

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

§2 Basic metadata: Names, speaker numbers, endangerment, country

§3 Classification

§4 Maps

§5 Bibliography

§6 CLDF data: Lexicon (ASJP), Grammar (Grambank), Phonology (PHOIBLE)

§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 levels
 - ▶ 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

```
1 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.

```
1 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

```
1 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.

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

- Yields

```
1 Gheg Albanian aln gheg1238
```

§1: Glottolog Identifiers and Levels

- The level attribute holds the level of a languoid

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

- Yields

```
1 Gheg Albanian aln LanguoidLevel(ordinal=2, id='language',  
2     description='defined by mutual non-intelligibility')  
2 Indo-European None LanguoidLevel(ordinal=1, id='family',  
3     description='sub-grouping of languoids above the language  
3     level')  
3 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.

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

- Yields

```
1 Kaishana language None
2 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:hwv: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)

```
1 >>> languoids = [1 for l in glottolog.languoids()]
2 >>> len(languoids)
3 27111
```

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

```
1 >>> languages = [l for l in glottolog.languoids() if l.
2     level.id == 'language']
3 >>> len(languages)
3 8618
```

- (Re efficiency/performance see details at

[https://pyglottolog.readthedocs.io/en/latest/api.html#
performance-considerations](https://pyglottolog.readthedocs.io/en/latest/api.html#performance-considerations))

§2: Basic Metadata: Canonical name

- Every language has one canonical name

```
1 >>> glottolog.languoid('gheg1238').name  
2 '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., fran ois, French, faransawiy, ...
- Extensive, but not complete, lists of (non-unique) alternative names aggregated from different sources are available

```
1 >>> glottolog.languoid('gheg1238').names
2 {'multitree': ['Albanesisch', 'Albanian', 'Albanian, Gheg', 'Arber', 'Arbresh', 'Arnaut', 'Geg', 'Gheg', 'Gheg Albanian', 'Guegue', 'Shgip', ' Shqipri', 'Shquipni', ' kip'], 'lexvo': ['Alban s guego [es]', 'Dialectul Gheg [ro]', 'Gegijski jezik [hr]', 'Gegisch [de]', 'Gegiska [sv]', 'Geg  [sq]', 'Gheg Albanian [en]', 'Gu gue [fr]', 'Г  [el]', '  [bg]', '  [ru]', '  [ja]'], 'hhbib_lgcode': ['Albanian of Zadrima', 'Albanian-Gheg', 'Gegskaja', 'Gheg', 'Malsia Madhe', 'NW Gheg', 'Northeast Geg Albanian from southern Kosovo', 'S dgegischen']}
```

§2: Basic Metadata: Countries

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

```
1 >>> glottolog.languoid('gheg1238').countries
2 [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

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

- Point can be the geographic, political, 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

```
1 >>> glottolog.languoid('gheg1238').endangerment
2 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).')
```



```
1 >>> glottolog.languoid('madi1261').endangerment
2 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

```
1 >>>glottolog.languoid('mbab1239').timespan #Mbabaram went  
2      extinct in 1985  
3 (1985, 1985)  
4 >>>glottolog.languoid('clas1249').timespan #Classical Armenian  
5      is attested between 400 AD - 1100 AD  
6 (400, 1100)  
7 >>>glottolog.languoid('mero1237').timespan #Meroitic is  
8      attested from 250 BC - 300 AD  
9 (-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

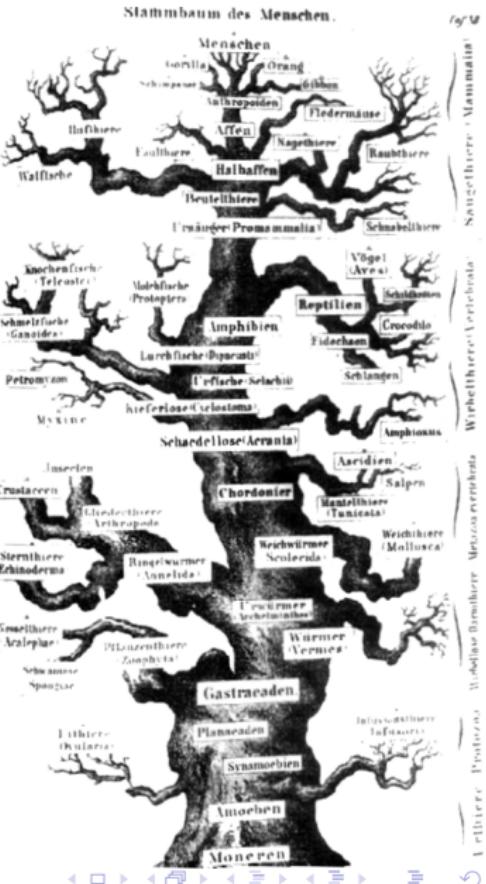
§2: Example Ethnologue Speaker Numbers

```
1 >>> import pandas
2 >>> df = pandas.read_csv('e28.tab', sep='\t', keep_default_na=
   False, index_col = "ISO 639-3") #Index by ISO 639-3, not
   glottocodes
3 >>> df.loc["kbr"]
4 Alternate Names      Caafiti, Caffino, Kaffa, Kaffinya, Kaf...
5 Classification      Afro-Asiatic, Omotic, North, Gonga-Gim...
6 Country              Ethiopia
7 Dialects             Kafa, Bosha (Garo). Bosha may be a dis...
8 ...
9 Location             Southern Nations, Nationalities, and P...
10 Other Comments       Traditional religion, Christian, Muslim.
11 Population           1,406,700, all users. L1 users: 1,360,...
12 Population Numeric  1360000
13 Population Numeric (L2) 46700
14 Typology             SOV.
15 Writing              Ethiopic script [Ethi], used in Church...
16 code+name            Kafa [kbr]
17 name                 Kafa
18 >>> int(df.loc["kbr"]["Population Numeric"])
19 1360000
```



§3 Classification

- Language classification is disputed and prone to update
- Ethnologue has a classification but classification is not a primary focus 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, ...

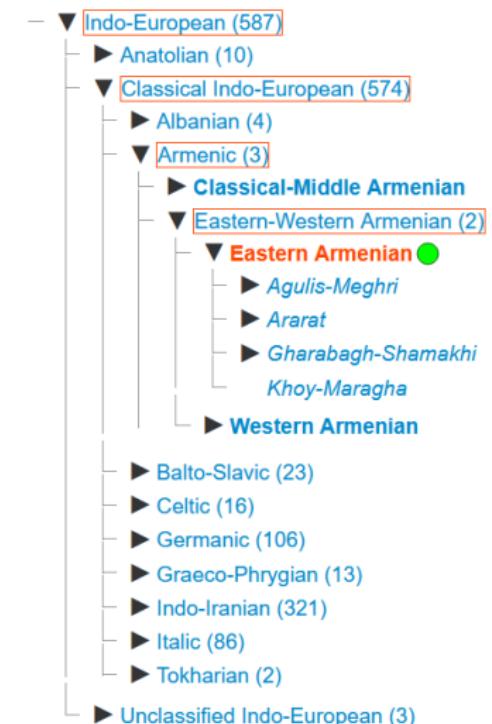


§3 Classification in Glottolog

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

```
1 >>> x = glottolog.languoid('hye') #  
2     Eastern Armenian  
3  
4 >>> x.parent #Gives the immediate  
5     parent  
6 <Family east2768>  
7 >>> x.children #Gives a list of the  
8     immediate children  
9 [  
10    <Dialect agul1245>, <Dialect  
11    arar1266>, <Dialect kara1458>, <  
12    Dialect khoi1249>]  
13 >>> x.ancestors #Gives an ordered  
14     list from the root up to (but not  
15     including) the current node  
16 [  
17    <Family indo1319>, <Family clas1257  
18    >, <Family arme1241>, <Family  
19    east2768>]
```

Classification



Comments on subclassification



§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

<i>Sign Language</i>	<i>Unclassifiable</i>
<i>Pidgin</i>	<i>Unattested</i>
<i>Artificial Language</i>	<i>Speech Register</i>
<i>Mixed Language</i>	<i>Bookkeeping</i>

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

```
1 >>> languages = [l for l in glottolog.languoids() if l.level.id
2   == 'language']
3 >>> len(languages)
4 8618
5 >>> canonical_languages = [l for l in languages if (l.ancestors
6   + [l])[0].name not in ['Sign Language', 'Unclassifiable',
7   'Pidgin', 'Unattested', 'Artificial Language', 'Speech
8   Register', 'Mixed Language', 'Bookkeeping']]
9 >>> len(canonical_languages)
10 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)



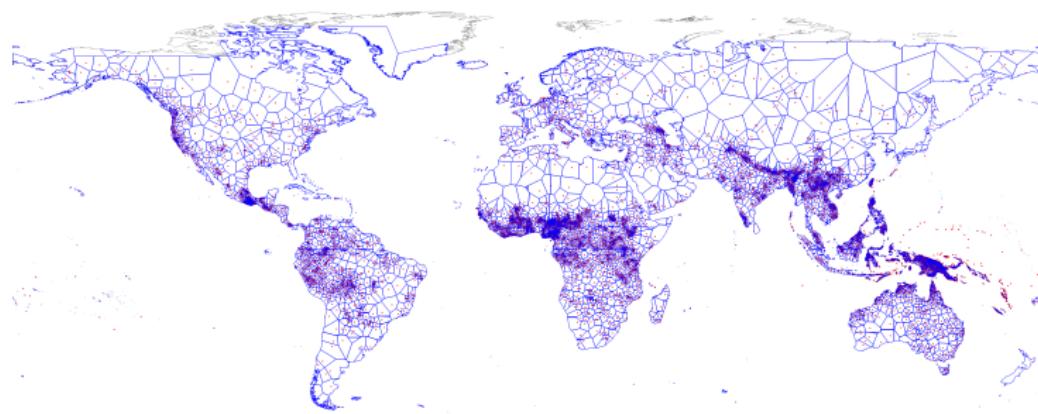
Experts believe the relationship between Western and Eastern Saharan is very deep ≈ 6000 years



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



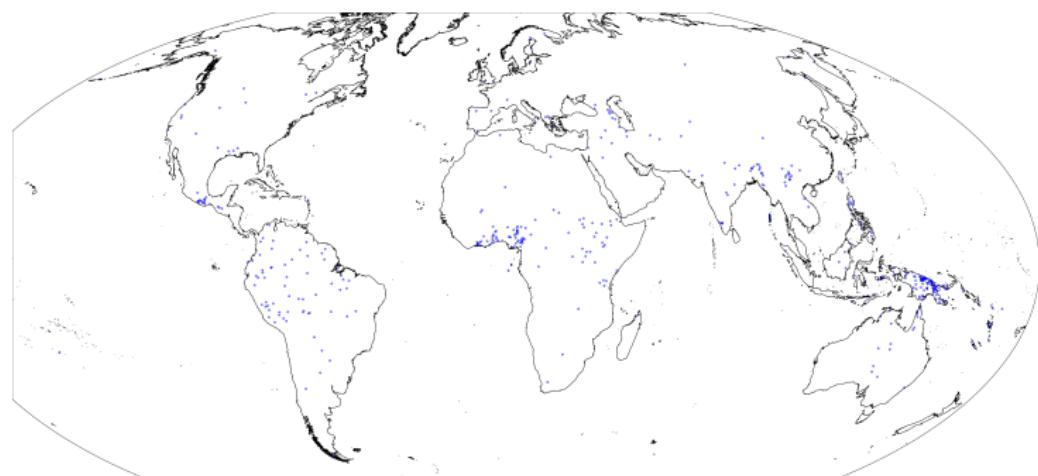
§4 Example Dot Map

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

```
1 #Select the languages
2 languages = [l for l in glottolog.languoids() if l.level.id ==
   'language']
3 canonical_languages = [l for l in languages if (l.ancestors +
   1)[0].name not in ['Sign Language', 'Unclassifiable', ''
   Pidgin', 'Unattested', 'Artificial Language', 'Speech
   Register', 'Mixed Language', 'Bookkeeping']]
4 languages_A = [l for l in canonical_languages if l.name.
   startswith("A")]
5
6 #Create the base map
7 import cartopy.crs as ccrs
8 import matplotlib.pyplot as plt
9 plt.figure(figsize=(36,18))
10 ax = plt.axes(projection=ccrs.Mollweide()) #Mollweide
11 #ax = plt.axes(projection=ccrs.PlateCarree()) #Mercator
12 ax.coastlines(resolution="10m") #Use coastlines
```

§4 Example Dot Map Result

```
1 #Plot the dots
2 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
3
4 #Save
5 plt.savefig('yerevan2026map_test.png', bbox_inches='tight',
   pad_inches=0, dpi = 600)
```



§5 Glottolog Bibliography

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

Classification

- Kwomtari-Nai (2)
 - Kwomtari
 - Etsie Yenali-Mangam
 - West Central Kwomtari
 - Nai

References

Details	Name	Title	Any Field	ca	Year	Pages	Doctype	ca	Provider	da
citation	Murray Honsberger and Carol Honsberger and Ian Tapper 2008	Kwomtari Phonology and Grammar Essentials	✓		2008	197	grammar	ausp2010, cdf_ni, impreve, sifis		
citation	Wetzer Beron 1963	Kwomtari Survey	✓		1963	42	overview, comparative, grammar_sketch	cdf_ni		
citation	Laycock, Donald C. no date	Notebook D16	✓				wordlist	nh		
citation	Laycock, Donald C. 1975	Sko, Kwomtari and Left May (Avai) Phyla	✓		1975	10	overview, comparative	nh		
citation	Laycock, Donald C. 1973	Sepik Languages: Checklist and Preliminary Classification	✓		1973	133	overview, comparative	cdf_ni, langsci, impreve		
citation	Richard Loving and Jack Bass 1964	Languages of the Amanab sub-district	✓		1964	8	overview, comparative	nh		
citation	Glaeton A. Lean 1986	Sandaun Province	✓		1986		overview, minimal	nh		
citation	Fyle, Andrew 2009	Gender, mobility and population history: exploring material culture distributions in the Upper Sepik and Central New Guinea	✓		2009	328	ethnographic, overview	nh		
citation	Honsberger, Murray and Honsberger, Carol and Tupper, Ian 2008	Introduction to the Kwomtari people and language	✓		2008	15	ethnographic	sifis		

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

§5 Glottolog Bibliography Example

```
1 >>>kwomtari = glottolog.languoid('nucl1593')
2 >>>kwomtari.sources
3 [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	$V_{l_1 p_1}$	$V_{l_1 p_2}$...
L_2	$V_{l_2 p_1}$	$V_{l_2 p_2}$...
...	

- 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)

ASJP Home Wordlists Meanings Sources

Wordlist Swedish

Compiled by Viveka Velupillai

Showing 1 to 40 of 40 entries

No.	Meaning	Concepticon	Word	Loan
1	I	⌚ I	yog	False
2	you	⌚ THOU	du	False
3	we	⌚ WE	vi	False
11	one	⌚ ONE	et	False
12	two	⌚ TWO	tv-o	False
18	person	⌚ PERSON	mEniSxE	False
19	fish	⌚ FISH	fisk	False
21	dog	⌚ DOG	h2nd-	False
22	louse	⌚ LOUSE	lus	False
23	tree	⌚ TREE	trEd	False
25	leaf	⌚ LEAF	lev	False
28	skin	⌚ SKIN	Sx-in	False
30	blood	⌚ BLOOD	bl-ud	False
31	bone	⌚ BONE	ben	False
34	horn	⌚ HORN (ANATOMY)	hun	False
39	ear	⌚ EAR	3rE	False

← Previous 1 Next →

GlottoCode: swed1254 ISO 639-3: swe

A map of Europe with a red dot indicating the location of Sweden. Labels include Suomi, Sverige, United Kingdom, Deutschland, and Belarus. A legend shows zoom controls (+, -, x), a scale bar, and OpenStreetMap contributors.

Coordinates: WGS84 60°N, 15°E
60.00, 15.00

number of speakers: 9,606,320

status: alive

Classification

WALS
IE > Germanic

Glottolog
Indo European > Germanic > Northwestgermanic > Northgermanic > Northscandinavian > East Centralswedic > Eastswedic

Ethnologue
Indo European > Germanic > North > Eastscandinavian > Danish
Swedish > Swedish

§6 Accessing ASJP using pycldf

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

```
1 import pycldf
2 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)

```
1 >>> for c in asjp.components:
2 >>>     print(c)
3 FormTable
4 LanguageTable
5 ParameterTable
```

§6 Accessing ASJP using pycldf: Example

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

```
1 >>> localid_to_iso = {row['ID']: row['ISO639P3code'] for row in
2     asjp['LanguageTable']}
3 >>> parameterid_to_gloss = {row['ID']: row['Name'] for row in
4     asjp['ParameterTable']}
5 >>> for row in asjp['FormTable']:
6     if localid_to_iso[row['Language_ID']] == 'hye':
7         if parameterid_to_gloss[row['Parameter_ID']] == '*stone':
8             print(row['Form'], row['Language_ID'])
9
10 kh~or ARMENIAN
11 kh~ar EASTERN_ARMENIAN
```

§6 CLDF data: Grammar (Grambank)

The screenshot shows the Grambank website at <https://grambank.clld.org>. The top navigation bar includes links for Home, Features, Languages and dialects, People, FAQ, Legal, Download, and Contact. The main content area features a "Welcome to Grambank" section with a brief description of the database. Below this is a "How to cite Grambank Online" section with citation details. A large circular graphic on the right illustrates various language features: *glottobank* (red), *grambank* (brown), *phonobank* (green), *lexibank* (blue), *parabank* (teal), and *numeralbank* (light blue). At the bottom, a "Statistics" box provides data: Languages 2,467, Features 195, and Datapoints 441,663 (362,025 excl. "not known").

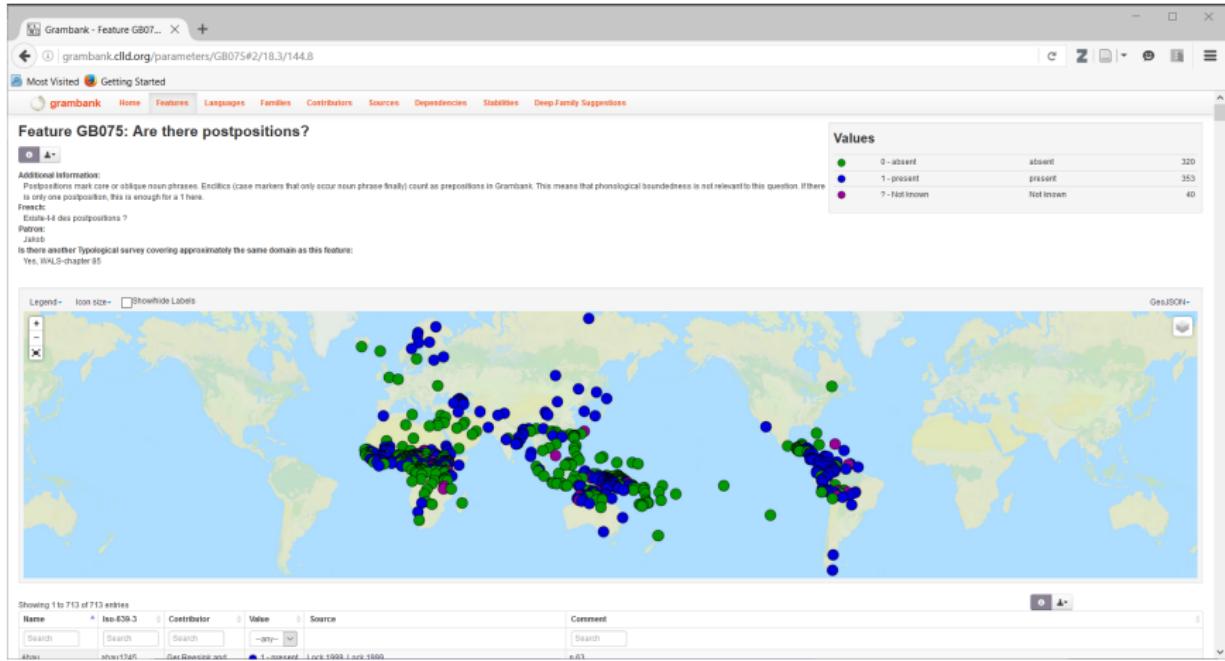
<https://grambank.clld.org/> (Skirgård et al. 2023)

§6 The Grambank Database: 195 Features

The screenshot shows a web browser window titled "Grambank - Features". The address bar contains the URL "grambank.clld.org/parameters". The page header includes links for "Home", "Features", "Languages", "Families", "Contributors", "Sources", "Dependencies", "Stabilities", and "Deep-Family Suggestions". Below the header, a section titled "Features" displays a table of 195 entries. The table has columns for "Id", "Feature", "Morphosyntactic unit", "Form", "Function", "Languages", and "Details". Each row contains a "Search" button for that specific column. The "Languages" column shows the count of languages for each feature, and the "Details" column contains a "Values" button.

Id	Feature	Morphosyntactic unit	Form	Function		
					Languages	Details
GB020	Are there definite or specific articles?	article, NP, demonstrative	declension, particle, paradigm	definiteness, deixis	708	<button>Values</button>
GB021	Do indefinite nominals commonly have indefinite articles?	article, NP, demonstrative	declension, particle, paradigm	definiteness, deixis	673	<button>Values</button>
GB022	Are there prenominal articles?	article, NP, demonstrative	word order, declension, particle, paradigm, preflection	definiteness, deixis	665	<button>Values</button>
GB023	Are there postnominal articles?	article, NP, demonstrative	word order, declension, particle, paradigm, sufflation	definiteness, deixis	671	<button>Values</button>
GB024	What is the order of numeral and noun in the NP?	numeral, NP	word order	quantification, attribution	670	<button>Values</button>
GB025	What is the order of adnominal demonstrative and noun?	demonstrative, NP	word order	deixis, attribution	677	<button>Values</button>
GB026	Can adnominal property words occur discontinuously?	adjective, clause, NP	word order, special construction	attribution	639	<button>Values</button>
GB027	Are nominal conjunction and comitative expressed by different elements?	adposition, case marker, NP	declension, fagging	argument marking, oblique, conjunction, coordination, case marking, comitative	673	<button>Values</button>
GB028	Is there an inclusive/exclusive distinction?	pronoun, index, verb	declension, paradigm, indexing	clusivity, deixis, person, number	724	<button>Values</button>
GB030	Is there a gender distinction in independent 3rd person pronouns?	pronoun	declension, paradigm	person, number, gender	728	<button>Values</button>
GB031	Is there a dual or unit augmented form (in addition to plural or augmented) for all person categories in the pronoun system?	pronoun	declension, paradigm	person, number	683	<button>Values</button>
GB035	Are there three or more distinct contrasts in demonstratives?	demonstrative	declension, paradigm	deixis	691	<button>Values</button>
GB036	Do demonstratives show an elevation distinction?	demonstrative	declension, paradigm	deixis	723	<button>Values</button>
GB037	Do demonstratives show a visible-nonvisible distinction?	demonstrative	declension, paradigm	deixis	602	<button>Values</button>
GB038	Are there demonstrative classifiers?	demonstrative, classifier, NP	declension, paradigm, special construction, particle	deixis, nominal classification, attribution	692	<button>Values</button>
GB039	Is there nonphonological allomorphy of noun number markers?	noun, NP	declension	number	668	<button>Values</button>

§6 Grambank Example Feature



§6 Grambank Data Provenance

- Data collected by humans from descriptive grammars

Name	Iso-639-3	Contributor	Value	Source	Comment
Abau	abau1245	Ger Reesink and Ruth Singer Michael Dunn	● 1 - present	Lock 1999; Lock 1999	p 63
Abé	abee1242	Roberto Herrera	● 0 - absent	Gibry (1987)	
Abul	abul1241	Ger Reesink and Ruth Singer Michael Dunn	● 0 - absent	Kratzschwil 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 infected for person, e.g. Irish) however, they also may combine with aspectual inflection which why I describe them as verbs within Abul system
Achagua	acha1250	Swintha Danielsen	● 1 - present	Wilson 1992; Wilson 1992	
Achéron	ache1245	Sören Peiper	● 0 - absent	Alamín (2012:1-86)	
Adzera	adze1240	Ger Reesink and Ruth Singer Michael Dunn	● 0 - absent		
Aghwan	aghw1237	Natalia Neschcheret	● 1 - present	Gippert (2008:9-41)	
Aiton	aito1238	Jeremy Coller	● 0 - absent	Morey 2002:240; Morey 2002	
Aja (Benin)	ajab1235	Richard Kowalki	● 1 - present	Morey 2010:92; Morey 2010	More postpos than prepos.
Aja (Sudan)	ajaas1235	Jakob Lesage	● 0 - absent	Santandrea (1976: 83)	
Akai-Jenu	akaj1239	Harald Hammarström	● 1 - present	Abbi 2013; Abbi 2013	
Akateko	west6355	Roberto Herrera	● 0 - absent	Zavala (1992); Schüle (2000)	
Akic	mos1247	Natalia Neschcheret	● 0 - absent	Henne et al. (2015:1-175)	
Alagwa	alag1248	Suzanne van der Meer	● 0 - absent		
Alambatik	alam1246	Ger Reesink and Ruth Singer Michael Dunn	● 1 - present	Bruce 1964; Bruce 1974; Bruce 1974; Bruce 1984	
Ana (Sudan)	amas1236	Richard Kowalki	● 1 - present	Stevenson 1938:99; Stevenson 1938	
Amblás	ambu1247	Ger Reesink and Ruth Singer Michael Dunn	● 1 - present	Wilson 1980; Wilson 1980	p. 115ff
Amharc	amha1245	Hedvig Skjærstad	● 1 - present	Hopos; Desalegn (2014 p.c.)	

§6 Grambank with pycldf

```
1 import pycldf
2 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)

```
1 >>> for c in grambank.components:
2       >>>     print(c)
3 ValueTable
4 LanguageTable
5 ParameterTable
6 CodeTable
```

§6 Accessing Grambank using pycldf: Example

- Get some typological features for Eastern Armenian [nucl1235]

```
1 >>> parameterid_to_name = {row['ID']: row['Name'] for row in
2     grambank['ParameterTable']}
3
4 >>> for row in grambank['ValueTable']:
5     if row['Language_ID'] == 'nucl1235':
6         print(parameterid_to_name[row['Parameter_ID']],
7               row['Value'])
8
9 Are there definite or specific articles? 1
10 Do indefinite nominals commonly have indefinite articles? 1
11 Are there prenominal articles? 0
12 Are there postnominal articles? 1
13 What is the order of numeral and noun in the NP? 1
14 ...
15
```

§6 CLDF data: Phonology (PHOIBLE)

The screenshot shows the PHOIBLE website. At the top, there's a navigation bar with links for Home, Contributors, Inventories, Languages, Segments, Sources, Conventions, and FAQ. Below the navigation is a search bar containing the text '['fɔɪ.bɛ]'. To the right of the search bar are links for Legal, Download, About, Credits, and Contact. A 'Cite' sidebar is visible on the right, containing citation information for Moran & McCloy (2019) and Jena: Max Planck Institute for the Science of Human History, along with a 'cite' button.

Welcome to PHOIBLE

PHOIBLE is a repository of cross-linguistic phonological inventory data, which have been extracted from source documents and tertiary databases and compiled into a single searchable convenience sample. Release 2.0 from 2019 includes 3020 inventories that contain 3183 segment types found in 2186 distinct languages.

A bibliographic record is provided for each source document; note that some languages in PHOIBLE have multiple entries based on distinct sources that disagree about the number and/or identity of that language's phonemes.

Two principles guide the development of PHOIBLE, though it has proved challenging both theoretically and technologically to abide by them:

1. Be faithful to the language description in the source document (now often called 'doculect', for reasons indicated above)
2. Encode all character data in a consistent representation in Unicode IPA

In addition to phoneme inventories, PHOIBLE includes distinctive feature data for every phoneme in every language. The feature system used was created by the PHOIBLE developers to be descriptively adequate cross-linguistically. In other words, if two phonemes differ in their graphemic representation, then they necessarily differ in their featural representation as well (regardless of whether those two phonemes coexist in one known doculect). The feature system is loosely based on the feature system in Hayes 2009 with some additions drawn from Moisik & Esling 2011.

However, the final feature system goes beyond both of these sources, and is potentially subject to change as new languages are added in subsequent editions of PHOIBLE.

The data set also includes additional genealogical and geographical information about each language from [Glottolog](#).

The PHOIBLE project also integrates the theoretical model of distinctive features from an extended phonological feature set based on International Phonetic Alphabet ([Association 2005](#)) and on [Hayes 2009](#). This is accomplished by creating a mapping relationship from each IPA segment to a set of features ([Moran 2012](#)). In this way, the IPA is a pivot for interoperability across all resources in PHOIBLE because their contents are encoded in Unicode IPA.

<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 Ghango Ate, Hannah Gibson, Hans-Philipp

Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliaia 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. . .