Understanding the Language Data Landscape to Make Informed Decisions

Conrad Nied

slides

demo

TRANSLATION COMMONS

Abstract

Understanding the Language Data Landscape to Make Informed Decisions

When expanding into new markets or adapting to shifting demographics, choosing which languages to invest in is critical—but often confusing. Language data can be hard to find, difficult to interpret, and easy to misread. Common pitfalls include conflating dialects with macrolanguages, or assuming that spoken use implies digital readiness. This talk breaks down these nuances and offers practical strategies for making informed decisions. We’ll also share tools—including our own, Language Navigator—to help reduce research time and boost confidence in your language strategy.

Meet the speaker

Conrad Nied

Software engineer with background in linguistics, anthropology.

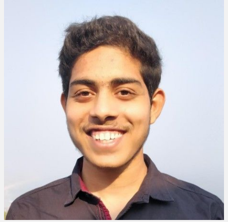
Freelancing with Unicode, Translation Commons, & Stanford

7 years at Facebook/Meta Platforms

Conrad Nied & his cat Pistacchio

Translation Commons

TRANSLATION COMMONS



and more…

https://translationcommons.org/

Org contributors

TRANSLATION COMMONS

hlt

UNICODE

unesco

Presentation Outline

1. Problem Statement
2. Our Journey to Language Navigator
3. Interactive Demo
4. What comes next?

Problem Statement

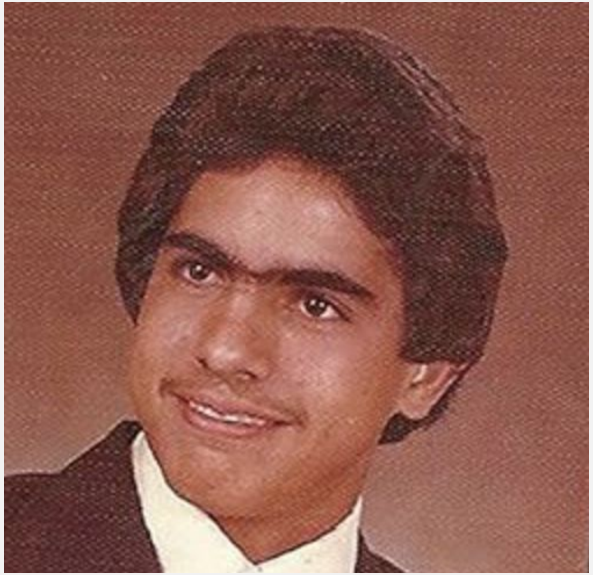
Making decisions on which language to support

Language Support – Why? Example 1

The family of Willie Ramirez rushed Willie to the hospital saying he was…

"Intoxicado"

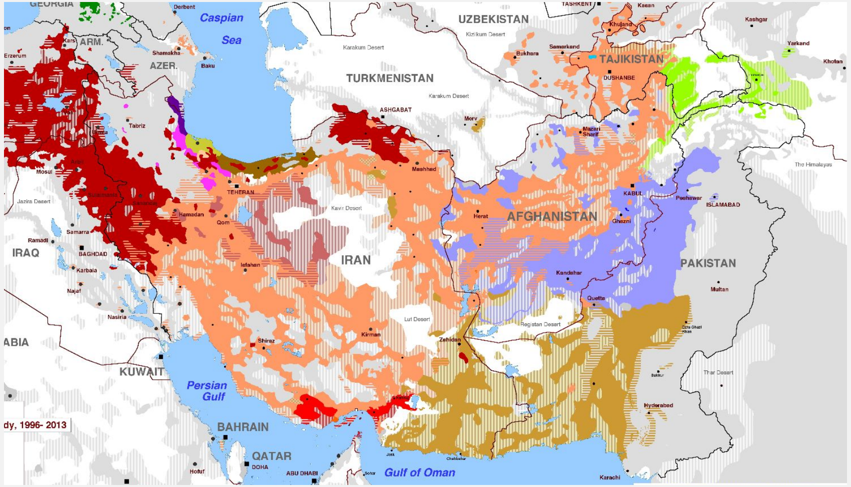
What does intoxicado mean?
What should the medical staff treat?



Language Support – Why? Example 2

Persian, Farsi, Dari, Tajik

All names for a language group on the a continuum of spectrum of languages. Colored peach in the map.

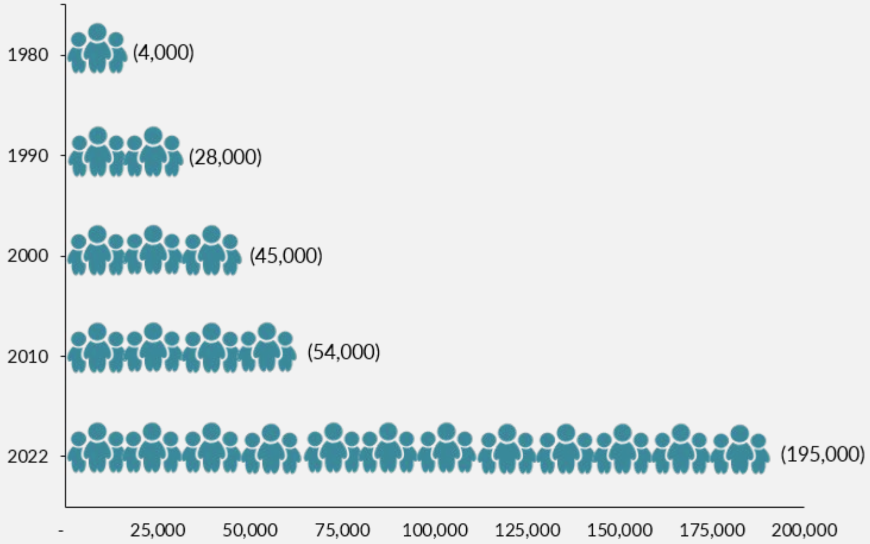


Map of Iranic languages
Dr. Michael Izady for the Atlas of the Islamic World and Vicinity
(New York, Columbia University, Gulf 2000 Project: 2006-present)

Language Support – Why? Example 2

With tens of thousands of Afghans in the US, hospitals and governments need to support Afghan language.

How do you determine which languages to support?



https://www.migrationpolicy.org/article/afghan-immigrants-united-states

Sources: Data from U.S. Census Bureau 2010 and 2022 American Community Surveys (ACS), and Campbell J. Gibson and Kay Jung, "Historical Census Statistics on the Foreign-Born Population of the United States: 1850-2000" (Working Paper no. 81, U.S. Census Bureau, Washington, DC, February 2006), available online.

Language Support – Why? Example 2

So you get this data. Is it easy to understand?

California has hired lots of Persian speakers – but they are trained in Iranian Persian not Dari – they are intelligible (and something is always better than nothing) but key words may differ.

Language	Speakers in Afghanistan
Dari	28,251,300
Persian	21,022,500
Pashto	18,079,350
Southern Pashto	6,000,000
Southern Uzbek	2,910,000
Hazaragi	2,480,655
English	2,201,400

Data from CLDR and Ethnologue, compiled in the Language Navigator

https://translation-commons.github.io/lang-nav/data?territoryFilter=AF&view=Table&objectType=Locale

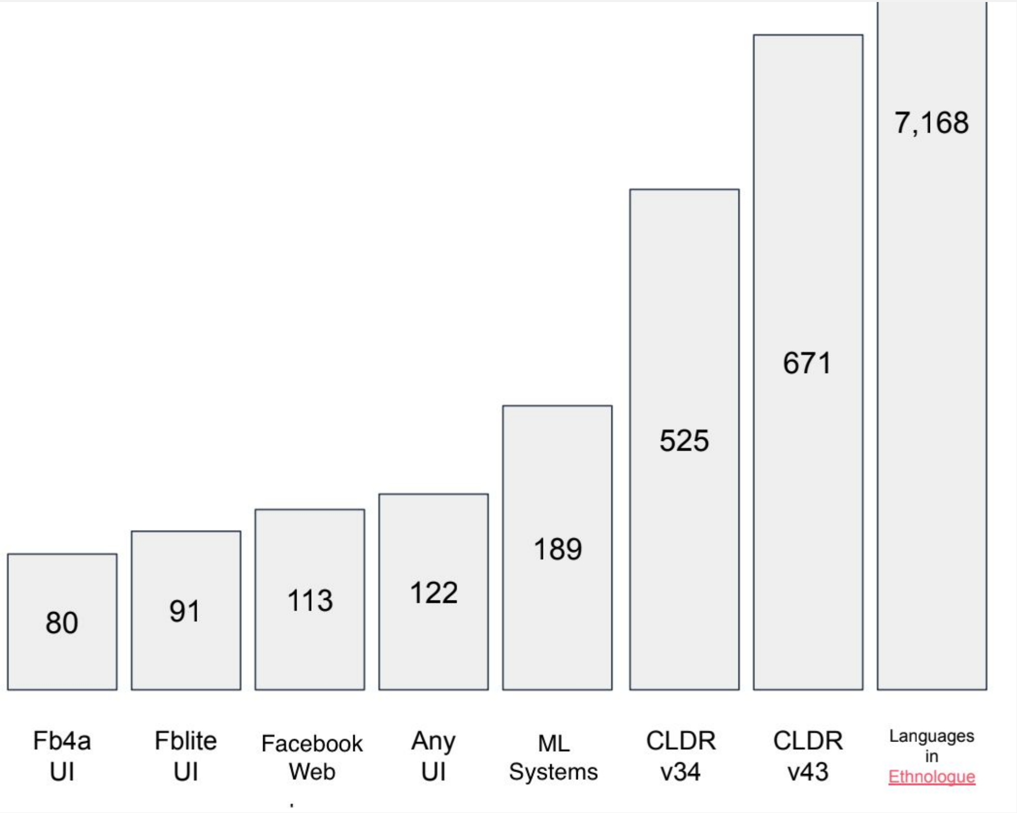
Our journey

To find language data

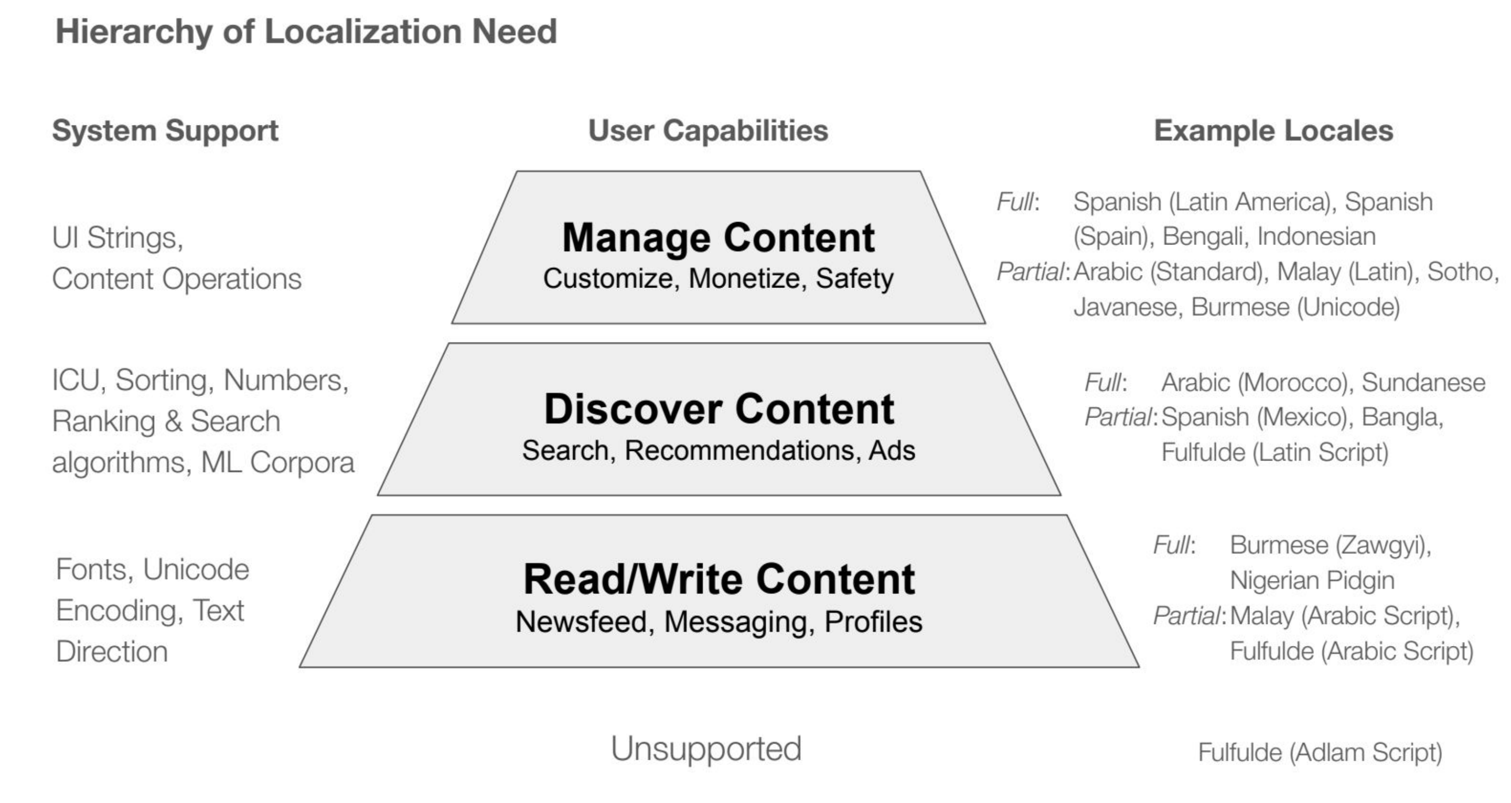
Locales in Meta

At Meta, I launched dozens of languages in the User Interface.

Not all launches were successful though.



Beyond the user interface



Hierarchy of Localization Need

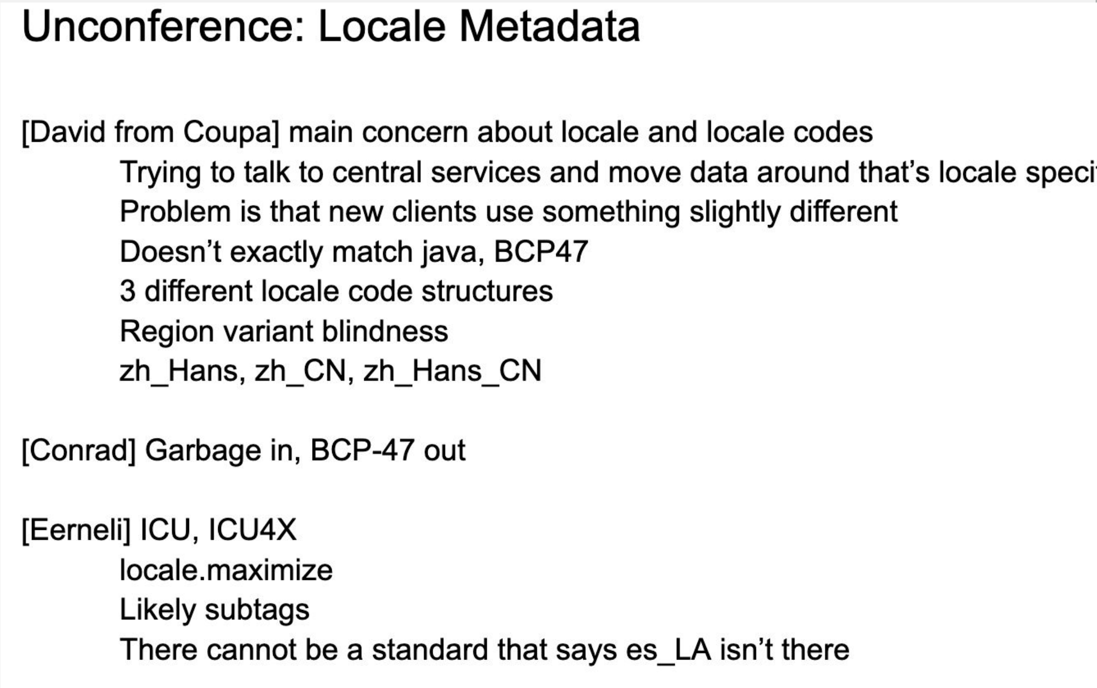
System Support	User Capabilities	Example Locales
UI Strings, Content Operations	**Manage Content** Customize, Monetize, Safety	*Full*: Spanish (Latin America), Spanish (Spain), Bengali, Indonesian *Partial*: Arabic (Standard), Malay (Latin), Sotho, Javanese, Burmese (Unicode)
ICU, Sorting, Numbers, Ranking & Search algorithms, ML Corpora	**Discover Content** Search, Recommendations, Ads	*Full*: Arabic (Morocco), Sundanese *Partial*: Spanish (Mexico), Bangla, Fulfulde (Latin Script)
Fonts, Unicode Encoding, Text Direction	**Read/Write Content** Newsfeed, Messaging, Profiles	*Full*: Burmese (Zawgyi), Nigerian Pidgin *Partial*: Malay (Arabic Script), Fulfulde (Arabic Script)
	Unsupported	Fulfulde (Adlam Script)

UTW Last Year

Takeaways:

1. Hard to find the data.
2. Hard to find language codes.
3. Hard to choose the right languages.

Meeting notes



Unconference: Locale Metadata

[David from Coupa] main concern about locale and locale codes

Trying to talk to central services and move data around that's locale speci
Problem is that new clients use something slightly different
Doesn't exactly match java, BCP47
3 different locale code structures
Region variant blindness
zh_Hans, zh_CN, zh_Hans_CN

[Conrad] Garbage in, BCP-47 out

[Eerneli] ICU, ICU4X

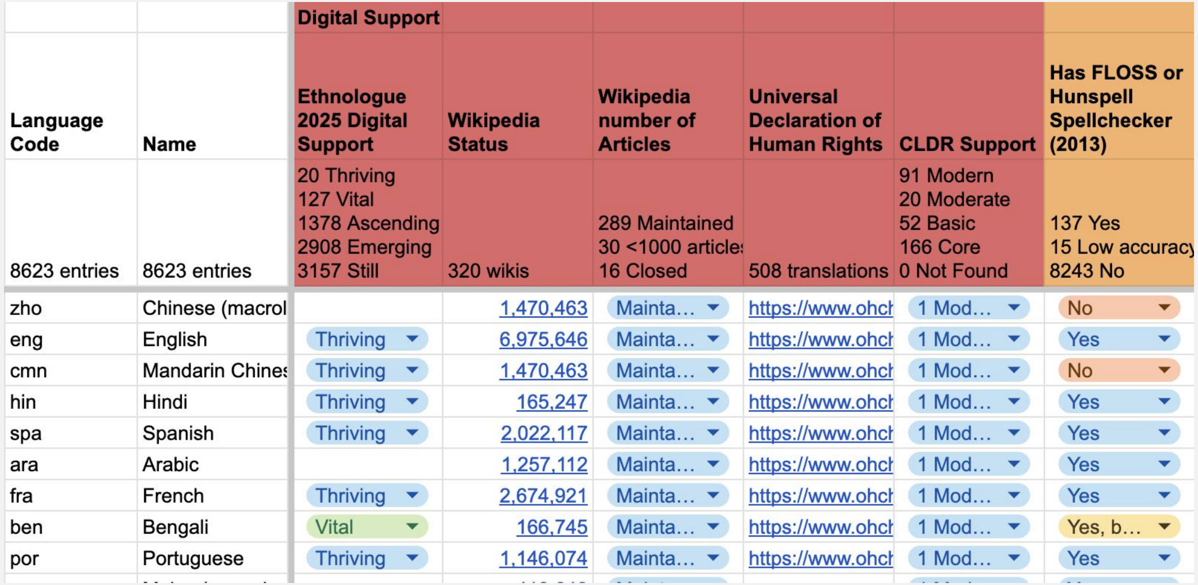
locale.maximize
Likely subtags
There cannot be a standard that says es_LA isn't there

So we made a Spreadsheet

8655 languages

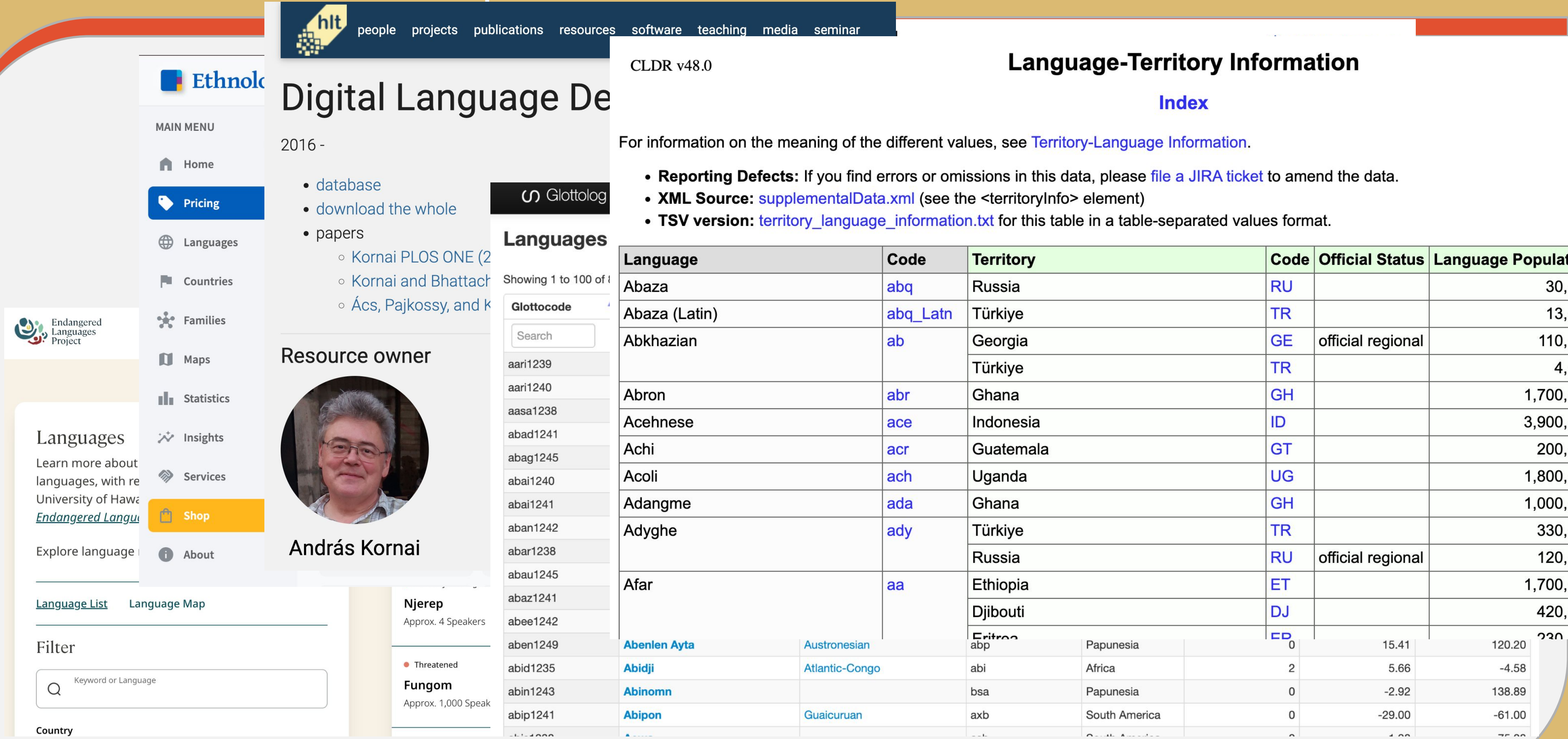
10077 locales

11042 population datapoints



		Digital Support					
Language Code	Name	Ethnologue 2025 Digital Support	Wikipedia Status	Wikipedia number of Articles	Universal Declaration of Human Rights	CLDR Support	Has FLOSS or Hunspell Spellchecker (2013)
8623 entries	8623 entries	20 Thriving 127 Vital 1378 Ascending 2908 Emerging 3157 Still	320 wikis	289 Maintained 30 <1000 articles 16 Closed	508 translations	91 Modern 20 Moderate 52 Basic 166 Core 0 Not Found	137 Yes 15 Low accuracy 8243 No
zho	Chinese (macrol		1,470,463	Mainta…	https://www.ohch	1 Mod…	No
eng	English	Thriving	6,975,646	Mainta…	https://www.ohch	1 Mod…	Yes
cmn	Mandarin Chines	Thriving	1,470,463	Mainta…	https://www.ohch	1 Mod…	No
hin	Hindi	Thriving	165,247	Mainta…	https://www.ohch	1 Mod…	Yes
spa	Spanish	Thriving	2,022,117	Mainta…	https://www.ohch	1 Mod…	Yes
ara	Arabic		1,257,112	Mainta…	https://www.ohch	1 Mod…	Yes
fra	French	Thriving	2,674,921	Mainta…	https://www.ohch	1 Mod…	Yes
ben	Bengali	Vital	166,745	Mainta…	https://www.ohch	1 Mod…	Yes, b…
por	Portuguese	Thriving	1,146,074	Mainta…	https://www.ohch	1 Mod…	Yes

Went to lots of places



Endangered Languages Project

Languages

Learn more about languages, with re University of Hawa *Endangered Langu*

Explore language

Language List Language Map

Filter

Keyword or Language

Country

Njerep
Approx. 4 Speakers

Threatened
Fungom
Approx. 1,000 Speak

Ethnolo

MAIN MENU

- Home
- Pricing
- Languages
- Countries
- Families
- Maps
- Statistics
- Insights
- Services
- Shop
- About

hlt people projects publications resources software teaching media seminar

Digital Language De

2016 -

- database
- download the whole
- papers
  - Kornai PLOS ONE (2
  - Kornai and Bhattach
  - Ács, Pajkossy, and K

Resource owner

András Kornai

Glottolog

Languages

Showing 1 to 100 of

Search

Glottocode							
aari1239							
aari1240							
aasa1238							
abad1241							
abag1245							
abai1240							
abai1241							
aban1242							
abar1238							
abau1245							
abaz1241							
abee1242							
aben1249	Abenlen Ayta	Austronesian	abp	Papunesia	0	15.41	120.20
abid1235	Abidji	Atlantic-Congo	abi	Africa	2	5.66	-4.58
abin1243	Abinomn		bsa	Papunesia	0	-2.92	138.89
abip1241	Abipon	Guaicuruan	axb	South America	0	-29.00	-61.00

CLDR v48.0

Language-Territory Information

Index

For information on the meaning of the different values, see Territory-Language Information.

- **Reporting Defects:** If you find errors or omissions in this data, please file a JIRA ticket to amend the data.
- **XML Source:** supplementalData.xml (see the <territoryInfo> element)
- **TSV version:** territory_language_information.txt for this table in a table-separated values format.

Language	Code	Territory	Code	Official Status	Language Populat
Abaza	abq	Russia	RU		30,
Abaza (Latin)	abq_Latn	Türkiye	TR		13,
Abkhazian	ab	Georgia	GE	official regional	110,
		Türkiye	TR		4,
Abron	abr	Ghana	GH		1,700,
Acehnese	ace	Indonesia	ID		3,900,
Achi	acr	Guatemala	GT		200,
Acoli	ach	Uganda	UG		1,800,
Adangme	ada	Ghana	GH		1,000,
Adyghe	ady	Türkiye	TR		330,
		Russia	RU	official regional	120,
Afar	aa	Ethiopia	ET		1,700,
		Djibouti	DJ		420,
		Eritrea	ER		230,

All common problems

Hard to find

Hard to trust

Hard to understand

Hard to compare

Good data has Paywalls

Guiding Principles

Free and Open

Actionable Insights

Inclusive

TRANSLATION COMMONS

Language Navigator

Demo through concepts

Live demos

1. Language Details
2. Language Families
3. Language Vitality Map
4. Languages in a Country, Censuses
5. Other language data

Link:
https://translation-commons.github.io/lang-nav/

Feedback, mailto:
conrad@translationcommons.org

Kinds of feedback

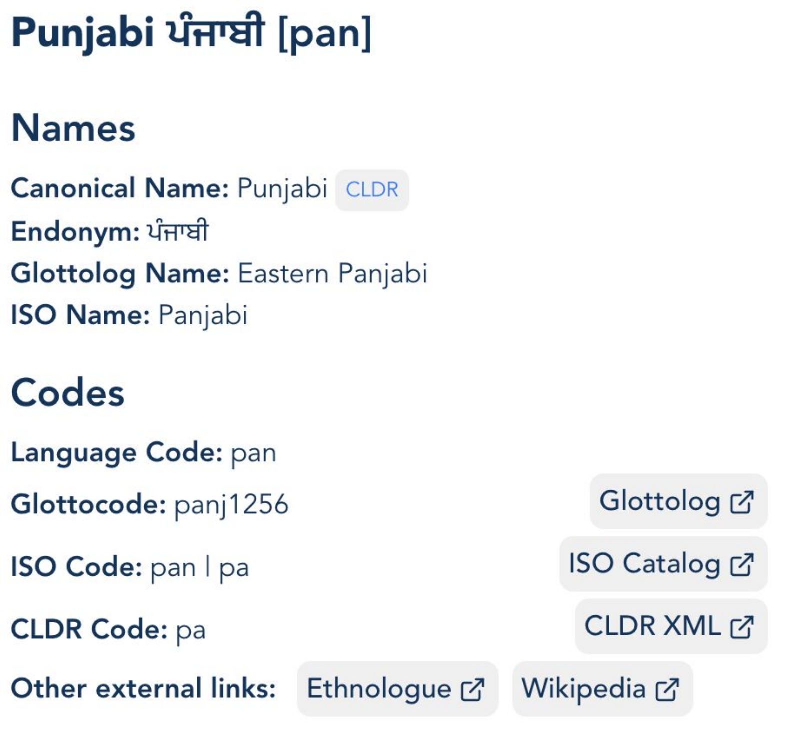
1. X is wrong!
2. Here's a link to a reputable source
3. Interaction ideas

Common Pitfalls

1. What’s a Language
   a. Families
   b. Macro-languages
   c. Dialects, Orthographic differences
2. Incompatible Source Methodologies
   a. Spoken, Written, Used at Home
   b. Definition of “vitality”
3. Language Identification
   a. Language codes
   b. Language names

Start with languages

link



Punjabi ਪੰਜਾਬੀ [pan]

Names

Canonical Name: Punjabi CLDR

Endonym: ਪੰਜਾਬੀ

Glottolog Name: Eastern Panjabi

ISO Name: Panjabi

Codes

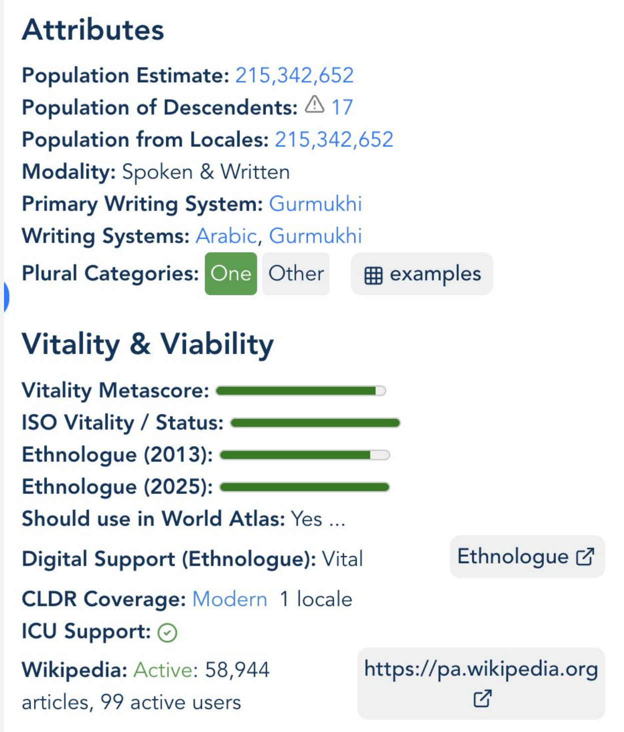
Language Code: pan

Glottocode: panj1256 Glottolog

ISO Code: pan | pa ISO Catalog

CLDR Code: pa CLDR XML

Other external links: Ethnologue Wikipedia



Attributes

Population Estimate: 215,342,652

Population of Descendents: ⚠ 17

Population from Locales: 215,342,652

Modality: Spoken & Written

Primary Writing System: Gurmukhi

Writing Systems: Arabic, Gurmukhi

Plural Categories: One Other ▦ examples

Vitality & Viability

Vitality Metascore:

ISO Vitality / Status:

Ethnologue (2013):

Ethnologue (2025):

Should use in World Atlas: Yes ...

Digital Support (Ethnologue): Vital Ethnologue

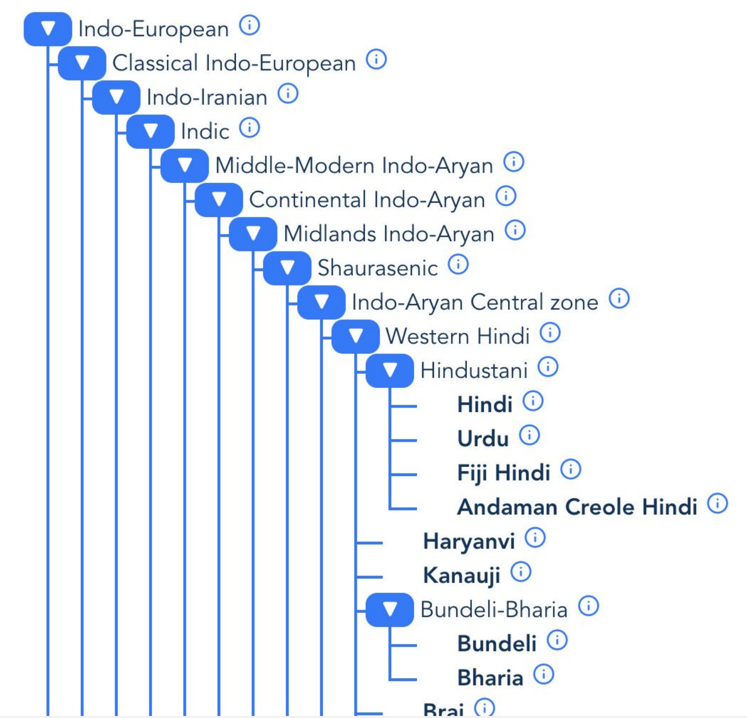
CLDR Coverage: Modern 1 locale

ICU Support: ✓

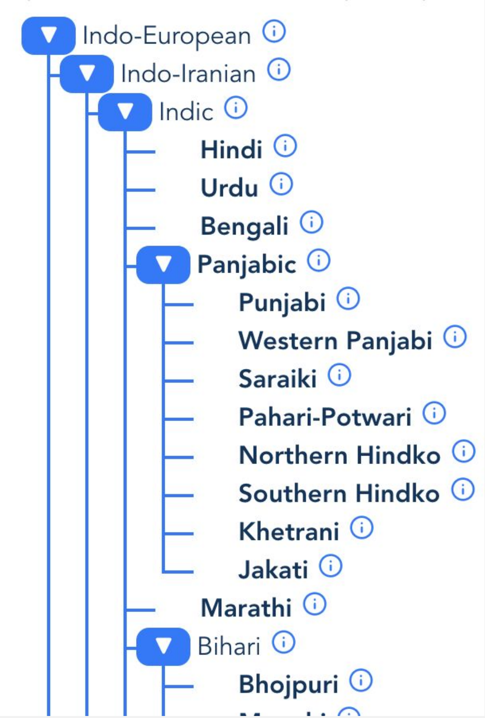
Wikipedia: Active: 58,944 articles, 99 active users https://pa.wikipedia.org

Language Families

link



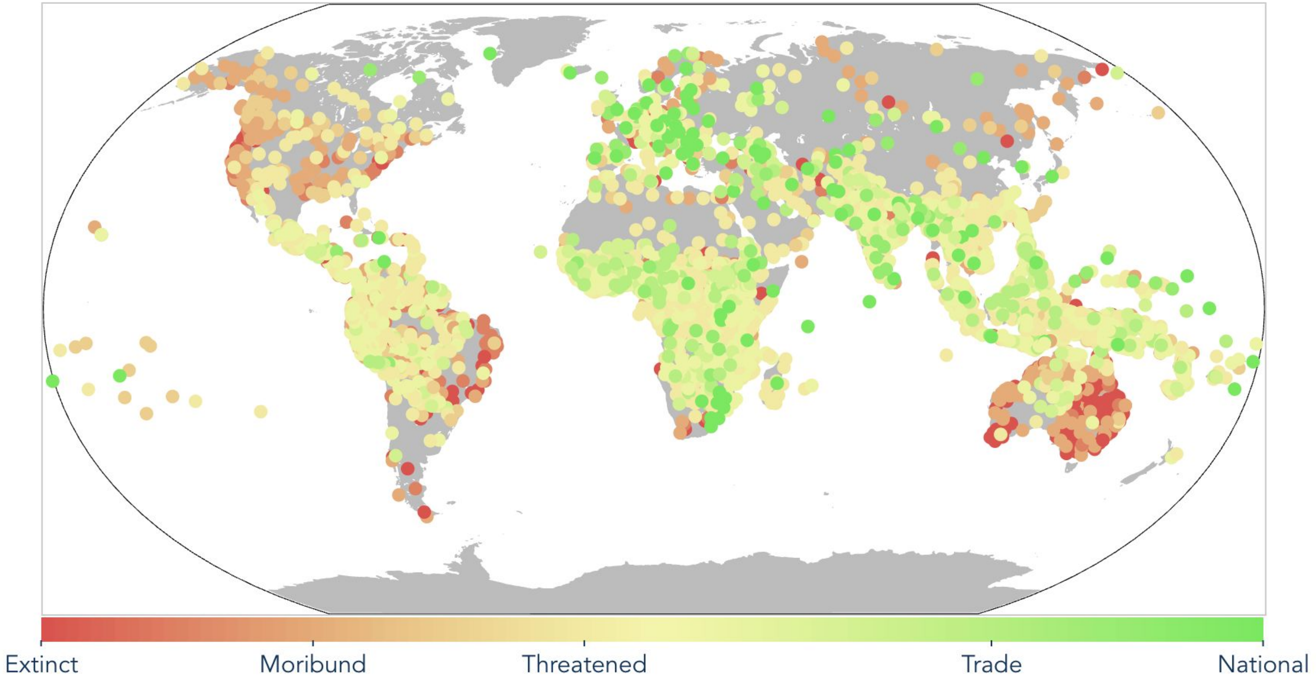
- Indo-European
  - Classical Indo-European
    - Indo-Iranian
      - Indic
        - Middle-Modern Indo-Aryan
          - Continental Indo-Aryan
            - Midlands Indo-Aryan
              - Shaurasenic
                - Indo-Aryan Central zone
                  - Western Hindi
                    - Hindustani
                      - **Hindi**
                      - **Urdu**
                      - **Fiji Hindi**
                      - **Andaman Creole Hindi**
                    - **Haryanvi**
                    - **Kanauji**
                    - Bundeli-Bharia
                      - **Bundeli**
                      - **Bharia**
                    - **Braj**



- Indo-European
  - Indo-Iranian
    - Indic
      - **Hindi**
      - **Urdu**
      - **Bengali**
      - **Panjabic**
        - **Punjabi**
        - **Western Panjabi**
        - **Saraiki**
        - **Pahari-Potwari**
        - **Northern Hindko**
        - **Southern Hindko**
        - **Khetrani**
        - **Jakati**
      - **Marathi**
      - Bihari
        - **Bhojpuri**

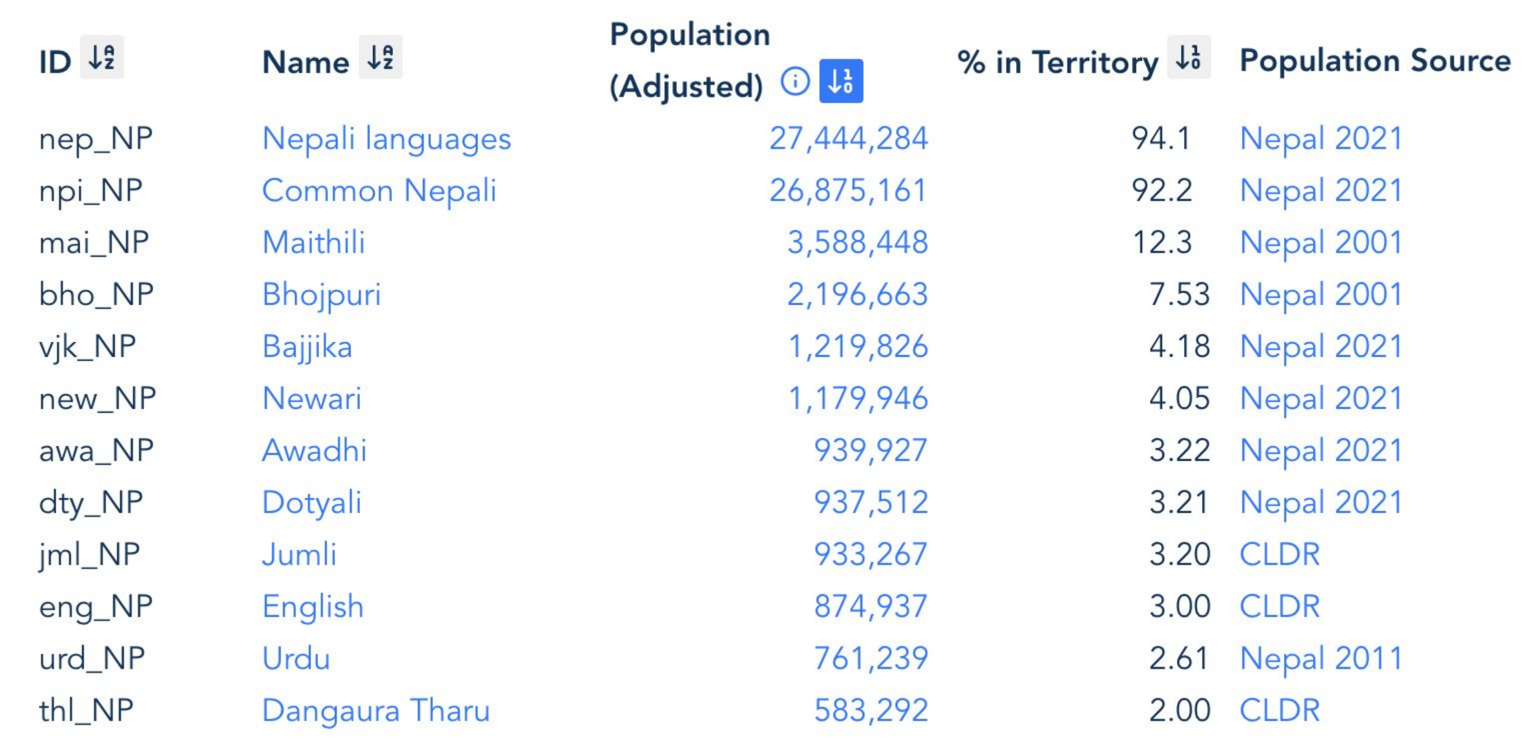
Visualize Language Vitality

link



link

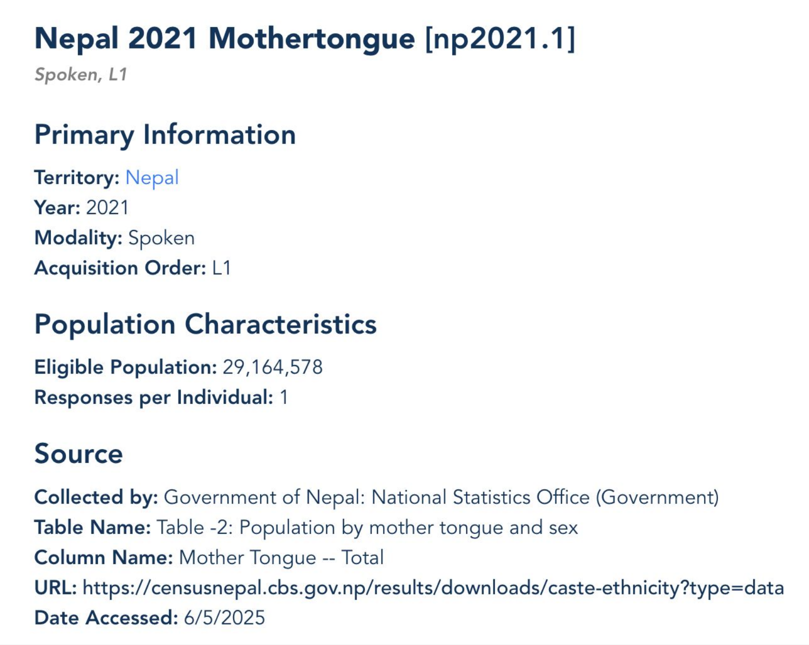
Determining the languages of a country



ID	Name	Population (Adjusted)	% in Territory	Population Source
nep_NP	Nepali languages	27,444,284	94.1	Nepal 2021
npi_NP	Common Nepali	26,875,161	92.2	Nepal 2021
mai_NP	Maithili	3,588,448	12.3	Nepal 2001
bho_NP	Bhojpuri	2,196,663	7.53	Nepal 2001
vjk_NP	Bajjika	1,219,826	4.18	Nepal 2021
new_NP	Newari	1,179,946	4.05	Nepal 2021
awa_NP	Awadhi	939,927	3.22	Nepal 2021
dty_NP	Dotyali	937,512	3.21	Nepal 2021
jml_NP	Jumli	933,267	3.20	CLDR
eng_NP	English	874,937	3.00	CLDR
urd_NP	Urdu	761,239	2.61	Nepal 2011
thl_NP	Dangaura Tharu	583,292	2.00	CLDR

link

Censuses



Nepal 2021 Mothertongue [np2021.1]

Spoken, L1

Primary Information

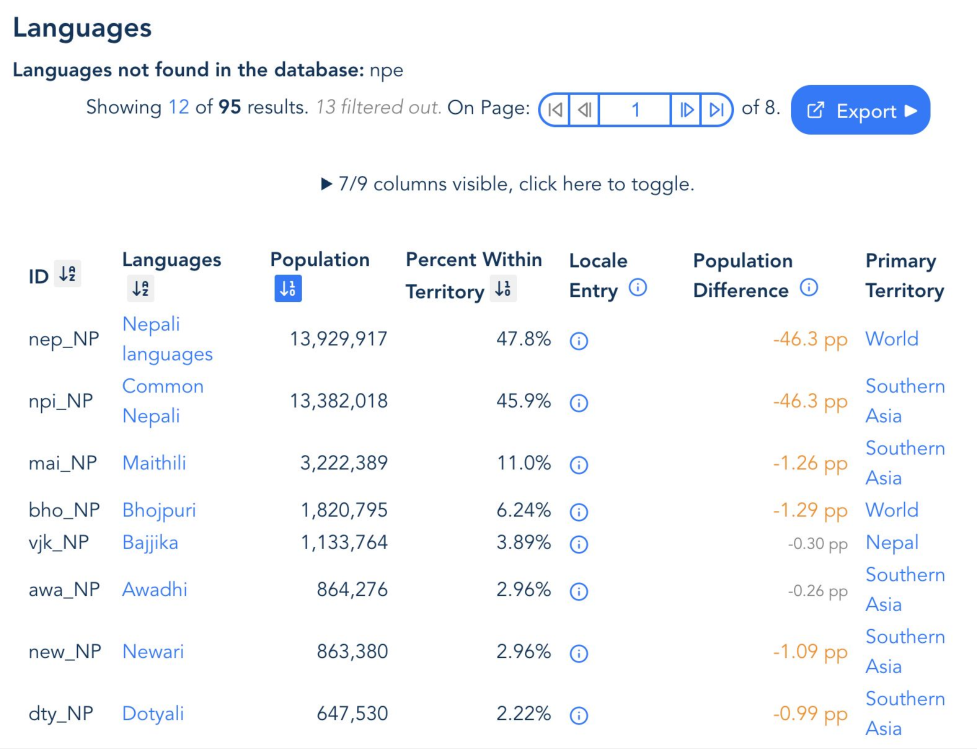
Territory: Nepal
Year: 2021
Modality: Spoken
Acquisition Order: L1

Population Characteristics

Eligible Population: 29,164,578
Responses per Individual: 1

Source

Collected by: Government of Nepal: National Statistics Office (Government)
Table Name: Table -2: Population by mother tongue and sex
Column Name: Mother Tongue -- Total
URL: https://censusnepal.cbs.gov.np/results/downloads/caste-ethnicity?type=data
Date Accessed: 6/5/2025



Languages

Languages not found in the database: npe

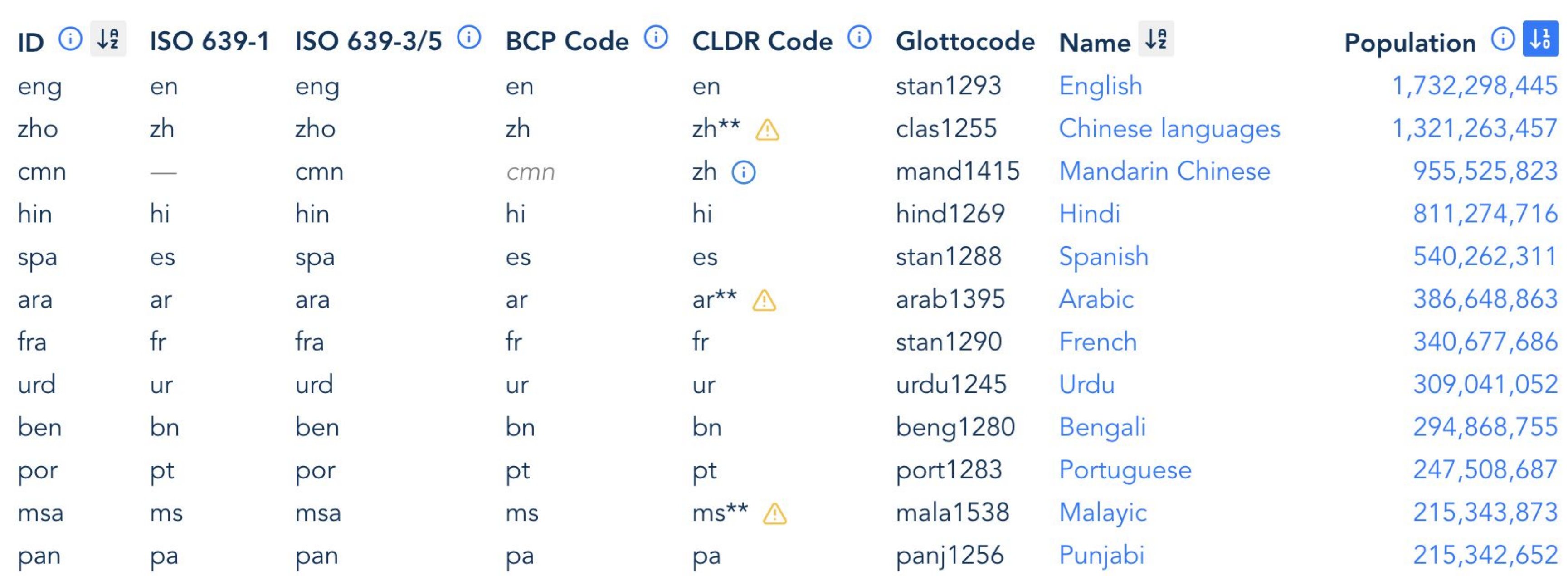
Showing 12 of **95** results. *13 filtered out.* On Page: 1 of 8.

▶ 7/9 columns visible, click here to toggle.

ID	Languages	Population	Percent Within Territory	Locale Entry ⓘ	Population Difference ⓘ	Primary Territory
nep_NP	Nepali languages	13,929,917	47.8%	ⓘ	-46.3 pp	World
npi_NP	Common Nepali	13,382,018	45.9%	ⓘ	-46.3 pp	Southern Asia
mai_NP	Maithili	3,222,389	11.0%	ⓘ	-1.26 pp	Southern Asia
bho_NP	Bhojpuri	1,820,795	6.24%	ⓘ	-1.29 pp	World
vjk_NP	Bajjika	1,133,764	3.89%	ⓘ	-0.30 pp	Nepal
awa_NP	Awadhi	864,276	2.96%	ⓘ	-0.26 pp	Southern Asia
new_NP	Newari	863,380	2.96%	ⓘ	-1.09 pp	Southern Asia
dty_NP	Dotyali	647,530	2.22%	ⓘ	-0.99 pp	Southern Asia

link

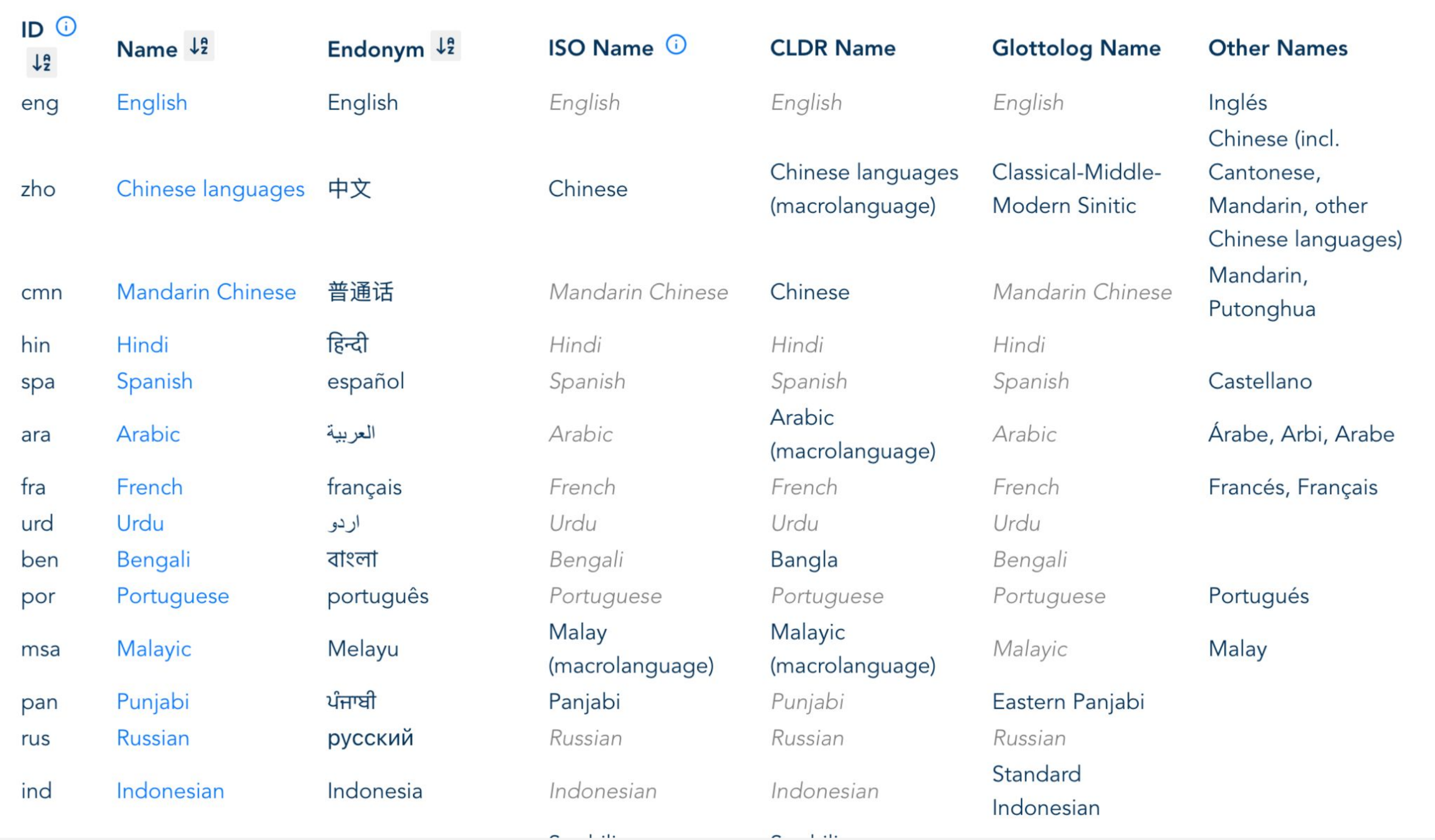
Language Codes



ID	ISO 639-1	ISO 639-3/5	BCP Code	CLDR Code	Glottocode	Name	Population
eng	en	eng	en	en	stan1293	English	1,732,298,445
zho	zh	zho	zh	zh** ⚠	clas1255	Chinese languages	1,321,263,457
cmn	—	cmn	*cmn*	zh ⓘ	mand1415	Mandarin Chinese	955,525,823
hin	hi	hin	hi	hi	hind1269	Hindi	811,274,716
spa	es	spa	es	es	stan1288	Spanish	540,262,311
ara	ar	ara	ar	ar** ⚠	arab1395	Arabic	386,648,863
fra	fr	fra	fr	fr	stan1290	French	340,677,686
urd	ur	urd	ur	ur	urdu1245	Urdu	309,041,052
ben	bn	ben	bn	bn	beng1280	Bengali	294,868,755
por	pt	por	pt	pt	port1283	Portuguese	247,508,687
msa	ms	msa	ms	ms** ⚠	mala1538	Malayic	215,343,873
pan	pa	pan	pa	pa	panj1256	Punjabi	215,342,652

Language names

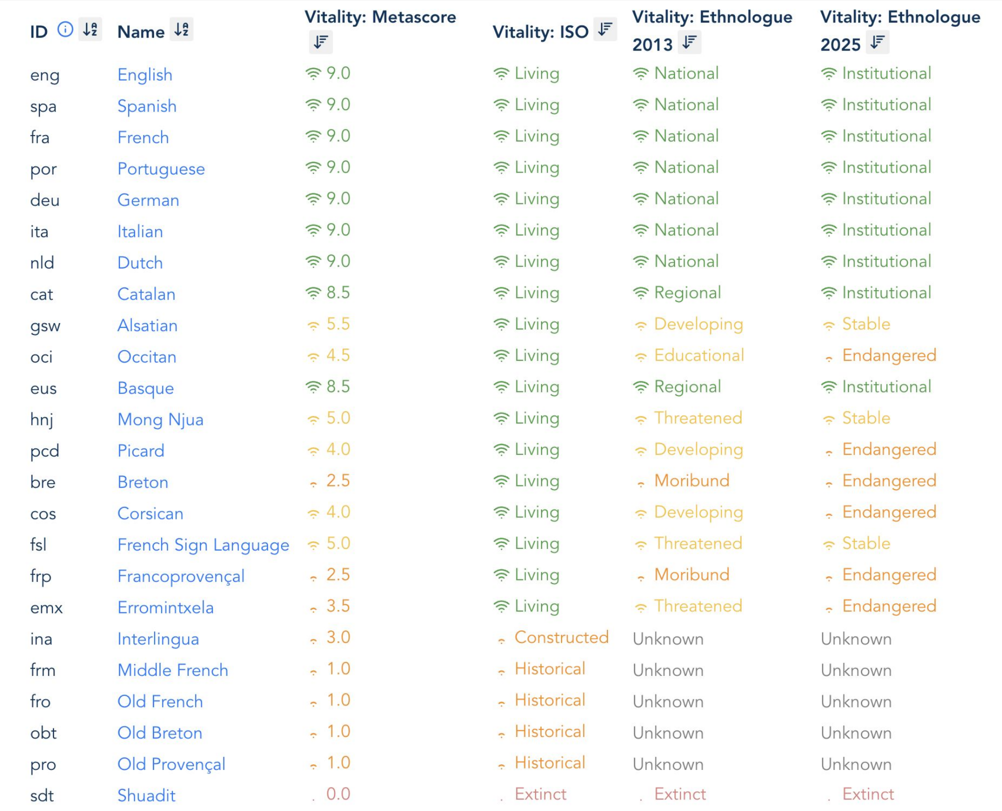
link



ID	Name	Endonym	ISO Name	CLDR Name	Glottolog Name	Other Names
eng	English	English	*English*	*English*	*English*	Inglés
zho	Chinese languages	中文	Chinese	Chinese languages (macrolanguage)	Classical-Middle-Modern Sinitic	Chinese (incl. Cantonese, Mandarin, other Chinese languages)
cmn	Mandarin Chinese	普通话	*Mandarin Chinese*	Chinese	*Mandarin Chinese*	Mandarin, Putonghua
hin	Hindi	हिन्दी	*Hindi*	*Hindi*	*Hindi*	
spa	Spanish	español	*Spanish*	*Spanish*	*Spanish*	Castellano
ara	Arabic	العربية	*Arabic*	Arabic (macrolanguage)	*Arabic*	Árabe, Arbi, Arabe
fra	French	français	*French*	*French*	*French*	Francés, Français
urd	Urdu	اردو	*Urdu*	*Urdu*	*Urdu*	
ben	Bengali	বাংলা	*Bengali*	Bangla	*Bengali*	
por	Portuguese	português	*Portuguese*	*Portuguese*	*Portuguese*	Portugués
msa	Malayic	Melayu	Malay (macrolanguage)	Malayic (macrolanguage)	*Malayic*	Malay
pan	Punjabi	ਪੰਜਾਬੀ	Panjabi	*Punjabi*	Eastern Panjabi	
rus	Russian	русский	*Russian*	*Russian*	*Russian*	
ind	Indonesian	Indonesia	*Indonesian*	*Indonesian*	Standard Indonesian	

Language vitality

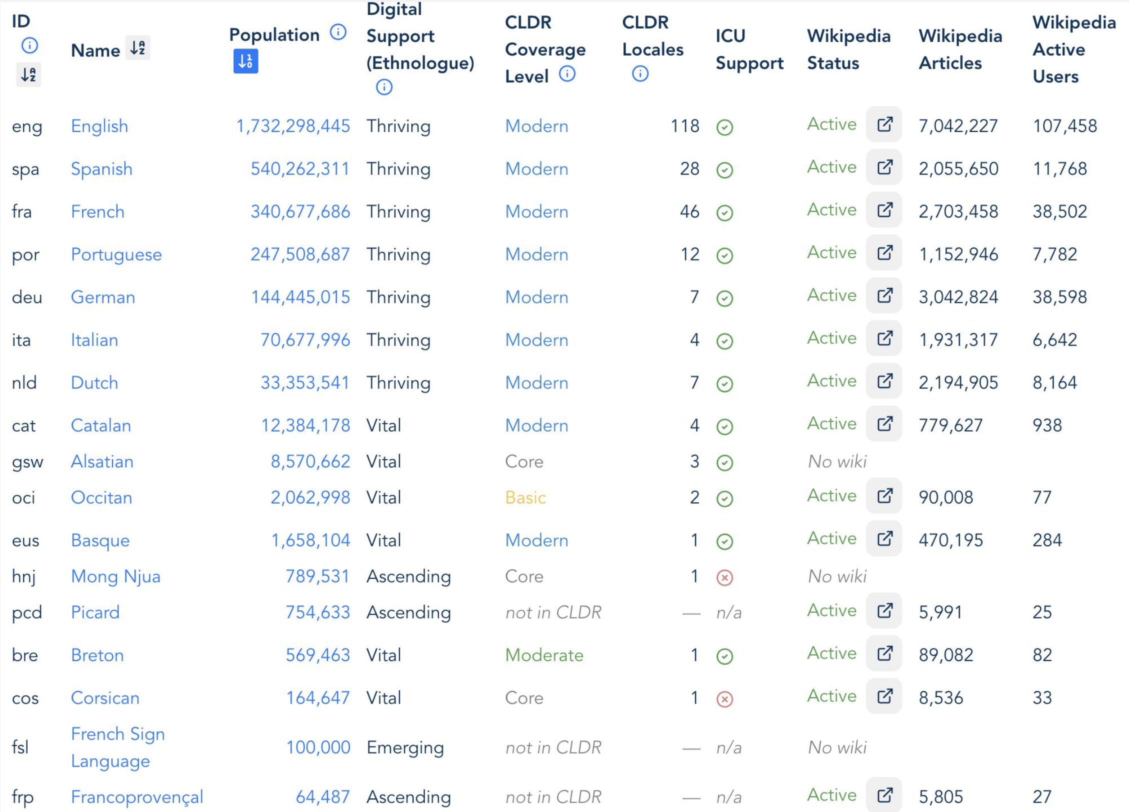
link



ID ⓘ	Name	Vitality: Metascore	Vitality: ISO	Vitality: Ethnologue 2013	Vitality: Ethnologue 2025
eng	English	9.0	Living	National	Institutional
spa	Spanish	9.0	Living	National	Institutional
fra	French	9.0	Living	National	Institutional
por	Portuguese	9.0	Living	National	Institutional
deu	German	9.0	Living	National	Institutional
ita	Italian	9.0	Living	National	Institutional
nld	Dutch	9.0	Living	National	Institutional
cat	Catalan	8.5	Living	Regional	Institutional
gsw	Alsatian	5.5	Living	Developing	Stable
oci	Occitan	4.5	Living	Educational	Endangered
eus	Basque	8.5	Living	Regional	Institutional
hnj	Mong Njua	5.0	Living	Threatened	Stable
pcd	Picard	4.0	Living	Developing	Endangered
bre	Breton	2.5	Living	Moribund	Endangered
cos	Corsican	4.0	Living	Developing	Endangered
fsl	French Sign Language	5.0	Living	Threatened	Stable
frp	Francoprovençal	2.5	Living	Moribund	Endangered
emx	Erromintxela	3.5	Living	Threatened	Endangered
ina	Interlingua	3.0	Constructed	Unknown	Unknown
frm	Middle French	1.0	Historical	Unknown	Unknown
fro	Old French	1.0	Historical	Unknown	Unknown
obt	Old Breton	1.0	Historical	Unknown	Unknown
pro	Old Provençal	1.0	Historical	Unknown	Unknown
sdt	Shuadit	0.0	Extinct	Extinct	Extinct

link

Language digital support



ID	Name	Population	Digital Support (Ethnologue)	CLDR Coverage Level	CLDR Locales	ICU Support	Wikipedia Status	Wikipedia Articles	Wikipedia Active Users
eng	English	1,732,298,445	Thriving	Modern	118	✓	Active	7,042,227	107,458
spa	Spanish	540,262,311	Thriving	Modern	28	✓	Active	2,055,650	11,768
fra	French	340,677,686	Thriving	Modern	46	✓	Active	2,703,458	38,502
por	Portuguese	247,508,687	Thriving	Modern	12	✓	Active	1,152,946	7,782
deu	German	144,445,015	Thriving	Modern	7	✓	Active	3,042,824	38,598
ita	Italian	70,677,996	Thriving	Modern	4	✓	Active	1,931,317	6,642
nld	Dutch	33,353,541	Thriving	Modern	7	✓	Active	2,194,905	8,164
cat	Catalan	12,384,178	Vital	Modern	4	✓	Active	779,627	938
gsw	Alsatian	8,570,662	Vital	Core	3	✓	*No wiki*		
oci	Occitan	2,062,998	Vital	Basic	2	✓	Active	90,008	77
eus	Basque	1,658,104	Vital	Modern	1	✓	Active	470,195	284
hnj	Mong Njua	789,531	Ascending	Core	1	✗	*No wiki*		
pcd	Picard	754,633	Ascending	*not in CLDR*	—	*n/a*	Active	5,991	25
bre	Breton	569,463	Vital	Moderate	1	✓	Active	89,082	82
cos	Corsican	164,647	Vital	Core	1	✗	Active	8,536	33
fsl	French Sign Language	100,000	Emerging	*not in CLDR*	—	*n/a*	*No wiki*		
frp	Francoprovençal	64,487	Ascending	*not in CLDR*	—	*n/a*	Active	5,805	27

What comes next?

The Future of Language Navigator

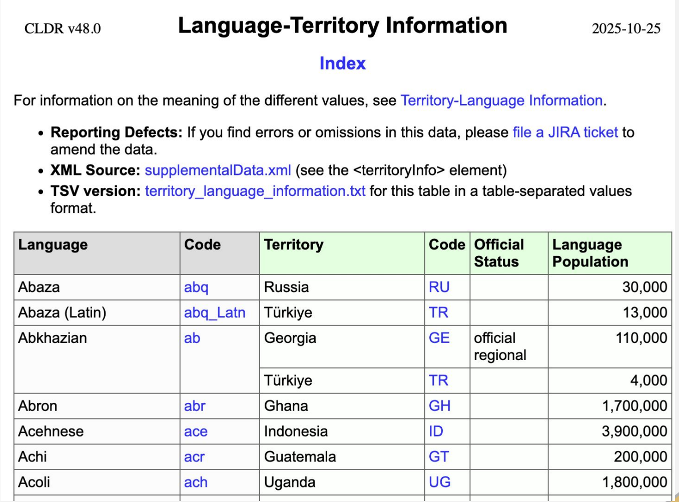
Roadmap

	Alpha June 2025	Beta Nov 2025	v1.0 2026
Data	Languages Languoids Writing Systems Territories	Vitality Digital Support Censuses Variants	Keyboards Lexical similarity Monolingualism In Education More Sources Fill in gaps
Interactions	Hovercards Table Hierarchy Simple Search	Search by Names Map Filters Export	Language Decoding Decision Trees Feedback API, Integrations Database-backed

Future

Partnering with UNESCO on their World Atlas of Languages

Integrating with Unicode CLDR Language x Territory Data



CLDR v48.0 **Language-Territory Information** 2025-10-25

Index

For information on the meaning of the different values, see Territory-Language Information.

- **Reporting Defects:** If you find errors or omissions in this data, please file a JIRA ticket to amend the data.
- **XML Source:** supplementalData.xml (see the <territoryInfo> element)
- **TSV version:** territory_language_information.txt for this table in a table-separated values format.

Language	Code	Territory	Code	Official Status	Language Population
Abaza	abq	Russia	RU		30,000
Abaza (Latin)	abq_Latn	Türkiye	TR		13,000
Abkhazian	ab	Georgia	GE	official regional	110,000
		Türkiye	TR		4,000
Abron	abr	Ghana	GH		1,700,000
Acehnese	ace	Indonesia	ID		3,900,000
Achi	acr	Guatemala	GT		200,000
Acoli	ach	Uganda	UG		1,800,000

Your contributions

Send feedback, provide data, join the project

conrad@translationcommons.org

https://github.com/Translation-Commons/lang-nav/

slides

demo

Thank you

conrad@translationcommons.org

TRANSLATION
COMMONS

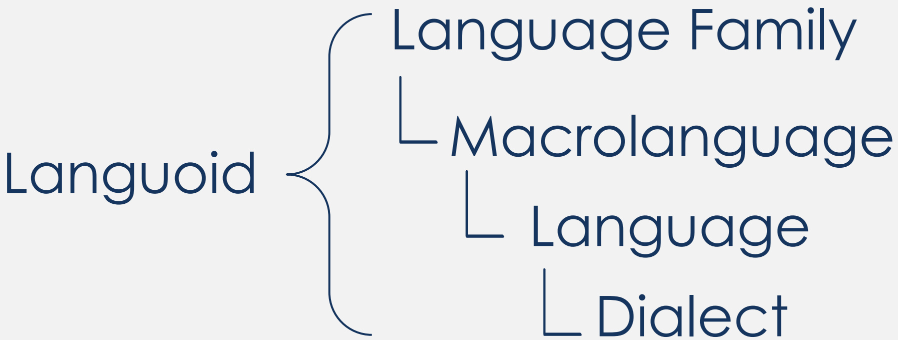
Extra slides

Common Pitfalls

1. What’s a Language
   a. Families
   b. Macro-languages
   c. Dialects, Orthographic differences
2. Incompatible Source Methodologies
   a. Spoken, Written, Used at Home
3. Language Identification
   a. Language Codes
   b. Language names

Pitfall 1: What’s a Language?

Depending on the source, data presented may not refer to just 1 mutually-intelligible, single-dictionary, single-spelling, single-pronunciation way of communicating. Different



Languoid

Language Family

Macrolanguage

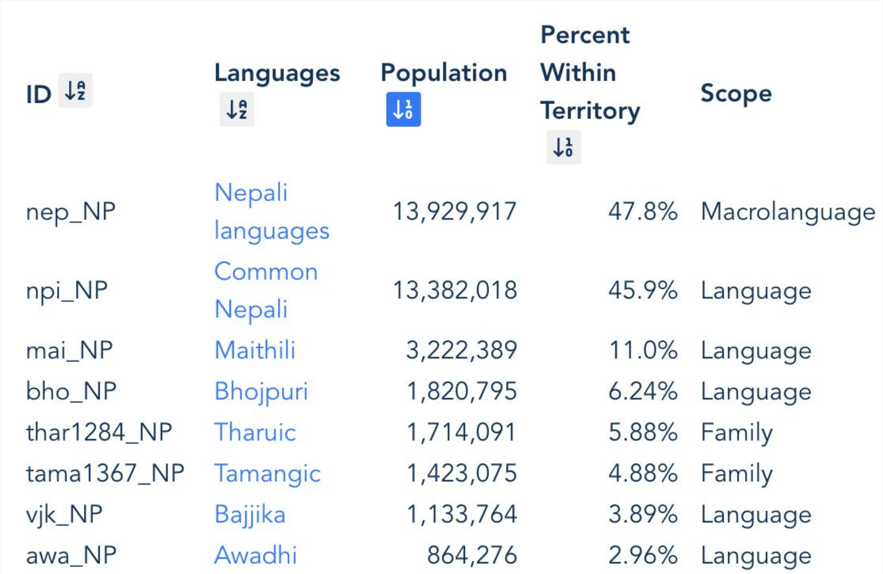
Language

Dialect

Pitfall 1a: Language Families

When getting language data, rows of data may refer to specific languages but may also reference language families.

For example, Tharu and Tamang are line items Nepal census 2021 but are considered language families. [link](#)



ID	Languages	Population	Percent Within Territory	Scope
nep_NP	Nepali languages	13,929,917	47.8%	Macrolanguage
npi_NP	Common Nepali	13,382,018	45.9%	Language
mai_NP	Maithili	3,222,389	11.0%	Language
bho_NP	Bhojpuri	1,820,795	6.24%	Language
thar1284_NP	Tharuic	1,714,091	5.88%	Family
tama1367_NP	Tamangic	1,423,075	4.88%	Family
vjk_NP	Bajjika	1,133,764	3.89%	Language
awa_NP	Awadhi	864,276	2.96%	Language

Pitfall 1b: Macrolanguages

Macrolanguage grouping of semi-mutually intelligible languages, mostly from a shared history. This does not mean you can translate one and you

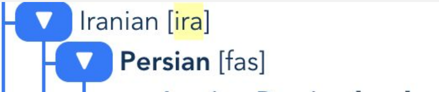
fa Persian formally includes **pes** Iranian Persian & **prs** Dari but not **tgk** Tajik.

See the Malayic languages and Chinese languages.

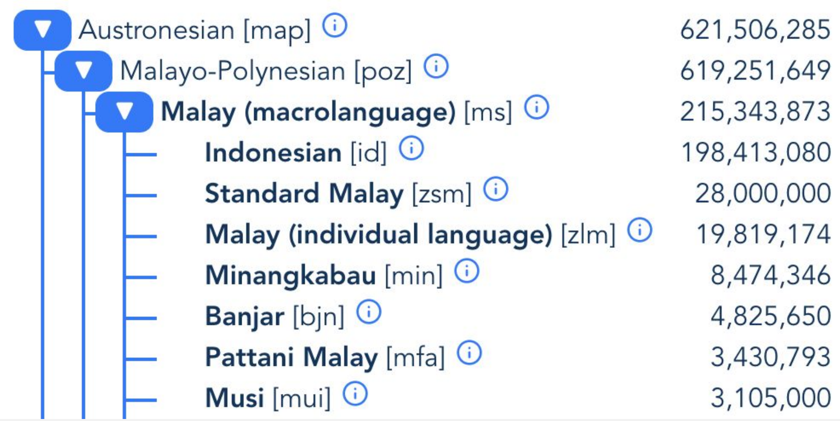
In CLDR…

zh ≠ Chinese → **cmn** Mandarin.

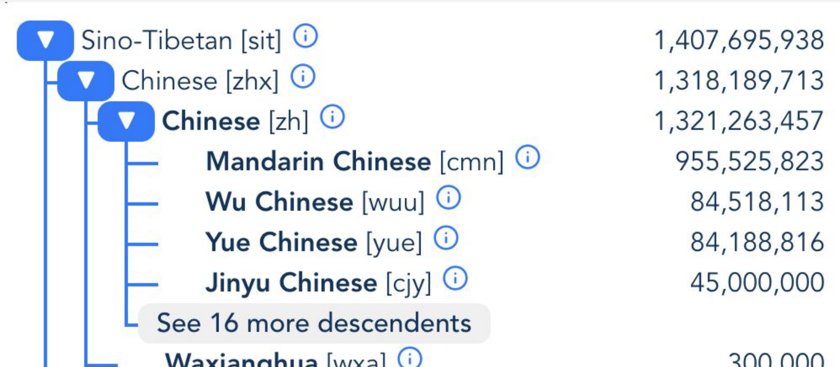
ms ≠ Chinese → **zsm** Standard Malay.



Iranian [ira]
Persian [fas]



Austronesian [map] ⓘ	621,506,285
Malayo-Polynesian [poz] ⓘ	619,251,649
Malay (macrolanguage) [ms] ⓘ	215,343,873
Indonesian [id] ⓘ	198,413,080
Standard Malay [zsm] ⓘ	28,000,000
Malay (individual language) [zlm] ⓘ	19,819,174
Minangkabau [min] ⓘ	8,474,346
Banjar [bjn] ⓘ	4,825,650
Pattani Malay [mfa] ⓘ	3,430,793
Musi [mui] ⓘ	3,105,000



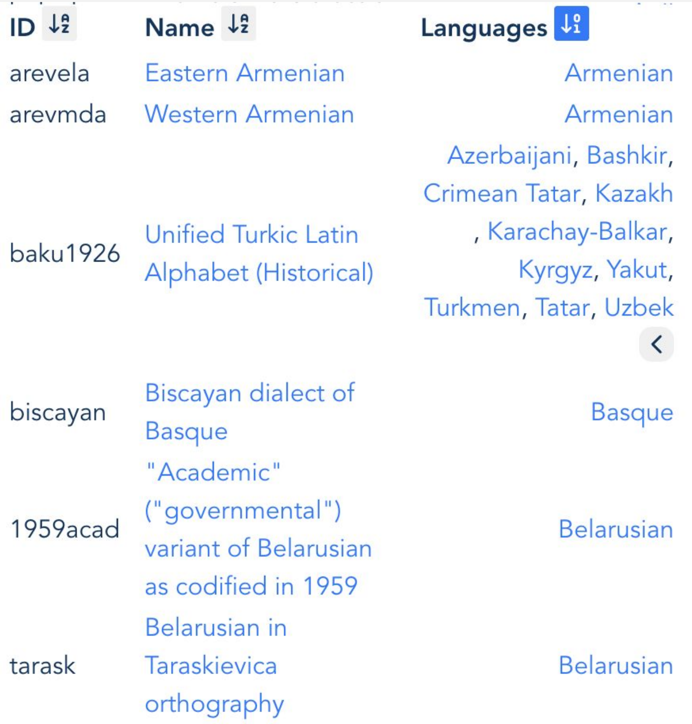
Sino-Tibetan [sit] ⓘ	1,407,695,938
Chinese [zhx] ⓘ	1,318,189,713
Chinese [zh] ⓘ	1,321,263,457
Mandarin Chinese [cmn] ⓘ	955,525,823
Wu Chinese [wuu] ⓘ	84,518,113
Yue Chinese [yue] ⓘ	84,188,816
Jinyu Chinese [cjy] ⓘ	45,000,000
See 16 more descendents	
Waxianghua [wxa] ⓘ	300,000

Pitfall 1c: Dialects & Orthographic Variants

Even in 1 language, there could be significant differences in standard usage.

Does this impact business languages? Usually not, pick the standard (if such exists). But it impacts personal usage, spelling, search results,

Many are registered in IANA language subtag registry but coverage is spotty, all requests are manual. You can explore them here.



ID	Name	Languages
arevela	Eastern Armenian	Armenian
arevmda	Western Armenian	Armenian
baku1926	Unified Turkic Latin Alphabet (Historical)	Azerbaijani, Bashkir, Crimean Tatar, Kazakh, Karachay-Balkar, Kyrgyz, Yakut, Turkmen, Tatar, Uzbek
biscayan	Biscayan dialect of Basque	Basque
1959acad	"Academic" ("governmental") variant of Belarusian as codified in 1959	Belarusian
tarask	Belarusian in Taraskievica orthography	Belarusian

What kind of service is it?

Writing messages to friends and family

Reading currencies

Searching for information

Reporting urgent information

Exploring an interface

Reading technical documentation

Concepts

Which language ID scheme?

- ISO 639-3: 7,920 entries, (website)
  - Used by most major Tech companies, Ethnologue
  - English [eng], Italian [ita]
- BCP-47: 7,920 languages, >80 million possible combinations
  - ISO codes but uses the ISO 639-1 codes when available
  - Eg. English: [en], Italian [it]
  - Includes mechanism to combine script, region, & variant data to express more languages
- Glottolog (Glottocode): 8,605 entries (website)
  - Eg. English [stan1293], Italian [ital1282]
- Linguist List: ???

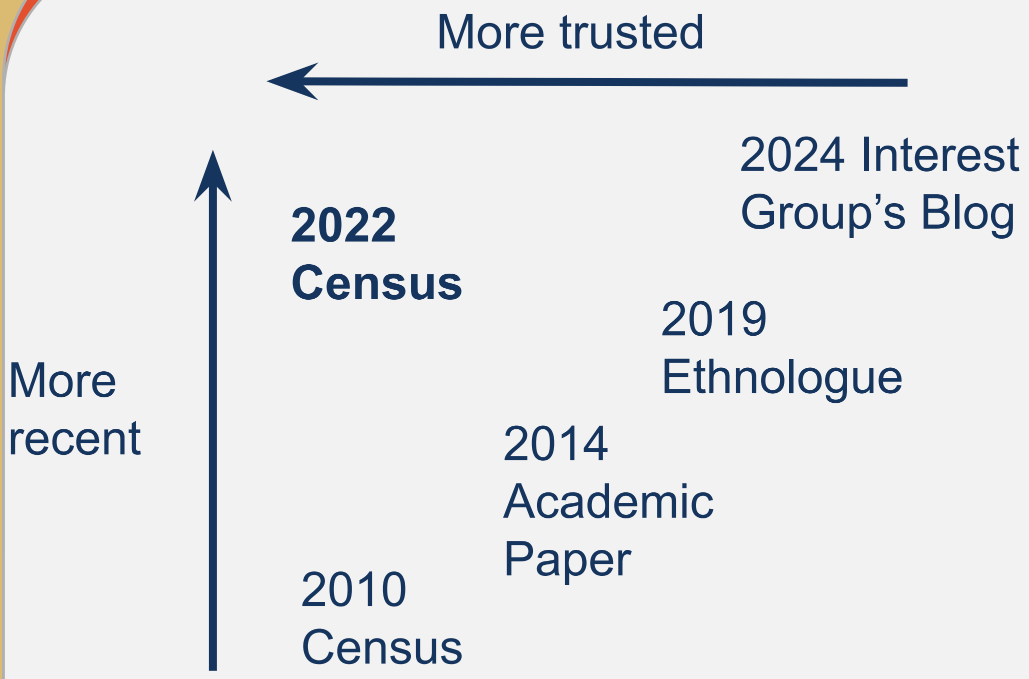
Locale Codes

Most locales are combinations of an ISO 639 language + ISO 3166 territory but there often are cases where we use unconventional codes. All ISO languages have 3-letter codes but if there's a 2-letter ISO code we use that (en, not eng).

Generally, we consider this locale format the "BCP-47" standard (Best Common Practices).

Special difference	**Examples**
Typical locale ISO-639 & ISO-3166	**en_GB:** English in the United Kingdom **de_CH**: High German in Switzerland **gsw_CH**: Swiss German in Switzerland
UN M49 Region	**eo_001**: Esperanto across the world **es_419**: Spanish, Latin America
ISO 15924 Script Code	**zh_Hant_TW**: Chinese in Traditional Han writing for Taiwan **zh_Hans_CN**: Chinese in Simplified Han writing for China
IANA-registered variant code	**rm_CH_VALLEDAR**: Rumantsch in Switzerland in the Valledar dialect **ca_ES_VALENCIA**: Catalan in Spain in the Valencian dialect
Other variant codes, see https://unicode.org/reports/tr35/#unicode-bcp-47-u-extension	

Picking the best language record



More trusted

More recent

2024 Interest Group’s Blog

2022 Census

2019 Ethnologue

2014 Academic Paper

2010 Census

When picking the best record, strike a balance between recency and the quality of the source.

Sometimes government sources are missing, missing the language you are trying to find, or only categorize the ethnic group not the actual language users. In those cases it may be better to go with a non-governmental source.

Always aim to get 1) a primary source and 2) a source without an agenda to make their numbers look a particular way.

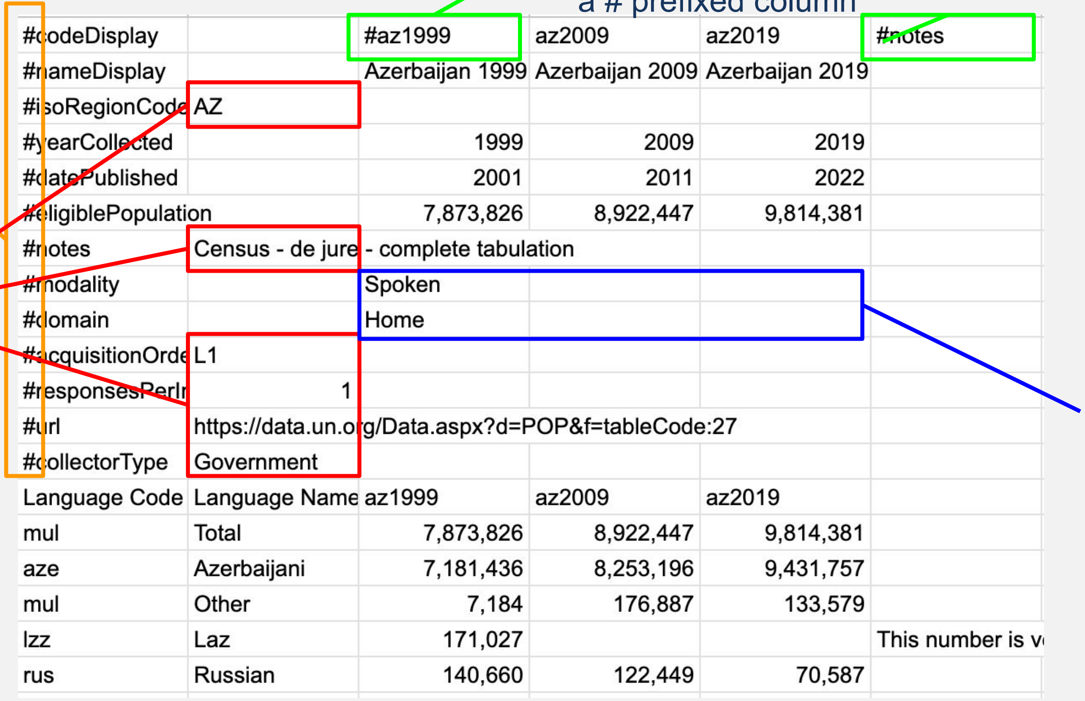
the Azerbaijan metadata

data.un.org

Census Metadata breakdown

Census table columns prefixed with # are not imported but good to leave in for context. Optional: Add a notes column with a # prefixed column

Metadata fields all start with a #

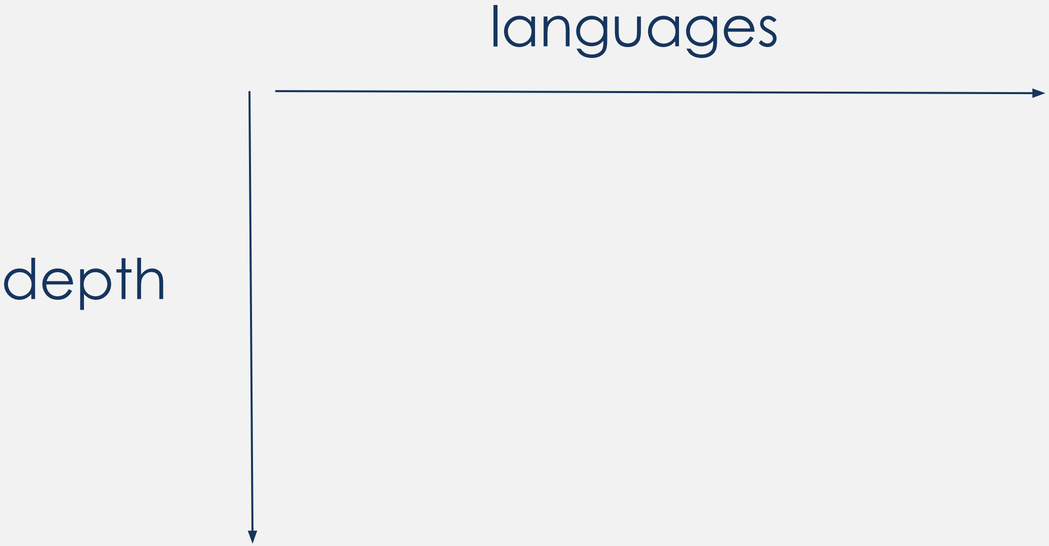


#codeDisplay		#az1999	az2009	az2019	#notes
#nameDisplay		Azerbaijan 1999	Azerbaijan 2009	Azerbaijan 2019	
#isoRegionCode	AZ				
#yearCollected		1999	2009	2019	
#datePublished		2001	2011	2022	
#eligiblePopulation		7,873,826	8,922,447	9,814,381	
#notes	Census - de jure - complete tabulation				
#modality		Spoken			
#domain		Home			
#acquisitionOrde	L1				
#responsesPerl	1				
#url	https://data.un.org/Data.aspx?d=POP&f=tableCode:27				
#collectorType	Government				
Language Code	Language Name	az1999	az2009	az2019	
mul	Total	7,873,826	8,922,447	9,814,381	
aze	Azerbaijani	7,181,436	8,253,196	9,431,757	
mul	Other	7,184	176,887	133,579	
lzz	Laz	171,027			This number is v
rus	Russian	140,660	122,449	70,587	

Fields that are the same for every table should be in column 2

Metadata that's missing is not included, eg. we don't know the "age" of people interviewed by this census so the metadata row is left out.

Leave cells empty when you don't know. Here you can see the language is "Spoken at Home" in 1999 but we don't know for the data from 2009 or 2019



languages

depth