Databases of World's Languages

Zdeněk Žabokrtský, Magda Ševčíková, Anna Nedoluzhko

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

A warm-up quiz (pen-and-paper)

- Please write down a list of names of as many languages as you can.

How many languages are there in the world?

- some 7k living languages
- Although the major languages belong to only a few language families, the total diversity of all living languages is enormous, way behind what a human individual can grasp → We need data!
- We also need simplified perspectives on languages to navigate their abundance and diversity

What counts as a language?

How many languages exist, exactly?

- Let us consult a comprehensive reference source:
  - Ethnologue (27th ed., 2024): 7,164 languages
  - Ethnologue (26th ed., 2023): 7,168 languages
  - Ethnologue (25th ed., 2022): 7,151 languages
  - Ethnologue (24th ed., 2021): 7,139 languages
  - Ethnologue (23rd ed., 2020): 7,117 languages
  - Ethnologue (22nd ed., 2019): 7,111 languages
  - Ethnologue (21st ed., 2018): 7,097 languages
  - . . .
  - Ethnologue (16th ed., 2009): 6,909 languages
  - . . .
  - Ethnologue (7th ed., 1969): 4,493 languages
  - . . .
  - Encyclopedia Britannica (1911): app. 1,000 lang.
- differences not due to a language number increase, but due to a finer-grained view and intensive documentation efforts (field linguistics etc.)

What counts as a language?

- The central question: a language, or a dialect?
- A mixture of criteria can be applied...
  - Linguistic criteria
    - mutual intelligibility
    - phonological/grammatical/lexical similarities
  - Sociopolitical criteria
    - cultural and political identity
    - standardization
  - Historical developments...
- ... but in fact, the boundary between a language and a dialect often not really linguistically grounded
- Max Weinreich: "A language is a dialect with an army and navy."

Dialects, or separate languages?

- A gradual scale rather than dichotomy, examples:
- Obviously dialects: European Portuguese vs. Brazilian Portuguese
- Obviously separate languages: Czech vs. Japanese
- Very close language varieties that are usually considered separate languages:
  - Czech and Slovak - high mutual intelligibility, esp. in written form
  - Hindi and Urdu - high mutual intelligibility, only in spoken form
  - Macedonian and Bulgarian - high mutual intelligibility in both written and spoken forms
- Relatively distant language varieties that are, however, often considered dialects:
  - Mandarin Chinese and Cantonese - very low intelligibility in spoken form
  - Northern vs. Southern Italian Dialects

Speaker populations

L1 vs. L2

- A first language, native language, native tongue, or mother tongue (L1)
  - is a language which a person acquires first in her/his life, usually naturally as a child
  - full fluency expected
  - a (simultaneous) bilingualism possible - growing up with two L1 languages

- A second language (L2)
  - is a language which a person acquires next to her/his first language
  - a neighbouring language, another language of the speaker's home country, or a foreign language.
  - polyglots - even more than 10 L2 languages (but to varying degrees of fluency)

Top L1 vs L1+L2 speaker population sizes, in millions

1. Mandarin Chinese: 990
2. Spanish: 484
3. English: 390
4. Hindi: 345
5. Portuguese: 250
6. Bengali: 242
7. Russian: 145
8. Japanese: 124
9. Vietnamese: 86
10. Turkish: 85
11. Yue Chinese (incl. Cantonese): 85
12. Egyptian Arabic: 84
13. Marathi: 83
14. Telugu: 83
15. Wu Chinese: 83
16. Korean: 81
17. Tamil: 79
18. Urdu: 78
19. Standard German: 76
20. Indonesian: 75

1. English: 1528
2. Mandarin Chinese: 1184
3. Hindi: 609
4. Spanish: 558
5. Modern Standard Arabic (without dialects): 335
6. French: 312
7. Bengali: 284
8. Portuguese: 267
9. Russian: 253
10. Indonesian: 252
11. Urdu: 246
12. Standard German: 134
13. Japanese: 126
14. Nigerian Pidgin: 121
15. Egyptian Arabic: 119
16. Marathi: 99
17. Vietnamese: 97
18. Telugu: 96
19. Hausa: 94
20. Turkish: 91

Dialects vs. separate languages, revisited

- Obvious from the previous slide: unclear (or conficting) criteria for the boundaries between dialects and languages are definitely rare

Speaker population sizes, the opposite end of the scale

- A living language - at least one living L1 speaker
- An extinct language - a language that no longer has any native speakers
- A dead language - no longer the native language of any community, but is still in use, e.g. Latin in classical studies or in religion

- (however, “a dead language” and “an extinct language” are sometimes used interchangeably)

Expanded Graded Intergenerational Disruption Scale

- EGIDS - a 13-level scale that assesses the vitality and endangerment of languages based on their transmission across generations, institutional support, and societal use

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

Endangered languages

- critically endangered languages - e.g. only a handful of elderly speakers
- on average, a language disappears in every two weeks
- half of the world's languages could disappear by the end of the 21st century if current trends continue

Geographical distribution of languages

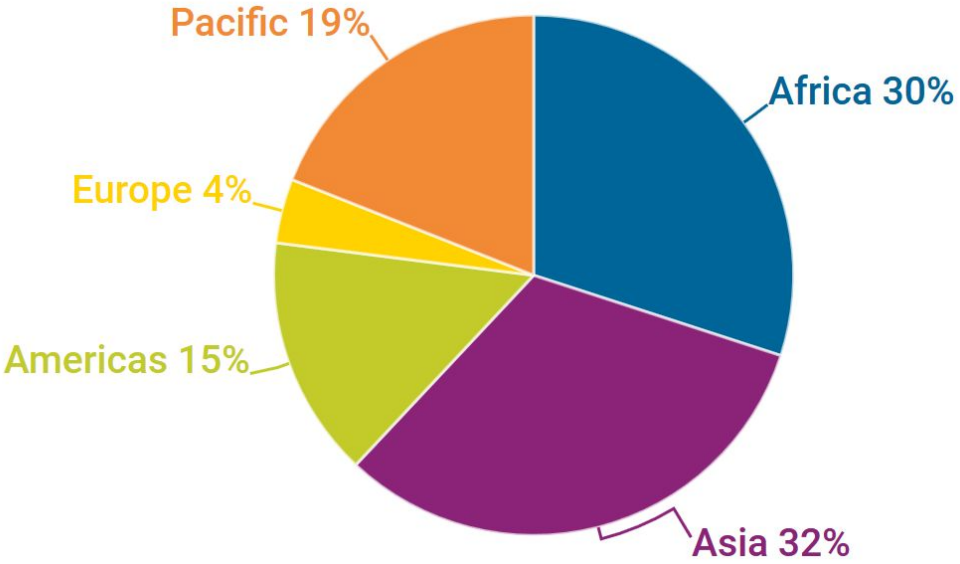
Regions of origin of languages

Each language is located in its primary country, i.e. it is shown just once; e.g. English in the United Kingdom, Esperanto in Poland.

Languages by region of origin

Ethnologue (2018):

- 2,300 languages (out of 7,097) are from Asia
- 2,143 languages are concentrated in Africa
- 288 languages belong to Europe



Languages by region and by L1 speaker population size

English listed as a language belonging to Europe. Therefore, all L1 speakers of English (e.g. incl. English speakers living in the USA) categorized under Europe.

L1 top languages

- nearly 7,900,000,000 speakers around the world
- 40 % of speakers use one of just 5 languages as L1
    - Chinese, Spanish, English, Hindi, Arabic

- nearly 90 % of speakers use languages from Asia or Europe as L1

Standardized Language Lists

ISO 639

- ISO 639 is a set of standards from the International Organization for Standardization; a naming convention
- approved in 1967
- main parts of ISO 639:
    - ISO 639-1 – two-letter codes for languages and language groups (macrolanguages); ‘cs’ for Czech
    - ISO 639-2 – two slightly different sets of three-letter codes (639-2/T and 639-2/B, ‘ces’ and ‘cze’, respectively)
    - ISO 639-3 – three-letter codes (‘ces’);
    - ISO 639-4 – rather documentation of the principles, no language code specs,
    - ISO 639-5 – three-letter codes for language families
    - ISO 639-6 – four-letter codes to represent language variants, including dialects; withdrawn

• the individual standards designed to work together (no naming collisions)

Wait - can two-letter codes be sufficient?

ISO 639-1

- 184 codes for “world’s major languages”
- e.g. ‘cs’ for Czech, ‘de’ for German
- ‘no’ for Norwegian, which is considered a macrolanguage covering both Bokmål (‘nb’) and Nynorsk (‘nn’)

ISO 639-2

- 488 languages and language groups
  - ISO 639-2/T: three-letter codes, for the same languages as 639-1
  - ISO 639-2/B: three-letter codes, mostly the same as 639-2/T, but with some codes derived from English names of the languages
- an example of a difference: Czech: ‘ces’ in 639-2/T, while ‘cze’ in 639-2/B

ISO 639-3

- aim to cover all known languages
- over 7,000 languages/language varieties
- extension based on Ethnologue
- special values such as ‘und’ (undetermined) or ‘mul’ (multiple languages)

A few examples from ISO 639-1, 2/T, 2/B, and 3

Language	ISO 369-1	639-2/T	ISO 639-2/B	ISO 639-3
Czech	cs	ces	cze	ces
Dutch	nl	nld	dut	nld
German X	de	deu	ger	deu
Greek	el	ell	gre	ell
Russian	ru	rus	rus	rus

(Recall A. Tanenbaums's quote "The nice thing about standards is that there are so many of them to choose from.")

ISO 639-5

- three-letter codes for language families and groups
- Examples:
  - ine – Indo-European languages
  - ine:sla – Slavic languages
  - ine:sla:zlw – West Slavic languages
  - ine:sla:zlw:wen – Sorbian languages

Glottolog

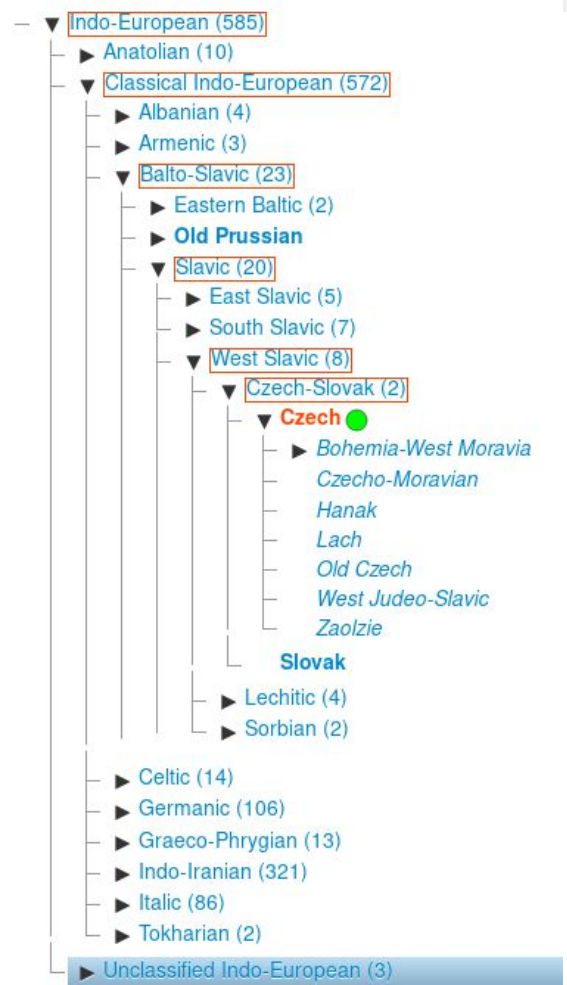
- a database that catalogues the world's languages
- maintained by the Max Planck Institute for Evolutionary Anthropology in Leipzig
- umbrella term 'languoids' – languages, dialects, and families of the world
- currently 25,900 languoids:
  - 8,533 language-level
  - 4,571 family-level
  - 12,796 dialect-level

Glottolog's Glottocodes

- each languoid has a unique identifies – a glottocode
- four alphanumeric characters and four decimal digits
- Examples:
  - stan1295 German
  - midd1343 Middle High German
  - oldh1241 Old High German (ca. 750-1050)
  - berl1235 Berlin German
  - penn1240 Pennsylvania German
  - germ1288 German-Yiddish-Romani-Rotwelsch
  - germ1281 German Sign Language
  - swis1240 Swiss-German Sign Language

Glottolog's hierarchical grouping of languages

- around 240 top-level families, plus around 180 isolates
- Example: the position of Czech:

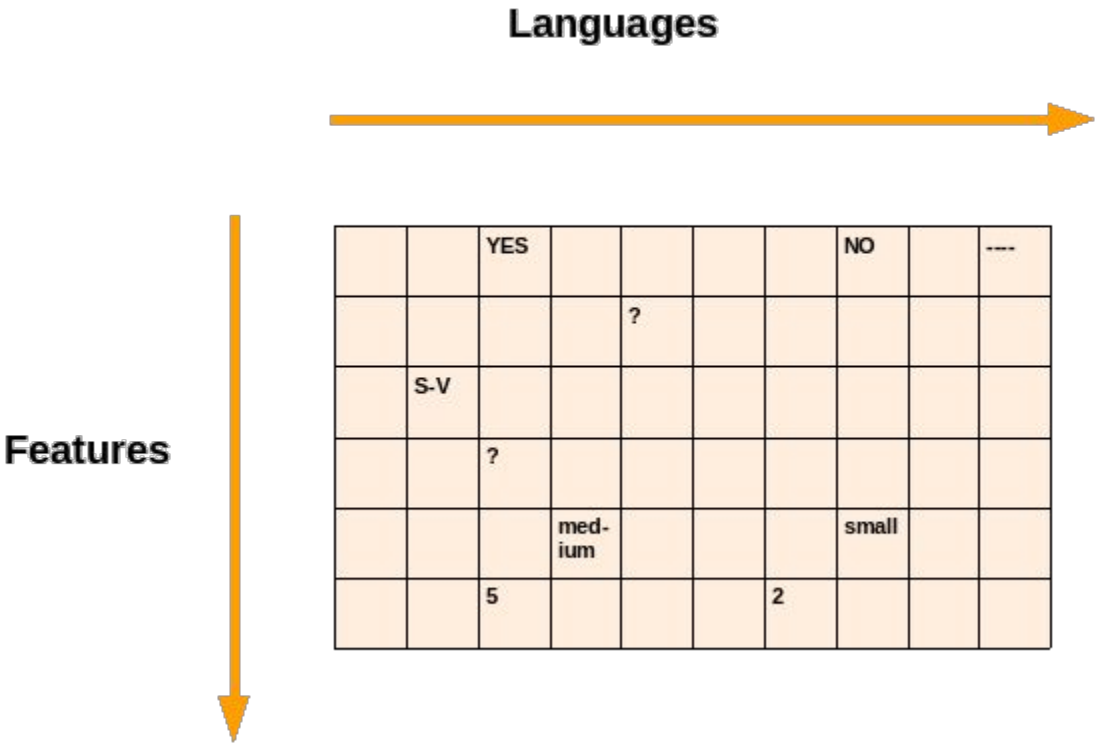


Time for a demo

- https://glottolog.org/glottolog/language

Databases of Language Features

Databases of languages and their properties



Layers of language description

- different "levels of abstraction”, helping linguists and language learners to understand and describe how a language functions
- A possible inventory of layers
  - Phonetics: The Layer of Sounds
  - Phonetics: The Layer of Sounds
  - Morphology: The Layer of Word Structure
  - Syntax: The Layer of Sentence Structure
  - Semantics: The Layer of Meaning
  - Pragmatics: The Layer of Contextual Meaning
  - ...
- We can study differences and similarities among languages from all these perspectives
- Language features typically organized along such layers too
- No broad consensus as for the exact layers’ content, however, comparing languages as wholes seems not mentally tractable

The World Atlas of Language Structures – WALS

- location, affiliation and typological (phonological, lexical, and grammatical) properties of languages
  - 2,662 languages
  - 192 features
  - geographical distribution of a feature's values on – a map for each feature

WALS feature example

Feature 33A: Coding of Nominal Plurality

This feature is described in the text of chapter 33 Coding of Nominal Plurality by Matthew S. Dryer cite

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

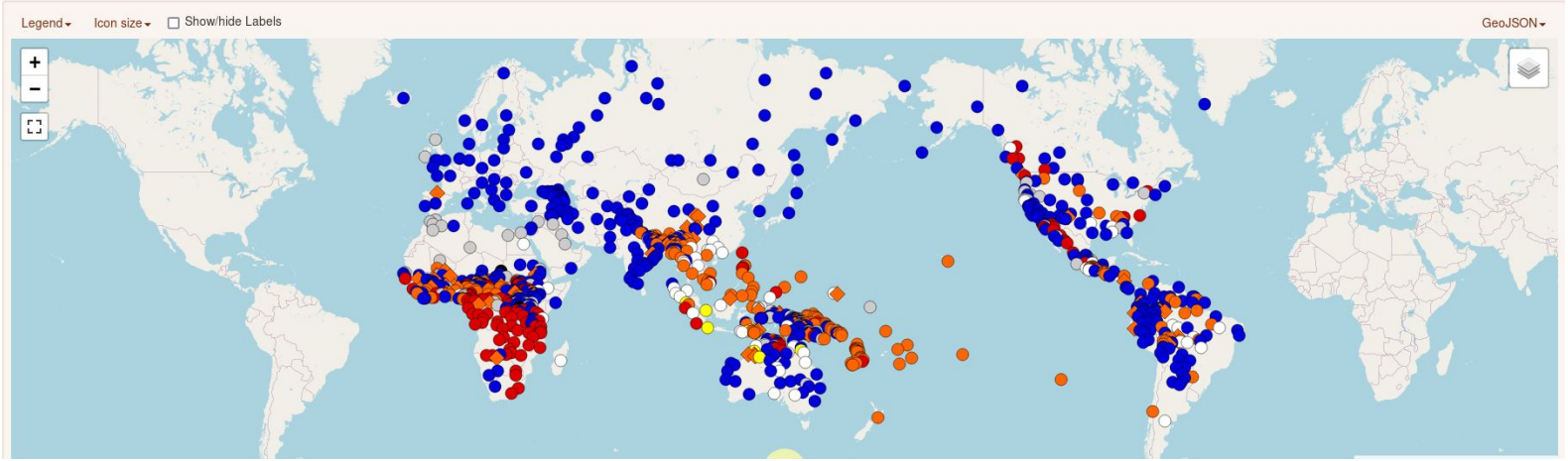
× 33A: Coding of Nominal Plurality

Submit

Values

Plural prefix	126
Plural suffix	513
Plural stem change	6
Plural tone	4
Plural complete reduplication	8
Mixed morphological plural	60
Plural word	170
Plural clitic	81
No plural	98

reload



Feature areas in WALS

- Phonology
  - e.g. Consonant Inventories (values: Small, Moderately Small, ..., Large)
- Morphology
  - e.g. Inflectional Synthesis of a Verb (values: 0-1 category per word, ..., 12-13 categories per word)
- Nominal Categories
  - e.g. Definite Article (values: Definite word distinct from demonstrative, Definite affix, No definite or indefinite article...)
- Word Order
  - e.g. Order of Subject and Verb (values: SV, VS, No dominant order)
- Lexicon
  - e.g. Hand and Arm (values: Identical, Different)

Time for a demo

- https://wals.info/

Glottobank's Grambank

- Grambank is a part of a larger project called Glottobank, together with
  - Lexibank (lexicons)
  - Parabank (paradigms)
  - Numeralbank (numerals)
  - Phonobank (phonetic changes)

Grambank

- 2,467 language varieties (in 215 families + 101 isolates)
- 195 features

Random examples of Grambank features (mostly the expectable ones)

- GB022 Are there prenominal articles?
- GB030 Is there a gender distinction in independent 3rd person pronouns?
- GB044 Is there productive morphological plural marking on nouns?
- GB075 Are there postpositions?
- GB122 Is verb compounding a regular process?
- GB134 Is the order of constituents the same in main and subordinate clauses?
- GB328 Can the relative clause precede the noun?
- GB415 Is there a politeness distinction in 2nd person forms?
- GB172 Can an article agree with the noun in gender/noun class?

Random examples of Grambank features (less expectable ones)

- GB054 Is there a gender/noun class system where plant status is a factor in class assignment?
- GB320 Is paucal number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB301 Is there an inclusory construction?
- GB266 Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?
- GB099 Can verb stems alter according to the person of a core participant?
- GB109 Is there verb suppletion for participant number?
- GB155 Are causatives formed by affixes or clitics on verbs?

Time for a demo

- https://grambank.clld.org/

Summary

Some databases of languages and their features

- inventories of languages (incl. tree-shaped hierarchies on top of the inventories):
  - ISO 639-3: some 7 k languages/language varieties/macrolanguages
  - Glottolog: some 26 k languoids (languages/dialects/families)
  - WALS: 2.6 k languages
  - Grambank: 2.5 k languages (in 215 families, plus isolates)
  - Ethnologue: 7 k languages

- inventories of features
  - ISO 639: only basic classification (living/extinct/artificial... languages)
  - Glottolog: only basic classifications (sign/pidgin/artificial... , endangered/non-endangered)
  - WALS: 192 features, plus language genus, family, and macroarea
  - Grambank: 195 features (and other types of information available in the umbrella Glottolog project)

Databases of languages and their features

- an obvious and natural trade-off: either many languages, or many features
- non-trivial factor: differences in correctness* and completeness of feature values

  *: genealogical hierarchies as well as language feature inventories (and values) are often subjected to interpretation

- many phenomena that do not fit the languages×features scheme nicely: language continua, code switching …
- keep in mind that there is often no obvious ground truth

Homework assignment

HW1 specification

- Task: Using the WALS or Glottolog or Grambank data (or any combination of them), write a Python code that does something interesting with the data.
- For instance, you can
  - try to identify “language universals” in the form of implications or statistical correlations among typological features,
  - or given a set of typological features for a set of languages (and possibly also its position in a genealogical tree), try to predict values of some other feature,
  - or given a set of typological features for a set of language, try to induce a genealogical tree
  - or try to identify errors/inconsistencies/outliers inside any resource, or differences between any two resources.
- Write a short report (0.5 - 1 A4 page) about your findings.

HW1 submission

- Deadline: see this course's main web page
- Submission via gitlab, like in NPFL070, NPFL124, NPFL125…
  - Log in at https://gitlab.mff.cuni.cz/
  - Create a repository named 'NPFL150', identifier 'npfl150'
  - Leave visibility level at 'Private'
  - Give access to Zdeněk Žabokrtký (role 'Reporter'), click 'Invite'
  - Create directory 'hw1' and upload (commit+push) your solution, ideally in a form of a Python code executed from a Makefile (don't upload the data, as they should be downloaded by the Makefile)
  - upload also the short report (a PDF file)