World Atlas of Language Structures

Daniel Zeman, Rudolf Rosa

February 22, 2024




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

unless otherwise stated

Multilingual Natural Language Processing

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University


Daniel Zeman, Rudolf Rosa, Ondřej Bojar

zeman@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/courses/npfl120



Variability of Languages in Time and Space



- NPFL100
- Sister course of this one
  - You have attended ⇒ advantage
  - You haven't ⇒ no disaster... but take it next year :-)
- They: more linguistics, less computation
- We: less linguistics, more computation
  - ... today is an exception :-)

Why Multilingual Processing?

- A blatantly incomplete study:
  - ACL main conference proceedings
  - Paper title contains “parsing”
- ACL-COLING 1998 (Montréal, Canada) 
  - 9 papers
  - 3 languages: English (4×), Spanish (1×), German (1×)
  - 4× no evaluation/language
  - English often implicitly, without mentioning it!

Why Multilingual Processing?

- A blatantly incomplete study:
  - ACL main conference proceedings
  - Paper title contains “parsing”
- ACL-COLING 1998 (Montréal, Canada) 
  - 9 papers
  - 3 languages: English (4×), Spanish (1×), German (1×)
  - 4× no evaluation/language
  - English often implicitly, without mentioning it!
- ACL 2007 (Praha, Czechia) 
  - 12 papers
  - 13 languages: en (7×), de (3×); ar, cs, da, eu, ja, nl, pt, sl, sv, zh
  - Max 8 langs/paper; average 1.9 langs/paper

Why Multilingual Processing?

- A blatantly incomplete study:
  - ACL main conference proceedings
  - Paper title contains "parsing"
- ACL-COLING 1998 (Montréal, Canada) 🇨🇦
  - 9 papers
  - 3 languages: English (4×), Spanish (1×), German (1×)
  - 4× no evaluation/language
  - English often implicitly, without mentioning it!
- ACL 2007 (Praha, Czechia) 🇨🇿
  - 12 papers
  - 13 languages: en (7×), de (3×); ar, cs, da, eu, ja, nl, pt, sl, sv, zh
  - Max 8 langs/paper; average 1.9 langs/paper
- ACL 2016 (Berlin, Germany) 🇩🇪
  - 24 papers
  - 24 languages: en (18×), de (6×), zh (5×); ar, bg, ca, cs, da, el, es, eu, fr, he, hu, it, ja, ko, ml, nl, pl, pt, sl, sv, tr
  - Max 18 langs/paper; average 3.1 langs/paper

Why Multilingual Processing?

- Trend:
  - No evaluation on data
  - Evaluation on English (usually Penn Treebank)
    - Rarely something else
    - But usually one language per paper
  - Evaluation on multiple languages
    - Still skewed towards a few families
    - "Big languages" of Eurasia
    - Indo-European, Uralic, Turkic, Semitic, Chinese, Japanese, Korean
  - Resource-poor languages

Why Multilingual Processing?

- Trend:
  - No evaluation on data
  - Evaluation on English (usually Penn Treebank)
    - Rarely something else
    - But usually one language per paper
  - Evaluation on multiple languages
    - Still skewed towards a few families
    - "Big languages" of Eurasia
    - Indo-European, Uralic, Turkic, Semitic, Chinese, Japanese, Korean
  - Resource-poor languages

- Is my algorithm language-independent?
  - Not likely!
  - Test on 4 IE languages does not prove it!
  - Many families missing or underrepresented
  - Some with hundreds of millions of speakers (Austronesian, Niger-Congo)
  - Those languages behave quite differently!

How Many Languages?

- Often cited: 7000 (Ethnologue / SIL)
  - Criticized (Dixon): SIL's aim is translating the Bible
  - Language vs. dialect? Living vs. extinct?
- More realistic: about 4000?
- Many of them endangered

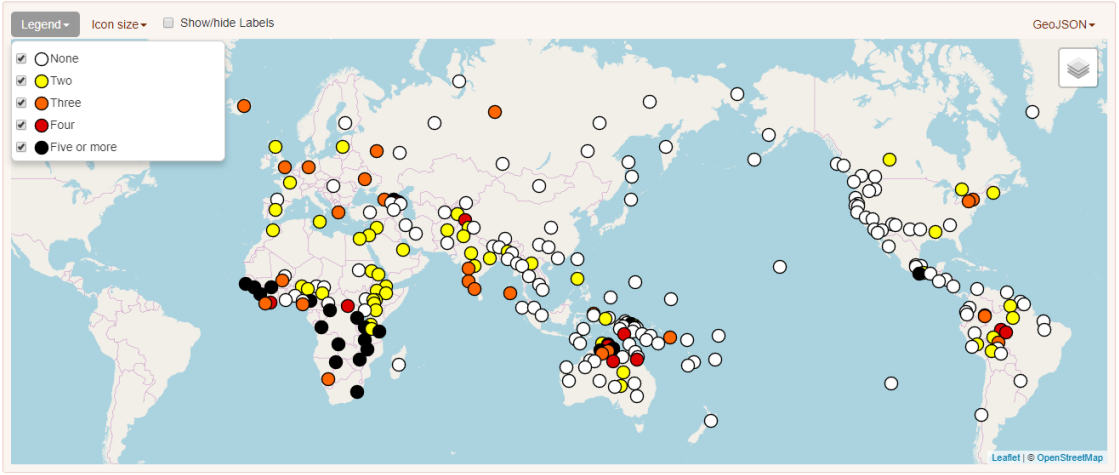
Language Codes

- ISO standard (paid; but unofficial lists are easily obtainable)
- ISO 639-1: two-letter; only major languages
- ISO 639-2: three-letter; more languages; a mess, don't use :-)
  - T-codes: ces, deu, fra, nld, zho, ...
  - B-codes: cze, ger, fre, dut, chi, ...
  - group codes: sla (Slavic), ine (Indo-European), ...
- ISO 639-3: three-letter
  - copy from 639-2/T if exists
  - for other languages: Ethnologue
  - special: mul (multiple langs), mis (langs without code), und (undetermined/unknown), zxx (no linguistic content, e.g. animal sounds)
- Some people/tools use always 639-3
- RFC4646: use 639-1 if available, use three-letter otherwise (e.g. Wiki)

- Glottolog codes: four letters + four digits
  - 8475 entries (`http://glottolog.org/glottolog/language`)

WALS: World Atlas of Language Structures

Number of Genders



WALS: Is It Useful for NLP?

- Yes!
- Database of language features is downloadable
  - Currently 192 features (WALS chapters)
- Similar languages – needed in cross-lingual projection
- But not all features are helpful everywhere!
  - We process text
  - Features 1A to 19A are about phonology
    - E.g. 1A: Consonant Inventories = Moderately small
  - Features 129 to 138 are about lexicon
  - Those that matter may not all have the same weight
- Some features are useful but sparsely annotated
  - Writing system: only indicated for 5 languages

Gender in WALS

- Lexical category of nouns
- Agreement or cross-reference elsewhere:
  - Pronouns
  - Adjectives, determiners (inflection)
  - Verbs (inflection)
  - … or a subset thereof

- Data:
  - Ukrainian and Russian: 3 genders (not 4, with animacy)
  - Czech and Slovak not shown at all
  - English: 3 genders; although only in pronouns!
  - 2 is more similar to 4 than 0 is to 2

Potentially Important Features

- Word order features (18)
- Verbal person marking (4)
- Locus of marking (head marking vs. dependent marking)
- Case (7)
- Endemic function words
  - Copula
  - Question particles in polar questions

SIGTYP 2020 Shared Task

- **Prediction of typological features**
- `https://sigtyp.github.io/st2020.html`

- ⇒ ÚFAL team (from this course) won the task!
- `https://www.aclweb.org/anthology/2020.sigtyp-1.4/`

Other Typological Databases

- Grambank (`https://grambank.clld.org/`)
- … to be expanded …