Linguistic Typology for NLP researchers: Methods and Resources in the 21st century

Harald Hammarström
harald.hammarstrom@lingfil.uu.se

22 Jan 2026 Yerevan

Today: Hands-on Working with Language Diversity Resources

§1: Language Identifiers

- Most databases are indexed by ISO 639-3 three letter **language** identifiers popularized by Ethnologue, e.g., `eng`, `aln`, …
- Glottolog has 8-char alphanumeric glottocodes, e.g., `stan1293`, `gheg1238` which index **families**, **languages**, **dialects** alike
  - Glottocodes thus have three different **level**s
  - The level-agnostic term in Glottolog is **languoid**
  - The ISO 639-3 codes and language-level glottocodes are nearly 1-1 and Glottolog serves both
- (Both id:s have some mnemonic value but are in principle arbitrarily coined)
- Both the ISO 639-3 and Glottocode inventories come in yearly updates — you can argue with both entities
  - Updates to ISO 639-3 are handled by SIL `https://iso639-3.sil.org/`
  - Updates to Glottocodes are handled by the Glottolog editors

§1: Identifiers Examples

- Gheg Albanian [aln]:
  `https://www.ethnologue.com/language/aln`
- Gheg Albanian [gheg1238]:
  `https://glottolog.org/resource/languoid/id/gheg1238` see ISO 639-3 equivalent in the top-right corner
- But no ISO 639-3 equivalent for, e.g.,
  - Family:
    `https://glottolog.org/resource/languoid/id/indo1319`
  - Dialect:
    `https://glottolog.org/resource/languoid/id/erev1240`

§1: Pyglottolog

- Pyglottolog is a Python-interface to Glottolog
  `https://github.com/glottolog/pyglottolog`
- Install `pyglottolog`
  - `pip install pyglottolog`
  - download a copy of the glottolog data from `https://glottolog.org/meta/downloads` (and place it e.g., in a directory called `glottolog/`)
- Now the following should give you an instance of Glottolog

```
1 from pyglottolog import Glottolog
2 glottolog = Glottolog('glottolog') # 'glottolog' is the name of
      the directory where you placed your downloaded copy
```

§1: Identifiers in Pyglottolog

- The language Gheg Albanian has ISO 639-3 `aln` and Glottocode `gheg1238`
- Use the identifier to grab a specific languoid

```
x = glottolog.languoid('gheg1238') # x = glottolog.languoid('aln') would also work but is slower
```

- This object has attributes for various items of metadata information, e.g.

```
print(x.name, x.iso, x.glottocode)
```

§1: Identifiers in Pyglottolog

- The language Gheg Albanian has ISO 639-3 `aln` and Glottocode `gheg1238`
- Use the identifier to grab a specific languoid

```
x = glottolog.languoid('gheg1238') # x = glottolog.languoid('aln') would also work but is slower
```

- This object has attributes for various items of metadata information, e.g.

```
print(x.name, x.iso, x.glottocode)
```

- Yields

```
Gheg Albanian aln gheg1238
```

§1: Glottolog Identifiers and Levels

- The `level` attribute holds the level of a languoid

```
for glottocode in ['gheg1238', 'indo1319', 'erev1240']:
    x = glottolog.languoid(glottocode)
    print(x.name, x.iso, x.level)
```

- Yields

```
Gheg Albanian aln LanguoidLevel(ordinal=2, id='language',
    description='defined by mutual non-intellegibility')
Indo-European None LanguoidLevel(ordinal=1, id='family',
    description='sub-grouping of languoids above the language
    level')
Erevan None LanguoidLevel(ordinal=3, id='dialect', description=
    'any variety which is not a language')
```

§1: Glottolog non-ISO 639-3 languages

- In a small amount of cases, a language-level glottocode has no ISO 639-3 equivalent
- This means that there "should be" an ISO 639-3 code but there isn't, and `ethnologue_comment` has an explanation why, e.g.

```
x = glottolog.languoid('kais1242')
print(x.name, x.level.id, x.iso)
print(x.ethnologue_comment)
```

- Yields

```
Kaishana language None
EthnologueComment(isohid='NOCODE_Kaishana', comment_type='
    missing', ethnologue_versions=['E16', 'E17', 'E18', 'E19',
    'E20', 'E21', 'E22', 'E23', 'E24', 'E25', 'E26', 'E27', '
    E28'], comment='Kaishana, a presumed extinct Northern
    Arawakan (**hh:hvw:Ramirez:Arawak**) language, is missing
    from E16/E17/E18/E19/E20/E21/E22/E23/E24/E25/E26/E27/E28
    (**hh:hw:Martius:Brasiliens**, **hh:hw:Nimuendaju:
    MakusiWapicanaIpurinaKapisana**, **hh:w:Hanke:Kaisana**).')
```

§1: Glottolog Traversing

- The **languoids()** method gives an iterator for all languoids (families, languages and dialects alike)

```
>>> languoids = [l for l in glottolog.languoids()]
>>> len(languoids)
27111
```

- If you want only languages, filter on the language level

```
>>> languages = [l for l in glottolog.languoids() if l.
    level.id == 'language']
>>> len(languages)
8618
```

- (Re efficiency/performance see details at `https://pyglottolog.readthedocs.io/en/latest/api.html#performance-considerations`)

§2: Basic Metadata: Canonical name

- Every language has one canonical name

```
>>> glottolog.languoid('gheg1238').name
'Gheg Albanian'
```

- This name is (forced to be) unique in the Glottolog canonical namespace, so there are disambiguators whenever two different languages happen to have the same name in nature

Canonical name	Glottocode	ISO 639-3
Madi	jama1261	jaa
Madi (Papua New Guinea)	madi1261	grg

- NB: The canonical name of a language may change between different ISO 639-3/Glottolog editions but the **id**, i.e. glottocode/ISO 639-3 code does not!

§2: Basic Metadata: Further names

- Languages have a multitude of names
  - Different endonyms, e.g., Spanish/Castilian
  - Different exonyms, e.g., saksalainen, allemand, nemetskij, …
  - Different renderings in different languages, e.g., françois, French, faransawiy, …
- Extensive, but not complete, lists of (non-unique) alternative names aggregated from different sources are available

```
>>> glottolog.languoid('gheg1238').names
{'multitree': ['Albanesisch', 'Albanian', 'Albanian, Gheg', '
    Arber', 'Arbresh', 'Arnaut', 'Geg', 'Gheg', 'Gheg Albanian'
    , 'Guegue', 'Shgip', 'ēShqipri', 'Shquipni', 'Škip'], '
    lexvo': ['Albanés guego [es]', 'Dialectul Gheg [ro]', '
    Gegijski jezik [hr]', 'Gegisch [de]', 'Gegiska [sv]', 'Gegë
     [sq]', 'Gheg Albanian [en]', 'Guègue [fr]', 'Γ          [el
    ]', '   [bg]', '                  [ru]', '   [ja]'], '
    hhbib_lgcode': ['Albanian of Zadrima', 'Albanian-Gheg', '
    Gegskaja', 'Gheg', 'Malsia Madhe', 'NW Gheg', 'Northeast
    Geg Albanian from southern Kosovo', 'Südgegischen']}
```

§2: Basic Metadata: Countries

- List of countries where a language is spoken available in Glottolog

```
>>> glottolog.languoid('gheg1238').countries
[Country(id='AL', name='Albania'), Country(id='BG', name='
    Bulgaria'), Country(id='ME', name='Montenegro'),
    Country(id='MK', name='North Macedonia'), Country(id='
    RO', name='Romania'), Country(id='RS', name='Serbia')]
```

- To be "spoken in a country" isn't really well-defined (how many speakers? how long? still spoken there?) — some kind of common-sense definition in practice
- In Ethnologue there is a "primary country" for each language
- In Glottolog, there is a centre-point coordinate for each language

```
>>> (glottolog.languoid('gheg1238').latitude, glottolog.
    languoid('gheg1238').longitude)
(42.317, 21.3837)
```

- Point can be the geographic, politial, historical, … centre
- The country the coordinate is in can be considered the primary country according to Glottolog

§2: Basic Metadata: Endangerment

- Ethnologue curates its own (unsourced) endangerment data
- Glottolog reports an Agglomerated Endangerment Scale (AES) combining ElCat, UNESCO and Ethnologue (details see Hammarström et al. 2018) for each language (not dialects)

AES status	# of languages	% of languages
not endangered	2643	34.13%
threatened	1595	20.59%
shifting	1805	23.31%
moribund	422	5.45%
nearly extinct	299	3.86%
extinct	981	12.67%
total:	7745	

- Imperfect quality: There are many questionable endangerment assessments …

§2: Basic Metadata: Endangerment Examples

```
>>> glottolog.languoid('gheg1238').endangerment
Endangerment(status=AES(ordinal=1, id='safe', name='not
    endangered', egids='<=6a', unesco='safe', elcat='at risk/
    safe', reference_id='hh:hel:Hammarstrom:Visualization'),
    source=AESSource(id='E28', name='Ethnologue 28', url='https
    ://www.ethnologue.com/', reference_id='hh:h:Ethnologue:28')
    , comment='Albanian, Gheg (aln-aln) = 4 (Educational).')
```

```
>>> glottolog.languoid('madi1261').endangerment
Endangerment(status=AES(ordinal=3, id='definite', name='
    shifting', egids='7', unesco='definitely endangered', elcat
    ='threatened/endangered', reference_id='hh:hel:Hammarstrom:
    Visualization'), source=AESSource(id='ElCat', name='The
    Catalogue of Endangered Languages (ELCat)', url='http://
    endangeredlanguages.com', reference_id='hh:hel:Campbell:
    ElCat'), comment='Madi (4241-grg) = Endangered (20 percent
    certain, based on the evidence available) [Lewis 2009](cldf
    :lewis:ed:09)')
```

§2: Basic Metadata: Timespan for extinct languages

- All languages marked as extinct (in endangerment) have an attribute **timespan** which is year of extinction or range of attestation

```
>>>glottolog.languoid('mbab1239').timespan #Mbabaram went extinct in 1985
(1985, 1985)
>>>glottolog.languoid('clas1249').timespan #Classical Armenian is attested between 400 AD - 1100 AD
(400, 1100)
>>>glottolog.languoid('mero1237').timespan #Meroitic is attested from 250 BC - 300 AD
(-250, 300)
```

§2: Basic Metadata: Speaker Numbers

- Glottolog does not have speaker numbers
- Ethnologue has speaker numbers, but
  - Speaker numbers difficult to obtain (for anyone) and there are many old/erroneous/questionable numbers
  - Ethnologue often, but not always, gives a source with a year (not necessarily year of observation)
  - Year of edition vs year of source lags behind

Ed.	Year	Avg. year of pop. source	Δ	# missing year	# lgs
13	1996	1987.31	-8.69	1279	6988
14	2000	1988.08	-11.92	1007	7148
15	2005	1993.14	-11.86	695	7299
16	2009	1995.85	-13.15	769	7357
17	2013	1998.91	-14.09	836	7561
18	2015	1999.71	-15.29	746	7532
…	…	…	…	…	…
27	2024	2006.83	-17.17	489	7679
28	2025	2007.90	-17.10	494	7681

  - Still, far better than nothing

§2: Basic Metadata: Speaker Numbers from Ethnologue

- Ethnologue is behind a paywall, so a static copy `e28.txt`, `e18.tab` is handed out for this course
- Ethnologue gives population information as a string

Name:	Kafa
ISO 639-3:	kbr
Population:	1,406,700, all users. L1 users: 1,360,000 (2022). L2 users: 46,700. 445,000 monolinguals (1994 census). Ethnic population: 1,480,000 (2022).

- Total number of L1 speakers is extracted (by HH) into a field `Population Numeric`

Population Numeric: 1360000

§2: Example Ethnologue Speaker Numbers

```
>>> import pandas
>>> df = pandas.read_csv('e28.tab', sep='\t', keep_default_na=
    False, index_col = "ISO 639-3") #Index by ISO 639-3, not
    glottocodes
>>> df.loc["kbr"]
Alternate Names      Caafiti, Caffino, Kaffa, Kaffinya, Kaf...
Classification       Afro-Asiatic, Omotic, North, Gonga-Gim...
Country                                                Ethiopia
Dialects             Kafa, Bosha (Garo). Bosha may be a dis...
...
Location             Southern Nations, Nationalities, and P...
Other Comments       Traditional religion, Christian, Muslim.
Population           1,406,700, all users. L1 users: 1,360,...
Population Numeric                                      1360000
Population Numeric (L2)                                   46700
Typology                                                    SOV.
Writing              Ethiopic script [Ethi], used in Church...
code+name                                          Kafa [kbr]
name                                                     Kafa
>>> int(df.loc["kbr"]["Population Numeric"])
1360000
```

§3 Classification

- Language classification is disputed and prone to update
- Ethnologue has a classification but classification is not a primary focus a so inconsistencies, unresolved subclassifications, no sources, …
- Glottolog has a principled classification and subclassification, see `https://glottolog.org/glottolog/glottologinformation`
  - Reference given for each decision for justification, further information, …

§3 Classification in Glottolog

- Use the attributes **parent**, **children**, **ancestors**

```
>>> x = glottolog.languoid('hye') # Eastern Armenian
>>> x.parent #Gives the immediate parent
<Family east2768>
>>> x.children #Gives a list of the immediate children
[<Dialect agul1245>, <Dialect arar1266>, <Dialect kara1458>, <Dialect khoi1249>]
>>> x.ancestors #Gives an ordered list from the root up to (but not including) the current node
[<Family indo1319>, <Family clas1257>, <Family arme1241>, <Family east2768>]
```

§3 Glottolog Families and Pseudofamilies

- Glottolog has 421 “normal” language families/isolates
- Then there are 8 pseudo-families, which are really categories of alleged languages which cannot be classified

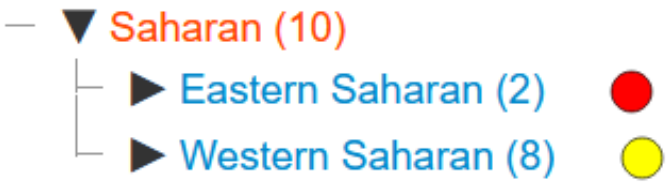
Sign Language	*Unclassifiable*
Pidgin	*Unattested*
Artificial Language	*Speech Register*
Mixed Language	*Bookkeeping*

- For many purposes you want to exclude the pseudo-families and their “languages”

```
>>> languages = [l for l in glottolog.languoids() if l.level.id == 'language']
>>> len(languages)
8618
>>> canonical_languages = [l for l in languages if (l.ancestors + [l])[0].name not in ['Sign Language', 'Unclassifiable', 'Pidgin', 'Unattested', 'Artificial Language', 'Speech Register', 'Mixed Language', 'Bookkeeping']]
>>> len(canonical_languages)
7675
```

§3 Desiderata on Families **Not** Available

- Can you get a number (like 63% etc) on the **certainty** for a given family/subfamily?
  - Difficult to achieve in a consistent and scientific way
- Can you get the **time-depth** of a given family/subfamily?
  - Datings based on archaeology, written history, expert intuition only rarely available
  - Some kind of lexical distance could be computed — this has not been incorporated in Glottolog (or Ethnologue)

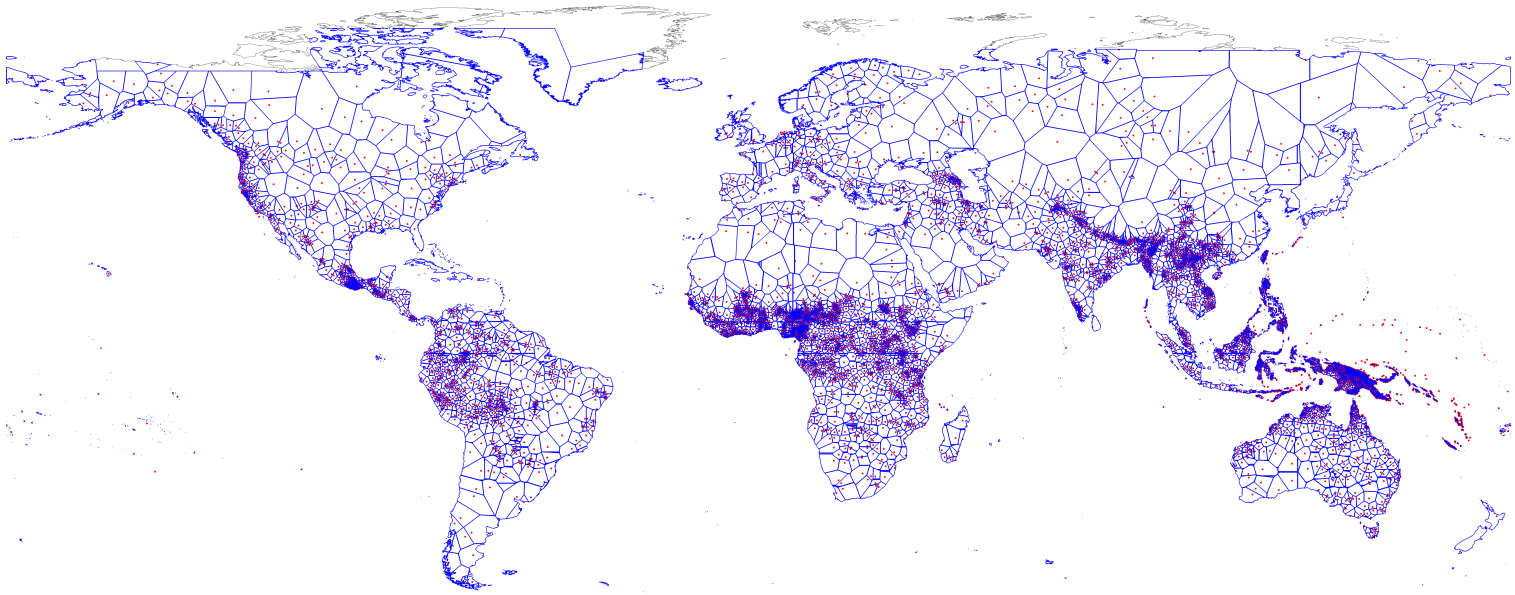


Experts believe the relationship between Western and Eastern Saharan is very deep ≈ 6000 years

Experts believe the relationship between Sakhalin and Hokkaido-Kuril Ainu is very shallow ≈ 1000 years

§4 Maps

- Unfortunately no open database with empirical Polygon data for geospatial language distributions (but coming soon, `https://github.com/glottography/asher2007world`)
- But Glottolog has centre-point coordinates
- Voronoi regions from centre-point coordinates offer some approximation of the full spatial distribution



§4 Example Dot Map

- Let's make a map with a dot for all the languages whose name begins with A

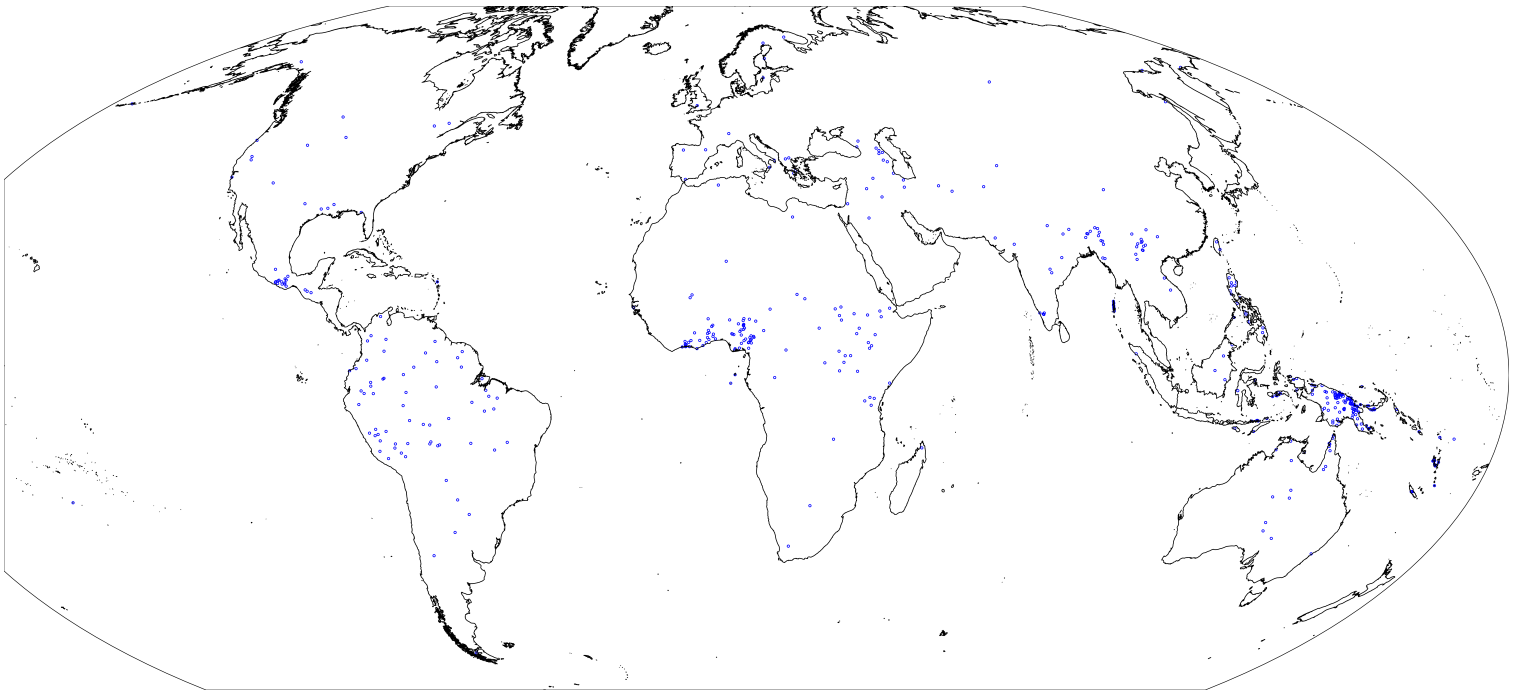
```
#Select the languages
languages = [l for l in glottolog.languoids() if l.level.id == 'language']
canonical_languages = [l for l in languages if (l.ancestors + [l])[0].name not in ['Sign Language', 'Unclassifiable', 'Pidgin', 'Unattested', 'Artificial Language', 'Speech Register', 'Mixed Language', 'Bookkeeping']]
languages_A = [l for l in canonical_languages if l.name.startswith("A")]

#Create the base map
import cartopy.crs as ccrs
import matplotlib.pyplot as plt
plt.figure(figsize=(36,18))
ax = plt.axes(projection=ccrs.Mollweide()) #Mollweide
#ax = plt.axes(projection=ccrs.PlateCarree()) #Mercator
ax.coastlines(resolution="10m") #Use coastlines
```

§4 Example Dot Map Result

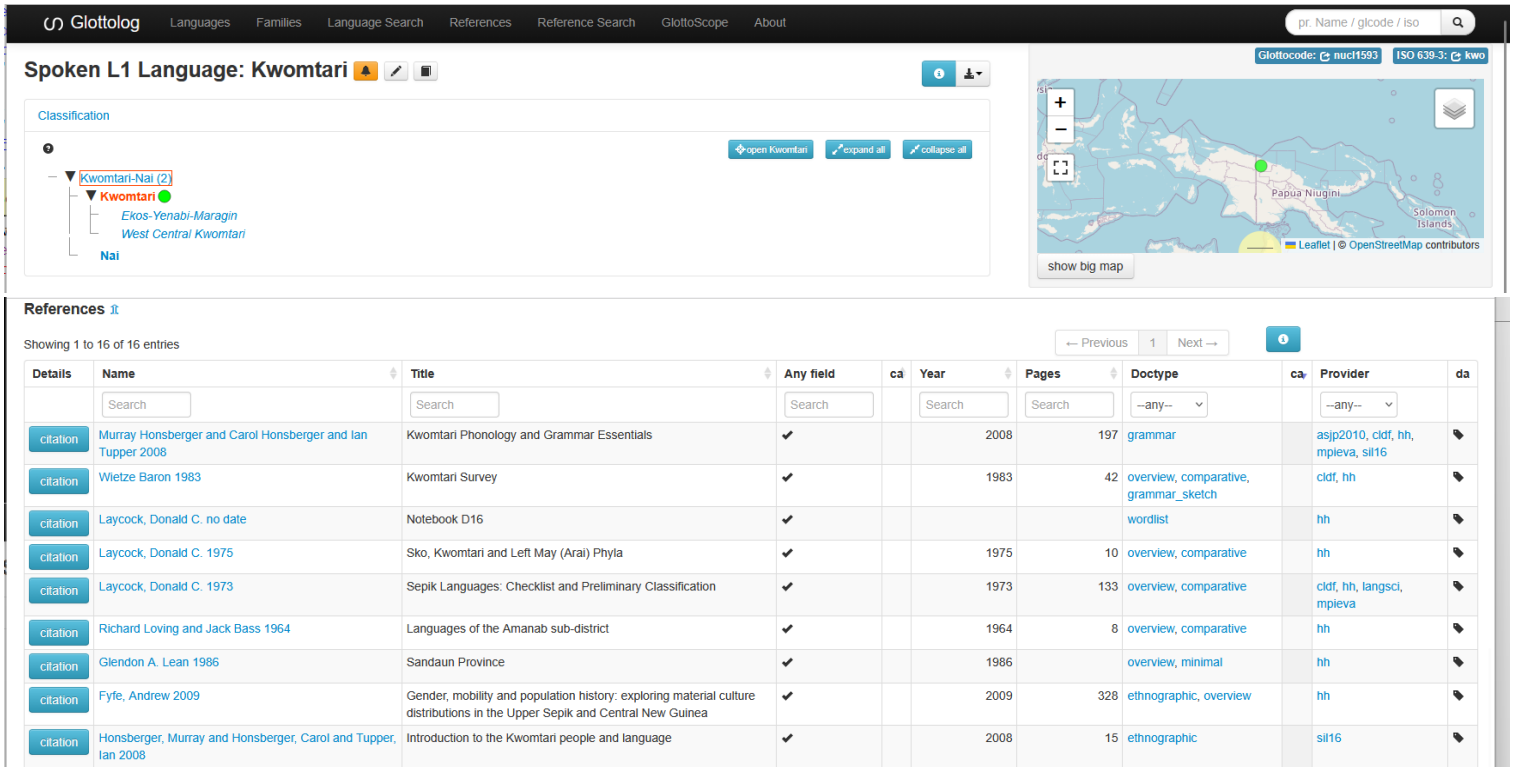
```
#Plot the dots
plt.scatter([l.longitude for l in languages_A], [l.latitude for
     l in languages_A], transform=ccrs.PlateCarree(),
    facecolors='none', edgecolors = "blue", s = 10) #s is
    thickness

#Save
plt.savefig('yerevan2026map_test.png', bbox_inches='tight',
    pad_inches=0, dpi = 600)
```



§5 Glottolog Bibliography

- Glottolog has a very big biliography $>$ 400,000 references
- Glottolog includes the most extensive grammatical description for **every** language



Details	Name	Title	Any field	ca	Year	Pages	Doctype	ca	Provider	da
	Search	Search	Search		Search	Search	--any--		--any--	
citation	Murray Honsberger and Carol Honsberger and Ian Tupper 2008	Kwomtari Phonology and Grammar Essentials	✔		2008	197	grammar		asjp2010, cldf, hh, mpieva, sil16	
citation	Wietze Baron 1983	Kwomtari Survey	✔		1983	42	overview, comparative, grammar_sketch		cldf, hh	
citation	Laycock, Donald C. no date	Notebook D16	✔				wordlist		hh	
citation	Laycock, Donald C. 1975	Sko, Kwomtari and Left May (Arai) Phyla	✔		1975	10	overview, comparative		hh	
citation	Laycock, Donald C. 1973	Sepik Languages: Checklist and Preliminary Classification	✔		1973	133	overview, comparative		cldf, hh, langsci, mpieva	
citation	Richard Loving and Jack Bass 1964	Languages of the Amanab sub-district	✔		1964	8	overview, comparative		hh	
citation	Glendon A. Lean 1986	Sandaun Province	✔		1986		overview, minimal		hh	
citation	Fyfe, Andrew 2009	Gender, mobility and population history: exploring material culture distributions in the Upper Sepik and Central New Guinea	✔		2009	328	ethnographic, overview		hh	
citation	Honsberger, Murray and Honsberger, Carol and Tupper, Ian 2008	Introduction to the Kwomtari people and language	✔		2008	15	ethnographic		sil16	

- Can be accessed and used for various statistics (relevant to CL?)

§5 Glottolog Bibliography Example

```
>>>kwomtari = glottolog.languoid('nucl1593')
>>>kwomtari.sources
[Reference(key='cldf:hammarstroem:forkel:etal:14', pages=None,
    trigger=None, endtag='**'), Reference(key='cldf:lewis:paul:
    etal:ed:15', pages=None, trigger=None, endtag='**'),
    Reference(key='cldf:spencer:08', pages=None, trigger=None,
    endtag='**'), Reference(key='hh:g:Honsbergeretal:Kwomtari',
     pages=None, trigger=None, endtag='**'), Reference(key='hh:
    he:Fyfe:Upper-Sepik', pages=None, trigger=None, endtag='**'
    ), Reference(key='hh:hld:Lean:Sandaun', pages=None, trigger
    =None, endtag='**'), Reference(key='hh:hv:BassLoving:Amanab
    ', pages=None, trigger=None, endtag='**'), Reference(key='
    hh:hv:Laycock:Sepik', pages=None, trigger=None, endtag='**'
    ), Reference(key='hh:hv:Laycock:SkoKwomtariLeftMay', pages=
    None, trigger=None, endtag='**'), Reference(key='hh:hvs:
    Baron:Kwomtari', pages=None, trigger=None, endtag='**'),
    Reference(key='hh:w:Laycock:D16', pages=None, trigger=None,
     endtag='**'), Reference(key='sil16:50993', pages=None,
    trigger=None, endtag='**'), Reference(key='sil16:50994',
    pages=None, trigger=None, endtag='**'), Reference(key='
    sil16:50995', pages=None, trigger=None, ...]
```

§6 CLDF: Cross-Linguistic Data Formats

- CLDF: Cross-Linguistic Data Format is an emerging standard for shareable cross-linguistic databases, see Forkel et al. (2018) and `https://cldf.clld.org/`
- All indexed by ISO 639-3 codes and/or Glottocodes
- In the typical case we have a (Language, Parameter, Value) triplets, i.e., a matrix with languages and parameters with a cell value

	P_1	P_2	…
L_1	$V_{l_1p_1}$	$V_{l_1p_2}$	…
L_2	$V_{l_2p_1}$	$V_{l_2p_2}$	…
…	…	…	

- I'll briefly introduce three

  Lexicon: **ASJP** 40-word basic lexicon
  Grammar: **Grambank** 195 typological features (morphosyntax)
  Phonology: **PHOIBLE** segmental inventories

§6 CLDF data: Lexicon (ASJP)

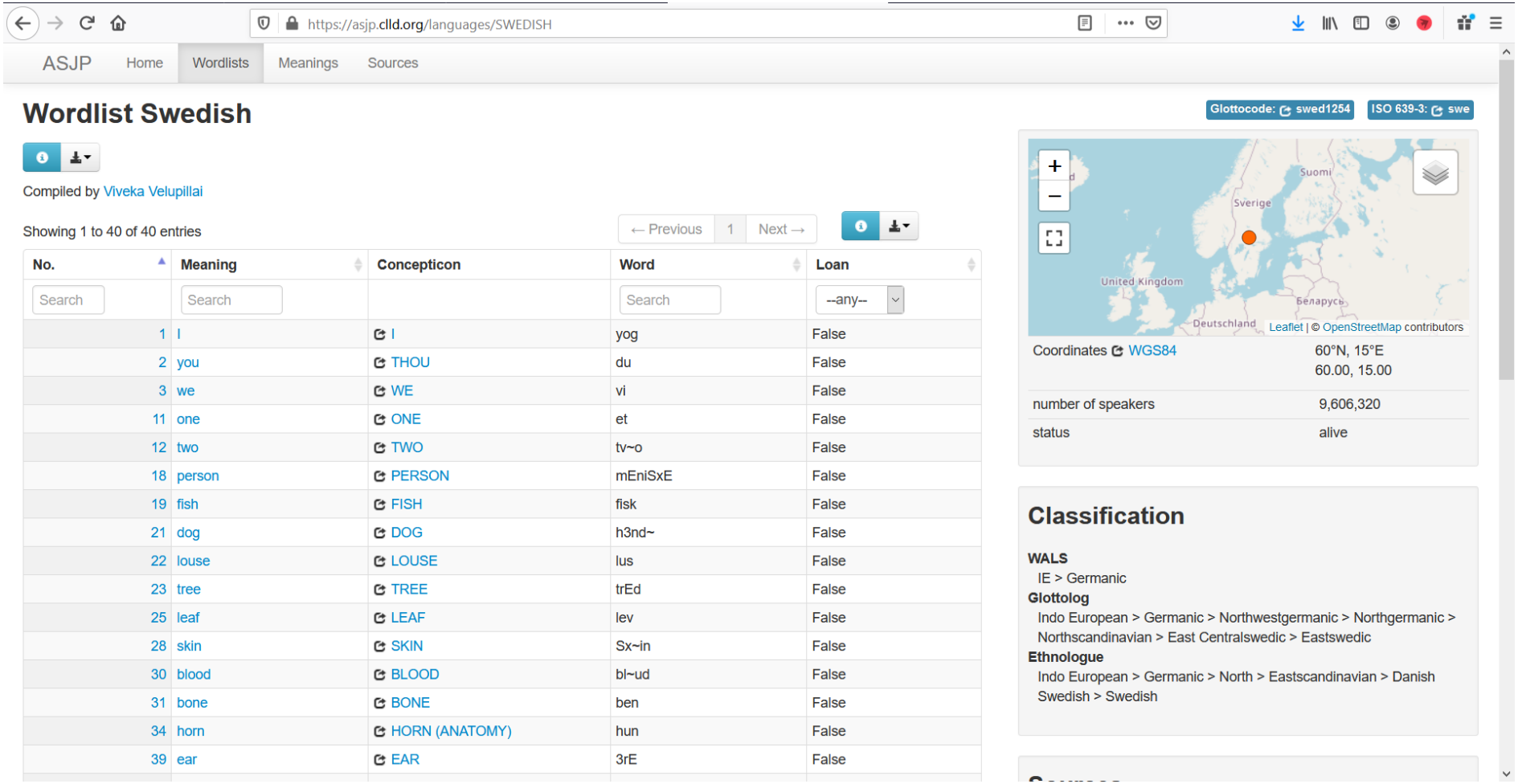
- Only 40 words, but **many** languages

  ASJP Database v 21 (2025) `https://asjp.clld.org/`

# Wordlists	11 540
# ISO 639-3 languages	6 135
# Meanings per language	40
Total # words	568 820

Transcription impoverished (compared to IPA) with only 7 vowels and 34 consonants, no tone

§6 ASJP Example (Swedish)



§6 Accessing ASJP using pycldf

- Install pycldf, download a local copy of ASJP, read with pycldf

```
import pycldf
asjp = pycldf.Dataset.from_metadata(r'dbs\asjp\asjp21_lexibank-asjp-0127953\cldf\cldf-metadata.json') #Specify where you placed your downloaded copy
```

- It has components Language, Parameter (= here meaning), Value (= here form)

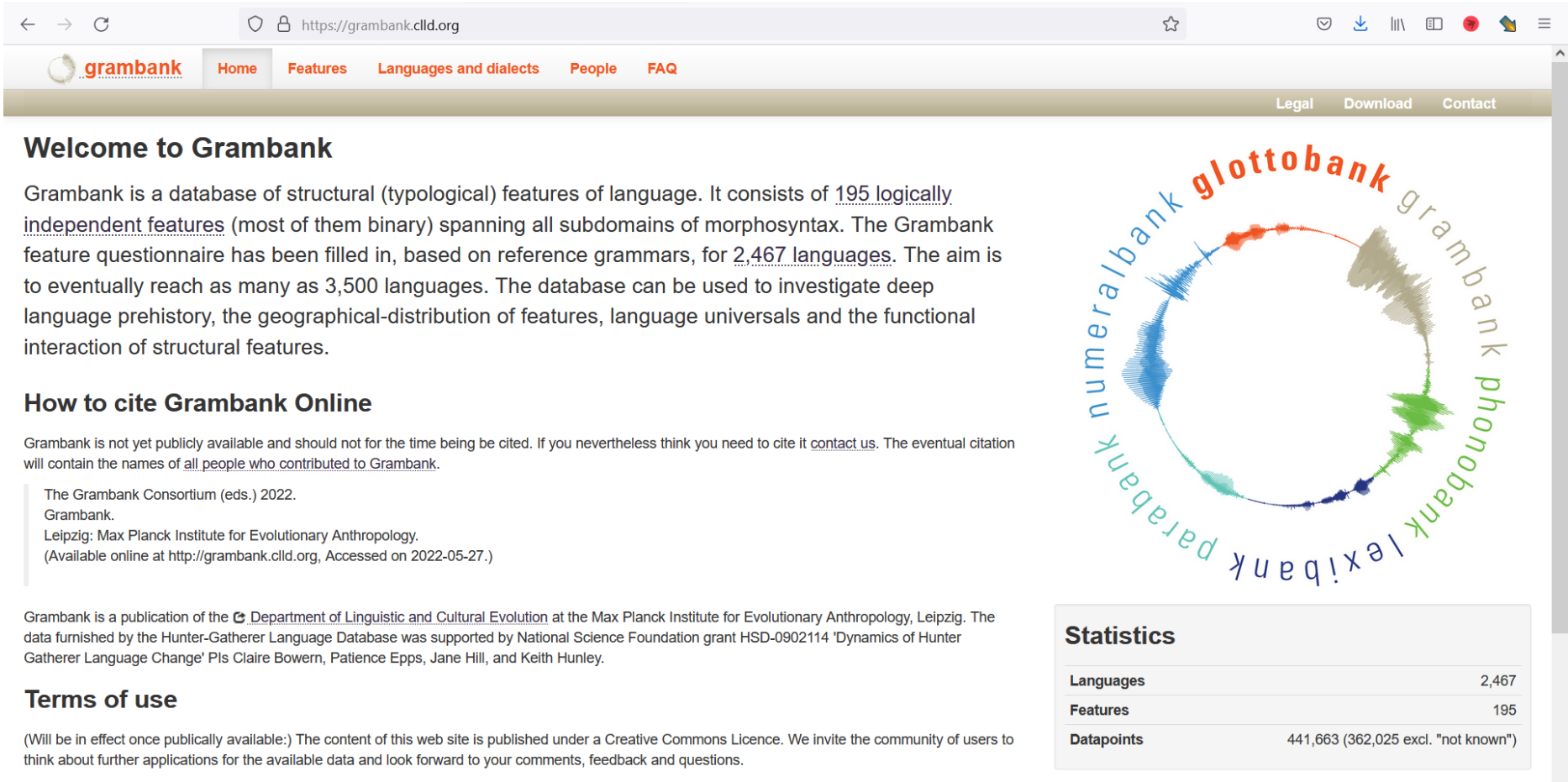
```
>>> for c in asjp.components:
>>>     print(c)
FormTable
LanguageTable
ParameterTable
```

§6 Accessing ASJP using pycldf: Example

- Get the form for 'man' in any list tagged as Eastern Armenian [hye]

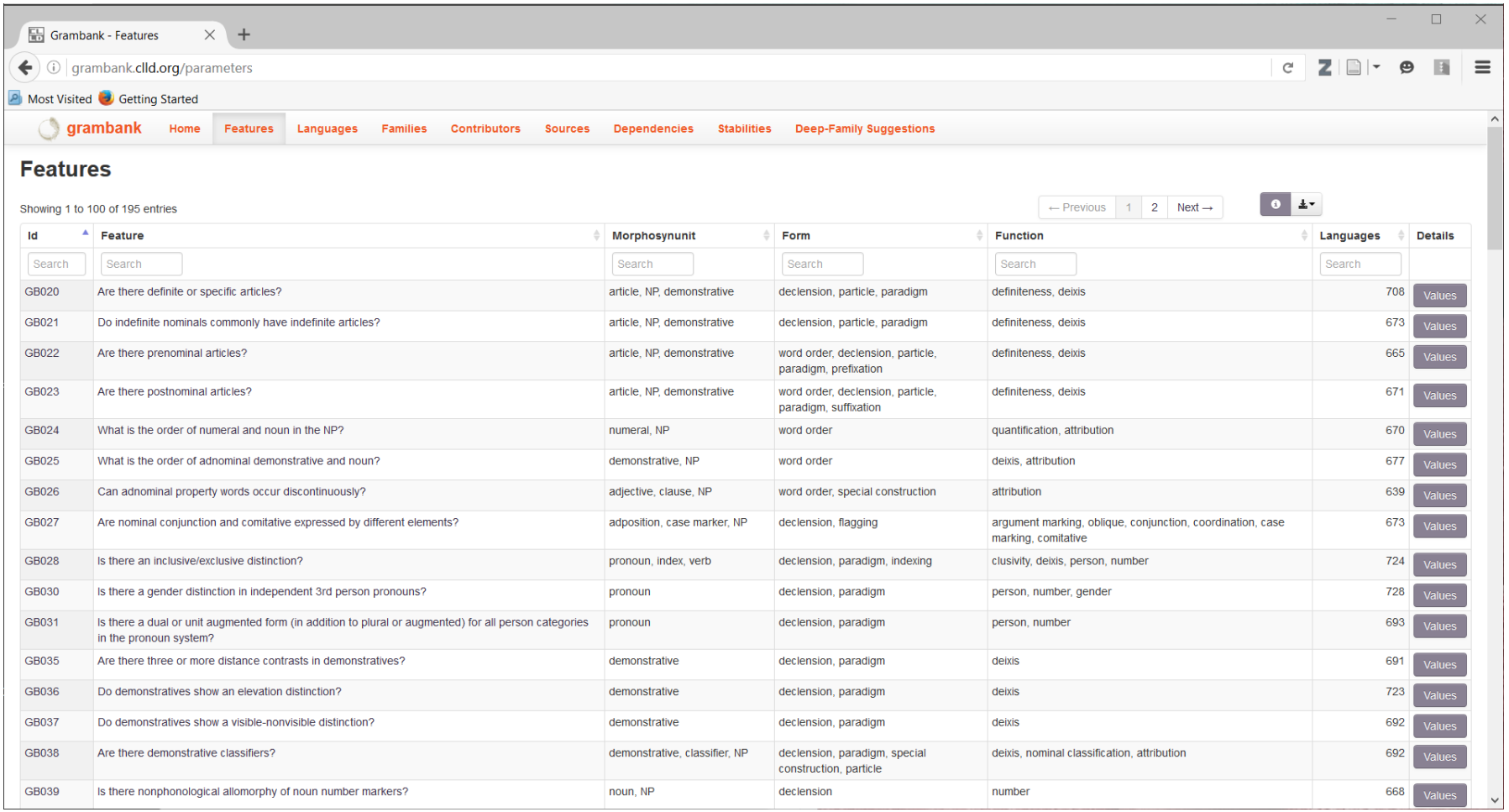
```
>>> localid_to_iso = {row['ID']: row['ISO639P3code'] for row in asjp['LanguageTable']}
>>> parameterid_to_gloss = {row['ID']: row['Name'] for row in asjp['ParameterTable']}
>>> for row in asjp['FormTable']:
>>>     if localid_to_iso[row['Language_ID']] == 'hye':
>>>         if parameterid_to_gloss[row['Parameter_ID']] == 'stone':
>>>             print(row['Form'], row['Language_ID'])
kh~or ARMENIAN
kh~ar EASTERN_ARMENIAN
```

§6 CLDF data: Grammar (Grambank)

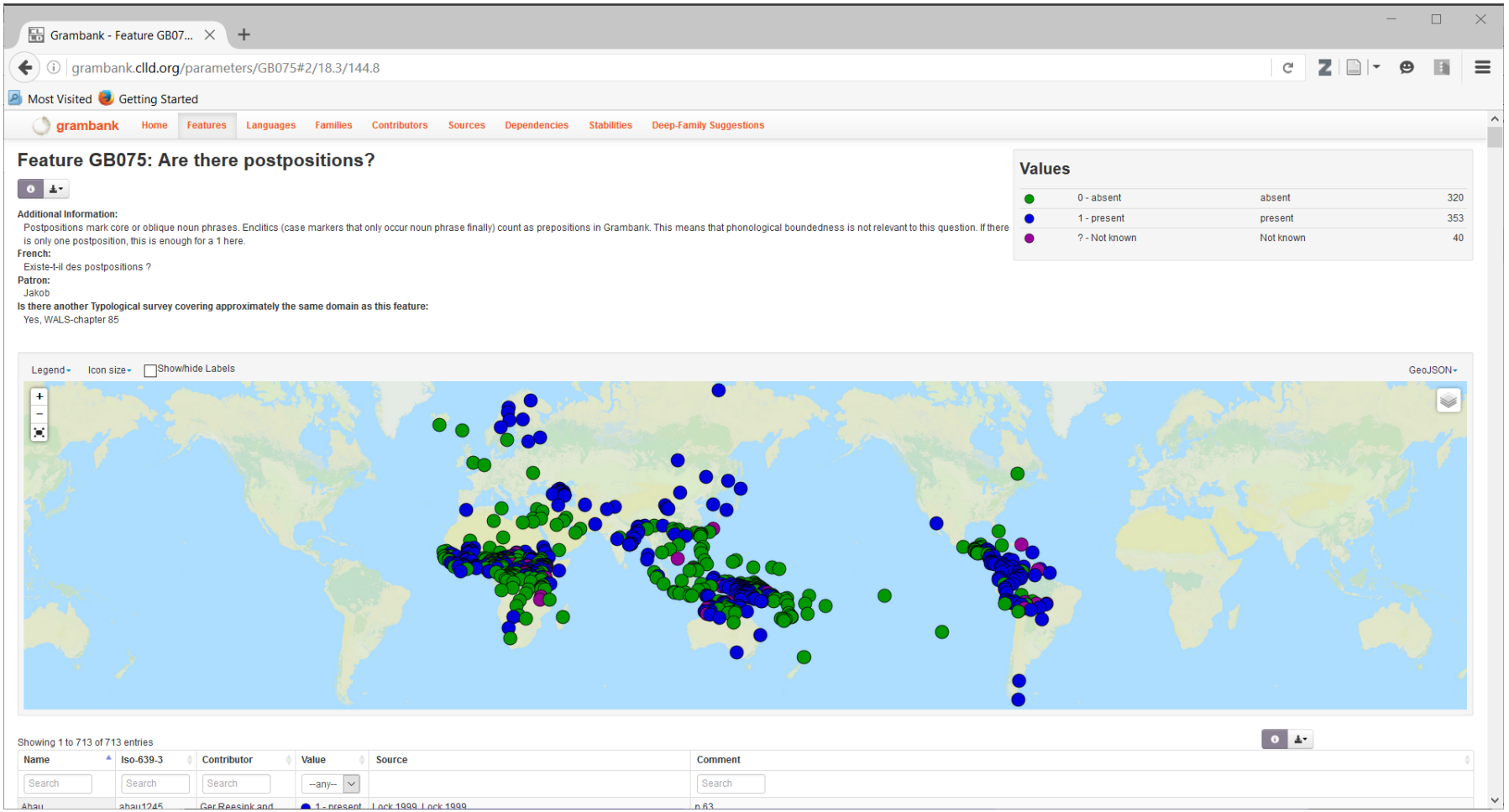


`https://grambank.clld.org/` (Skirgård et al. 2023)

§6 The Grambank Database: 195 Features

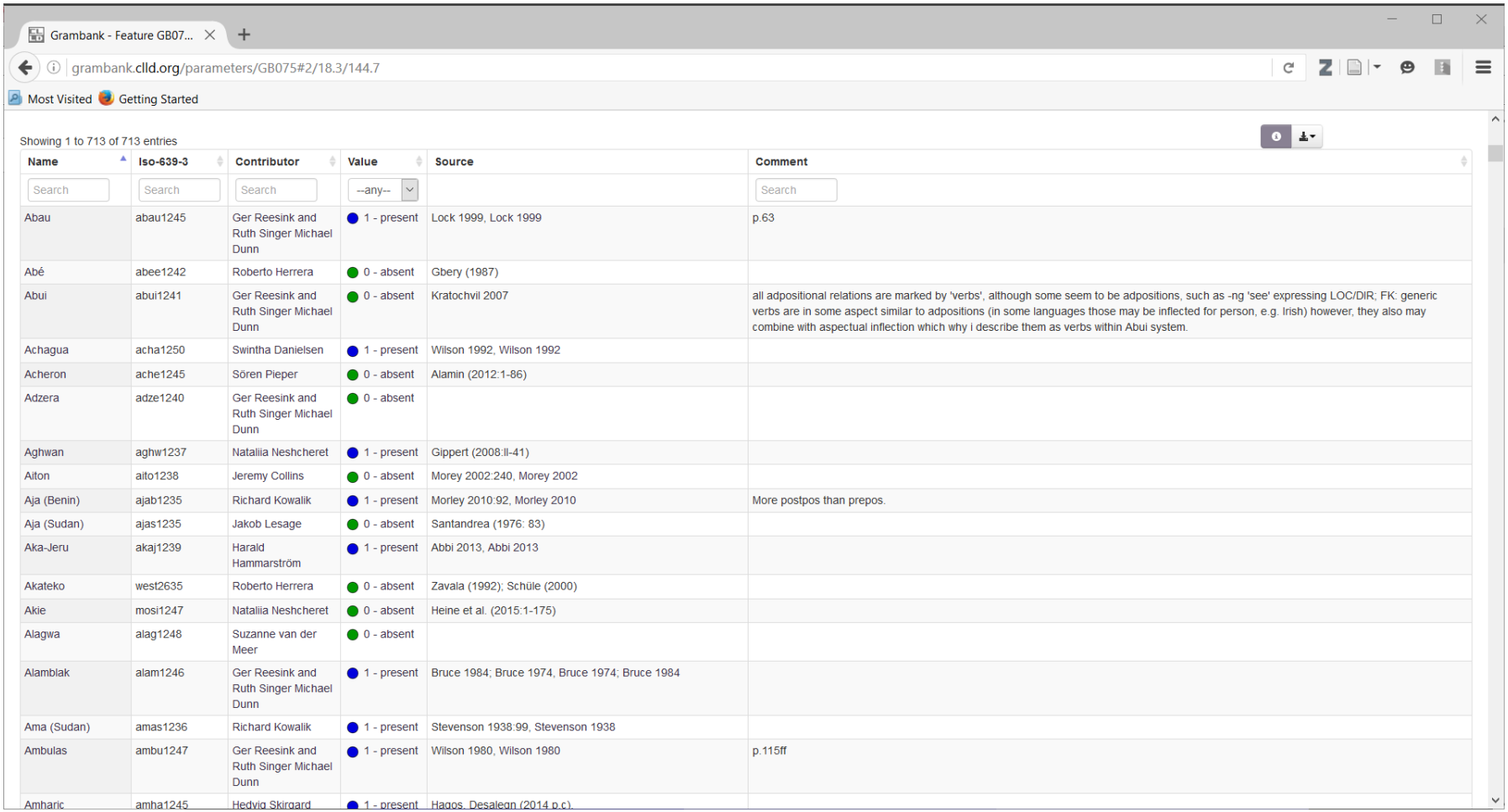


§6 Grambank Example Feature



§6 Grambank Data Provenance

- Data collected by humans from descriptive grammars



Grambank - Feature GB07...

grambank.clld.org/parameters/GB075#2/18.3/144.7

Most Visited Getting Started

Showing 1 to 713 of 713 entries

Name	Iso-639-3	Contributor	Value	Source	Comment
Search	Search	Search	--any--		Search
Abau	abau1245	Ger Reesink and Ruth Singer Michael Dunn	1 - present	Lock 1999, Lock 1999	p.63
Abé	abee1242	Roberto Herrera	0 - absent	Gbery (1987)	
Abui	abui1241	Ger Reesink and Ruth Singer Michael Dunn	0 - absent	Kratochvil 2007	all adpositional relations are marked by 'verbs', although some seem to be adpositions, such as -ng 'see' expressing LOC/DIR; FK: generic verbs are in some aspect similar to adpositions (in some languages those may be inflected for person, e.g. Irish) however, they also may combine with aspectual inflection which why i describe them as verbs within Abui system.
Achagua	acha1250	Swintha Danielsen	1 - present	Wilson 1992, Wilson 1992	
Acheron	ache1245	Sören Pieper	0 - absent	Alamin (2012:1-86)	
Adzera	adze1240	Ger Reesink and Ruth Singer Michael Dunn	0 - absent		
Aghwan	aghw1237	Nataliia Neshcheret	1 - present	Gippert (2008:II-41)	
Aiton	aito1238	Jeremy Collins	0 - absent	Morey 2002:240, Morey 2002	
Aja (Benin)	ajab1235	Richard Kowalik	1 - present	Morley 2010:92, Morley 2010	More postpos than prepos.
Aja (Sudan)	ajas1235	Jakob Lesage	0 - absent	Santandrea (1976: 83)	
Aka-Jeru	akaj1239	Harald Hammarström	1 - present	Abbi 2013, Abbi 2013	
Akateko	west2635	Roberto Herrera	0 - absent	Zavala (1992); Schüle (2000)	
Akie	mosi1247	Nataliia Neshcheret	0 - absent	Heine et al. (2015:1-175)	
Alagwa	alag1248	Suzanne van der Meer	0 - absent		
Alamblak	alam1246	Ger Reesink and Ruth Singer Michael Dunn	1 - present	Bruce 1984; Bruce 1974, Bruce 1974; Bruce 1984	
Ama (Sudan)	amas1236	Richard Kowalik	1 - present	Stevenson 1938:99, Stevenson 1938	
Ambulas	ambu1247	Ger Reesink and Ruth Singer Michael Dunn	1 - present	Wilson 1980, Wilson 1980	p.115ff
Amharic	amha1245	Hedvig Skirgard	1 - present	Hagos, Desalegn (2014 p.c.)	

§6 Grambank with pycldf

```
import pycldf
grambank = pycldf.Dataset.from_metadata(r'dbs\grambank\grambank-grambank-7ae000c\cldf\StructureDataset-metadata.json') # Specify where you placed your downloaded copy
```

- It has our same three components Language, Parameter (= here feature), Value (= here absense/presence)

```
>>> for c in grambank.components:
>>>     print(c)
ValueTable
LanguageTable
ParameterTable
CodeTable
```

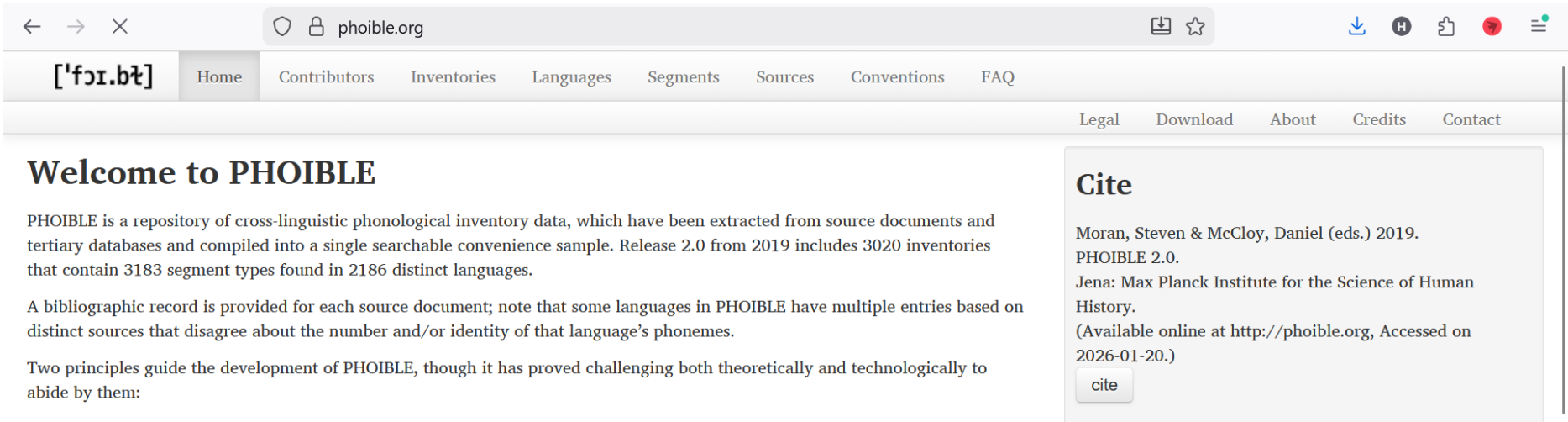
§6 Accessing Grambank using pycldf: Example

- Get some typological features for Eastern Armenian [nucl1235]

```
>>> parameterid_to_name = {row['ID']: row['Name'] for row in grambank['ParameterTable']}
>>> for row in grambank['ValueTable']:
>>>     if row['Language_ID'] == 'nucl1235':
>>>         print(parameterid_to_name[row['Parameter_ID']], row['Value'])

Are there definite or specific articles? 1
Do indefinite nominals commonly have indefinite articles? 1
Are there prenominal articles? 0
Are there postnominal articles? 1
What is the order of numeral and noun in the NP? 1
...
```

§6 CLDF data: Phonology (PHOIBLE)



https://phoible.org/

Summary: Resources

- Language inventory (codes, names, countries, …): Ethnologue/Glottolog
- Family classification: Ethnologue/Glottolog
- Speaker numbers: Ethnologue
- Bibliographical references: Glottolog
- Centre-point geographical coordinates: Glottolog
- CLDF Data: ASJP, Grambank, PHOIBLE, …

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Nature Scientific Data* 5(180205). 1–10.

Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg & Bettina Speckmann. 2018. Simultaneous Visualization of Language Endangerment and Language Description. *Language Documentation & Conservation* 12. 359–392.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp

Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023.

Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. . .