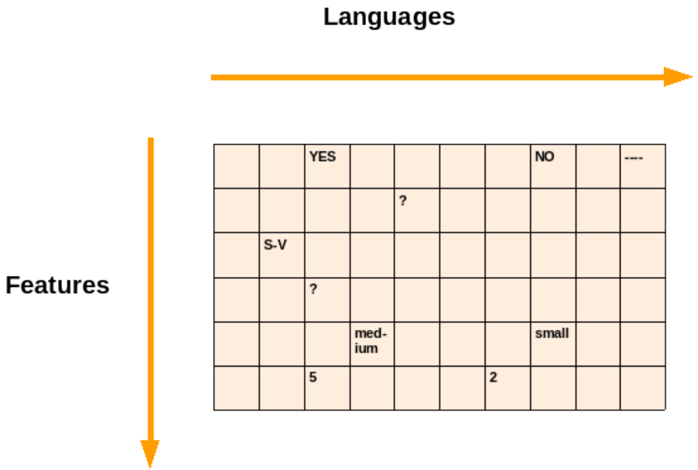
Databases of languages and their properties

Zdeněk Žabokrtský

October 23, 2024

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



Languages

Features

		YES					NO		----
				?					
	S-V								
		?							
			med- ium				small		
		5				2			

Outline

ISO 639

Glottolog

WALS

Grambank

Summary

Homework

Exercise

- Please list the names of as many languages as possible.

ISO 639

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  - ISO 639-6 – four-letter codes to represent language variants, including dialects; withdrawn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

Exercise

- Could you suggest (at least) two reasons why it could be problematic to group languages into tree-shaped hierarchies?

Exercise

- Could you suggest (at least) two reasons why it could be problematic to group languages into tree-shaped hierarchies?
  - (to some extent) arbitrary granularity
  - (to some extent) arbitrary branching factor within the trees
  - gradual changes – dialect continua, language continua
  - interplay of vertical and horizontal transmission (Sprachbunds, mixed languages, network-like modes, wave model...)

Glottolog

Glottolog

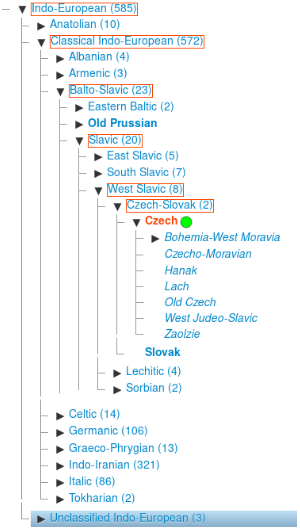
- a database that catalogues the world's languages
- maintained by the Max Planck Institute for Evolutionary Anthropology in Leipzig
- umbrella term 'languoids' – languages, dialects, and families of the world
- currently 25,900 languoids:
  - 8,533 language-level
  - 4,571 family-level
  - 12,796 dialect-level

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

Hierarchical grouping of languages

- around 240 top-level families, plus around 180 isolates



- ▼ Indo-European (585)
  - ▶ Anatolian (10)
  - ▼ Classical Indo-European (572)
    - ▶ Albanian (4)
    - ▶ Armenic (3)
    - ▼ Balto-Slavic (23)
      - ▶ Eastern Baltic (2)
      - ▶ **Old Prussian**
      - ▼ Slavic (20)
        - ▶ East Slavic (5)
        - ▶ South Slavic (7)
        - ▼ West Slavic (8)
          - ▼ Czech-Slovak (2)
            - ▼ **Czech** ●
              - ▶ *Bohemia-West Moravia*
              - *Czecho-Moravian*
              - *Hanak*
              - *Lach*
              - *Old Czech*
              - *West Judeo-Slavic*
              - *Zaolzie*
            - **Slovak**
          - ▶ Lechitic (4)
          - ▶ Sorbian (2)
    - ▶ Celtic (14)
    - ▶ Germanic (106)
    - ▶ Graeco-Phrygian (13)
    - ▶ Indo-Iranian (321)
    - ▶ Italic (86)
    - ▶ Tokharian (2)
  - ▶ Unclassified Indo-European (3)

Time for a demo

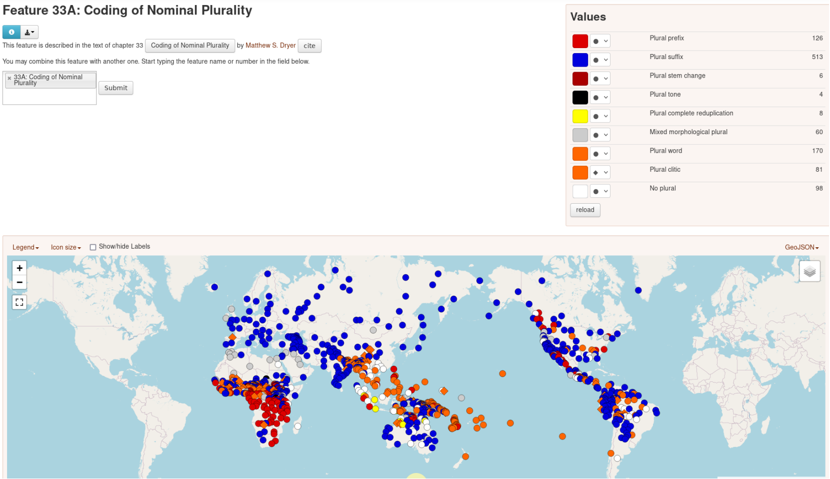
- https://glottolog.org/glottolog/language

WALS

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

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

- 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

Grambank

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 expected ones)

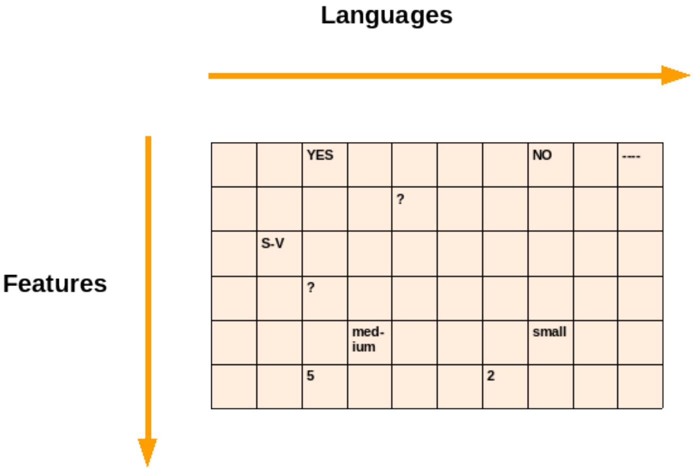
- GB054 Is there a gender/noun class system where plant status is a factor in class assignment?
- GB320 Is paucal number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB301 Is there an inclusory construction?
- GB266 Is there a comparative construction that employs 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

Summary



Languages

Features

		YES					NO		----
				?					
	S-V								
		?							
			medium				small		
		5				2			

Summary

- inventories of languages (plus 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)
- 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)

Summary, cont.

- 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

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.

Alternative HW1 spec, only for non-programmers

- import some of the data resources into a spreadsheet editor, and try to identify some patterns (such as correlations among feature values) using functions of the spreadsheet editor
- write a short report (0.5 - 1 A4 page) about your findings

HW1 submission

- Submission via gitlab, like in NPFL070, NPFL124, NPFL125...
  - Log in at https://gitlab.mff.cuni.cz/
  - Create a repository named 'NPFL100', identifier 'npfl100'
  - 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)
- Deadline: see this course's main web page