Google

LinguaMeta: Unified metadata for thousands of languages

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, Emily Drummond
{sandyritchie, dvanesch, uokonkwo, shikharv, emilydrummond}@google.com
LREC-COLING | 20-25 May, 2024

Why does metadata matter?

- Out of 7,000+ world languages, only ~100 of them have widely available language technologies.
- **Reliable, open-source metadata** is necessary to develop new technologies.
    - We need to define languages and know key information about them, such as number of speakers or the writing system(s) they use.
- Metadata can help us with language **definition**, **representation**, and **scoping** for tech expansion.

LinguaMeta

- Language metadata was previously **scattered** across open-source and proprietary repositories
- We expanded our previous research (van Esch et al. 2022) from 2.8K languages to **7.5K+ languages**
  - **Compiled** and **standardized** all metadata
  - Conducted **additional research**
  - Marked each data point with its **source** for better traceability and backlinking
- Available in JSON and TSV format

Categories and sources

- 20+ metadata categories from 8 sources:
  - Language codes (BCP-47, ISO 639-3…)
  - Language names
  - Number of speakers
  - Writing systems
  - Locales
  - Geographic coordinates
  - Official status
  - Endangerment status

Categories and sources

Metadata type	Source(s)	Example for Romanian
ISO 639-3 code	ISO 639	ron
BCP-47 code	IETF	ro
ISO 639-2b code	ISO 639	rum
Deprecated BCP-47 codes	IETF	mo
Glottocode	Glottolog	roma1327
Wikidata code	Glottolog	Q7913
English name	CLDR, Glottolog, Google, IETF, ISO 639, Wikidata, Wiktionary	Romanian
Endonym	CLDR, Glottolog, Wikidata, Google	română
Names in other languages	CLDR, Glottolog, Google, Wikidata	roumain [fr], Rumänisch [de], román nyelv [hu]...
Estimated number of speakers	CLDR, Google, Wikipedia	21,100,000
Writing system(s)	Google, Wikidata, Wiktionary, assumed by locale, GlotScript*	Latin
Locale(s)	Glottolog, Google	Romania, Moldova
Regions	n/a, derived from locales	region: Europe subregion: Eastern Europe
Coordinates	Glottolog, Google	latitude: 46.3913 longitude: 24.2256
Official status	CLDR, Google	official in Romania, Moldova
Endangerment	Glottolog	safe
Scope	ISO 639	individual language
Macrolanguage BCP-47 code	ISO 639	
Individual language BCP-47 codes	ISO 639	
Description	Wikidata	Eastern Romance language

Methodology

- Many metadata categories include data pulled from multiple sources, which were **compiled**, **standardized**, and often **manually curated**.
- We highlight three categories here:
  - Population statistics
  - Writing systems
  - Language names

Population statistics

- Helpful for **prioritizing** technology development
- Includes **total speaker population** estimates and populations **broken down by locale**
- We provide **rounded estimates** (by order of magnitude) to avoid a false sense of exactness
- Required significant manual interventions
  - Issues with counting L1 vs. L2 speakers, multilingual speakers, breakdown by country, etc.

Writing systems

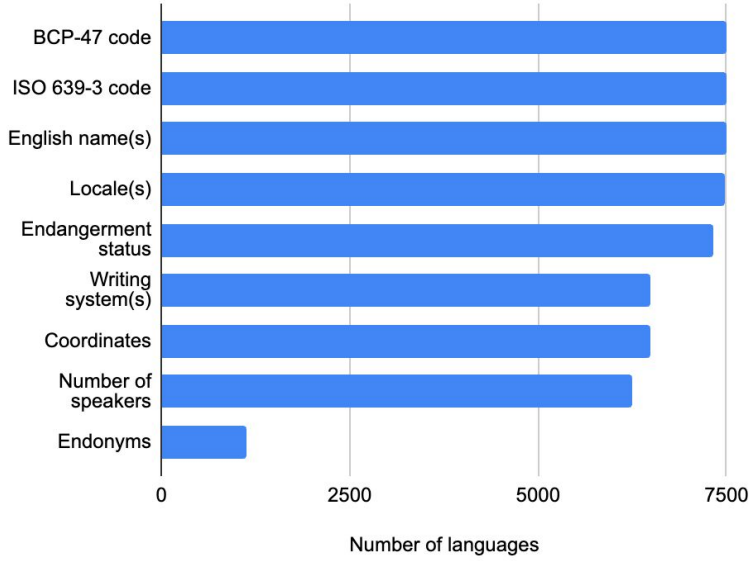
- Required for language technologies that have a **textual component** (which is most of them)
- Many languages are written with **different scripts** in different contexts
  - Scripts can **vary by locale** (e.g. Pakistan vs. India)
  - Developed a **set of tags** for additional contexts of use (e.g. official use, transliteration)
- We supplemented our data with **locale-based assumptions** where Latin is the sole dominant script

Language names

- Extensive name data allows us to develop more accurate and well-localized **user interfaces**
- LinguaMeta includes **many names** for every language
  - Marks **one name as “canonical”** to facilitate technical applications, but also includes alternative names
  - Includes over 2200 **endonyms**
- Many are **tagged with the script** they appear in

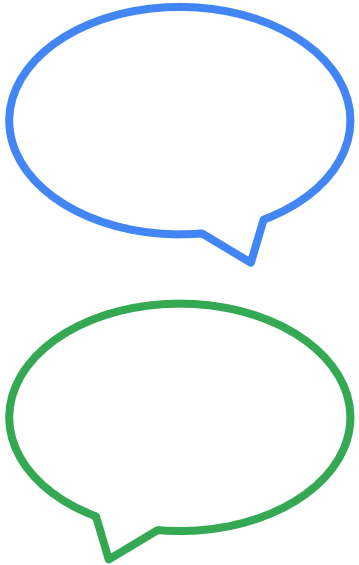
Coverage

- We analyze % coverage for a number of categories where full coverage is theoretically possible
  - **~100% coverage:** codes, English names, locales, and endangerment status
  - **~85% coverage:** writing systems, coordinates, and number of speakers
  - Only **15% coverage for endonyms**, which are largely manually curated



Extensions and future work

- Improve quality through **feedback** from research and speaker communities
- Add and improve several categories
  - Add **language varieties** beyond ISO standard
  - Refine and collect new **writing system** information, using e.g. Glotscript (Kargaran et al. 2023)
  - Improve **population statistics by locale,** which may require more detailed census data



Thank you