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

Steven Moran¹, Christian Bentz², Ximena Gutierrez-Vasques³, Olga Sozinova³, Tanja Samardzic³

University of Neuchatel¹
University of Tübingen²
URPP Language and Space, University of Zürich³

LREC 2022
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)
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

Endangerment Status  ○ endangered  ● extinct  ● safe  ○ vulnerable

[Map showing distribution of endangered languages around the world, with markers indicating different endangerment statuses.]
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

Mode

- Written
- Spoken

Genre (broad)

- Fiction
- Non-fiction
- Conversation
- Professional
- Technical
- Grammar

Genre (narrow)

- Fiction: General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction
- Conversation: Face-to-face Conversations, Telephone Conversations, Interviews, Spontaneous Speeches
- Professional: Hobbies, Official Documents, Academic Prose, Professional Letters
Data collection and curation

Mode
- Written
- Spoken

Genre (broad)
- Conversation
- Professional
- Technical
- Grammar
- Fiction
- Non-fiction

Genre (narrow)
- General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction
- Face-to-face Conversations, Telephone Conversations, Interviews, Spontaneous Speeches
- Hobbies, Official Documents, Academic Prose, Professional Letters
## Data collection and curation

<table>
<thead>
<tr>
<th>Genre</th>
<th>Langs*</th>
<th>Tokens</th>
<th>Scripts†</th>
</tr>
</thead>
<tbody>
<tr>
<td>conversation</td>
<td>10</td>
<td>15,835</td>
<td>1</td>
</tr>
<tr>
<td>fiction</td>
<td>12</td>
<td>36,811,339</td>
<td>7</td>
</tr>
<tr>
<td>grammar</td>
<td>5</td>
<td>1271</td>
<td>1</td>
</tr>
<tr>
<td>nonfiction</td>
<td>73</td>
<td>101,588,748</td>
<td>13</td>
</tr>
<tr>
<td>professional</td>
<td>40</td>
<td>80,092</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>89</strong></td>
<td><strong>ca. 138 million</strong></td>
<td><strong>16</strong></td>
</tr>
</tbody>
</table>

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

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

Sources

- UDHR
- Manual transcription formats
- PBC
- Open Subtitles
- Gutenberg
- Bibles

Targets

- SQLite DB
- RData
Data availability

Sources
- UDHR
- Manual transcription formats
- PBC
- Open Subtitles
- Gutenberg
- Bibles

Transform

Targets
- SQLite DB
- RData
Data availability

Source text files available directly in the repo

Generate these locally using the scripts in the repo

Source:
- UDHR
- Manual transcription formats
- PBC
- Open Subtitles
- Gutenberg
- Bibles

Targets:
- SQLite DB
- RData
Source text files available directly in the repo

Generate these locally using the scripts in the repo

*Database dumps are also available online
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

CC BY-NC-SA 4.0 license*

*The particular copyright is given in the metadata header for each text

* “takedown principle”, i.e., we can remove such material if contacted by people aggrieved by it (Seyfeddinipur et al., 2019, p. 554)
DEMO
Thank you!