Introductions

Chair for Multilingual Computational Linguistics (MCL) at University of Passau

Where to find us: `https://www.geku.uni-passau.de/en/mcl/`

Teaching

- Comparative linguistics, computational linguistics, historical linguistics, cognitive linguistics.
- Teaching of this year included e.g. Language History and Historical Language Description, Culture documentation and fieldwork, Language diversity

Research

- Investigation into evolutionary, typological, and cognitive aspects of linguistic variation
- Ongoing projects on e.g. computational approaches to fieldwork documentation, language processing, polysemy, object naming, phonological vectors, etc pp

Tools

- Development of data and tools for computer-assisted linguistic analysis such as CLDF, LingPy, Edictor and more (discussed in part 2)

Jessica

- PhD 2021 in General Linguistics @ Heinrich-Heine University in Düsseldorf
  Passau since July 2023
- Research focus on Maltese, language processing and distributional semantic representations
- Area of linguistics: psycholinguistics, morphology, phonetics & phonology, computational linguistics

Kellen

- MA @ Tsing Hua University 國立清華大學 Taiwan
  PhD @ La Trobe University Australia
  Postdoc @ Universität Zürich until 2023
  Asst. Prof @ Passau since October 2023
  Adjunct research fellow @ La Trobe since 2019
- Research: linguistic fieldwork, reconstruction of linguistic history (phonological reconstruction, migration, probabilistic phylogenies, maps), anthropological linguistics, tone systems
- Minority Sinitic languages (Wu, Datian Min, Hakka); Tibeto-Burman languages of NW Myanmar & NE India

Schedule

Day 1 - data collection & data structures
Day 2 - data continued + hands on practice
Day 3 - linguistic phylogenies. what they are, how they're done, how to read them
Day 4 - language processing

Part 1 - Data collection & structuring

SSOL 2024, České Budějovice

Jessica Nieder & Kellen Parker van Dam
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau, Germany

21 August 2024

Data collection

Informal survey

What languages do you work with?

Informal survey

What languages do you work with?

Where does the data come from?

Informal survey

What languages do you work with?

Where does the data come from?

What format is it in?

Informal survey

What languages do you work with?

Where does the data come from?

What format is it in?

Do you do any first-hand data collection? (show of hands)

Data sources

Lots of data are available in some conventional sources (old published word lists etc)

Plenty of these have been digitised, and are readily available in computer-friendly formats.

Languages without many data sources

- Wolam Ngio (max 6000 speakers in 2011)
  - dictionary, descriptive grammar, phonological description, account of tonogenesis imagefor it's sub-branch. full documentation.
    with Ms. Keen Thaam
- Gongvanpounyiu (maybe 1500 speakers?)
  - sketch grammar, phonology, tone work
    with Ms. Methiam Thangjiu
- Muishaung (around 2000 speakers?)
  - descriptive grammar, dictionary, etc. full documentation
    with Mr. Wanglung Keluim
- Kaisan (maybe 1000 speakers?)
  - sketch grammar, oral history corpus
    with Mr. L. Bahnkong Kaisan, Ms. Nora Muheim

search 'Patkaian' on Glottolog for a bigger tree image

Data sources

Lots of data are available in some conventional sources (old published word lists etc)

Plenty of these have been digitised, and are readily available in computer-friendly formats.

For major languages, published sources are often sufficient, but what if you're not after English or German data? Or after non-standard varieties?

question:

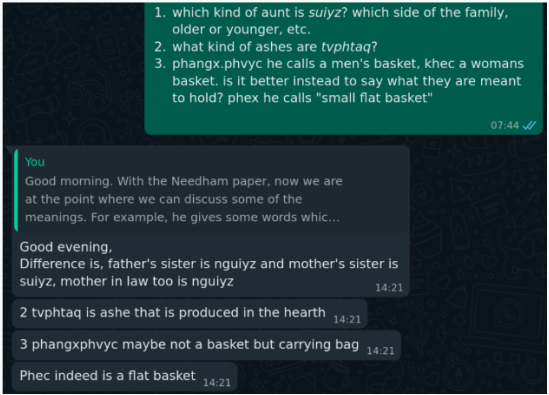
What are some potential non-traditional data sources which you may use?

Some less conventional sources

WhatsApp discussions with native-speakers

Social media plays a huge part in how people interact with their languages, whether they have standard spellings or not.

Even basic concept data in published word lists can be severely lacking, as languages don't always divide the lived experience in the same way, or certain terms carry additional semantics that can be easily missed.



Data structuring

Structuring data

Computer-assisted language processing begins with good data.
What does that look like?

1. machine readable **text**
    - **OCR**'d text if scans
    - **Unicode** or something comparable

Great! our WhatsApp chats are good computational data then?

Structuring data

Computer-assisted language processing begins with ~~good~~ well structured data.
What does that look like?

1. machine readable **text**
    - OCR'd text if scans
    - Unicode or something comparable
2. machine readable **structure**
    - flat (two-dimensional) format
    - standardised, ideally

What we'll go over today

1. Data format comparison
2. tabular data
   2.1 why it's good
   2.2 keeping it flat
3. CLDF (cross-linguistic data format)
   3.1 benefits/costs
   3.2 tools
   3.3 future-proofing you work in case you some day need it
4. wrapping up

Examples

Data structures differ from application to application. If you've dealt with legacy projects or inherited data, or maybe even if not, you'll have seen some of these.
Let's quickly see some examples.

- Toolbox / Shoebox
- FLeX
- ELAN
- Google Sheets (or equivalent)

Toolbox (& Shoebox, Lexique Pro)

A standard linear textfile used for the discontinued **Shoebox** and **Toolbox** programs. Each field has a name, with custom fields available.

Each pre-set field has a standard meaning, e.g. \lx for lexeme, and custom fields can be added.

Files are read top to bottom, with `\lx` starting each new entry.

```
\lx kkq
\ph kAk4
\ps v.
\de to keep one portion
\xv kkqwjmYigqwj
\xr kAk4 wai4 maiN1 wai4
\so phk_kaak4wai4maiN1wai4.wav
\dt 16/Jan/2022

\lx kkD
...
```

Hailowng, Morey & van Dam (in press) Tai Phake dictionary

Toolbox - Phake Dictionary_Ailot.txt

File Edit Database Project Tools Checks View Window Help

[no filter]

Phake Dictionary_Ailot.txt:1

\lx Lexeme	\ph Phonetic form	\hm Homonym numbe	\ps Part of spe	\ge Gloss (E)	\de Definition	\pl Couplet	\pd Couplet form	\se Subentry	\notes *	\dt Date (last edited)
ကဆက်	kāk^{1}	2	*n.*	spoon	bamboo or	*empty*	*empty*	ကဆက်ခီ	ailot	07/Apr/2016
ကဆက်	kak^{1}	5	*v.*	stammer	stammer, s	ထိမ်ကြောင	lin^{4} kak^{1}	ကဆက်ကျ	ailot	07/Apr/2016
ကဆိ	kāk^{3} kāk^{3}	1	*v.*	caws of cro	Onom -cav	*empty*	*empty*	*no field*	ailot	13/Apr/2016
ကဆက်	kāk^{4}	3	*v.*	keep.portior	to keep on	ကဆက်ဝေ	kāk^{4} wai^{4}	ကဆက်ရှိ	ailot	14/Apr/2016
ကဆက်ဝေ	kāk^{3} sɔ̄^{2}	1	*n.*	lock	a lock	*empty*	*empty*	*no field*	ailot	07/Apr/2016
ကင်	kaŋ^{4}	3	*n.*	administrativ	an adminis	ကင်ဘ	kāŋ^{4} taïi^{3}	*no field*	ailot	11/Feb/2016
ကင်	kaŋ^{3}	20								04/Apr/2016
ကင်	kaŋ^{3}	21								04/Apr/2016
ကင်	kāŋ^{4}	1								07/Apr/2016
ကင်	kāŋ^{2}	10								07/Apr/2016
ကင်	kāŋ^{2}	11								07/Apr/2016
ကင်	kāŋ^{2}	12								07/Apr/2016
ကင်	kāŋ^{2}	14								07/Apr/2016
ကင်	kāŋ^{2}	15								07/Apr/2016
ကင်	kaŋ^{3}	16								07/Apr/2016

Phake Dictionary_Ailot.txt:2

\lx Lexeme	ကဆက်ဝေ
. \le Lexeme Alternative Spelling	
. . \ph Phonetic form	kāk^{3} sɔ̄^{2}
. . . . \so Source	phk_kaak3sQ2.wav
. \hm Homonym number	1
. . \ps Part of speech	*n.*
. . . \sn Sense number	1
. . . . \de Definition (E)	a lock
. . . . \ge Gloss (E)	lock
. . . . \pc1 Picture 1	1_phk_kAk3sQ2.jpg
. . . . \pc2 Picture 2	2_phk_kAk3sQ2.jpg
. . . . \pl Couplet form	
. . . . . \pd Couplet form phonetic	
. . . . . \pde Couplet form English	
. . . . . \pdn Couplet form Assamese	
. . . . \dn Definition Assamese	
. . . . \rf Reference	
. . . . . \xv Example Phake	
. . . . . \xr Example Phonetic	

Toolbox (& Shoebox, Lexique Pro)

pro – Works well for smaller projects or when the user is incredibly consistent.

con –Any inconsistency majorly complicates one's ability to convert to other formats, such as 2D tabular data.

> *e.g.: inconsistent use of examples, sub-entries, erroneous placement in the wrong field, etc.*

con – one dimensional

```
\lx kkq
\ph kAk4
\ps v.
\de to keep one portion
\xv kkqwjmYigqwj
\xr kAk4 wai4 maiN1 wai4
\so phk_kaak4wai4maiN1wai4.wav
\dt 16/Jan/2022

\lx kkD
...
```

Hailowng, Morey & van Dam (in press) Tai Phake dictionary

FLeX - FieldWorks Language Explorer

XML (extensible markup language) formatted documents, as a mostly standard format.

Interoperability with other recent SIL software forces an internal standard.

```
<FreeTranslation>
    <AStr ws="en">
        <Run ws="en">'I the Wihu singer of
        the original place.'</Run>
    </AStr>
</FreeTranslation>
<Reference>
    <Str>
        <Run ws="en">Tangsa_Mossang:
        nst-mos_20130218_12:142</Run>
    </Str>
</Reference>
```

Morey, S. (2013) unpublished Wihu transcripts

FLeX - FieldWorks Language Explorer

pro – Wide range of capabilities

con – Not very human readable, difficult to programmatically deal with, **but** many export options are available in FLeX.

con – Multi-dimensional file format means difficulty in using for computer-assisted processing.

con – Windows-only SIL software

```
<FreeTranslation>
    <AStr ws="en">
        <Run ws="en">'I the Wihu singer of
        the original place.'</Run>
    </AStr>
</FreeTranslation>
<Reference>
    <Str>
        <Run ws="en">Tangsa_Mossang:
        nst-mos_20130218_12:142</Run>
    </Str>
</Reference>
```

Morey, S. (2013) unpublished Wihu transcripts

Kalaba - FieldWorks Language Explorer

File Edit View Data Insert Format Tools Parser Window Help

Kalaba

Texts & Words

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

Texts

Title

Show All

My Green Mat

Text

Title Kal

Eng My Green Mat

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

1.1 Word	pus	yalola	
Morphemes	pus	yalo	-la
Lex. Entries	pus$_1$	yalo	-la
Lex. Gloss	green	mat	1SgPoss
Lex. Gram. Info.	adj	N (I)	N:(Possessor)
Word Gloss	green	my mat	
Word Cat.	mod	N	

nihimbilira

ni-	him-	*bili	-ra
ni-	hiN-	*bili	-ra
1SgSubj	3SgObj	to.see	Pres
V:(Subject)	V:Object	trans (1)	sta:Tense
I see			
V			

Free I see my green mat.

Lexicon

Texts & Words

Grammar

Notebook

Lists

25/Mar/2011 Queue: (-/-/-) No Parser Loaded Sorted by Title 1/1

ELAN - FieldWorks Language Explorer

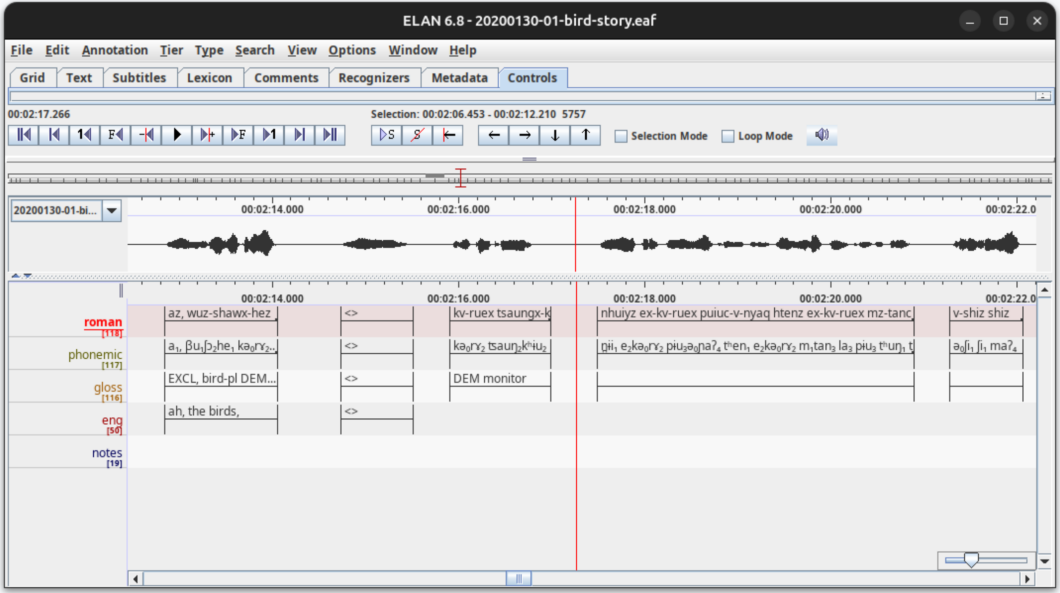
XML format, mostly focused on time-aligned transcripts but can be used to generate dictionaries as well.

ELAN documents are not consistently structured, and very susceptible to individual users' changing habits. Difficult to automate data extraction for.

Same **pros/cons** as FLeX but not Windows-only.

```
    <TIME_SLOT TIME_SLOT_ID="ts839" TIME_VALUE="315587"/>
    <TIME_SLOT TIME_SLOT_ID="ts840" TIME_VALUE="315587"/>
</TIME_ORDER>
<TIER ANNOTATOR="Kellen" LINGUISTIC_TYPE_REF="default-lt"
    PARTICIPANT="Ngunxkhuingz" TIER_ID="roman">
    <ANNOTATION>
        <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1"
            TIME_SLOT_REF1="ts4" TIME_SLOT_REF2="ts6">
            <ANNOTATION_VALUE>
                ngvyz muingc kuex Ngunxkhuingz
            </ANNOTATION_VALUE>
        </ALIGNABLE_ANNOTATION>
    </ANNOTATION>
    <ANNOTATION>
        ...
```

van Dam, K. P. (2018) unpublished Muishaung transcript


ELAN 6.8 - 20200130-01-bird-story.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls
00:02:17.266
Selection: 00:02:06.453 - 00:02:12.210 5757
Selection Mode
Loop Mode
20200130-01-bi...
00:02:14.000
00:02:16.000
00:02:18.000
00:02:20.000
roman [118]
az, wuz-shawx-hez
kv-ruex tsaungx-k
nhuiyz ex-kv-ruex puiuc-v-nyaq htenz ex-kv-ruex mz-tanc
v-shiz shiz
phonemic [117]
gloss [116]
EXCL, bird-pl DEM...
DEM monitor
eng [50]
ah, the birds,
notes [19]

JSON - Javascript Object Notation

Multi-dimensional plaintext file format used for a few difference purposes.

This example is its use in an ELAN-like text transcription, but with a different data structure.

The same issues with automating work based on inconsistent structures apply.

```
"stories": [
    {
        "title": "méiyǒu wǎng shàng zǒu de kōngjiān",
        "date": "20120315"
    "transcript": {
        "5dFouFnv": {
            "time": "0",
            "roman": "Nà, nǐ dìyīfèn gōngzuò shì shénme?
            Shì zhèige ma?",
            "english": "So, what was your first job? Was
            it this one?",
        },
        "yecqGhsk": {
            "time": "3.7",
            "roman": "Bù shì. Wǒ dì yī fèn gōngzuò shì..."
```

van Dam, K. P. (2018) unpublished Muishaung transcript

Dimensionality of the data?

2-dimensional data is preferred because

- It's what most tools expect
- It ensures that the program/script reads it properly

Any more dimensions require more work to ensure that things are being read with the correct dependencies, and that we're not missing anything.

It's not necessary for everything, but for computer assisted language comparison, staring with 2D data is necessary.

Let's look at what might be an easy solution but which can still be complicated: spreadsheets

Spreadsheets

Spreadsheets may seem 2D, with rows and columns. but it's easy to add more without meaning to.

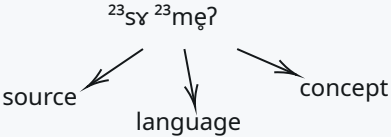
To be 2D, only pairs of data should be linked. Take the example on the right from one of my spreadsheets.

The data value **^{23}sɤ ^{23}mę̰ʔ** corresponds to a language (Thang), a source (Weidert97), and a concept (THREE).

Weidert, Alfons. "Tibeto-Burman Tonology." (1987): 1-530.

source	Weidert87	Thaam24	Wayesha10
language	Thang	Wolam	Lainong
three	^{23}sɤ ^{23}mę̰ʔ	ha.meʔ	ʃiam^{53}
four	^{12}bɤ^{2}lę̰	pə.le	ba^{21}li^{33}
five	^{12}bɤ ^{2}ŋo̰u	pə.ŋu	bə^{21}ŋəu^{33}

adapted from van Dam (in preparation)



Spreadsheets

We're already at three dimensions here.

We've put in the source in a new row, but computational tools will not like this.

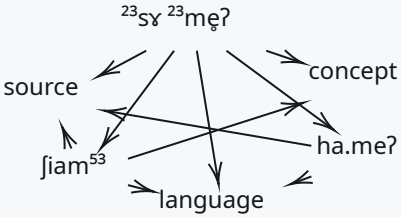
But actually our data value also links to every other data value in the same row.

To be properly 2-dimensional, each column should be exactly one type of data, and each row should be exactly one piece of data.

Not to mention spreadsheet comments/notes!

source	Weidert87	Thaam24	Wayesha10
language	Thang	Wolam	Lainong
three	^{23}sɤ ^{23}mę̰ʔ	ha.meʔ	ʃiam^{53}
four	^{12}bɤ^{2}lę̰	pə.le	ba^{21}li^{33}
five	^{12}bɤ ^{2}ŋo̰u	pə.ŋu	bə^{21}ŋəu^{33}

adapted from van Dam (in preparation)



TSVs & CSVs

We can talk about spreadsheets, but what we're really talking about is tabular data.

Simply put, this is data separated by new lines and tabs (or commas).

This has the additional benefit that its **plain text**. You don't need an `.xlsx` compatible program to read the data.

```
source,Weidert87,Thaam24,Wayesha10
language,Thang,Wolam,Lainong
three,˨˩sʊ ˨˩mu̯ʔ,ha.meʔ,ʃiam˥˧
four,˧˨bʊ˨˩lu̯,pə.le,ba˨˩li˧˧
five,˧˨bʊ ˨˩ŋu̯u,pə.ŋu,bə˨˩ŋəu˧˧
```

comma-separated values (CSV)

Flattinging the data

To flatten our data, we need to separate out the values so that each row refers to exactly one linguistic value, in this case our word forms. So we need to modify this...

source	Weidert87	Thaam24	Wayesha10
language	Thang	Wolam	Lainong
three	^{23}sɤ ^{23}mę?	ha.me?	ʃiam^{53}
four	^{12}bɤ^{2}lę	pə.le	ba^{21}li^{33}
five	^{12}bɤ ^{2}ŋǫu	pə.ŋu	bə^{21}ŋəu^{33}

Thaam, K and K. P. van Dam (2024). Wolam Ngiopit dictionary
Wayesha, A. J. (2010). A phonological description of Leinong Naga. Chiang Mai.

concept	form	language	source	comment
three	^{23}sɤ ^{23}mę?	Thang	Weidert87	
three	ha.me?	Wolam	Thaam24	*s>h
three	ʃiam^{53}	Lainong	Wayesha10	
four	^{12}bɤ^{2}lę	Thang	Weidert87	
four	pə.le	Wolam	Thaam24	
four	ba^{21}li^{33}	Lainong	Wayesha10	
five	^{12}bɤ ^{2}ŋǫu	Thang	Weidert87	
five	pə.ŋu	Wolam	Thaam24	
five	bə^{21}ŋəu^{33}	Lainong	Wayesha10	

... into something like this

Notice we can also include relevant comments here that may have been in the spreadsheet.

Why is this helpful?

It can be less intuitive to read, and I often start out less flat. But this is where we should be trying to end up for programmatical uses.

Now we can pull the whole table into Python, R, Julia or whatever else we're using, split it by lines and tabs (or commas), and we know that column 0 will have the concept, 1 the form, 2 the language, 3 the source and 4 the comments.

concept	form	language	source	comment
three	^{23}sɤ ^{23}mę̥ʔ	Thang	Weidert87	
three	ha.meʔ	Wolam	Thaam24	*s>h
three	ʃiam^{53}	Lainong	Wayesha10	
four	^{12}bɤ^{2}lę̥	Thang	Weidert87	
four	pə.le	Wolam	Thaam24	
four	ba^{21}li^{33}	Lainong	Wayesha10	
five	^{12}bɤ ^{2}ŋo̯u	Thang	Weidert87	
five	pə.ŋu	Wolam	Thaam24	
five	bə^{21}ŋəu^{33}	Lainong	Wayesha10	

Doing the same with the original structure is much harder, and requires a lot more conditional coding to handle the data.

Why is this helpful?

This also lets us easily handle multiple forms per language. Imagine something like this with two values for FIVE.

source	Weidert87	Thaam24	Wayesha10
language	Thang	Wolam	Lainong
three	^{23}sɤ ^{23}mɛ̱?	ha.me?	ʃiam^{53}
four	^{12}bɤ^{2}lɛ̱	pə.le	ba^{21}li^{33}
five	^{12}bɤ ^{2}ŋo̱u	pə.ŋu; pa.ŋu	bə^{21}ŋəu^{33}

We can simply have two rows to account for this, rather than worrying about searching for the semicolon.

concept	form	language	source	...
three	^{23}sɤ ^{23}mɛ̱?	Thang	Weidert87	
three	ha.me?	Wolam	Thaam24	
three	ʃiam^{53}	Lainong	Wayesha10	
four	^{12}bɤ^{2}lɛ̱	Thang	Weidert87	
four	pə.le	Wolam	Thaam24	
four	ba^{21}li^{33}	Lainong	Wayesha10	
five	^{12}bɤ ^{2}ŋo̱u	Thang	Weidert87	
five	pə.ŋu	Wolam	Thaam24	
five	pa.ŋu	Wolam	Thaam24	
five	bə^{21}ŋəu^{33}	Lainong	Wayesha10	

Why is this helpful?

There are a lot of other reasons to do this, including better version control (`git`) when making changes, guaranteed **future access*** and **interoperability** between users.

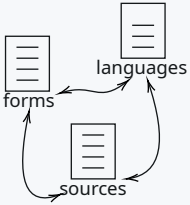
If you've ever struggled to open a data format from someone else, you will hopefully see the appeal of plain text.

concept	form	language	source	...
three	^{23}sɤ ^{23}mę?	Thang	Weidert87	
three	ha.meʔ	Wolam	Thaam24	
three	ʃiam^{53}	Lainong	Wayesha10	
four	^{12}bɤ^{2}lę	Thang	Weidert87	
four	pə.le	Wolam	Thaam24	
four	ba^{21}li^{33}	Lainong	Wayesha10	
five	^{12}bɤ ^{2}ŋo̯u	Thang	Weidert87	
TSVs five	pə.ŋu	Wolam	Thaam24	
five	bə^{21}ŋəu^{33}	Lainong	Wayesha10	

Why else is this helpful?

If you set up your data in this way, you're also creating a database.

Most of the world runs on relational databases like this, from blogs and news sites to login credentials in social media, to your class registrations.



concept	form	language	source	...
three	²³sɤ ²³mę̰?	Thang	Weidert87	...
three	ha.me?	Wolam	Thaam24	

table for linguistic forms

ID	glottocode	group	family
Thang	than1260	Sal	Sino-Tibetan
Wolam	wola1254	Sal	Sino-Tibetan

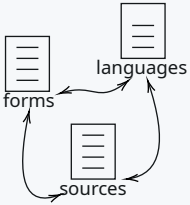
table for languages

ID	title	author	year
Weidert87	Tibeto-Bur...	A. Weidert	1987
Thaam24	Wolam Ng...	Thaam & van D..	2024

table for sources

Why else is this helpful?

While this can seem like a lot to deal with, you can ensure more consistency in the data by using consistent keys for repeated references, such as language names, sources, or anything else



concept	form	language	source	...
three	^{23}sɤ ^{23}mę?	Thang	Weidert87	...
three	ha.me?	Wolam	Thaam24	

table for linguistic forms

ID	glottocode	group	family
Thang	than1260	Sal	Sino-Tibetan
Wolam	wola1254	Sal	Sino-Tibetan

table for languages

ID	title	author	year
Weidert87	Tibeto-Bur...	A. Weidert	1987
Thaam24	Wolam Ng...	Thaam & van D..	2024

table for sources

One option: CLDF

To help ensure consistency of structure, many linguistic databases use a format called Cross-Linguistic Data Formats (CLDF), based on a few important principles:

- Data should be both editable ”by hand” and amenable to writing and reading via software
- Data should be encoded as Unicode text files.
- Referencing existing data preferred over repeating data (e.g. language names)
- Compatibility with existing tools, standards and practices

https://cldf.clld.org/

Forkel, R. et al. Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. Sci. Data. 5:180205 doi: 10.1038/sdata.2018.205 (2018).

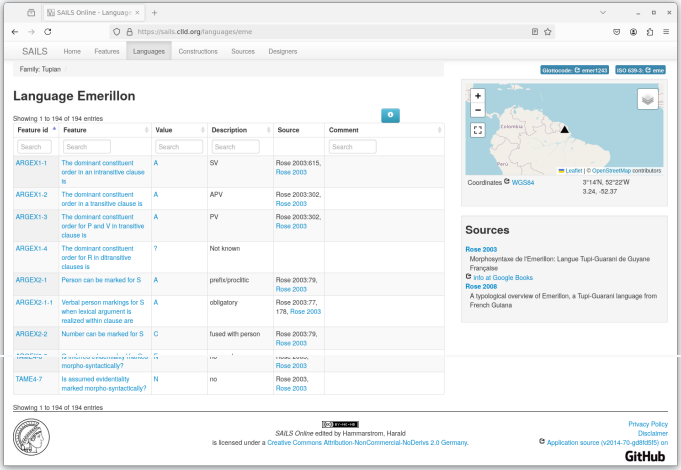
Forkel, R. et al. Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. Sci. Data. 5:180205 doi: 10.1038/sdata.2018.205 (2018)

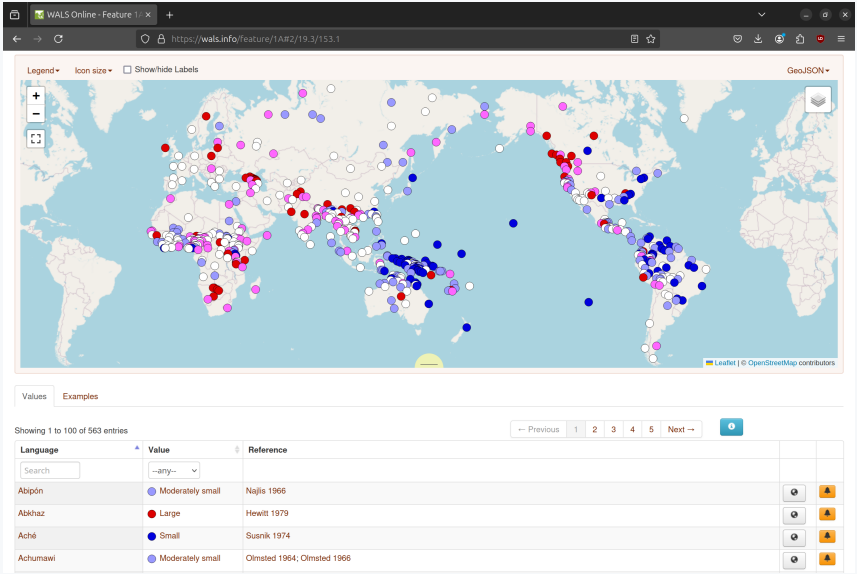
Introduction to CLDF

CLDF acts as a set of standards to ensure consistency. Used by:

- World Atlas of Language Structures (WALS) `http://wals.info`
- Glottolog `http://glottolog.org`
- South American Indigenous Language Structures (SAILS)
- PHOIBLE
- Intercontinental Dictionary Series (IDS)
- World Loanword Database (WOLD)
- Lexibank
- *etc etc*

Basically by any website of linguistic data that looks kinda like this →




WALS Online - Feature 1A
https://wals.info/feature/1A#2/19.3/153.1
Legend
Icon size
Show/hide Labels
GeoJSON
Leaflet | © OpenStreetMap contributors
Values
Examples
Showing 1 to 100 of 563 entries
← Previous 1 2 3 4 5 Next →
Language
Value
Reference
Search
--any--
Abipón
Moderately small
Najlis 1966
Abkhaz
Large
Hewitt 1979
Aché
Small
Susnik 1974
Achumawi
Moderately small
Olmsted 1964; Olmsted 1966

Glottolog 5.0 - Mandarin

https://glottolog.org/resource/languoid/id/mand1415

Glottolog | Languages | Families | Language Search | References | Reference Search | GlottoScope | About | pr. Name / glcode / iso

Spoken L1 Language: Mandarin Chinese

Glottocode: mand1415 | ISO 639-3: cmn

Classification

open Mandarin Chinese | expand all | collapse all

- Sino-Tibetan (514)
  - Bodic (84)
    - Bodish (54)
    - Kaike-Ghale-Tamangic (13)
    - Tshanglic (2)
      - **Kalaktang Monpa**
      - **Tshangla**
    - West Himalayish (15)
  - Brahmaputran (42)
  - Burmo-Qiangic (158)
  - Dhimal-Lhokpu-Toto (3)
  - Digarish (2)
  - **Gongduk**
  - Himalayish (46)
  - Karenic (20)
  - Kho-Bwa (7)
  - Kman-Meyor (2)
  - Kuki-Chin-Naga (93)
  - Macro-Bai (6)
  - Macro-Tani (12)
  - Miji (2)
  - Mruic (2)
  - Nungish (3)
  - **Olekha**
  - Raji-Raute (3)
  - Sinitic (26)
  - Unclassified Sino-Tibetan (1)

Comments on subclassification

Zev Handel 2016 Jerry Norman 2016

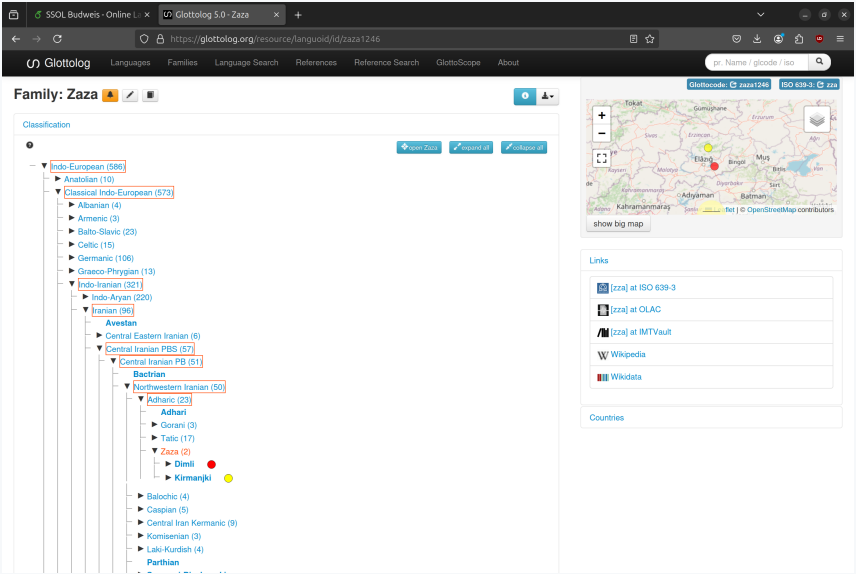
show big map

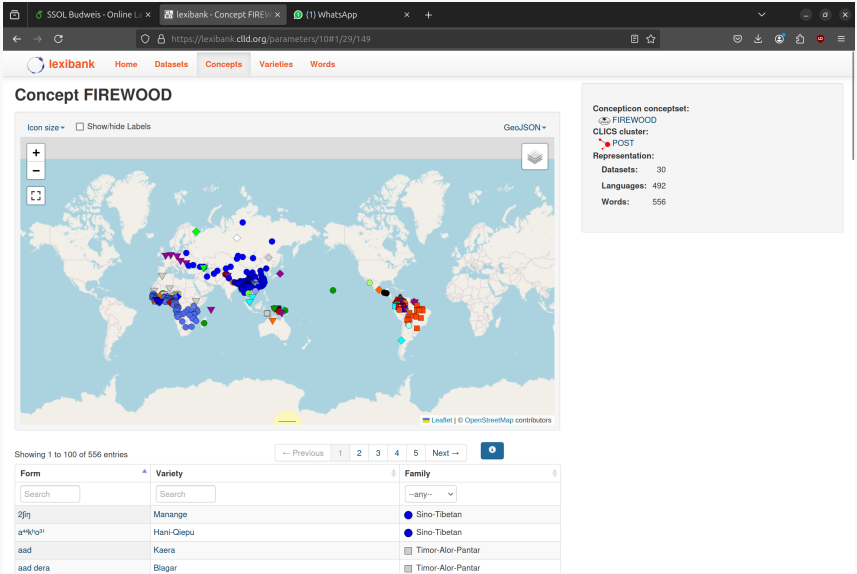
Links

- [cmn] at ISO 639-3
- [cmn] at OLAC
- [cmn] at IMTVault
- Kunming at WALS
- Mandarin at WALS
- Mandarin Chinese at Grambank
- Wikipedia
- PHOIBLE
- Wikidata

Alternative names

Countries


SSOL Budweis - Online La
Glottolog 5.0 - Zaza
https://glottolog.org/resource/languoid/id/zaza1246
Glottolog
Languages
Families
Language Search
References
Reference Search
GlottoScope
About
pr. Name / glcode / iso
Family: Zaza
Glottocode: zaza1246
ISO 639-3: zza
Classification
open Zaza
expand all
collapse all
Indo-European (586)
Anatolian (10)
Classical Indo-European (573)
Albanian (4)
Armenic (3)
Balto-Slavic (23)
Celtic (15)
Germanic (106)
Graeco-Phrygian (13)
Indo-Iranian (321)
Indo-Aryan (220)
Iranian (96)
Avestan
Central Eastern Iranian (6)
Central Iranian PBS (57)
Central Iranian PB (51)
Bactrian
Northwestern Iranian (50)
Adharic (23)
Adhari
Gorani (3)
Taliṣ (17)
Zaza (2)
Dimli
Kirmanjki
Balochic (4)
Caspian (5)
Central Iran Kermanic (9)
Komisenian (3)
Laki-Kurdish (4)
Parthian
Tokat
Gümüşhane
Erzurum
Erzincan
Sivas
Elâzığ
Muş
Bingöl
Bitlis
Malatya
Adıyaman
Batman
Diyarbakır
Siirt
Kahramanmaraş
Osmaniye
Adana
Van
Leaflet | © OpenStreetMap contributors
show big map
Links
[zza] at ISO 639-3
[zza] at OLAC
[zza] at IMTVault
Wikipedia
Wikidata
Countries


SSOL Budweis - Online La
lexibank - Concept FIREW
(1) WhatsApp
https://lexibank.clld.org/parameters/10#1/29/149
lexibank
Home
Datasets
Concepts
Varieties
Words
Concept FIREWOOD
Icon size
Show/hide Labels
GeoJSON
Leaflet | © OpenStreetMap contributors
Conceptlcon conceptset:
FIREWOOD
CLICS cluster:
POST
Representation:
Datasets: 30
Languages: 492
Words: 556
Showing 1 to 100 of 556 entries
← Previous
1
2
3
4
5
Next →
Form
Variety
Family
Search
Search
--any--
2ʃŋ
Manange
Sino-Tibetan
a⁴⁴k⁵⁵ɔ³¹
Hani-Qiepu
Sino-Tibetan
aad
Kaera
Timor-Alor-Pantar
aad dera
Blagar
Timor-Alor-Pantar

Hindu Kush Areal Typolo: ×

https://hindukush.clld.org/languages

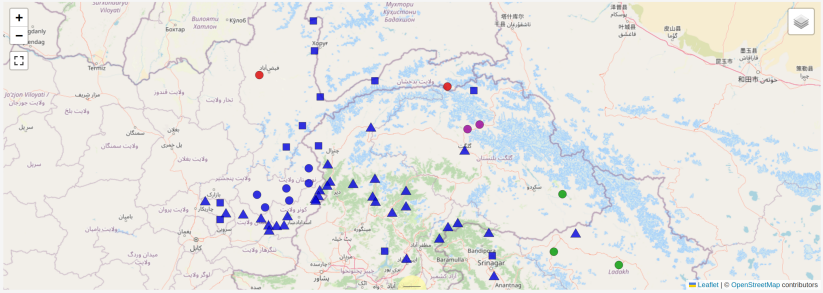
Hindu Kush Areal Typology | Languages | Features | Wordlist | References | About

The Region and its Languages

Map | Table

Hindu Kush, or the Greater Hindu Kush, in this context is really a shorthand for the remote region where the ranges of the Hindu Kush, the Karakoram, the Pamirs and the westernmost extension of the Himalayas meet (Lijegren 2014: 134–138; Bashir 2016: 264). These northwestern outskirts of the subcontinent are inhabited by at least 50 distinct ethnolinguistic communities (Hammarström, Forkel & Haspelmath 2017; Lewis, Simons & Fennig 2016). The Hindu Kush, in this sense, is part of the territories of several countries -- primarily Afghanistan, Pakistan and India. The geographically most salient feature is its mountainous environment, especially vis-à-vis the Indo-Gangetic plains situated south of it. While being a transit zone of sorts between the cultural spheres of South Asia, Central Asia, West Asia and the Himalayas, this is simultaneously the easternmost extension of Iranian languages, the northernmost extension of Indo-Aryan languages as well as the westernmost extension of Sino-Tibetan. Apart from those three phylogenetic components, the region is also home to Nuristani, at least two Turkic language enclaves and the language isolate Burushaski. These six major phylogenies of the linguistic landscape of the Hindu Kush will be introduced briefly.

Legend | Icon size | Show/hide Labels | GeoJSON



Indo-Aryan (which along with Iranian and Nuristani belongs to the larger Indo-Iranian branch of Indo-European) is the largest phylogenetic component, making up at least half of the languages in the Hindu Kush region, relatively evenly distributed in a southern belt stretching from east to west. Those can be grouped into at least nine relatedness clusters or groups (Pashai, Kunar, Chitral, Kohistani, Shina, Kashmiri, Western Punjabi, Rajasthani, and Central), although the exact placement of a few of them remains uncertain (Strand 1973: 207–208; 2001: 251; Bashir 2003). In the past, the label "Dardic" was collectively applied to languages belonging to the six first-mentioned groups, all of them Northwestern Indo-Aryan languages, with a longstanding presence in the region. That label is, however, no longer relevant as a classificatory entity (Morgenstierne 1961). The region's Western Punjabi varieties (such as Pahari-Pothwari and Hindko) are really part of a larger Punjabi continuum with an extension far south of the region, and as such probably have more in common with the closest main Indo-Aryan languages of the Indo-Pakistani plains than

Introduction to CLDF

In addition to the flat text files, CLDF uses a `JSON` file to help keep track of important metadata about the tables.

This includes which types of text are allowed in which field, as well as keeping track of more standard field types such as language names, IPA forms, etc.

As an example, here's part of the **languages** table's metadata from WALS.info

```
"dc:conformsTo": "http://cldf.clld.org/v1.0/terms.rd...",
"dc:description": "WALS' languages and language grou...",
"tableSchema": {
  "columns": [
    {
      "datatype": {
        "base": "decimal",
        "minimum": "-90", "maximum": "90"
      },
      "name": "Latitude", "propertyUrl": "http://cldf..."
    },
    {
      "datatype": {
        "base": "string",
        "format": "[a-z0-9]{4}[1-9][0-9]{3}"
      },
      "propertyUrl": "http://cldf.clld.org/v1.0/term...",
      "valueUrl": "http://glottolog.org/resource/lan...",
      "name": "Glottocode"
    }
```

Conclusion

Conclusion For now, it's most important to understand what's behind these standards that make for good and usable data.

- flat structures
- plain text but Unicode compliant
- data that is machine readable and human readable
- data which can be machine- and human editable
- version-control ready

Tomorrow: We will put this into practice with the EDICTOR software to see some of the benefits of well-structured data.

We're going to work with some parallel data in the form of Swadesh lists. This will take you to a Google spreadsheet containing a number of lists. Go here and copy one of the tabs (just one) to a spreadsheet of your own.

https://tinyurl.com/ssol-swadesh

Part 2 - Practice

SSOL 2024, České Budějovice

Jessica Nieder & Kellen Parker van Dam
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau, Germany

22 August 2024

Morris Swadesh

- Doctoral student of Edward Sapir
- 1933 thesis on Nuučaan̓uł
- developed the list named after him of 207 **concepts** meant to be used in lexicostatistics - an old system of quantifying linguistic similarity as an effort to determine relatedness

Cross-linguistics concept lists have a long history, but none are as famous as those of Swadesh, namely the 100-concept version and the 207-concept version.
The Google Sheet from yesterday is based on the 207 word list.

Universality of concepts

Swadesh's list/s were intended toward basicness.

Basic concept lists exist in many forms, many attempting to adjust Swadesh lists for better cultural fitting.

Importantly, cross-linguistic concept lists make an effort toward universal concepts.

What might be a universal concept?

Universality of concepts

Why do we care about universals, anyway?

CLLD Concepticon 3.2.0 -

https://concepticon.clld.org

Concepticon | Home | Concepts | Concept sets | Concept lists | Languages | Compilers | Sources

About | Legal | Download | Contact

Welcome to the Concepticon

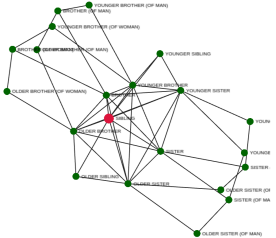
This resource presents an attempt to link the large amount of different concept lists which are used in the linguistic literature, ranging from Swadesh lists in historical linguistics to naming tests in clinical studies and psycholinguistics.

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.

Note that the most important contribution by the Concepticon project are not the definitions given for individual concept sets, but the judgments which individual elicitation glosses to assign to the same concept set. As a result, the definitions may sometimes look less than optimal. We appreciate any help in improving the definitions, but we recommend users to check the list of assigned elicitation glosses first, since these assignments should inform the definition, and not vice versa.

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. For details about the relation between Concepticon and NoRaRe, refer to Tjuka, Forkel, and List (2023) "Curating and extending data for language comparison in Concepticon and NoRaRe"



arbitraire

Cite

List, Johann Mattis & Tjuka, Annika & van Zantwijk, Mathilda & Blum, Frederic & Ugarte, Carlos Barrientos & Rzymski, Christoph & Greenhill, Simon & Forkel, Robert (eds.) 2024.
CLLD Concepticon 3.2.0 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7298022

DOI 10.5281/zenodo.7298022

cite

Version

concepticon.clld.org serves the latest released version of data curated at concepticon/concepticon-data. Older released version are accessible via DOI: 10.5281/zenodo.586412 on ZENODO as well.

CLLD Concepticon 3.2.0 edited by List, Johann Mattis & Tjuka, Annika & van Zantwijk, Mathilda & Blum, Frederic

Privacy Policy
Disclaimer

CLLD Concepticon 3.2.0 -

https://concepticon.clld.org/values

Concepticon | Home | Concepts | Concept sets | Concept lists | Languages | Compilers | Sources

Concepts

Showing 1 to 100 of 131,335 entries

← Previous 1 2 3 4 5 Next →

Id	Description in source	Concept set
Search	Search	Search
BeijingDaxue-1964-905-312	算盘 [chinese]	ABACUS
List-2016-180-50	算盘 [chinese]; abacus [english]	ABACUS
Nicholas-1989-60-60	Abacus [english]	ABACUS
Calero-2002-15-15	Abacus [english]; Ábaco [spanish]	ABACUS
Hale-1973-1798-762	abandon [english]	ABANDON
Alpher-1999-151-48	to leave it [english]	ABANDON
Hale-1961-100-40	to leave it [english]	ABANDON
OGrady-1969-100-48	leave it [english]	ABANDON
Nagano-2013-1256-637	give up (v) [english]	ABANDON
Snider-2004-1700-420	abandon [english]; abandoner [french]	ABANDON
Anonby-2018-1500-323	abandon [english]; ترک کردن [persian]	ABANDON
Lapesa-2014-772-1	abandon [english]	ABANDON
Gravina-2014-717-1	abandon [english]	ABANDON
Pallas-1789-285-65	potentia [latin]; Мочь [russian]	ABILITY
Pallas-1786-442-126	Faculté [french]; Krafft [german]; Potentia [latin]; Мочь [russian]	ABILITY
Pereira-2018-180-1	ability [english]	ABILITY
Xiao-2012-213-34	才 [chinese]	ABILITY
Hill-2015-999-580	ability [english]	ABILITY
Zalizniak-2020-2590-2567	ability [english]	ABILITY
Vulic-2020-2244-490	قدرة [arabic]; 能力 [cantonese]; 能力 [chinese]; ability [english]; võime [estonian]; kyky [finnish]; aptitude [french]; יכולת [hebrew]; umiejętność [polish]; способность [russian]; capacidad [spanish]; gallu [welsh]	ABILITY
Zalizniak-2024-4583-7	ability [english]	ABILITY
Vergallito-2020-1121-10	abortion [english]; aborto [italian]	ABORTION
Lai-2023-291-257	above [english]	ABOVE

CLICS³ - Colexified conce

https://clics.clld.org/edges/1277-1673

CLICS³ Home Datasets Varieties Concepts

Colexifications for "HAND" and "ARM"

Search:

Language	Family	Form for HAND	Gloss for HAND	Form for ARM	Gloss for ARM
Old High German	Indo-European	*hant*	hand	*hant*	arm
Cha'palaachi	Barbacoan	*ɬaapa*	hand	*ɬaapa*	arm
Dímina	Chibchan	*gúla*	hand	*gúla*	arm
Ika	Chibchan	*günni*	hand	*günni*	arm
Koreguaje	Tucanoan	*hĩñ̩*	hand	*hĩñ̩*	arm
Orejón	Tucanoan	*hĩɨ̃*	hand	*hĩɨ̃*	arm
Páez	Páez	*kuse*	hand	*kuse*	arm
Piratapuyo	Tucanoan	*õbõkã*	hand	*õbõkã*	arm
Siona	Tucanoan	*ãã-sada*	hand	*ãã-sada*	arm
Tsafiqui Pila	Barbacoan	*tede*	hand	*tede*	arm
Tucano	Tucanoan	*õbõkã*	hand	*õbõkã*	arm
Avar	Nakh-Daghestanian	*кеер*	hand	*кеер*	arm
Erzya Mordvin	Uralic	*kedʲ*	hand	*kedʲ*	arm
Khanty	Uralic	*yoś*	hand	*yoś*	arm
Komi	Uralic	*ki*	hand	*ki*	arm
Northern Saami	Uralic	*giehta*	hand	*giehta*	arm
Mansi	Uralic	*kaṭ*	hand	*kaṭ*	arm
Mari	Uralic	*ˈkið*	hand	*ˈkið*	arm
Nenets	Uralic	*ŋuda*	hand	*ŋuda*	arm

Graphs

This edge appears in cluster

ARM

and subgraphs

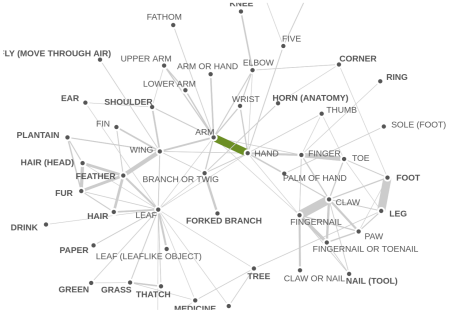
- Subgraph WING
- Subgraph FINGER
- Subgraph BRANCH OR TWIG
- Subgraph LOWER ARM
- Subgraph CLAW
- Subgraph THATCH
- Subgraph FINGERNAIL
- Subgraph HAND
- Subgraph GREEN
- Subgraph TOE
- Subgraph FORKED BRANCH
- Subgraph TOBACCO
- Subgraph THUMB
- Subgraph PALM OF HAND
- Subgraph WRIST
- Subgraph GRASS
- Subgraph LEAF (LEAFLIKE OBJECT)
- Subgraph UPPER ARM
- Subgraph ARM OR HAND
- Subgraph FATHOM
- Subgraph PAPER
- Subgraph CLAW OR NAIL
- Subgraph LEAF
- Subgraph SHOULDER
- Subgraph RING
- Subgraph FIN
- Subgraph FINGERNAIL OR TOENAIL
- Subgraph ARM
- Subgraph FIVE
- Subgraph HORN (ANATOMY)
- Subgraph ELBOW
- Subgraph FEATHER
- Subgraph CORNER
- Subgraph KNEE
- Subgraph NAIL (TOOL)

CLICS³ - Graph Subgraph

https://clics.clld.org/graphs/subgraph_1277

CLICS³ Home Datasets Varieties Concepts

Subgraph HAND

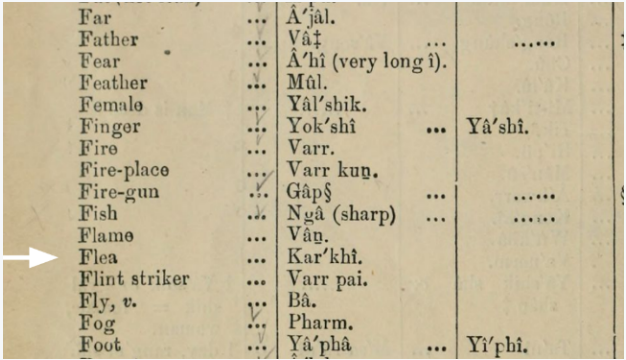


300 colexifications for "HAND" and "ARM":

Language	Family	Form
Gawwada	Afro-Asiatic	hargo
Hausa	Afro-Asiatic	hannu
Hausa	Afro-Asiatic	hannuu
Iraqw	Afro-Asiatic	dawa1
Polci	Afro-Asiatic	aam
Tarifiyt Berber	Afro-Asiatic	fus
Hokkaido Ainu	Ainu	tek
Kimochi.unn	Atlantic-Congo	owoko
Kiseri.unn	Atlantic-Congo	kuoko
Lema.unn	Atlantic-Congo	kuwoko
Machame.unn	Atlantic-Congo	woko
Siha.unn	Atlantic-Congo	oko
Swahili	Atlantic-Congo	mkono
KNB (a Pearic variety)	Austroasiatic	daj
Keme (Kemie variety)	Austroasiatic	thi53
Mang VN	Austroasiatic	eng6
Phong	Austroasiatic	si24
Samre	Austroasiatic	tia
Surin Khmer	Austroasiatic	daj
Vietnamese	Austroasiatic	tay
Alcorse, Alor-Besar	Austroasiatic	limano

Introduction to CLDF

One must still be cautious when it comes to published data, as we may not actually understand what's being meant by the glosses given.



Far	...	Â′jâl.		
Father	...	Vâ‡	...	.....
Fear	...	Â′hî (very long î).		
Feather	...	Mûl.		
Female	...	Yâl′shik.		
Finger	...	Yok′shî	...	Yâ′shî.
Fire	...	Varr.		
Fire-place	...	Varr kuṇ.		
Fire-gun	...	Gâp§	...	.....
Fish	...	Ṇgâ (sharp)	...	.....
Flame	...	Vâṇ.		
Flea	...	Kar′khî.		
Flint striker	...	Varr pai.		
Fly, v.	...	Bâ.		
Fog	...	Pharm.		
Foot	...	Yâ′phâ	...	Yî′phî.

Polysemy, colexification/dislexification

Features such as **polysemy** or **colexification** can only be discussed when talking about languages in comparison.

They can be useful windows into how cultures interpret and divide the lived experience.

Why might ******kun*** mean ‘twenty’ in many other Tibeto-Burman languages, but ‘all’ in Burmese?

Let’s look at the data

First things first

We're going to work with some parallel data in the form of Swadesh lists. This will take you to a Google spreadsheet containing a number of lists. Go here and copy one of the tabs (just one) to a spreadsheet of your own.

We'll use it later, but grab it now so we don't get too sidetracked later.

https://tinyurl.com/ssol-swadesh

Swadesh lists

If you've had a chance to look at the lists since yesterday, have you noticed anything interesting about them?

Swadesh list - Google Sheets

https://docs.google.com/spreadsheets/d/1hjkdo6IzojbYi6-ijdZ8QkyfM5rumJmN3-MEGNZLIKc/edit?gid=558029158#gid

Swadesh list

File Edit View Insert Format Data Tools Extensions Help

Share

100% View only

A1 Nº

	A	B	J	K	L	M	N	O	P	Q	R	S	T	U
1	Nº	English	Saraiki سرائیکی (sráĩki) edit (204)	Sindhi سنڌي edit (207)	Gujarati ગુજરાતી (guǰrātī) edit (207)	Marathi मराठी (marāṭhī) edit (207)	Konkani कोंकणी (konkṇī) edit (207)	Assamese অসমীয়া (axomia) edit (207)	Bengali বাংলা (baṅla) edit (207)	Odia ଓଡ଼ିଆ (oriā) edit (207)	Kashmiri کأشُر (kāśur) edit (193)	Sinhalese සිංහල (simhala) edit (207)	Dhivehi ދިވެހि (dhvehi) edit (206)	Romani romani chib edit (206)
49	48	louse	jū	jūa, jū	jū	ū	ův, ův	ukoni	ukun	ukuṇi		ukuṇā	ukunu, ukuṇu	ʒuv
50	49	snake	nāṅg	nāṅgu	sāp	sāp	sorop, jivāne	xap	śap	sāpa	saruph	ukuṇā	naṅnugati, haruf	sap
51	50	worm	kīṛā	kīro	kīḍo	alī	kīḍo	pelu	krimi	krumi, poka, kita	kyom	krimiyāwa, paṇu	fani, fani	kirmo
52	51	tree	draxt, vaṇ	vaṇu	jhāḍ, vṛkṣ	jhāḍ	jhāḍ, rūk	gos	gach	gacha, bruksha	kul	gasa	gas	rukh
53	52	forest	jaṅgal	jhaṅgal	jaṅgal, van	jaṅgal	rān	habi, zoṅghol, bi	bon	jaṅgala, baṇa	van	wana	valu	veś
54	53	stick	daṇḍā, soṭā	laṭhi	lākḍī	kāṭhī	dāṇḍo	bari, kathi	laṭhi	laṭhi, bāḍi	lōr	kōṭuwa	daʰḍi	kopal
55	54	fruit		mevā	phal	phaḷ	phaḷ	phol	phol	phaḷa	phal	palaturu	mēvāeḷuṅ	frukto
56	55	seed	bīj	bīju	bīj	bī	bī	guti	bici, bij	mañji		bīja	oṣ	sumburo
57	56	leaf	patr	panu	pāṇḍḍū	pān	pān	pat	pata	patra	پتہ, پان	kolē	faḷ	patrin
58	57	root	pār, mūdh	pāṛa	mūl	mūl	mūl	xipa	śikor	cera, muḷa		mula	mū	rikita
59	58	bark (of a tree)	chill, khal	choḍo	chāl	(jhāḍācī) sāl	(jhāḍāce) sāl	bakoli	chal	bakaḷa, cheli		potu	tosi	korca
60	59	flower	phul	gulu	phūl	phūl	phūl	phul	phul	phula		mal	mā	luluḍī
61	60	grass	ghāh	gāhu	ghās	gavat	taṇ	ghāh	ghaś	ghāsa		taṇakoḷa	vina	char
62	61	rope	rassā, rassī	raso	dordū	dorī, rassī	dorī	rosi	dori	rasi, daudi		kaʰbaya	rōnu, rōṇu	śelo
63	62	skin	khal, camṛī	khala	cāmḍī	tvacā	kāt	sal, samra	camṛa	camaṛā, chāla		ham	haṅ	morthl
64	63	meat	gośt	gośt	mās	mās	mās	moṅoh, maṅxo	maṅśo, gōśot	mārnsa		mas	mas	mas
65	64	blood	ratt, xūn, lahū	ratu	lohī	rakta	ragat	tez	rokto	rakta, lahu		lē	lē	rat
66	65	bone	haḍḍī	haḍo	hāḍkū	hāḍ	hāḍ	har	har	hāḍa, asthi		kaṭu	kasi	kōkalo
67	66	fat (noun)	mīṇh, carbī	carbī	carbī	carbī	vos	sorri, tel	corbi	moṭā, carbi		mēda	sarubī	thulo
68	67	egg	aṇḍā, ānā	āno	ṭūḍū	aṇḍe	tāṭī, aṇḍē	koni	dim	aṇḍā		bittaraya, aṇḍayi	bis	anro
69	68	horn	siṅgh	sīhu	śīgḍū	śiṅga	śiṅga	xiṅ	śiṅ	śiṅga		aʰga	daḷu	śing
70	69	tail	puch, puchar	puchu	pūchḍī	śepūṭ, śeptī	śempḍī	nez	lej	lāṅja		waligaya	nagū	pori
71	70	feather	khambh, par	khambhu	pīṅchū	pīs	pāk	pakhi	por	para		pihāṭu	dūnifaḷ	por
72	71	hair	vāl	vāra	vāḷ	kes	kes	suli	cul	bāḷa, keśa, loma		kes, koṇḍaya	istaṣi	bal
73	72	head	sir	matho	māthū	ḍoke	takḷī, mātte	mur, matha	matha	muṇḍa		oluwa, isa	bō	śero
74	73	ear	kann	kana	kān	kān	kān	kan	kan	kāna		kaṇa	kaṅfaḷ	kan
75	74	eye	akkh	akhi	ākha	ḍolā	ḍoḷo	soku	cōkh	ākhi		æsa	lō	jakh
76	75	nose	nakk	naku	nāk	nāk	nāṅka	nak	nak	nāka		nahaya	nēfaḷ	nakh
77	76	mouth	mūh	vātu	mō, mukh	toṇḍa	toṇḍa	mukh	mukh	pātī, mūha		kaṭa	aʰga	muj
78	77	tooth	dand	ḍanda	dāt	dāt	dānta	dāt	dāt	dānta		data	daī	dand

Germanic | Indo-Aryan | Bantu | Formosan | Romance | Slavic | Kra-Dai

Swadesh lists

If you've had a chance to look at the lists since yesterday, have you noticed anything interesting about them?

- gaps in coverage (missing words)
- multiple words given per concept
  - what's the difference?
  - which is more basic?

With the Swadesh data from before, pick a few languages and concepts, and try to flatten them.

Use tabs in a spreadsheet for the three tables for now.

Goals

1. have well structured data
2. bring it into EDICTOR
3. address cognate sets & alignment

We're going to go through this twice. Once with a sample German data set, and then again with data of your choosing.

EDICTOR

1. Open EDICTOR: https://edictor.org/
2. load German sample data

EDICTOR

1. Open EDICTOR: https://edictor.org/
2. load German sample data
3. add COGID, ALIGNMENT columns

ID	DOCULECT	CONCEPT	FORM
1	German	all	al
2	English	all	ɔːl
3	Danish	all	æˀl
4	Swedish	all	alː
5	Icelandic	all	at͡lɪr
6	Dutch	all	ɑlə
7	Norwegian	all	ɑlə
8	German	ashes	aʃə
9	English	ashes	æʃ
10	Danish	ashes	asg

NeighbourNets

With EDICTOR we can also export neighbour-joining split networks (NeighbourNets, seen right). These can be useful to visualise the similarity of data across your word list.

Note, however, that this is simply showing a distance based visualisation of data similarity, and may not actually tell you anything about linguistic genealogy. Use with caution.