Sign In" and "Sign Up" buttons. The main content area displays a dataset entry for "SAILS 2014". The publication date is listed as "03 April 2014". The dataset is categorized under "Dataset" and "Open access". A note indicates that the deposit contains both the data of SAILS and the software serving it, with a link to <http://sails.cld.org>. The file listing table shows one item: "sails-v2014.zip" uploaded on "05 Aug 2014" with a size of "1.1 MB". A "Download" button is provided next to the file name. To the right of the file table, a sidebar contains information about the dataset's availability in GitHub, its publication date (03 April 2014), DOI (10.5281/zenodo.11175), keyword(s) (linguistics), related publications and datasets, supplement information (link to GitHub repository), and collection details (Communities > Cross-Linguistic Linked Data).

Name	Date	Size
sails-v2014.zip	05 Aug 2014	1.1 MB

Available in
GitHub

Publication date:
03 April 2014

DOI
DOI 10.5281/zenodo.11175

Keyword(s):
linguistics

Related publications and datasets:
Supplement to:
<https://github.com/cld/sails/tree/v2014>, <http://sails.cld.org>

Collections:
Communities > Cross-Linguistic Linked Data

Figure 9: Archiving SAILS with ZENODO means longterm preservation and better citeability via DOI.

Standardization the Microsoft way?

- ▶ As demonstrated above, a standard software stack is useful.
- ▶ But software has a half-life of less than 10 years.
- ▶ Next step is essential: extract a **domain specific** API which can become standard.
 - ▶ Linked Data is still lacking in domain specificity.
 - ▶ Domain specific means semantic interoperability of linguistic concepts.

Towards a domain specific API: Decoupling database from visualization/analysis

- ▶ for OLAC there's OAI-PMH
- ▶ for mapping (i.e. leaflet, tilemill) there's GeoJSON
- ▶ but then there's RefLex
- ▶ and <http://phonotactics.anu.edu.au/>
- ▶ and the WALS Sunburst explorer
- ▶ ...

CLLD databases on OLAC

OLAC Archive Metrics		Comparative Archive Metrics								
(Click column headers to sort)										
Archive	Overall Rating	Number of Resources	Number of Resources Online	Distinct Languages	Distinct Linguistic Subfields	Distinct Linguistic Types	Distinct DCMI Types	Average Elements Per Record	Average Encoding Schemes Per Record	Average Metadata Quality Score
Glottolog 2.3	★★★★★	7684	7684	7664	0	1	1	10.0	7.0	9.3
Ethnologue: Languages of the World	★★★★	7480	7480	7479	0	0	1	10.0	7.0	8.3
SIL Language and Culture Archives	★★★★	28448	5467	3080	0	3	5	13.2	8.3	8.9
The LINGUIST List Language Resources	★★★★	2440	0	2430	0	0	1	11.0	7.0	8.4
WALS Online	★★★★★	2621	2621	2420	0	1	1	10.0	7.0	9.3
The Rosetta Project: A Long Now Foundation Library of Human Language	★★★★	6571	6571	2365	3	3	3	18.4	7.5	8.9
WALS Online RefDB	★★★★	7157	7157	2341	7	0	1	11.5	8.3	7.1
PHOIBLE Online	★★★★★	1672	1672	1668	1	1	1	11.0	8.0	9.5
Graduate Institute of Applied Linguistics Library	★★★★	8176	394	1335	23	3	5	14.3	7.2	7.8
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	★★★★★	9266	9189	839	4	3	3	26.7	12.3	9.0

Figure 10: 3 out of the top-ten of OLAC archives by number of distinct languages are based on CLLD datasets.

Visualization: Phonotactics

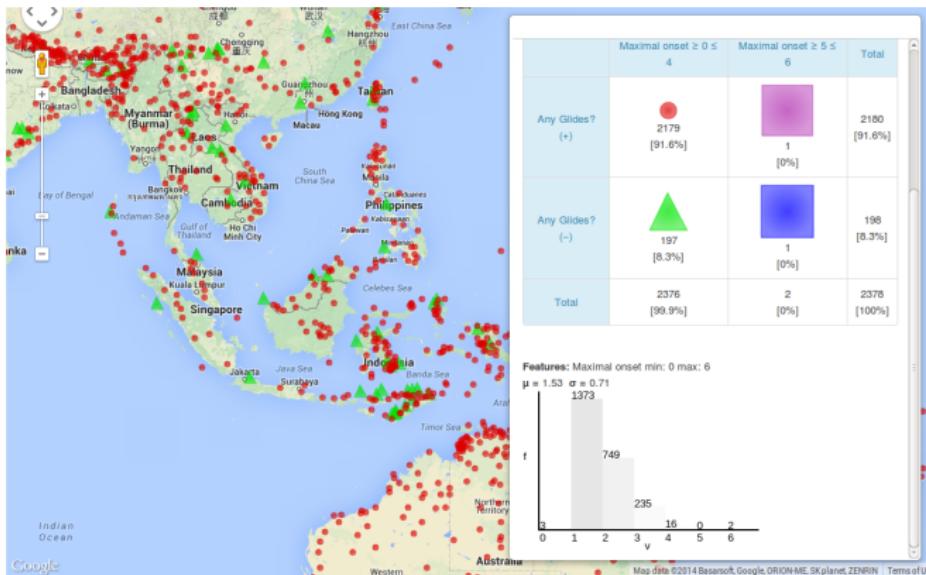


Figure 11: Configurable visualization of phonotactic features of the world's languages.

Visualization: WALS Sunburst Explorer

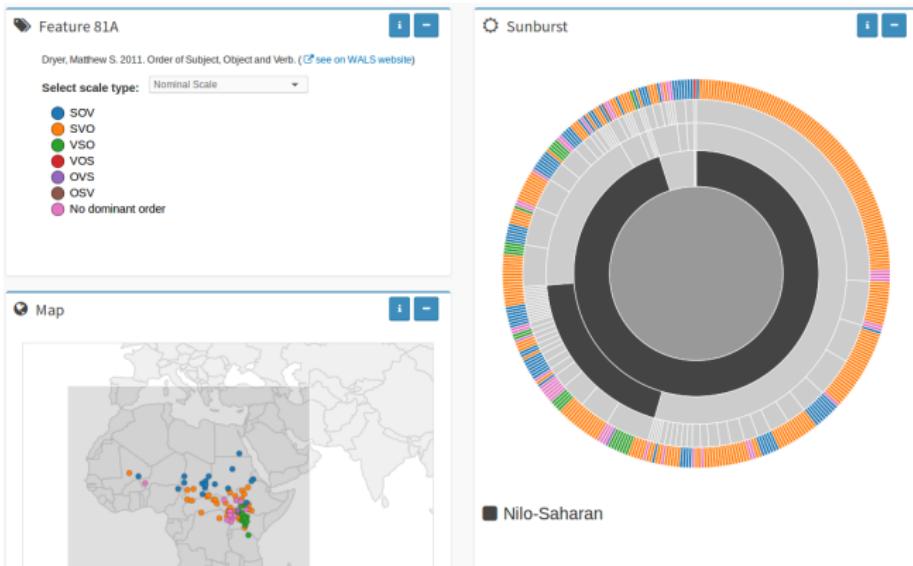
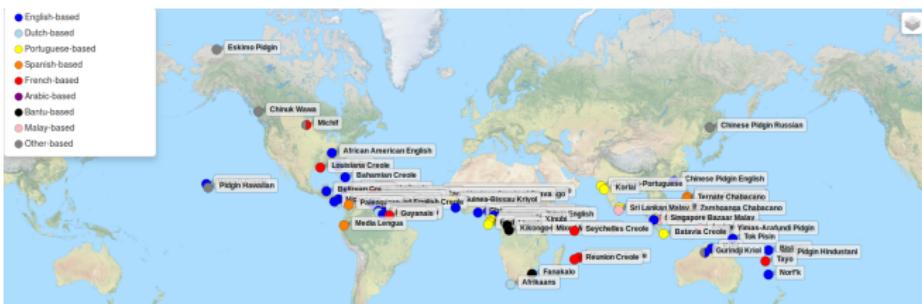


Figure 12: Combined visualization of geolocation, genealogy and coding for a WALS feature.

Semantic interoperability

- ▶ Being able to evaluate provenance data during the aggregation of a dataset is useful (e.g. in the ASJP project, some sources of wordlists are regarded as less trustworthy than others).
- ▶ Unambiguous identification of languages is required; Glottolog will help with that.
 - ▶ Being able to answer the question “which data do we have on a selected sample of languages?” as well as
 - ▶ “what sample of languages can we investigate given we need a certain selection of data (lexical, structural, etc.)?”
- ▶ For lexical data *lemon* can help to interpret the raw data, i.e. matching senses across languages (cf. Moran and Brümmer 2013).
- ▶ The requirements of statistical methods may lead to a standardisation of structural language parameters (features in the WALS sense), but we are not there yet.

Semantic interoperability: Language identification



The languages described in APiCS and eWAVE show that iso639-3 is insufficient for language identification.

Semantic interoperability: Limitations

- ▶ Generally, useful data formats will be dictated by the needs of the analysis tools (e.g. phylogenetic software),
- ▶ so doing analyses directly on the RDF model can not be expected.
- ▶ Example APiCS: Interoperability of typological resources is hampered by the difficulty of cross-linguistic categories.

Semantic interoperability: APiCS and WALS

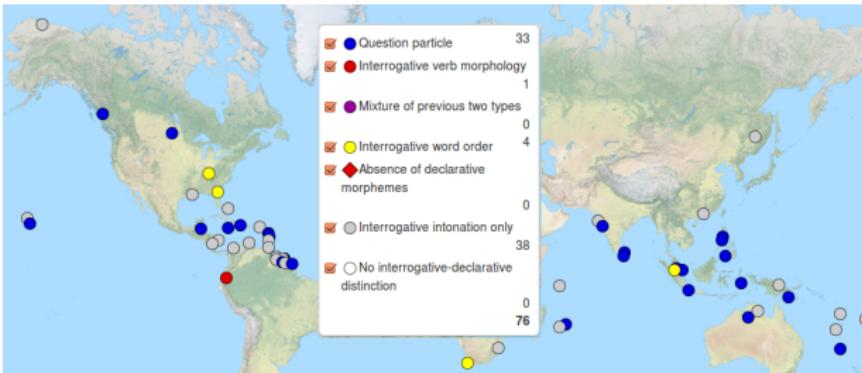
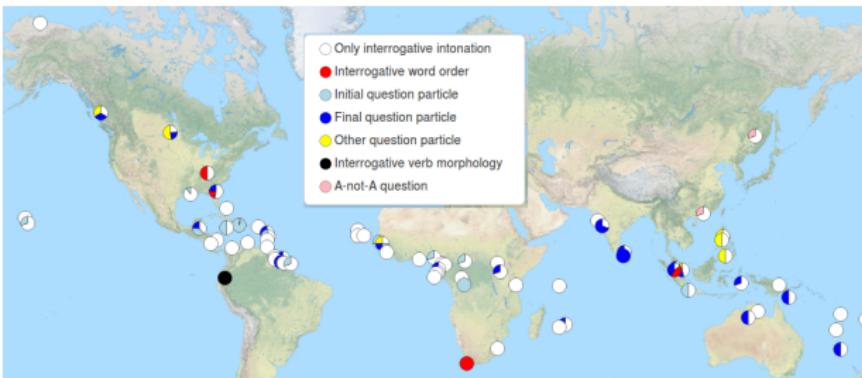


Figure 13: APiCS feature *Polar questions* – original and WALSified.

Towards a domain specific API

Roadmap:

1. "standardize" on software
2. determine what a proper API would look like (right now!)
 - ▶ collect use cases,
 - ▶ implement prototypes,
3. specify API – maybe ontologies, maybe RDF models, maybe ling-JSON ...

Hit the road . . .

<http://clld.org>

Thank you!