TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP

Steven Moran[1], Christian Bentz[2], **Ximena Gutierrez-Vasques**[3], Olga Sozinova[3], Tanja Samardzic[3]

University of Neuchatel[1]
University of Tübingen[2]
URPP Language and Space, University of Zürich[3]

LREC 2022

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

1. What is TeDDi Sample?

   Motivation
   Goal
   Data collection and curation
   Data availability

2. Demo

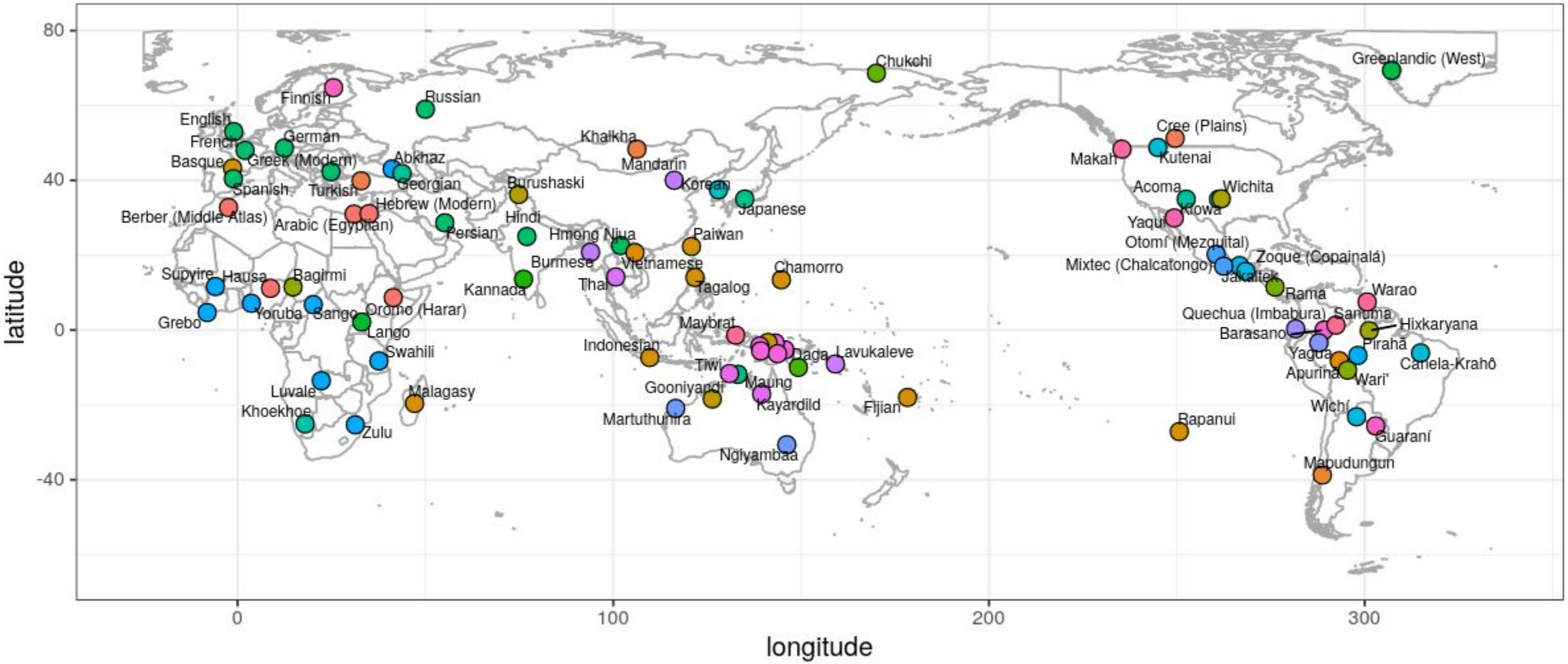
What is TeDDi Sample?

A diversity sample of text data. It features:

What is TeDDi Sample?

A diversity sample of text data. It features:

- Text corpora for languages on the typological diversity sample in the World Atlas of Language Structures (WALS)


80
40
0
-40
latitude
0
100
200
300
longitude
Chukchi
Greenlandic (West)
Finnish
Russian
English
French
German
Basque
Greek (Modern)
Spanish
Turkish
Abkhaz
Georgian
Hebrew (Modern)
Arabic (Egyptian)
Berber (Middle Atlas)
Persian
Khalkha
Mandarin
Korean
Japanese
Burushaski
Hindi
Hmong Njua
Paiwan
Burmese
Vietnamese
Thai
Kannada
Tagalog
Chamorro
Supyire
Hausa
Bagirmi
Grebo
Yoruba
Sango
Oromo (Harar)
Lango
Swahili
Luvale
Khoekhoe
Zulu
Malagasy
Indonesian
Maybrat
Tiwi
Gooniyandi
Martuthunira
Ngiyambaa
Maung
Kayardild
Daga
Lavukaleve
Fijian
Cree (Plains)
Makah
Kutenai
Acoma
Wichita
Kiowa
Yaqui
Otomí (Mezquital)
Mixtec (Chalcatongo)
Zoque (Copainalá)
Jakaltek
Rama
Warao
Quechua (Imbabura)
Sanuma
Barasano
Hixkaryana
Pirahã
Yagua
Apurinã
Wari'
Canela-Krahô
Wichí
Guaraní
Mapudungun
Rapanui

What is TeDDi Sample?

A diversity sample of text data. It features:

- Text corpora for languages on the typological diversity sample in the World Atlas of Language Structures (WALS)

- Currently 89 languages. More than 20k texts

What is TeDDi Sample?

A diversity sample of text data. It features:

- Text corpora for languages on the typological diversity sample in the World Atlas of Language Structures (WALS)

- Currently 89 languages. More than 20k texts

- Open-source corpus processing tools

Motivation

- Access to data from minority and under-resourced languages leads to deeper and more complete understanding of language

- Multilingual datasets are becoming increasingly important in NLP

Motivation

How to select languages?

Motivation

How to select languages?

Linguistics perspective

- More weight on representing a wide range of language families and areas, structural features

NLP Perspective

- Favors languages for which text data is readily available online.

Goal

The aim of TeDDi is to facilitate the use of text-based quantitative methods for analyzing linguistic diversity in both linguistic research and NLP:

Goal

The aim of TeDDi is to facilitate the use of text-based quantitative methods for analyzing linguistic diversity in both linguistic research and NLP:

- Approximate the diversity of languages across the world by means of text samples

- Complement the existing knowledge about the structure of languages, which mostly consists of high-level feature-value pairs stored in linguistic databases

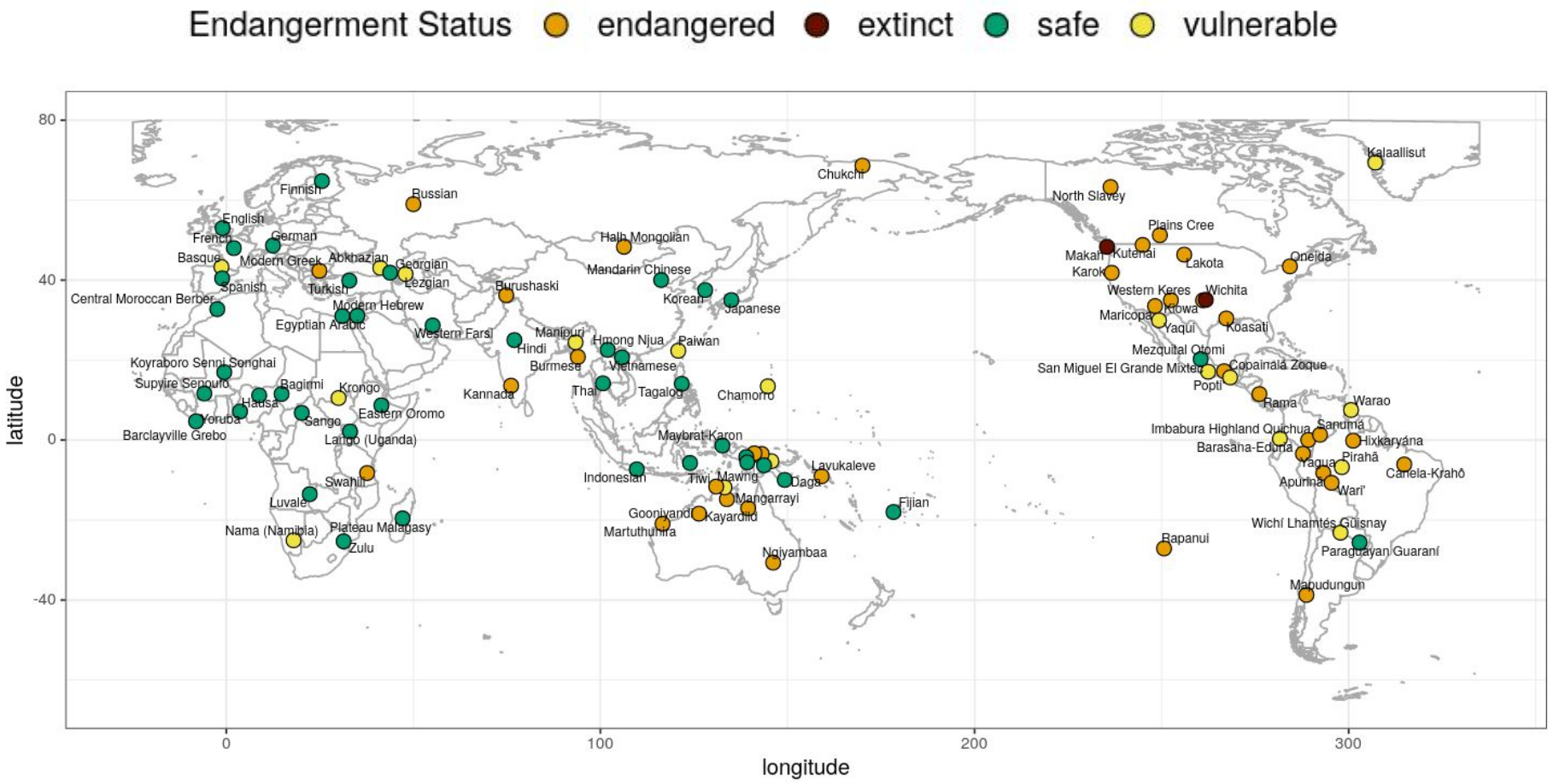
Data collection and curation

Based on WALS 100-language sample

Maximizes genealogical (language family) and areal (geographic) diversity

Selection is independent of text data availability

Data collection and curation



 TeDDi Sample. S. Moran, C. Bentz, X. Gutierrez, O. Sozinova and T. Samardzic.

Data collection and curation

Existing resources

- Project Gutenberg, Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights

Data collection and curation

Existing resources

- Project Gutenberg, Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights

Manually collected translations, transcriptions, and grammatical annotations

Data collection and curation

Existing resources

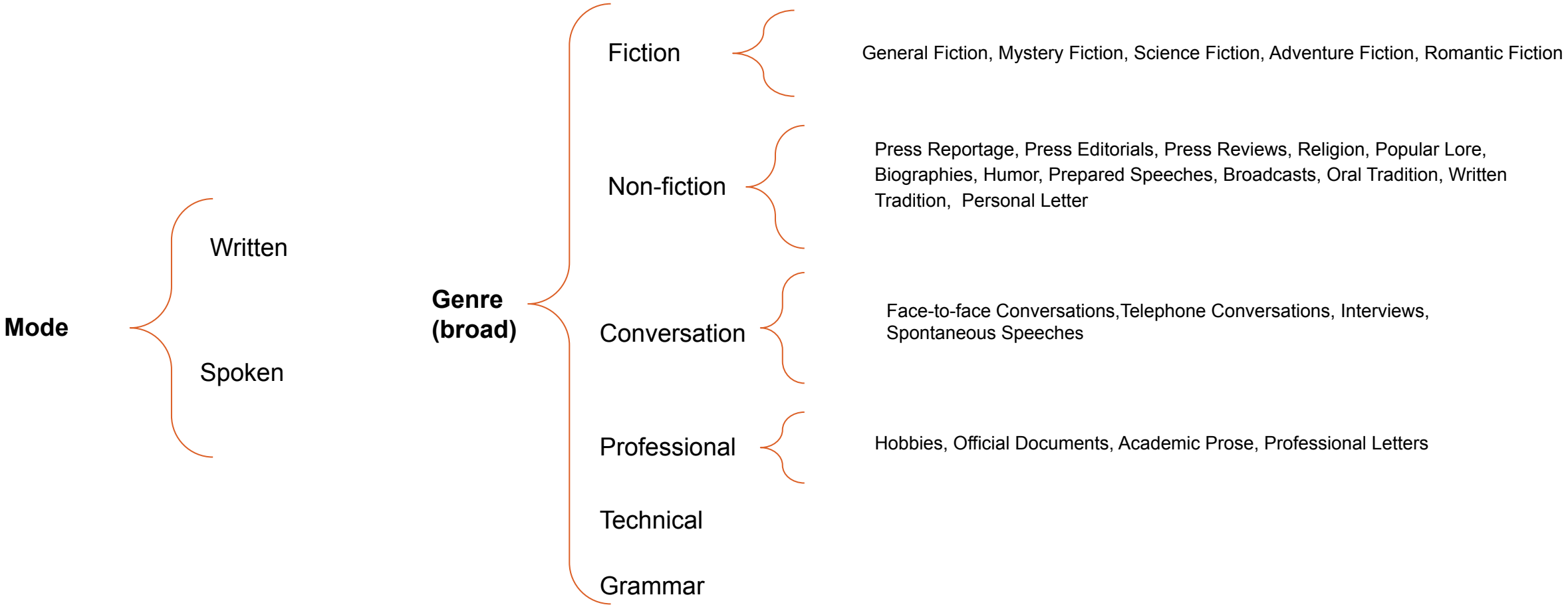
- Project Gutenberg, Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights

Manually collected translations, transcriptions, and grammatical annotations

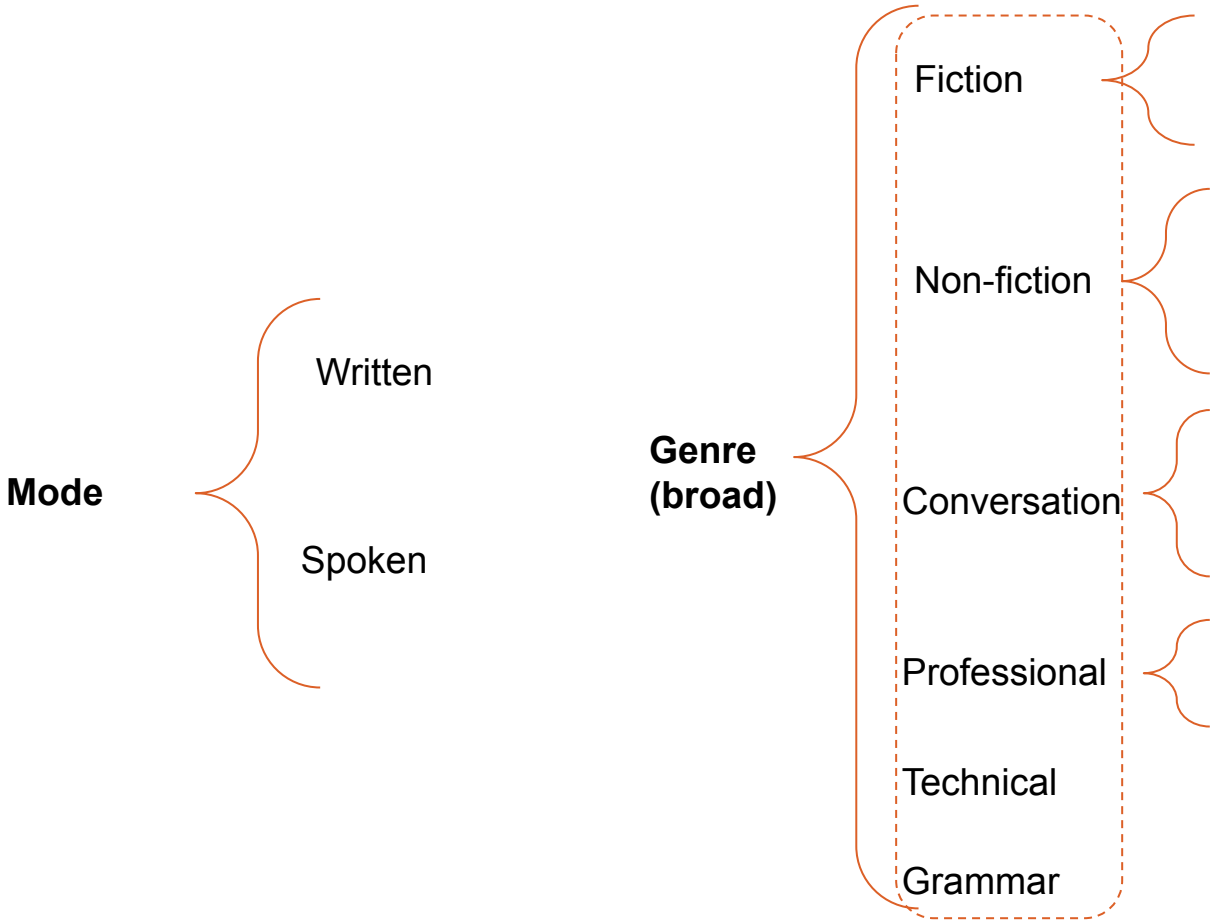
Sampling (for rich-resource languages)

- We do not include all of the texts available. Instead, we create smaller samples limiting the maximal size of a text unit to 50k tokens of contiguous text

Data collection and curation



Data collection and curation



Genre (narrow)

General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction

Press Reportage, Press Editorials, Press Reviews, Religion, Popular Lore, Biographies, Humor, Prepared Speeches, Broadcasts, Oral Tradition, Written Tradition, Personal Letter

Face-to-face Conversations,Telephone Conversations, Interviews, Spontaneous Speeches

Hobbies, Official Documents, Academic Prose, Professional Letters

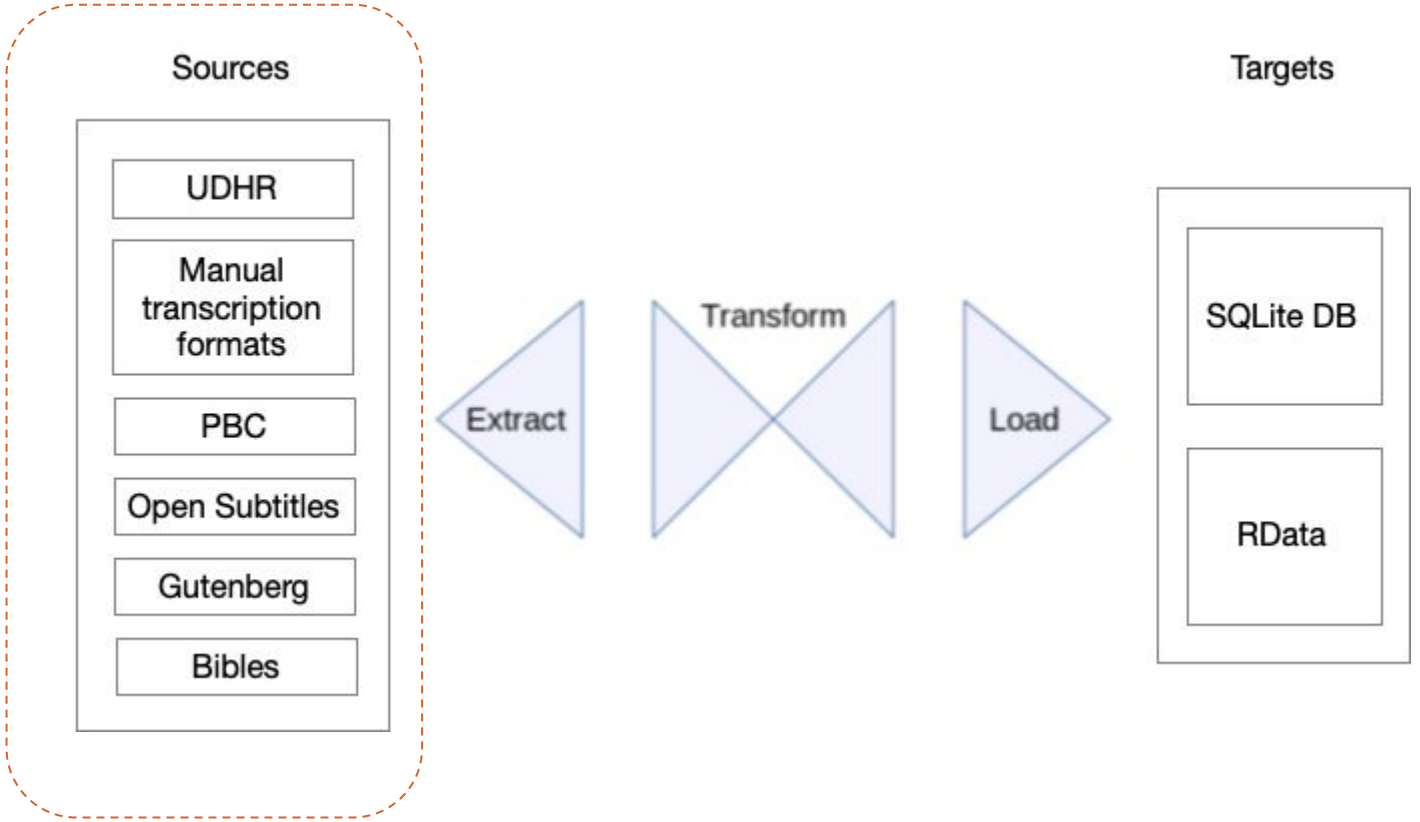
Data collection and curation

Genre	Langs*	Tokens	Scripts†
conversation	10	15,835	1
fiction	12	36,811,339	7
grammar	5	1271	1
nonfiction	73	101,588,748	13
professional	40	80,092	15
Total	**89**	**ca. 138 million**	**16**

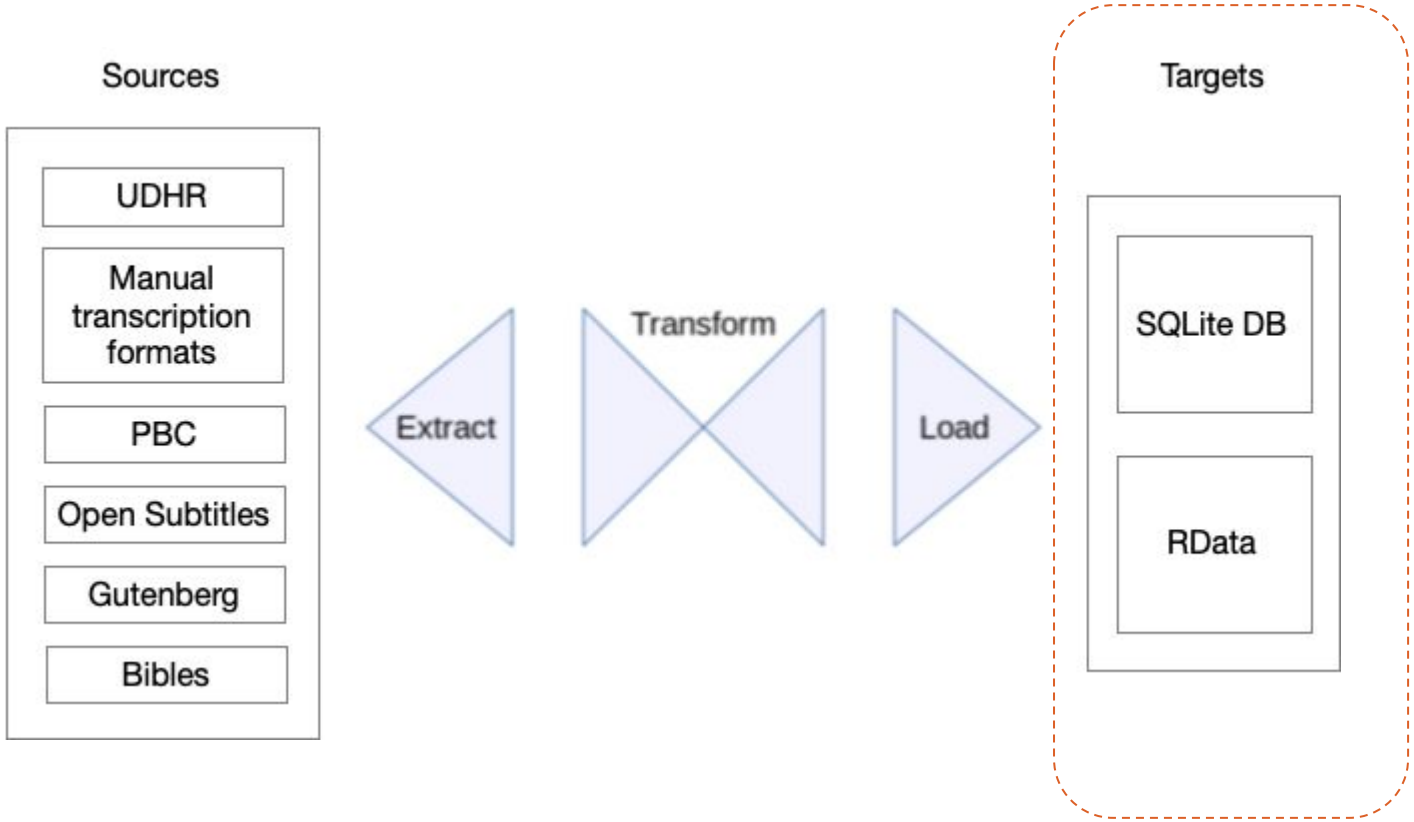
*According to ISO-639-3 codes.
†According to ISO-15924 codes.

- 89 different languages, stemming from 58 language families (according to WALS)

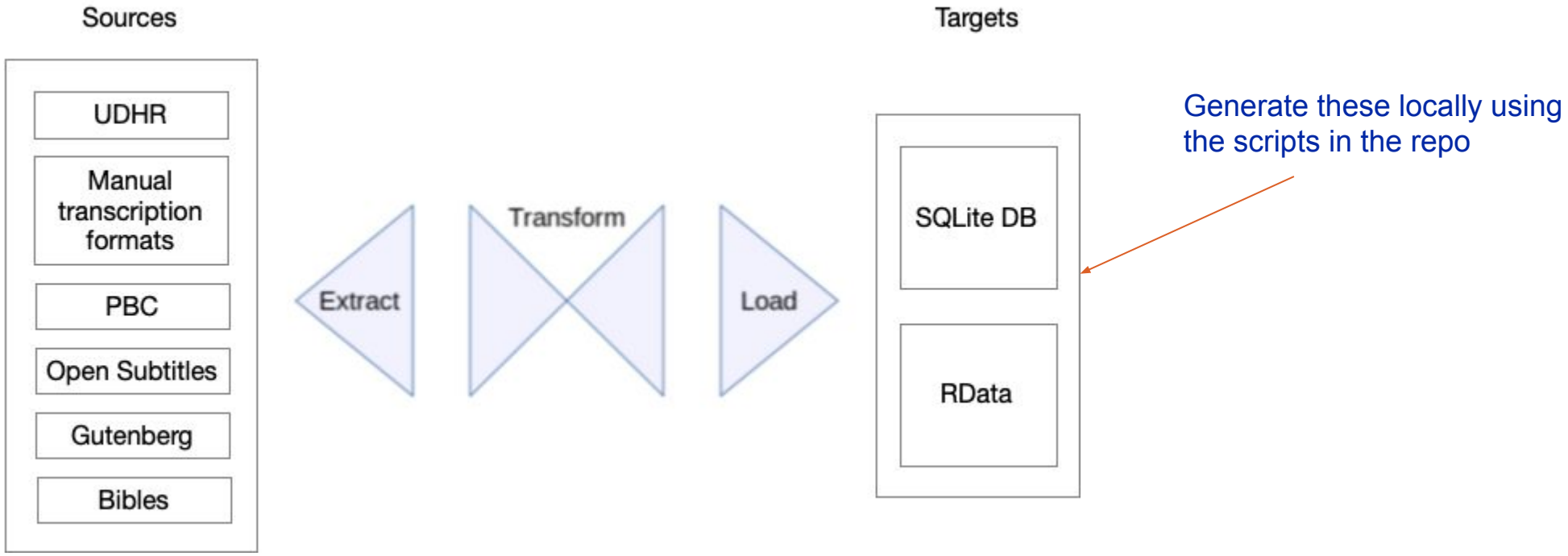
Data availability



Data availability



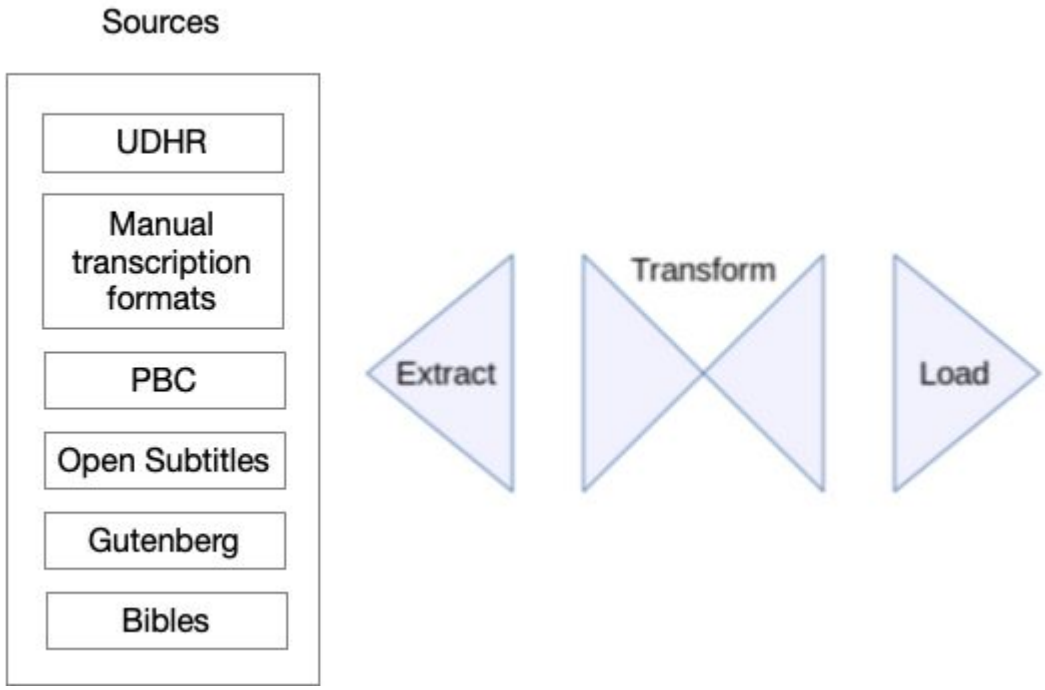
Data availability



Generate these locally using the scripts in the repo

Source text files available directly in the repo

Data availability

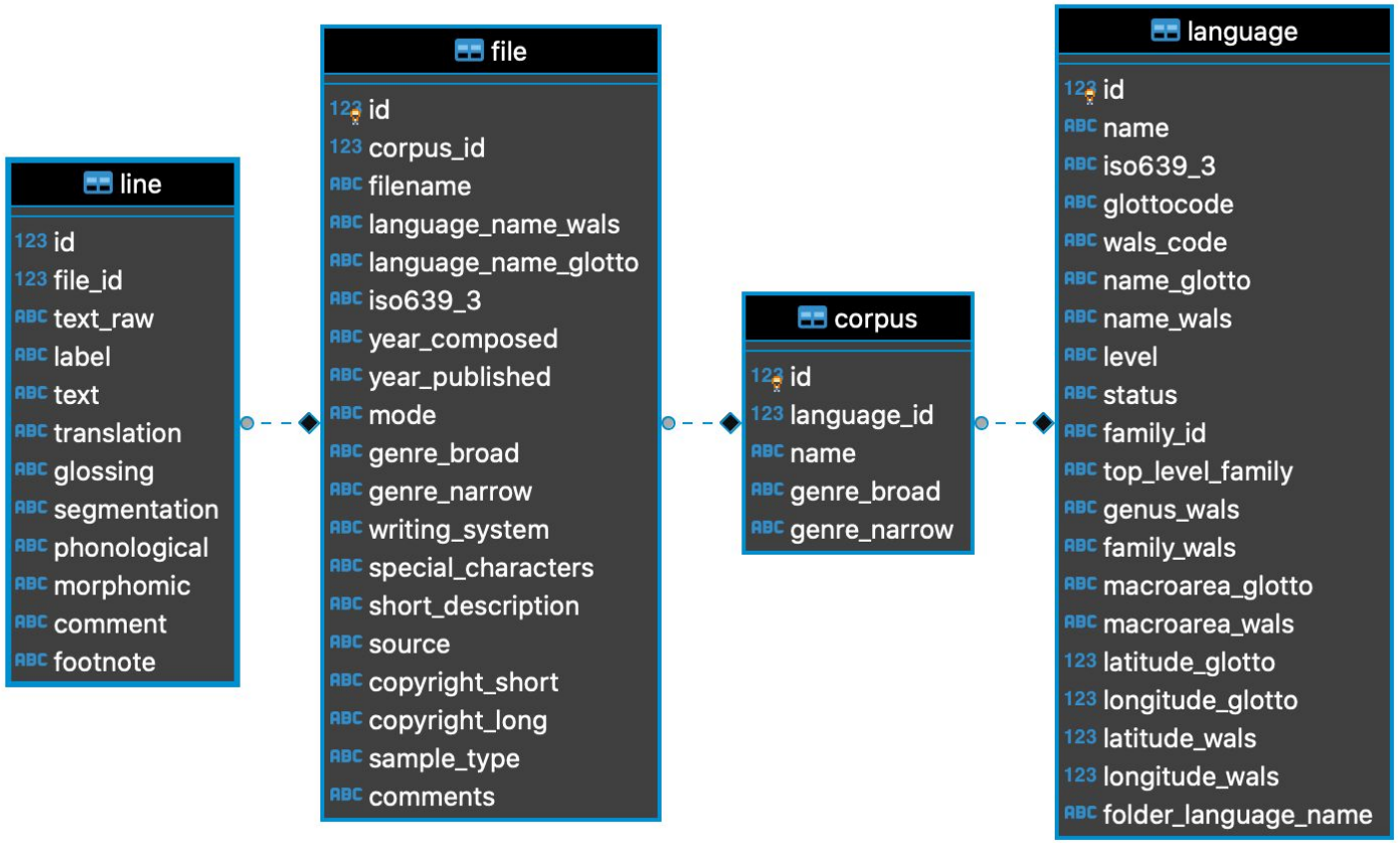


Generate these locally using the scripts in the repo

*Database dumps are also available online

Source text files available directly in the repo

Data availability



TeDDi's relational database schema

Data availability

Input corpora and the source code for processing them:

https://github.com/MorphDiv/TeDDi_sample

Data exported into CLDF (Cross-Linguistic Data Formats)

https://github.com/cldf-datasets/TeDDi_sample

DEMO

Thank you!