Typology: Lecture III
Databases and Sampling

Christian Bentz
University of Tübingen

May 3, 2017

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
Language Evolution: The Empirical Turn

Overview

- Phylogenetic
  - Ethnologue (20th edition)
  - Glottolog 3.0
- Structure
  - World Atlas of Language Structures (WALS)
  - AUTOTYP
- Phonetics/Phonology
  - UPSID
  - PHOIBLE
- Lexicon
  - Automated Similarity Judgement Program (ASJP)
  - World Loanword Database
- Others
  - D-PLACE
- Sampling Biases

Ethnologue (20th edition)

What is it?

- "A comprehensive reference work cataloging *all of the worlds known living languages*. Since 1951, the Ethnologue has been an active research project involving hundreds of linguists and other researchers around the world. It is widely regarded to be the most comprehensive source of information of its kind."
- **License**: Pay wall

[Simons, Gary F. and Charles D. Fennig (eds.). (2017). *Ethnologue: Languages of the World, Twentieth edition*. Dallas, Texas: SIL International]

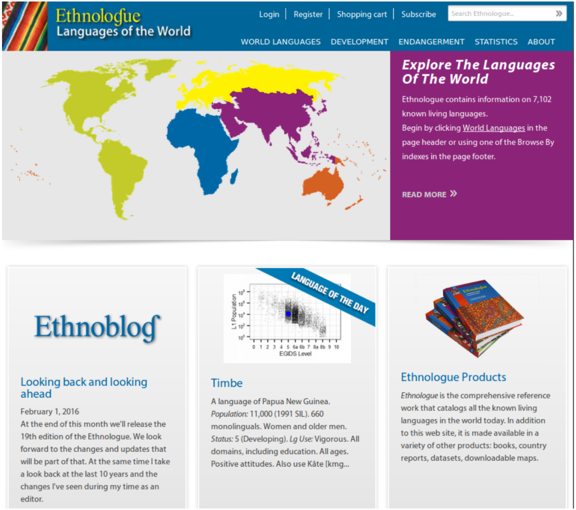
Ethnologue

What's in it?

- Languages: **7099**
- Information: Population sizes, language families and genera, endagerement status, etc.

[Simons, Gary F. and Charles D. Fennig (eds.). (2017). *Ethnologue: Languages of the World, Twentieth edition*. Dallas, Texas: SIL International]

Examples



Ethnologue
Languages of the World

Login | Register | Shopping cart | Subscribe
Search Ethnologue...
WORLD LANGUAGES DEVELOPMENT ENDANGERMENT STATISTICS ABOUT

Explore The Languages Of The World

Ethnologue contains information on 7,102 known living languages.
Begin by clicking World Languages in the page header or using one of the Browse By indexes in the page footer.

Ethnoblog

Looking back and looking ahead

February 1, 2016
At the end of this month we'll release the 19th edition of the Ethnologue. We look forward to the changes and updates that will be part of that. At the same time I take a look back at the last 10 years and the changes I've seen during my time as an editor.

Timbe

A language of Papua New Guinea. *Population:* 11,000 (1991 SIL). 660 monolinguals. Women and older men. *Status:* 5 (Developing). *Lg Use:* Vigorous. All domains, including education. All ages. Positive attitudes. Also use Kâte [kmg...

Ethnologue Products

Ethnologue is the comprehensive reference work that catalogs all the known living languages in the world today. In addition to this web site, it is made available in a variety of other products: books, country reports, datasets, downloadable maps.

Figure: https://www.ethnologue.com/

EXAMPLES

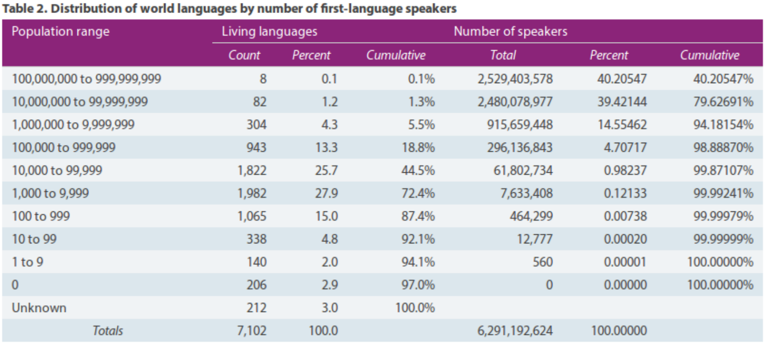
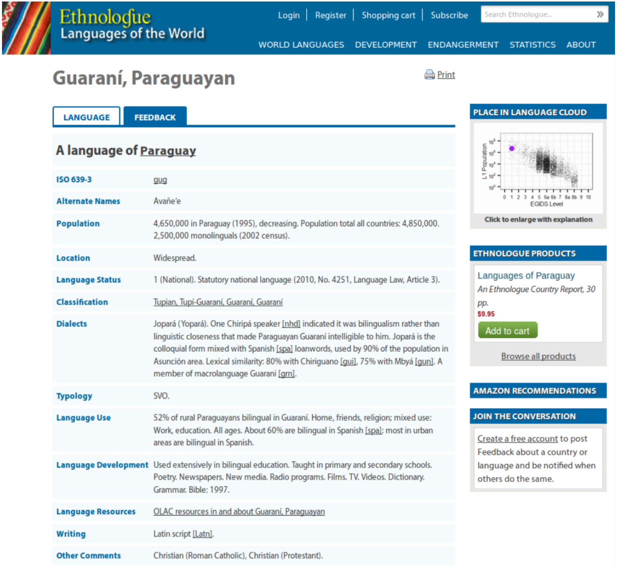


Table 2. Distribution of world languages by number of first-language speakers

Population range	Living languages			Number of speakers		
	Count	*Percent*	*Cumulative*	*Total*	*Percent*	*Cumulative*
100,000,000 to 999,999,999	8	0.1	0.1%	2,529,403,578	40.20547	40.20547%
10,000,000 to 99,999,999	82	1.2	1.3%	2,480,078,977	39.42144	79.62691%
1,000,000 to 9,999,999	304	4.3	5.5%	915,659,448	14.55462	94.18154%
100,000 to 999,999	943	13.3	18.8%	296,136,843	4.70717	98.88870%
10,000 to 99,999	1,822	25.7	44.5%	61,802,734	0.98237	99.87107%
1,000 to 9,999	1,982	27.9	72.4%	7,633,408	0.12133	99.99241%
100 to 999	1,065	15.0	87.4%	464,299	0.00738	99.99979%
10 to 99	338	4.8	92.1%	12,777	0.00020	99.99999%
1 to 9	140	2.0	94.1%	560	0.00001	100.00000%
0	206	2.9	97.0%	0	0.00000	100.00000%
Unknown	212	3.0	100.0%			
Totals	7,102	100.0		6,291,192,624	100.00000	

Figure: Ethnologue: Population Sizes

Examples



Ethnologue
Languages of the World

Login | Register | Shopping cart | Subscribe | Search Ethnologue...

WORLD LANGUAGES DEVELOPMENT ENDANGERMENT STATISTICS ABOUT

Guaraní, Paraguayan

Print

LANGUAGE FEEDBACK

A language of Paraguay

ISO 639-3	gug
Alternate Names	Avañe'e
Population	4,650,000 in Paraguay (1995), decreasing. Population total all countries: 4,850,000. 2,500,000 monolinguals (2002 census).
Location	Widespread.
Language Status	1 (National). Statutory national language (2010, No. 4251, Language Law, Article 3).
Classification	Tupian, Tupi-Guaraní, Guaraní, Guaraní
Dialects	Jopará (Yopará). One Chiripá speaker [nhd] indicated it was bilingualism rather than linguistic closeness that made Paraguayan Guaraní intelligible to him. Jopará is the colloquial form mixed with Spanish [spa] loanwords, used by 90% of the population in Asunción area. Lexical similarity: 80% with Chiriguano [gui], 75% with Mbyá [gun]. A member of macrolanguage Guarani [grn].
Typology	SVO.
Language Use	52% of rural Paraguayans bilingual in Guaraní. Home, friends, religion; mixed use: Work, education. All ages. About 60% are bilingual in Spanish [spa]; most in urban areas are bilingual in Spanish.
Language Development	Used extensively in bilingual education. Taught in primary and secondary schools. Poetry. Newspapers. New media. Radio programs. Films. TV. Videos. Dictionary. Grammar. Bible: 1997.
Language Resources	OLAC resources in and about Guaraní, Paraguayan
Writing	Latin script [Latn].
Other Comments	Christian (Roman Catholic), Christian (Protestant).

PLACE IN LANGUAGE CLOUD

Click to enlarge with explanation

ETHNOLOGUE PRODUCTS

Languages of Paraguay
An Ethnologue Country Report, 30 pp.
$9.95
Add to cart

Browse all products

AMAZON RECOMMENDATIONS

JOIN THE CONVERSATION

Create a free account to post Feedback about a country or language and be notified when others do the same.

Glottolog 3.0

What is it?

- “Comprehensive **reference information** for the world’s languages, especially the lesser known languages.”
- **License**: Open Access

[Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2017. Glottolog 3.0. Jena: Max Planck Institute for the Science of Human History.]

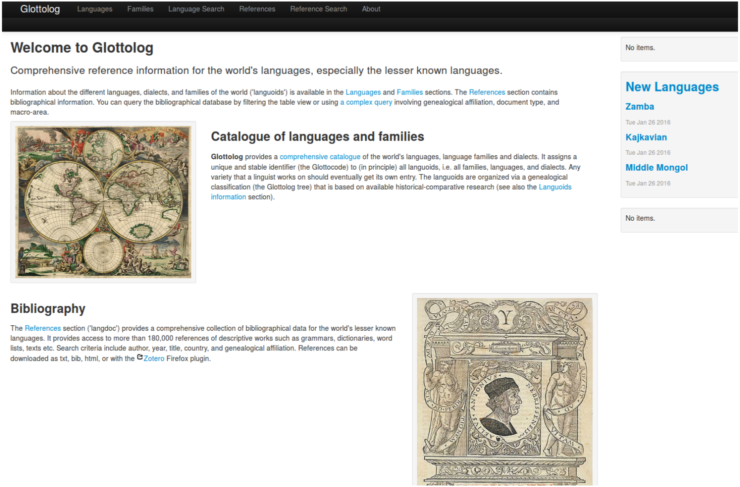
Glottolog 3.0

What's in it?

- Languages: **8444**
- Language Trees: **242** families, **188** isolates
- Bibliography: more than **180,000** references

[Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2017. Glottolog 3.0. Jena: Max Planck Institute for the Science of Human History.]

Examples



Welcome to Glottolog

Comprehensive reference information for the world's languages, especially the lesser known languages.

Information about the different languages, dialects, and families of the world ('languoids') is available in the Languages and Families sections. The References section contains bibliographical information. You can query the bibliographical database by filtering the table view or using a complex query involving genealogical affiliation, document type, and macro-area.

Catalogue of languages and families

Glottolog provides a comprehensive catalogue of the world's languages, language families and dialects. It assigns a unique and stable identifier (the Glottocode) to (in principle) all languoids, i.e. all families, languages, and dialects. Any variety that a linguist works on should eventually get its own entry. The languoids are organized via a genealogical classification (the Glottolog tree) that is based on available historical-comparative research (see also the Languoids information section).

Bibliography

The References section ('langdoc') provides a comprehensive collection of bibliographical data for the world's lesser known languages. It provides access to more than 180,000 references of descriptive works such as grammars, dictionaries, word lists, texts etc. Search criteria include author, year, title, country, and genealogical affiliation. References can be downloaded as txt, bib, html, or with the Zotero Firefox plugin.

No items.

New Languages

Zamba

Tue Jan 26 2016

Kajkavian

Tue Jan 26 2016

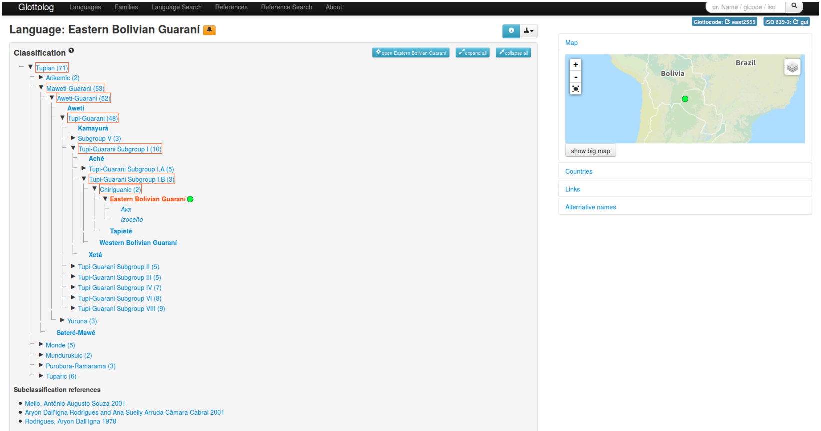
Middle Mongol

Tue Jan 26 2016

No items.

Figure: http://glottolog.org/

EXAMPLES



Language: Eastern Bolivian Guaraní

Classification

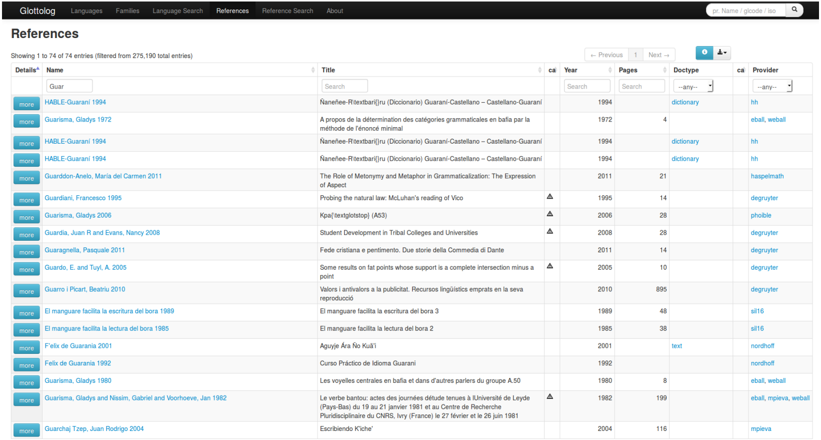
- ▼ Tupian (71)
  - ▶ Arikemic (2)
  - ▼ Maweti-Guarani (53)
    - ▼ Aweti-Guarani (52)
      - **Awetí**
      - ▼ Tupi-Guarani (48)
        - **Kamayurá**
        - ▶ Subgroup V (3)
        - ▼ Tupi-Guarani Subgroup I (10)
          - **Aché**
          - ▶ Tupi-Guarani Subgroup I.A (5)
          - ▼ Tupi-Guarani Subgroup I.B (3)
            - ▼ Chiriguanic (2)
              - ▼ **Eastern Bolivian Guaraní** ●
                - *Ava*
                - *Izoceño*
              - **Tapieté**
            - **Western Bolivian Guaraní**
          - **Xetá**
        - ▶ Tupi-Guarani Subgroup II (5)
        - ▶ Tupi-Guarani Subgroup III (5)
        - ▶ Tupi-Guarani Subgroup IV (7)
        - ▶ Tupi-Guarani Subgroup VI (8)
        - ▶ Tupi-Guarani Subgroup VIII (9)
      - ▶ Yuruna (3)
    - **Sateré-Mawé**
  - ▶ Monde (5)
  - ▶ Mundurukuic (2)
  - ▶ Purubora-Ramarama (3)
  - ▶ Tuparic (6)

Subclassification references

- Mello, Antônio Augusto Souza 2001
- Aryon Dall'Igna Rodrigues and Ana Suelly Arruda Câmara Cabral 2001
- Rodrigues, Aryon Dall'Igna 1978

Figure: Glottolog: Tree for Guarani

EXAMPLES



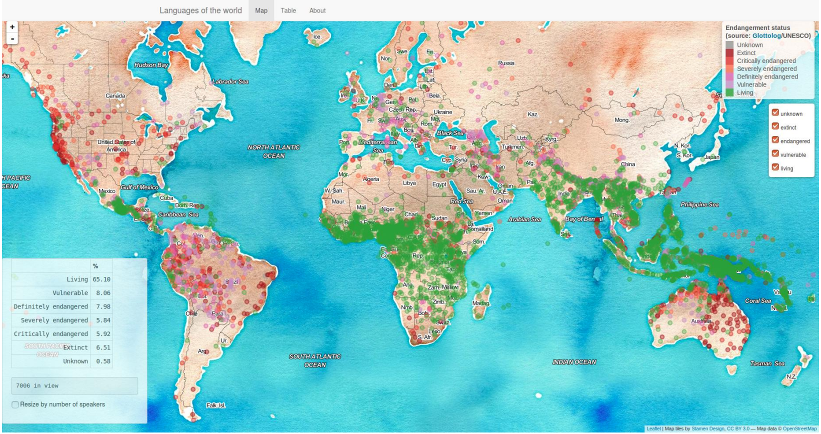
References

Showing 1 to 74 of 74 entries (filtered from 275,190 total entries)

Details	Name	Title	ca	Year	Pages	Doctype	ca	Provider
	Guar	Search		Search	Search	--any--		--any--
more	HABLE-Guaraní 1994	Ñaneñee-R\textbari{}ru (Diccionario) Guaraní-Castellano – Castellano-Guaraní		1994		dictionary		hh
more	Guarisma, Gladys 1972	A propos de la détermination des catégories grammaticales en bafia par la méthode de l'énoncé minimal		1972	4			eball, weball
more	HABLE-Guaraní 1994	Ñaneñee-R\textbari{}ru (Diccionario) Guaraní-Castellano – Castellano-Guaraní		1994		dictionary		hh
more	HABLE-Guaraní 1994	Ñaneñee-R\textbari{}ru (Diccionario) Guaraní-Castellano – Castellano-Guaraní		1994		dictionary		hh
more	Guarddon-Anelo, María del Carmen 2011	The Role of Metonymy and Metaphor in Grammaticalization: The Expression of Aspect		2011	21			haspelmath
more	Guardiani, Francesco 1995	Probing the natural law: McLuhan's reading of Vico	⚠	1995	14			degruyter
more	Guarisma, Gladys 2006	Kpa{\textglotstop} (A53)	⚠	2006	28			phoible
more	Guardia, Juan R and Evans, Nancy 2008	Student Development in Tribal Colleges and Universities	⚠	2008	28			degruyter
more	Guaragnella, Pasquale 2011	Fede cristiana e pentimento. Due storie della Commedia di Dante		2011	14			degruyter
more	Guardo, E. and Tuyl, A. 2005	Some results on fat points whose support is a complete intersection minus a point	⚠	2005	10			degruyter
more	Guarro i Picart, Beatriu 2010	Valors i antivalors a la publicitat. Recursos lingüístics emprats en la seva reproducció		2010	895			degruyter
more	El manguare facilita la escritura del bora 1989	El manguare facilita la escritura del bora 3		1989	48			sil16
more	El manguare facilita la lectura del bora 1985	El manguare facilita la lectura del bora 2		1985	38			sil16
more	F'elix de Guarania 2001	Aguyje Ára Ño Kuã'i		2001		text		nordhoff
more	Felix de Guarania 1992	Curso Práctico de Idioma Guarani		1992				nordhoff
more	Guarisma, Gladys 1980	Les voyelles centrales en bafia et dans d'autres parlers du groupe A.50		1980	8			eball, weball
more	Guarisma, Gladys and Nissim, Gabriel and Voorhoeve, Jan 1982	Le verbe bantou: actes des journées détude tenues à lUniversité de Leyde (Pays-Bas) du 19 au 21 janvier 1981 et au Centre de Recherche Pluridisciplinaire du CNRS, Ivry (France) le 27 février et le 26 juin 1981	⚠	1982	199			eball, mpieva, weball
more	Guarchaj Tzep, Juan Rodrigo 2004	Escribiendo K'iche'		2004	116			mpieva

Figure: Glottolog: References for Guarani

THE GLOTTOLOG DATA EXPLORER (HTTPS://CAINESAP.SHINYAPPS.IO/LANGMAP/)



☑ unknown
☑ extinct
☑ endangered
☑ vulnerable
☑ living

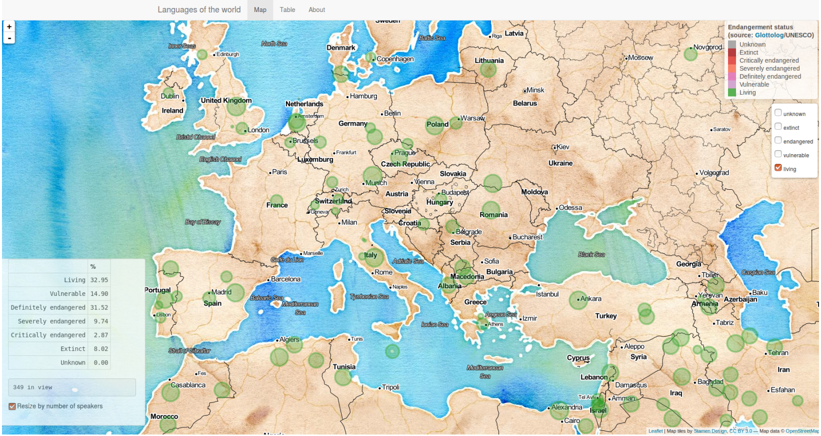
	%
Living	65.10
Vulnerable	8.06
Definitely endangered	7.98
Severely endangered	5.84
Critically endangered	5.92
Extinct	6.51
Unknown	0.58

7006 in view

□ Resize by number of speakers

[Caines, Andrew, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, and Paula Buttery (2016). The Glottolog Data Explorer: Mapping the worlds languages. Proceedings of the VisLR II Workshop at LREC'16.]

THE GLOTTOLOG DATA EXPLORER (HTTPS://CAINESAP.SHINYAPPS.IO/LANGMAP/)



Endangerment status
(source: Glottolog/UNESCO)

- ⬜ Unknown
- 🟥 Extinct
- 🟥 Critically endangered
- 🟧 Severely endangered
- 🟪 Definitely endangered
- 🟪 Vulnerable
- 🟩 Living

□ unknown
□ extinct
□ endangered
□ vulnerable
☑ living

	%
Living	32.95
Vulnerable	14.90
Definitely endangered	31.52
Severely endangered	9.74
Critically endangered	2.87
Extinct	8.02
Unknown	0.00

349 in view

☑ Resize by number of speakers

[Caines, Andrew, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, and Paula Buttery (2016). The Glottolog Data Explorer: Mapping the worlds languages. Proceedings of the VisLR II Workshop at LREC'16.]

WALS

What is it?

- "The **World Atlas of Language Structures (WALS)** is a large *database of structural (phonological, grammatical, lexical) properties of languages* gathered from descriptive materials (such as reference grammars) by a team of 55 authors"
- **License**: Open Access

[Dryer & Haspelmath (eds.), 2013]

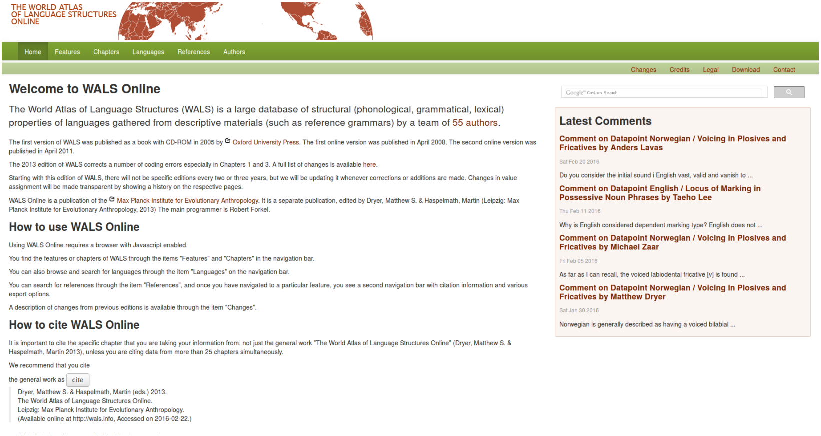
WALS

What's in it?

- Languages: **2679** (ca. 37% of the world's languages)
- Features: **192** (Phonology, Morphology, Word Order, etc.)
- Chapters: **151** (Consonant Inventories, Number of Cases, Verbal Inflection, etc.)

[Dryer & Haspelmath (eds.), 2013]

Examples



Welcome to WALS Online

The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors.

The first version of WALS was published as a book with CD-ROM in 2005 by Oxford University Press. The first online version was published in April 2008. The second online version was published in April 2011.

The 2013 edition of WALS corrects a number of coding errors especially in Chapters 1 and 3. A full list of changes is available here.

Starting with this edition of WALS, there will not be specific editions every two or three years, but we will be updating it whenever corrections or additions are made. Changes in value assignment will be made transparent by showing a history on the respective pages.

WALS Online is a publication of the Max Planck Institute for Evolutionary Anthropology. It is a separate publication, edited by Dryer, Matthew S. & Haspelmath, Martin (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013) The main programmer is Robert Forkel.

How to use WALS Online

Using WALS Online requires a browser with Javascript enabled.

You find the features or chapters of WALS through the items "Features" and "Chapters" in the navigation bar.

You can also browse and search for languages through the item "Languages" on the navigation bar.

You can search for references through the item "References", and once you have navigated to a particular feature, you see a second navigation bar with citation information and various export options.

A description of changes from previous editions is available through the item "Changes".

How to cite WALS Online

It is important to cite the specific chapter that you are taking your information from, not just the general work "The World Atlas of Language Structures Online" (Dryer, Matthew S. & Haspelmath, Martin 2013), unless you are citing data from more than 25 chapters simultaneously.

We recommend that you cite

the general work as

> Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013.
> The World Atlas of Language Structures Online.
> Leipzig: Max Planck Institute for Evolutionary Anthropology.
> (Available online at http://wals.info, Accessed on 2016-02-22.)

Latest Comments

Comment on Datapoint Norwegian / Voicing in Plosives and Fricatives by Anders Lavas

Sat Feb 20 2016

Do you consider the initial sound i English vast, valid and vanish to ...

Comment on Datapoint English / Locus of Marking in Possessive Noun Phrases by Taeho Lee

Thu Feb 11 2016

Why is English considered dependent marking type? English does not ...

Comment on Datapoint Norwegian / Voicing in Plosives and Fricatives by Michael Zaar

Fri Feb 05 2016

As far as I can recall, the voiced labiodental fricative [v] is found ...

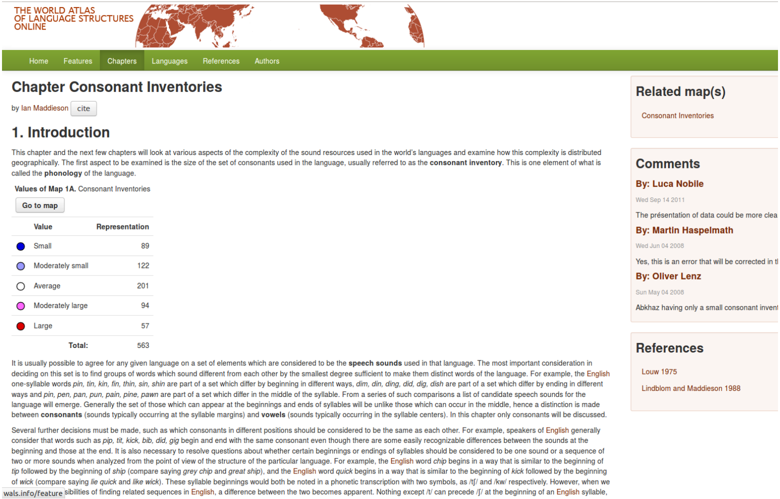
Comment on Datapoint Norwegian / Voicing in Plosives and Fricatives by Matthew Dryer

Sat Jan 30 2016

Norwegian is generally described as having a voiced bilabial ...

Figure: http://wals.info/

EXAMPLES



THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Chapter Consonant Inventories

by Ian Maddieson cite

1. Introduction

This chapter and the next few chapters will look at various aspects of the complexity of the sound resources used in the world's languages and examine how this complexity is distributed geographically. The first aspect to be examined is the size of the set of consonants used in the language, usually referred to as the **consonant inventory**. This is one element of what is called the **phonology** of the language.

Values of Map 1A. Consonant Inventories

Go to map

	Value	Representation
●	Small	89
●	Moderately small	122
○	Average	201
●	Moderately large	94
●	Large	57
	Total:	563

It is usually possible to agree for any given language on a set of elements which are considered to be the **speech sounds** used in that language. The most important consideration in deciding on this set is to find groups of words which sound different from each other by the smallest degree sufficient to make them distinct words of the language. For example, the English one-syllable words *pin, tin, kin, fin, thin, sin, shin* are part of a set which differ by beginning in different ways, *dim, din, ding, did, dig, dish* are part of a set which differ by ending in different ways and *pin, pen, pan, pun, pain, pine, pawn* are part of a set which differ in the middle of the syllable. From a series of such comparisons a list of candidate speech sounds for the language will emerge. Generally the set of those which can appear at the beginnings and ends of syllables will be unlike those which can occur in the middle, hence a distinction is made between **consonants** (sounds typically occurring at the syllable margins) and **vowels** (sounds typically occurring in the syllable centers). In this chapter only consonants will be discussed.

Several further decisions must be made, such as which consonants in different positions should be considered to be the same as each other. For example, speakers of English generally consider that words such as *pip, tit, kick, bib, did, gig* begin and end with the same consonant even though there are some easily recognizable differences between the sounds at the beginning and those at the end. It is also necessary to resolve questions about whether certain beginnings or endings of syllables should be considered to be one sound or a sequence of two or more sounds when analyzed from the point of view of the structure of the particular language. For example, the English word *chip* begins in a way that is similar to the beginning of *tip* followed by the beginning of *ship* (compare saying *grey chip* and *great ship*), and the English word *quick* begins in a way that is similar to the beginning of *kick* followed by the beginning of *wick* (compare saying *lie quick* and *like wick*). These syllable beginnings would both be noted in a phonetic transcription with two symbols, as /tʃ/ and /kw/ respectively. However, when we

wals.info/feature sibilities of finding related sequences in English, a difference between the two becomes apparent. Nothing except /t/ can precede /ʃ/ at the beginning of an English syllable,

Related map(s)

Consonant Inventories

Comments

By: Luca Nobile

Wed Sep 14 2011

The présentation of data could be more clea

By: Martin Haspelmath

Wed Jun 04 2008

Yes, this is an error that will be corrected in t

By: Oliver Lenz

Sun May 04 2008

Abkhaz having only a small consonant invent

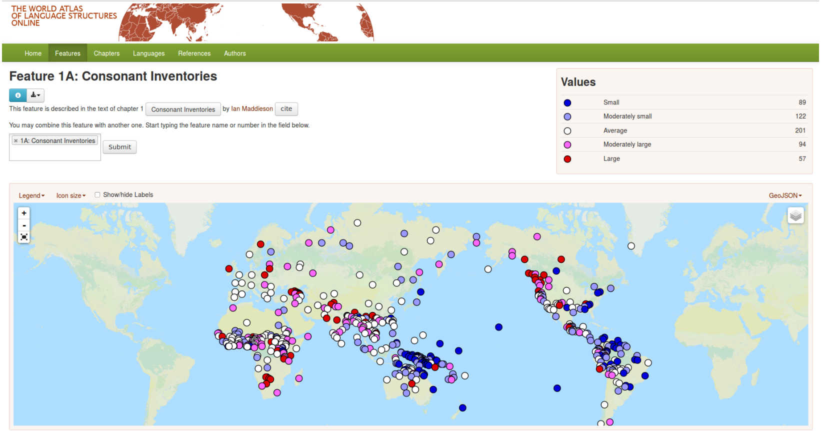
References

Louw 1975

Lindblom and Maddieson 1988

Figure: Chapter 1: Consonant Inventories

Examples



Feature 1A: Consonant Inventories

This feature is described in the text of chapter 1 Consonant Inventories by Ian Maddieson cite

You may combine this feature with another one. Start typing the feature name or number in the field below.

Values

●	Small	89
●	Moderately small	122
○	Average	201
●	Moderately large	94
●	Large	57

Figure: Feature 1A: Consonant Inventories

AUTOTYP

What is it?

- “AUTOTYP is a large-scale research program with goals in both quantitative and qualitative typology. In quantitative typology, we are interested in detecting and explaining geographical distributions of typological features and in producing statistical estimates of universal preferences as well as of genealogical inheritance and areal diffusion potentials.”
- **License**: Open Access

[Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. The autotyp genealogy and geography database: 2013 release. http://www.spw.uzh.ch/autotyp/.]

AUTOTYP

What's in it?

- Genealogical and Geographic Information for **2914** languages
- Phonological, morphological and syntactic information on several hundred languages (though most of this information is published in WALS as well)
- There is supposed to be a new release this year which should give more extensive grammatical information

[Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. The autotyp genealogy and geography database: 2013 release. http://www.spw.uzh.ch/autotyp/.]

SSWL

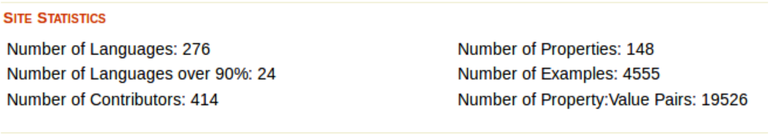
What is it?

- "The **Syntactic Structures of the World's Languages (SSWL)** is a searchable database that allows users to discover which properties (morphological, syntactic, and semantic) characterize a language, as well as how these properties relate across languages. This system is designed to be free to the public and open-ended. Anyone can use the database to perform queries."
- **License**: Open Access

http://sswl.railsplayground.net/

SSWL

What's in it?

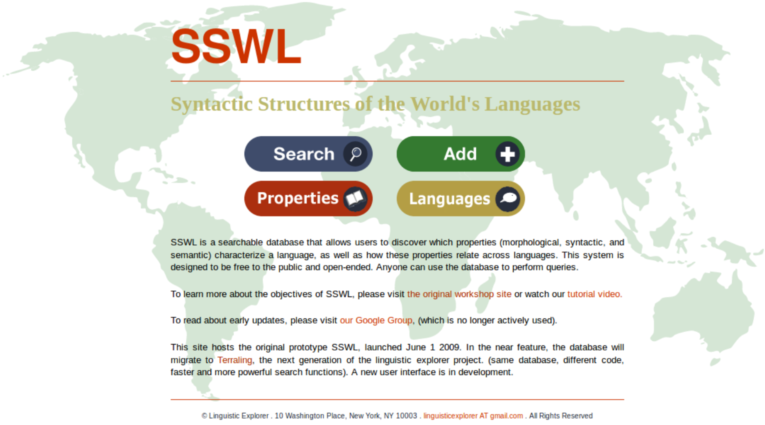


SITE STATISTICS

Number of Languages: 276
Number of Languages over 90%: 24
Number of Contributors: 414

Number of Properties: 148
Number of Examples: 4555
Number of Property:Value Pairs: 19526

Examples



SSWL

Syntactic Structures of the World's Languages

SSWL is a searchable database that allows users to discover which properties (morphological, syntactic, and semantic) characterize a language, as well as how these properties relate across languages. This system is designed to be free to the public and open-ended. Anyone can use the database to perform queries.

To learn more about the objectives of SSWL, please visit the original workshop site or watch our tutorial video.

To read about early updates, please visit our Google Group, (which is no longer actively used).

This site hosts the original prototype SSWL, launched June 1 2009. In the near feature, the database will migrate to Terraling, the next generation of the linguistic explorer project. (same database, different code, faster and more powerful search functions). A new user interface is in development.

© Linguistic Explorer . 10 Washington Place, New York, NY 10003 . linguisticexplorer AT gmail.com . All Rights Reserved

URIEL

What is it?

- "The URIEL knowledge base is a structured compendium of information on language typology and language universals"
- **License**: Open Access

[Littel, Patrick & Mortensen, David & Levin, Lori (eds.) 2016. URIEL Typological Database Pittsburgh: Carnegie Mellon University (Available online at http://www.cs.cmu.edu/ dmortens/uriel.html, Accessed at 2016-04-20).]

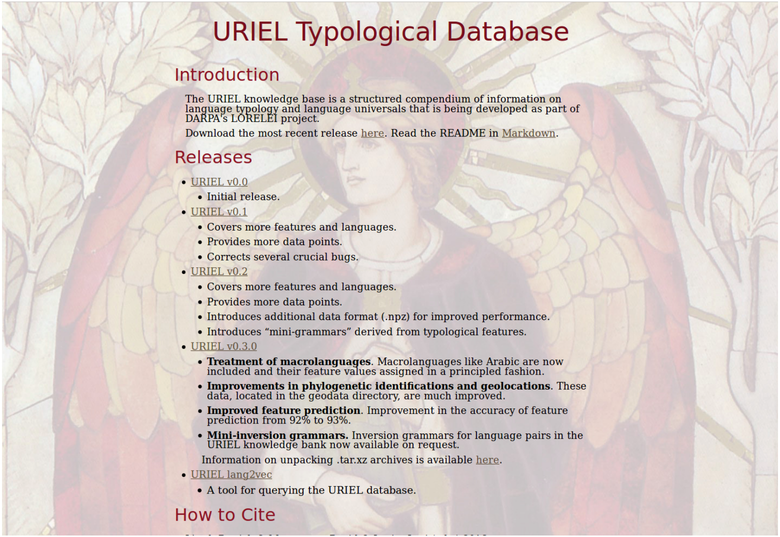
URIEL

What's in it?

- Geographic and genealogical information for **7971** languages
- Binary vectors of grammatical and phonological information from PHOIBLE, WALS, and SSWL

[Littel, Patrick & Mortensen, David & Levin, Lori (eds.) 2016. URIEL Typological Database Pittsburgh: Carnegie Mellon University (Available online at http://www.cs.cmu.edu/ dmortens/uriel.html, Accessed at 2016-04-20).]

Examples



URIEL Typological Database

Introduction

The URIEL knowledge base is a structured compendium of information on language typology and language universals that is being developed as part of DARPA's LORELEI project.

Download the most recent release here. Read the README in Markdown.

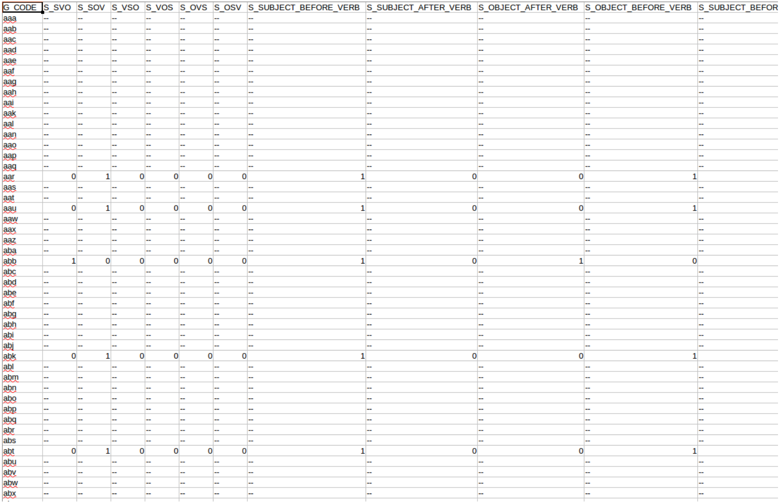
Releases

- URIEL v0.0
  - Initial release.
- URIEL v0.1
  - Covers more features and languages.
  - Provides more data points.
  - Corrects several crucial bugs.
- URIEL v0.2
  - Covers more features and languages.
  - Provides more data points.
  - Introduces additional data format (.npz) for improved performance.
  - Introduces “mini-grammars” derived from typological features.
- URIEL v0.3.0
  - **Treatment of macrolanguages**. Macrolanguages like Arabic are now included and their feature values assigned in a principled fashion.
  - **Improvements in phylogenetic identifications and geolocations**. These data, located in the geodata directory, are much improved.
  - **Improved feature prediction**. Improvement in the accuracy of feature prediction from 92% to 93%.
  - **Mini-inversion grammars.** Inversion grammars for language pairs in the URIEL knowledge bank now available on request.

  Information on unpacking .tar.xz archives is available here.
- URIEL lang2vec
  - A tool for querying the URIEL database.

How to Cite

EXAMPLES



G_CODE	S_SVO	S_SOV	S_VSO	S_VOS	S_OVS	S_OSV	S_SUBJECT_BEFORE_VERB	S_SUBJECT_AFTER_VERB	S_OBJECT_AFTER_VERB	S_OBJECT_BEFORE_VERB	S_SUBJECT_BEFOR
aaa	--	--	--	--	--	--	--	--	--	--	--
aab	--	--	--	--	--	--	--	--	--	--	--
aac	--	--	--	--	--	--	--	--	--	--	--
aad	--	--	--	--	--	--	--	--	--	--	--
aae	--	--	--	--	--	--	--	--	--	--	--
aaf	--	--	--	--	--	--	--	--	--	--	--
aag	--	--	--	--	--	--	--	--	--	--	--
aah	--	--	--	--	--	--	--	--	--	--	--
aai	--	--	--	--	--	--	--	--	--	--	--
aak	--	--	--	--	--	--	--	--	--	--	--
aal	--	--	--	--	--	--	--	--	--	--	--
aan	--	--	--	--	--	--	--	--	--	--	--
aao	--	--	--	--	--	--	--	--	--	--	--
aap	--	--	--	--	--	--	--	--	--	--	--
aaq	--	--	--	--	--	--	--	--	--	--	--
aar	0	1	0	0	0	0	1	0	0	1	
aas	--	--	--	--	--	--	--	--	--	--	--
aat	--	--	--	--	--	--	--	--	--	--	--
aau	0	1	0	0	0	0	1	0	0	1	
aaw	--	--	--	--	--	--	--	--	--	--	--
aax	--	--	--	--	--	--	--	--	--	--	--
aaz	--	--	--	--	--	--	--	--	--	--	--
aba	--	--	--	--	--	--	--	--	--	--	--
abb	1	0	0	0	0	0	1	0	1	0	
abc	--	--	--	--	--	--	--	--	--	--	--
abd	--	--	--	--	--	--	--	--	--	--	--
abe	--	--	--	--	--	--	--	--	--	--	--
abf	--	--	--	--	--	--	--	--	--	--	--
abg	--	--	--	--	--	--	--	--	--	--	--
abh	--	--	--	--	--	--	--	--	--	--	--
abi	--	--	--	--	--	--	--	--	--	--	--
abj	--	--	--	--	--	--	--	--	--	--	--
abk	0	1	0	0	0	0	1	0	0	1	
abl	--	--	--	--	--	--	--	--	--	--	--
abm	--	--	--	--	--	--	--	--	--	--	--
abn	--	--	--	--	--	--	--	--	--	--	--
abo	--	--	--	--	--	--	--	--	--	--	--
abp	--	--	--	--	--	--	--	--	--	--	--
abq	--	--	--	--	--	--	--	--	--	--	--
abr	--	--	--	--	--	--	--	--	--	--	--
abs	--	--	--	--	--	--	--	--	--	--	--
abt	0	1	0	0	0	0	1	0	0	1	
abu	--	--	--	--	--	--	--	--	--	--	--
abv	--	--	--	--	--	--	--	--	--	--	--
abw	--	--	--	--	--	--	--	--	--	--	--
abx	--	--	--	--	--	--	--	--	--	--	--

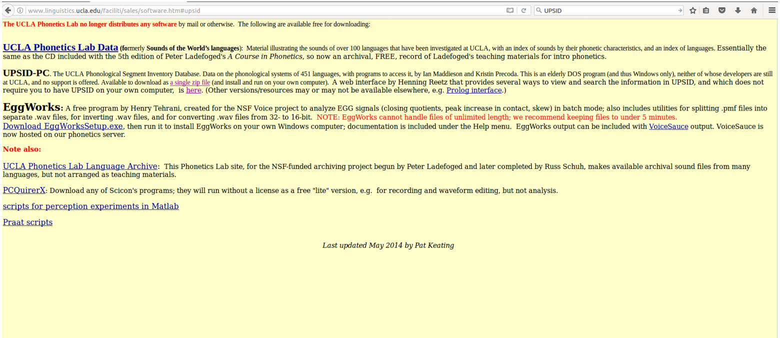
UPSID (UCLA Phonological Segment Inventory Database)

What is it?

- “This Database was compiled by Ian Maddieson and Kristin Precoda (cf. Maddieson, 1984) and contains information on the distribution of 919 different segments in 451 languages”
- **License**: Open Access

http://www.linguistics.ucla.edu/faciliti/sales/software.html

Examples



The UCLA Phonetics Lab no longer distributes any software by mail or otherwise. The following are available free for downloading:

UCLA Phonetics Lab Data (formerly **Sounds of the World's languages**): Material illustrating the sounds of over 100 languages that have been investigated at UCLA, with an index of sounds by their phonetic characteristics, and an index of languages. Essentially the same as the CD included with the 5th edition of Peter Ladefoged's *A Course in Phonetics*, so now an archival, FREE, record of Ladefoged's teaching materials for intro phonetics.

UPSID-PC. The UCLA Phonological Segment Inventory Database. Data on the phonological systems of 451 languages, with programs to access it, by Ian Maddieson and Kristin Precoda. This is an elderly DOS program (and thus Windows only), neither of whose developers are still at UCLA, and no support is offered. Available to download as a single zip file (and install and run on your own computer). A web interface by Henning Reetz that provides several ways to view and search the information in UPSID, and which does not require you to have UPSID on your own computer, is here. (Other versions/resources may or may not be available elsewhere, e.g. Prolog interface.)

EggWorks: A free program by Henry Tehrani, created for the NSF Voice project to analyze EGG signals (closing quotients, peak increase in contact, skew) in batch mode; also includes utilities for splitting .pmf files into separate .wav files, for inverting .wav files, and for converting .wav files from 32- to 16-bit. NOTE: EggWorks cannot handle files of unlimited length; we recommend keeping files to under 5 minutes.
Download EggWorksSetup.exe, then run it to install EggWorks on your own Windows computer; documentation is included under the Help menu. EggWorks output can be included with VoiceSauce output. VoiceSauce is now hosted on our phonetics server.

Note also:

UCLA Phonetics Lab Language Archive: This Phonetics Lab site, for the NSF-funded archiving project begun by Peter Ladefoged and later completed by Russ Schuh, makes available archival sound files from many languages, but not arranged as teaching materials.

PCQuirerX: Download any of Scicon's programs; they will run without a license as a free "lite" version, e.g. for recording and waveform editing, but not analysis.

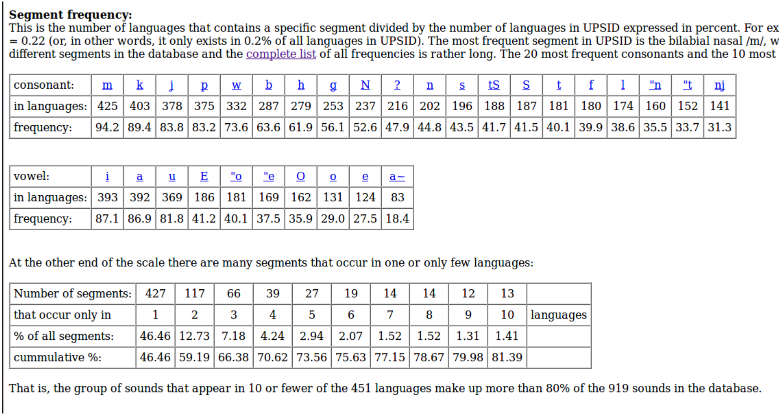
scripts for perception experiments in Matlab

Praat scripts

Last updated May 2014 by Pat Keating

http://www.linguistics.ucla.edu/faciliti/sales/software.html

Examples



Segment frequency:
This is the number of languages that contains a specific segment divided by the number of languages in UPSID expressed in percent. For ex = 0.22 (or, in other words, it only exists in 0.2% of all languages in UPSID). The most frequent segment in UPSID is the bilabial nasal /m/, w different segments in the database and the complete list of all frequencies is rather long. The 20 most frequent consonants and the 10 most

consonant:	m	k	j	p	w	b	h	g	N	?	n	s	tS	S	t	f	l	"n	"t	nj
in languages:	425	403	378	375	332	287	279	253	237	216	202	196	188	187	181	180	174	160	152	141
frequency:	94.2	89.4	83.8	83.2	73.6	63.6	61.9	56.1	52.6	47.9	44.8	43.5	41.7	41.5	40.1	39.9	38.6	35.5	33.7	31.3

vowel:	i	a	u	E	"o	"e	O	o	e	a~
in languages:	393	392	369	186	181	169	162	131	124	83
frequency:	87.1	86.9	81.8	41.2	40.1	37.5	35.9	29.0	27.5	18.4

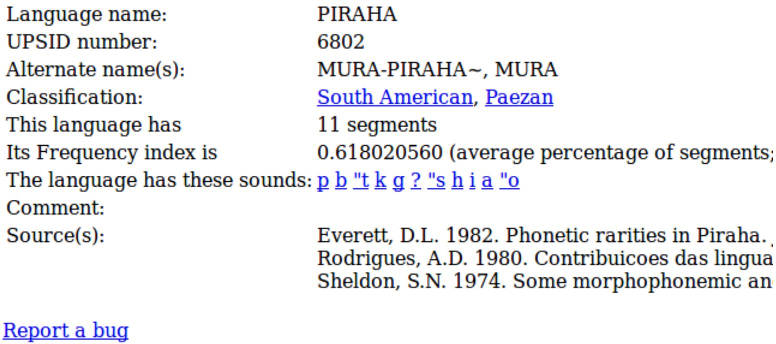
At the other end of the scale there are many segments that occur in one or only few languages:

Number of segments:	427	117	66	39	27	19	14	14	12	13	
that occur only in	1	2	3	4	5	6	7	8	9	10	languages
% of all segments:	46.46	12.73	7.18	4.24	2.94	2.07	1.52	1.52	1.31	1.41	
cummulative %:	46.46	59.19	66.38	70.62	73.56	75.63	77.15	78.67	79.98	81.39	

That is, the group of sounds that appear in 10 or fewer of the 451 languages make up more than 80% of the 919 sounds in the database.

Interface by Henning Reetz
(http://web.phonetik.uni-frankfurt.de/upsid_info.html)

Examples



Language name:	PIRAHA
UPSID number:	6802
Alternate name(s):	MURA-PIRAHA~, MURA
Classification:	South American, Paezan
This language has	11 segments
Its Frequency index is	0.618020560 (average percentage of segments;
The language has these sounds:	p b "t k g ? "s h i a "o
Comment:	
Source(s):	Everett, D.L. 1982. Phonetic rarities in Piraha. Rodrigues, A.D. 1980. Contribuicoes das lingua Sheldon, S.N. 1974. Some morphophonemic an

Report a bug

Interface by Henning Reetz
(http://web.phonetik.uni-frankfurt.de/upsid_info.html)

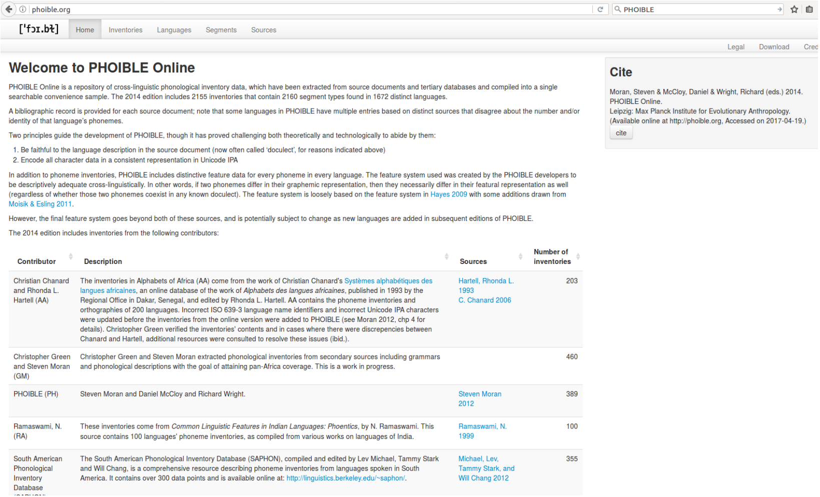
PHOIBLE

What is it?

- "PHOIBLE Online 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. The 2014 edition includes 2155 inventories that contain 2160 segment types found in 1672 distinct languages."
- **License**: Open Access

[Moran, Steven & McCloy, Daniel & Wright, Richard (eds.) 2014. PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://phoible.org, Accessed on 2017-04-21.)]

Examples



Welcome to PHOIBLE Online

PHOIBLE Online 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. The 2014 edition includes 2155 inventories that contain 2160 segment types found in 1672 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 any 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.

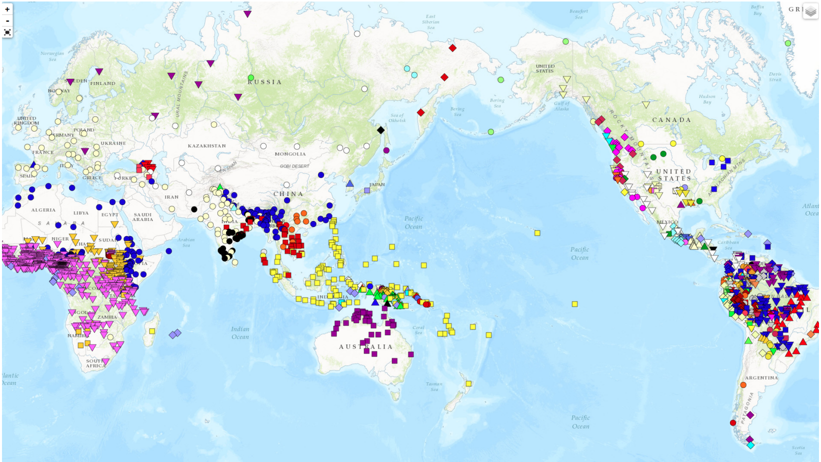
The 2014 edition includes inventories from the following contributors:

Contributor	Description	Sources	Number of inventories
Christian Chanard and Rhonda L. Hartell (AA)	The inventories in Alphabets of Africa (AA) come from the work of Christian Chanard's Systèmes alphabétiques des langues africaines, an online database of the work of *Alphabets des langues africaines*, published in 1993 by the Regional Office in Dakar, Senegal, and edited by Rhonda L. Hartell. AA contains the phoneme inventories and orthographies of 200 languages. Incorrect ISO 639-3 language name identifiers and incorrect Unicode IPA characters were updated before the inventories from the online version were added to PHOIBLE (see Moran 2012, chp 4 for details). Christopher Green verified the inventories' contents and in cases where there were discrepencies between Chanard and Hartell, additional resources were consulted to resolve these issues (ibid.).	Hartell, Rhonda L. 1993 C. Chanard 2006	203
Christopher Green and Steven Moran (GM)	Christopher Green and Steven Moran extracted phonological inventories from secondary sources including grammars and phonological descriptions with the goal of attaining pan-Africa coverage. This is a work in progress.		460
PHOIBLE (PH)	Steven Moran and Daniel McCloy and Richard Wright.	Steven Moran 2012	389
Ramaswami, N. (RA)	These inventories come from *Common Linguistic Features in Indian Languages: Phoentics*, by N. Ramaswami. This source contains 100 languages' phoneme inventories, as compiled from various works on languages of India.	Ramaswami, N. 1999	100
South American Phonological Inventory Database (SAPHON)	The South American Phonological Inventory Database (SAPHON), compiled and edited by Lev Michael, Tammy Stark and Will Chang, is a comprehensive resource describing phoneme inventories from languages spoken in South America. It contains over 300 data points and is available online at: http://linguistics.berkeley.edu/~saphon/.	Michael, Lev, Tammy Stark, and Will Chang 2012	355

Cite

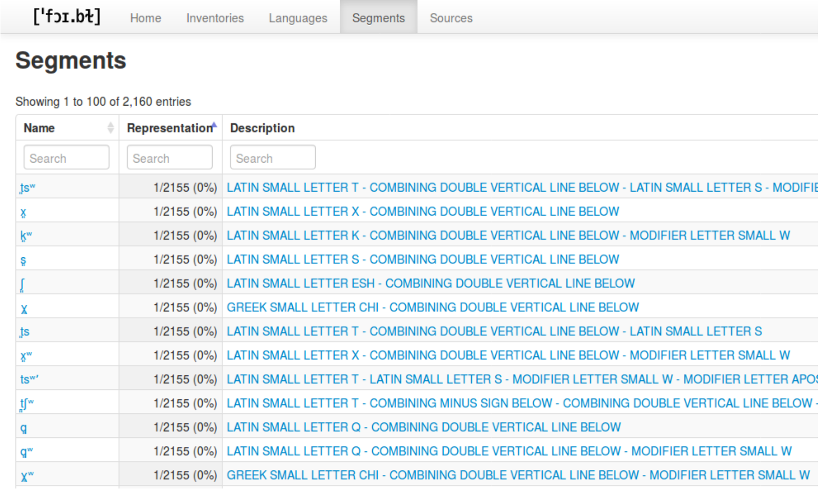
Moran, Steven & McCloy, Daniel & Wright, Richard (eds.) 2014.
PHOIBLE Online.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://phoible.org, Accessed on 2017-04-19.)

Examples



+
-
RUSSIA
SWEDEN
FINLAND
NORWAY
UNITED KINGDOM
GERMANY
POLAND
UKRAINE
FRANCE
SPAIN
ITALY
GREECE
TURKEY
KAZAKHSTAN
MONGOLIA
GOBI DESERT
CHINA
JAPAN
IRAN
SAUDI ARABIA
EGYPT
LIBYA
ALGERIA
SAHARA
NIGER
SUDAN
ZAMBIA
NAMIBIA
SOUTH AFRICA
AUSTRALIA
UNITED STATES
CANADA
MEXICO
ARGENTINA
PATAGONIA
ROCKY MOUNTAINS
APPALACHIAN MTNS
URAL MOUNTAINS
Norwegian Sea
East Siberian Sea
Beaufort Sea
Baffin Bay
Davis Strait
Hudson Bay
Sea of Okhotsk
Bering Sea
Gulf of Alaska
Pacific Ocean
Arabian Sea
Indian Ocean
Coral Sea
Tasman Sea
Caribbean Sea
Atlantic Ocean
Scotia Sea

EXAMPLES

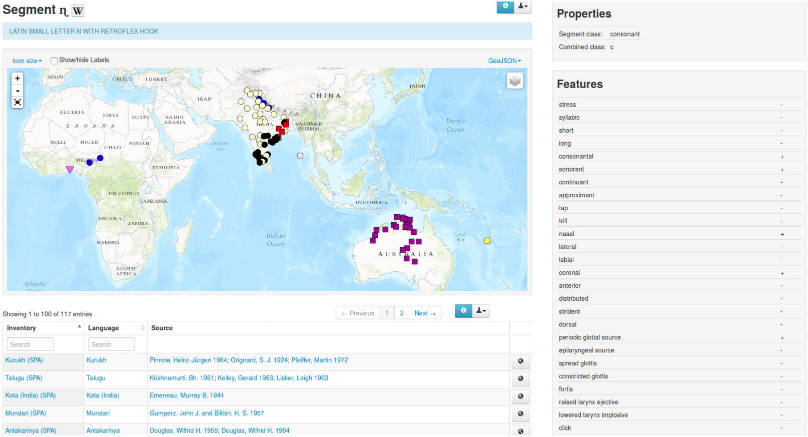


Segments

Showing 1 to 100 of 2,160 entries

Name	Representation	Description
Search	Search	Search
t͈sʷ	1/2155 (0%)	LATIN SMALL LETTER T - COMBINING DOUBLE VERTICAL LINE BELOW - LATIN SMALL LETTER S - MODIFIE
x͈	1/2155 (0%)	LATIN SMALL LETTER X - COMBINING DOUBLE VERTICAL LINE BELOW
k͈ʷ	1/2155 (0%)	LATIN SMALL LETTER K - COMBINING DOUBLE VERTICAL LINE BELOW - MODIFIER LETTER SMALL W
s͈	1/2155 (0%)	LATIN SMALL LETTER S - COMBINING DOUBLE VERTICAL LINE BELOW
ʃ͈	1/2155 (0%)	LATIN SMALL LETTER ESH - COMBINING DOUBLE VERTICAL LINE BELOW
χ͈	1/2155 (0%)	GREEK SMALL LETTER CHI - COMBINING DOUBLE VERTICAL LINE BELOW
t͈s	1/2155 (0%)	LATIN SMALL LETTER T - COMBINING DOUBLE VERTICAL LINE BELOW - LATIN SMALL LETTER S
x͈ʷ	1/2155 (0%)	LATIN SMALL LETTER X - COMBINING DOUBLE VERTICAL LINE BELOW - MODIFIER LETTER SMALL W
tsʷʼ	1/2155 (0%)	LATIN SMALL LETTER T - LATIN SMALL LETTER S - MODIFIER LETTER SMALL W - MODIFIER LETTER APOS
t̠͈ʃʷ	1/2155 (0%)	LATIN SMALL LETTER T - COMBINING MINUS SIGN BELOW - COMBINING DOUBLE VERTICAL LINE BELOW -
q͈	1/2155 (0%)	LATIN SMALL LETTER Q - COMBINING DOUBLE VERTICAL LINE BELOW
q͈ʷ	1/2155 (0%)	LATIN SMALL LETTER Q - COMBINING DOUBLE VERTICAL LINE BELOW - MODIFIER LETTER SMALL W
χ͈ʷ	1/2155 (0%)	GREEK SMALL LETTER CHI - COMBINING DOUBLE VERTICAL LINE BELOW - MODIFIER LETTER SMALL W

Examples



Segment ɳ W

LATIN SMALL LETTER N WITH RETROFLEX HOOK

Icon size▾ □ Show/hide Labels GeoJSON▾

Showing 1 to 100 of 117 entries

← Previous 1 2 Next →

Inventory	Language	Source	
Search	Search		
Kurukh (SPA)	Kurukh	Pinnow, Heinz-Jürgen 1964; Grignard, S. J. 1924; Pfeiffer, Martin 1972	
Telugu (SPA)	Telugu	Krishnamurti, Bh. 1961; Kelley, Gerald 1963; Lisker, Leigh 1963	
Kota (India) (SPA)	Kota (India)	Emeneau, Murray B. 1944	
Mundari (SPA)	Mundari	Gumperz, John J. and Bilibiri, H. S. 1957	
Antakarinya (SPA)	Antakarinya	Douglas, Wilfrid H. 1955; Douglas, Wilfrid H. 1964	

Properties

Segment class:	consonant
Combined class:	c

Features

stress	-
syllabic	-
short	-
long	-
consonantal	+
sonorant	+
continuant	-
approximant	-
tap	-
trill	-
nasal	+
lateral	-
labial	-
coronal	+
anterior	-
distributed	-
strident	-
dorsal	-
periodic glottal source	+
epilaryngeal source	-
spread glottis	-
constricted glottis	-
fortis	-
raised larynx ejective	-
lowered larynx implosive	-
click	-

ASJP

What is it?

- “The database of the **Automated Similarity Judgment Program (ASJP)** aims to contain 40-item word lists of all the world’s languages.”
- **License**: Open Access

[Wichmann, Brown, Holman, et al. (eds.), 2013]

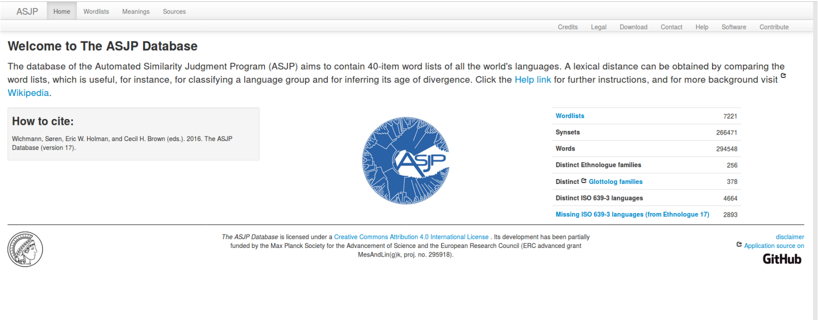
ASJP

What's in it?

- Languages: **4664** (ca. 62% of the world's languages)
- Word lists: **7221** (either 40 or 100 lexical items)

[Wichmann, Sren, Eric W. Holman, and Cecil H. Brown (eds.). (2016). The ASJP Database (version 17).

EXAMPLES



Welcome to The ASJP Database

The database of the Automated Similarity Judgment Program (ASJP) aims to contain 40-item word lists of all the world's languages. A lexical distance can be obtained by comparing the word lists, which is useful, for instance, for classifying a language group and for inferring its age of divergence. Click the Help link for further instructions, and for more background visit Wikipedia.

How to cite:

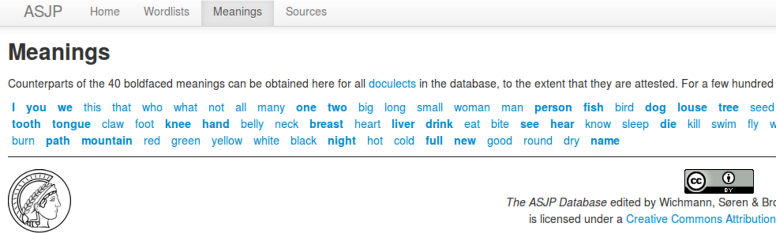
Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2016. The ASJP Database (version 17).

Wordlists	7221
Synsets	266471
Words	294548
Distinct Ethnologue families	256
Distinct Glottolog families	378
Distinct ISO 639-3 languages	4664
Missing ISO 639-3 languages (from Ethnologue 17)	2893

The ASJP Database is licensed under a Creative Commons Attribution 4.0 International License . Its development has been partially funded by the Max Planck Society for the Advancement of Science and the European Research Council (ERC advanced grant MesAndLin(g)k, proj. no. 295918).

disclaimer
Application source on GitHub

Examples



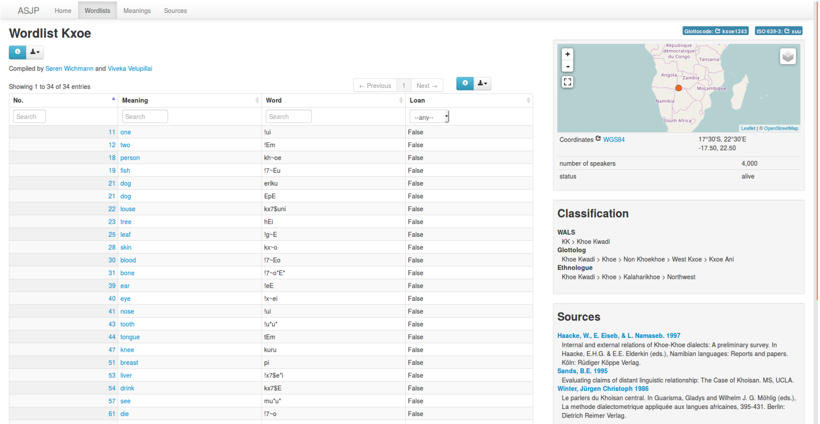
Meanings

Counterparts of the 40 boldfaced meanings can be obtained here for all doculects in the database, to the extent that they are attested. For a few hundred

I **you** **we** this that who what not all many **one** **two** big long small woman man **person** **fish** bird **dog** **louse** **tree** seed
tooth **tongue** claw foot **knee** **hand** belly neck **breast** heart **liver** **drink** eat bite **see** **hear** know sleep **die** kill swim fly w
burn **path** **mountain** red green yellow white black **night** hot cold **full** **new** good round dry **name**

The ASJP Database edited by Wichmann, Søren & Brc
is licensed under a Creative Commons Attribution

Examples



Wordlist Kxoe

Compiled by Søren Wichmann and Viveka Velupillai

Glottocode: kxoe1243 ISO 639-3: xuu

Showing 1 to 34 of 34 entries

← Previous 1 Next →

No.	Meaning	Word	Loan
Search	Search	Search	--any--
11	one	!ui	False
12	two	!Em	False
18	person	kh~oe	False
19	fish	!7~Eu	False
21	dog	eriku	False
21	dog	EpE	False
22	louse	kx7$uni	False
23	tree	hEi	False
25	leaf	!g~E	False
28	skin	kx~o	False
30	blood	!7~Eo	False
31	bone	!7~o*E*	False
39	ear	!eE	False
40	eye	!x~ei	False
41	nose	!ui	False
43	tooth	!u*u*	False
44	tongue	tEm	False
47	knee	kuru	False
51	breast	pi	False
53	liver	!x7$e*i	False
54	drink	kx7$E	False
57	see	mu*u*	False
61	die	!7~o	False

Coordinates WGS84	17°30'S, 22°30'E -17.50, 22.50
number of speakers	4,000
status	alive

Classification

WALS
KK › Khoe Kwadi
Glottolog
Khoe Kwadi › Khoe › Non Khoekhoe › West Kxoe › Kxoe Ani
Ethnologue
Khoe Kwadi › Khoe › Kalaharikhoe › Northwest

Sources

Haacke, W., E. Eiseb, & L. Namaseb. 1997
Internal and external relations of Khoe-Khoe dialects: A preliminary survey. In Haacke, E.H.G. & E.E. Elderkin (eds.), Namibian languages: Reports and papers. Köln: Rüdiger Köppe Verlag.
Sands, B.E. 1995
Evaluating claims of distant linguistic relationship: The Case of Khoisan. MS, UCLA.
Winter, Jürgen Christoph 1986
Le parlers du Khoisan central. In Guarisma, Gladys and Wilhelm J. G. Möhlig (eds.), La methode dialectometrique appliquée aux langues africaines, 395-431. Berlin: Dietrich Reimer Verlag.

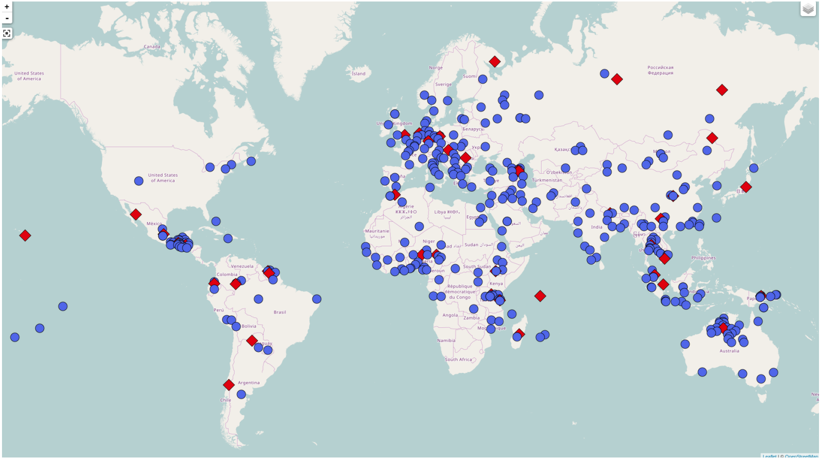
WOLD

What is it?

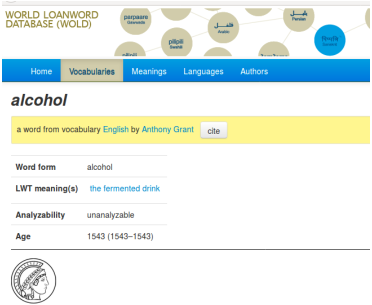
- "The **World Loanword Database (WOLD)** provides vocabularies (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word. It allows users to find loanwords, source words and donor languages in each of the 41 languages, but also makes it easy to compare loanwords across languages. "
- **License**: Open Access

[Haspelmath, Martin & Tadmor, Uri (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wold.clld.org, Accessed on 2017-04-21.)]

Examples


+
-
Canada
United States of America
Ísland
Norge
Suomi
Sverige
Российская Федерация
United Kingdom
Беларусь
Қазақстан
O'zbekiston
Türkmenistan
Türkiye
United States of America
México
Venezuela
Colombia
Brasil
Perú
Bolivia
Argentina
Chile
Algérie
Libya
Egypt
Mauritanie
Niger
Sudan
South Sudan
Kenya
République démocratique du Congo
Angola
Zambia
Namibia
South Africa
India
Philippines
Australia
Leaflet | © OpenStreetMap

EXAMPLES

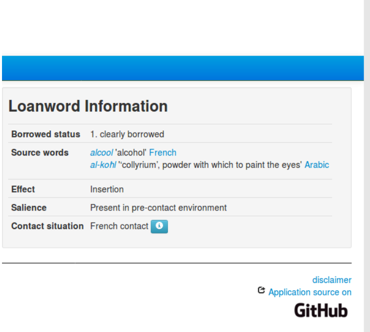


WORLD LOANWORD DATABASE (WOLD)

alcohol

a word from vocabulary English by Anthony Grant

Word form	alcohol
LWT meaning(s)	the fermented drink
Analyzability	unanalyzable
Age	1543 (1543–1543)



Loanword Information

Borrowed status	1. clearly borrowed
Source words	*alcool* 'alcohol' French *al-kohl* ''collyrium', powder with which to paint the eyes' Arabic
Effect	Insertion
Salience	Present in pre-contact environment
Contact situation	French contact

D-PLACE

What is it?

- "D-PLACE, which stands for Database of Places, Language, Culture, and Environment, represents an attempt to bring together this dispersed corpus of information [on cultural and climatic traits associated with societies]. It aims to make it easy for individuals to contrast their own cultural practices with those of other societies, and to consider the factors that may underlie cultural similarities and differences. "
- **License**: Open Access

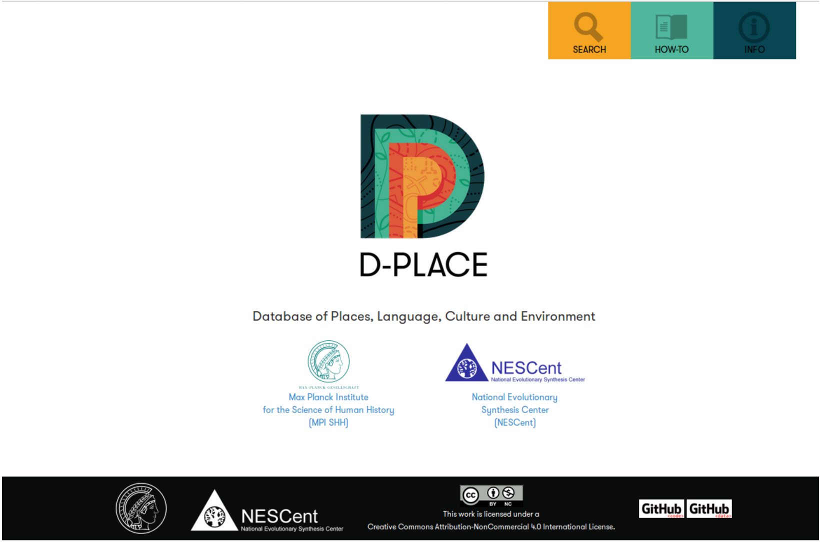
[Kirby, K.R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H-J, et al. (2016). D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. PLoS ONE, 11(7): e0158391. doi:10.1371/journal.pone.0158391.]

D-PLACE

What's in it?

- So far, D-PLACE contains cultural, linguistic, environmental and geographic information for over 1400 human societies. A society in D-PLACE represents a group of people in a particular locality, who often share a language and cultural identity.

Examples



D-PLACE

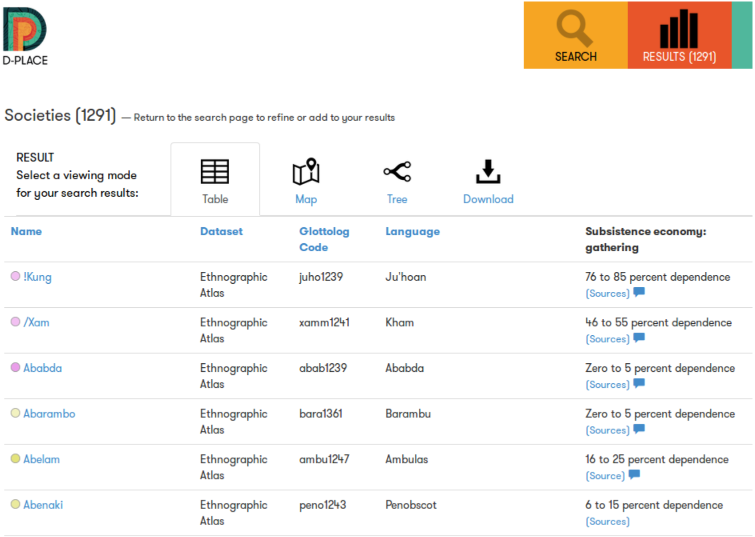
Database of Places, Language, Culture and Environment

Max Planck Institute
for the Science of Human History
(MPI SHH)

National Evolutionary
Synthesis Center
(NESCent)

This work is licensed under a
Creative Commons Attribution-NonCommercial 4.0 International License.

Examples

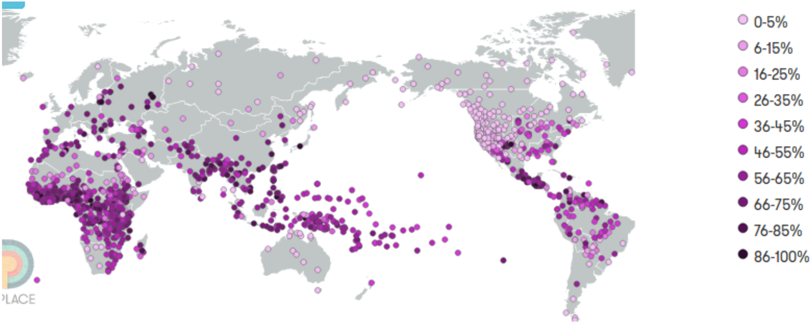


Societies (1291) — Return to the search page to refine or add to your results

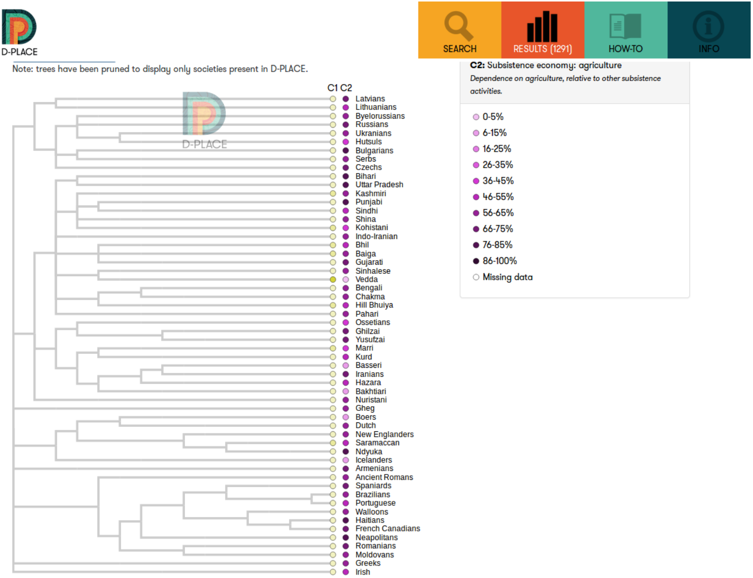
RESULT
Select a viewing mode
for your search results:

Name	Dataset	Glottolog Code	Language	Subsistence economy: gathering
● !Kung	Ethnographic Atlas	juho1239	Ju'hoan	76 to 85 percent dependence (Sources)
● /Xam	Ethnographic Atlas	xamm1241	Kham	46 to 55 percent dependence (Sources)
● Ababda	Ethnographic Atlas	abab1239	Ababda	Zero to 5 percent dependence (Sources)
● Abarambo	Ethnographic Atlas	bara1361	Barambu	Zero to 5 percent dependence (Sources)
● Abelam	Ethnographic Atlas	ambu1247	Ambulas	16 to 25 percent dependence (Source)
● Abenaki	Ethnographic Atlas	peno1243	Penobscot	6 to 15 percent dependence (Sources)

Examples



Examples



Note: trees have been pruned to display only societies present in D-PLACE.

C2: Subsistence economy: agriculture

Dependence on agriculture, relative to other subsistence activities.

- ● 0-5%
- ● 6-15%
- ● 16-25%
- ● 26-35%
- ● 36-45%
- ● 46-55%
- ● 56-65%
- ● 66-75%
- ● 76-85%
- ● 86-100%
- ○ Missing data

THE PROBLEM OF SAMPLING (VELUPILLAI, 2012)

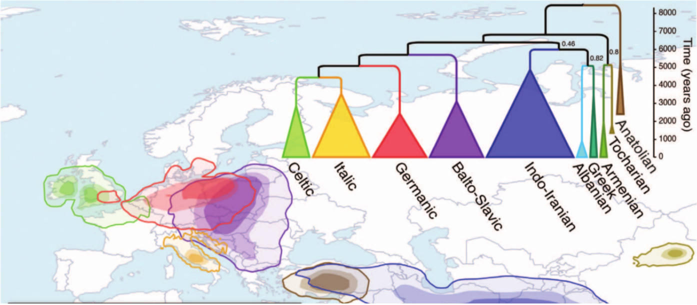
At **two levels of analysis** we have to deal with the problem of finding **representative samples**:

- **Within a language**:
  - **corpora**: should be **balanced**, i.e. represent a wide range of registers and styles
  - **experiments**: participants should **represent the population**, e.g. age range, educational background, gender, etc.
- **Across different languages**:
  - it is hard (impossible) to assess a feature across all 7000+ languages, hence we need a **balanced** and **unbiased** sample of them

The **second level** is most relevant for Typology, though note that the first level decides how we represent languages.

Sampling: Types of Biases

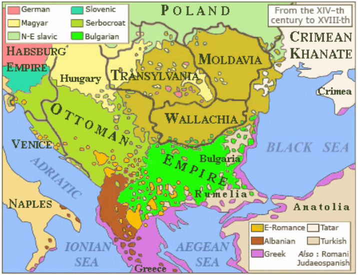
- **(Phylo-)genetic bias**
  languages are non-independent if they share common ancestry, i.e. a proto-language. A bias can result from over-representation of a specific **family or genus** (e.g. Indo-European, Germanic)



Bouckaert et al. (2012)

Sampling: Types of Biases

- **Areal bias**
  language contact leads to the spread of linguistic features. A bias can result when languages from a specific **geographic region** (e.g. Balkan Sprachbund) are over-represented



Sampling: Types of Biases

- **Typological bias**
  over-representation of specific **typological features** (e.g. languages with/without tone)
- **Cultural bias**
  certain cultures might put more emphasis on encoding specific kinds of information in their grammar (e.g. Korean and Japanese honorific markers)
- **Bibliographical bias** bias towards **well-described languages**. Not all languages are equally well described, and for many languages any kind of information is lacking all together (e.g. Sentinelese).

Sampling: Types of Samples

- **Variety sample**
  a sample covering **all the parameters of a linguistic variable under investigation**. For example, if we are dealing with word orders we would need to have a sample that covers all the logical possibilities SOV, SVO, OVS, etc. (if they all exist)
- **Convenience sample**
  Based on the **descriptions available** for a typological variable (relation to the biographical bias). Note that still an attempt can be made to balance the sample areally, genetically, etc.

Sampling: Types of Samples

- **Probability sample**
  A sample that does not have any of the biases named above, and hence represents only **fully independent languages**. Strictly speaking, only with such a sample is it possible to make valid **statistical judgements** about the probability of occurrence or co-occurrence of certain typological variables (e.g. how probable is it for a language to have tone?)
- **Random sample**
  Drawing languages **randomly without any sensitivity to biases**. Note that for small samples, a random sample can be highly biased. However, for bigger samples bias is increasingly unlikely.

Sampling: Problems

- **Variety sample:** only works if we know the number of logical possibilities per variable (e.g. for word order SOV, SVO, VSO, OSV, OVS, VOS). Does not work for variables that are open-ended (e.g. number of phonemes, case markers, etc.)
- **Convenience sample:** very likely to be biased in one way or another
- **Probability sample:** very hard (or impossible) to get. For example, Piantadosi & Gibson (2011) argue that to rectify a universal of the type *all languages have feature x*, we would need a sample of **500 independent** languages
- **Random sample:** small samples will be biased, for big samples there is the same problem as for probability samples