Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1150–1158
Marseille, 20-25 June 2022

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 Zurich[3]
Neuchatel, Switzerland[1], Tübingen, Germany[2], Zurich, Switzerland[3]
steven.moran@unine.ch, chris@christianbentz.de
{ximena.gutierrezvasques, olga.sozinova, tanja.samardzic}@uzh.ch

Abstract

We present the TeDDi sample, a diversity sample of text data for language comparison and multilingual Natural Language Processing. The TeDDi sample currently features 89 languages based on the typological diversity sample in the World Atlas of Language Structures. It consists of more than 20k texts and is accompanied by open-source corpus processing tools. The aim of TeDDi is to facilitate text-based quantitative analysis of linguistic diversity. We describe in detail the TeDDi sample, how it was created, data availability, and its added value through for NLP and linguistic research.

Keywords: Corpora, Quantitative Typology, Language Diversity, Language Documentation

1. Introduction

Following a long debate on the status of linguistic variation, the need to move beyond a limited set of WEIRD languages (Henrich et al., 2010; Majid and Levinson, 2010) is becoming widely recognized. A deeper and more complete understanding of language is being achieved through increased access to data from minority and low-resource languages. The same tendency is visible in NLP, where new multilingual datasets are currently released at a fast pace. These datasets, used for training and testing language models, have become especially interesting in the context of cross-linguistic transfer with few-shot and or even zero-shot learning.

The question of how to select languages to be included in multilingual samples is approached differently in the two fields. Linguists put more weight on representing a wide range of language families and areas, as well as structural features, collecting the data from grammars, and storing them in typological databases. Researchers in NLP, on the other hand, favor languages for which text data is readily available online.

Here, we present the TeDDi sample, which constitutes an intersection between the two approaches. Namely, it contains text samples for a selection of languages from the World Atlas of Language Structures (WALS) – spanning diverse families and areas. Since this selection is independent of text data availability, some languages in the sample have rich resources (e.g., English, Russian, Japanese), while others are only documented through fieldwork (e.g., Rama, Kayardild, Bagirmi). The challenge with resource-rich languages is how to select the texts to be included in the sample. The challenge with low-resource languages is entirely different, namely, finding, extracting and digitizing texts from low-resource sources, e.g., published grammars.

In the current version, our resulting sample consists of more than 20K texts. It is accompanied with a set of open-source processing tools. Our goal is to facilitate the use of text-based quantitative methods for analyzing linguistic diversity in both linguistic research and NLP.

We first present an overview of the language sample and data collection and curation processes in Section 2. In Section 3, we describe the TeDDi sample database development and data availability. Lastly, in Section 4, we discuss current and future research prospects using the TeDDi sample.

2. Data collection and curation

2.1. Language sample

The TeDDi sample aims to include text corpora for languages of the one hundred language sample provided in The World Atlas of Language Structures (WALS; Dryer and Haspelmath (2013)).[1] WALS is an atlas of worldwide linguistic diversity and it describes the structural features and geographic locations of 2676 languages. The WALS editors defined a core sample of one hundred languages which maximizes genealogical (language family) and areal (geographic) diversity. The aim was to minimize bias leading to a false picture of the relative frequency of different types of languages (Comrie et al., 2013).

While the 100 WALS sample aims to maximize areal, genealogical, and structural diversity, there are a few shortcomings (Comrie et al., 2013) which we briefly note here. First, given that the language sample is comprised of one hundred languages, it does not sample from each and every of the 427 known language families (Hammarström et al., 2021). Second, in some cases, editorial decisions were taken to include more than one data point from large language families. For

[1] https://wals.info/languoid/samples/100

Mode	Genre (broad)	Genre (narrow)	Source example
written	Fiction	General Fiction	OPUS: Books
written	Fiction	Mystery Fiction	-
written	Fiction	Science Fiction	-
written	Fiction	Adventure Fiction	-
written	Fiction	Romantic Fiction	-
written	Non-Fiction	Press Reportage	OPUS: GlobalVoices
written	Non-Fiction	Press Editorials	-
written	Non-Fiction	Press Reviews	OPUS: NewsCommentary
written	Non-Fiction	Religion	Bible Parallel Corpus
written	Non-Fiction	Popular Lore	Wikipedia Dumps
written	Non-Fiction	Biographies	-
written	Non-Fiction	Humor	-
written	Non-Fiction	Prepared Speeches	OPUS: OpenSubtitles2018
written	Non-Fiction	Broadcasts	-
spoken	Non-Fiction	Oral Tradition	-
written	Non-Fiction	Written Tradition	-
written	Non-Fiction	Personal Letters	-
spoken	Conversation	Face-to-face Conversations	-
spoken	Conversation	Telephone Conversations	-
spoken	Conversation	Interviews	-
spoken	Conversation	Spontaneous Speeches	SketchEngine: CHILDES
written	Professional	Hobbies	-
written	Professional	Official Documents	SketchEngine: Eur-Lex
written	Professional	Academic Prose	-
written	Professional	Professional Letters	-
written	Technical	-	OPUS: Ubuntu
spoken/written	Grammar	-	-

Table 1: Correspondence between genres, modes, potential sources (the current version of the data set does not include all listed sources).

instance, overall eight languages of the Austronesian language family are included in the sample – even though Austronesian is quite uniform in terms of its structural features across over 1000 languages. However, it spans a large geographic area, and having only one or two points in the Pacific would look sparse on a map. Third, a decision that all cross-linguistic resource compilers face is the availability of detailed grammatical descriptions. This is also known as the bibliographic bias in linguistic typology (Bakker, 2011; Moran, 2012). This hampers the inclusion of languages – in particular language isolates – for which there exist no texts or grammatical descriptions.

Therefore, the one hundred language sample contains well-known languages with many resources (e.g., English, French, Russian) as well as low-resource and endangered languages for which detailed linguistic descriptions and texts exist (e.g., conversational data from a grammar of Kayardild (Round, 2012); minority languages represented in the parallel bible translations (Mayer and Cysouw, 2014)).

The choice of using the WALS 100 language sample has two major benefits for capturing worldwide linguistic diversity. First, we target collections of texts that are not simply opportunistic and accessible, and as such, contribute to a growing amount of digital resources of low-resource and minority languages. Second, any analyses or measures derived from raw or linguistically annotated texts in the TeDDi sample are directly comparable to associated linguistic structures encoded in each of the 192 features in WALS. This enables direct comparisons of the relative frequency of tokens and types in text data to the cross-linguistic frequency of linguistic types as reported in the WALS across phonology, the lexicon, morphology, word order, etc. We discuss possible use cases in more detail in Section 4.

2.2. Text sources

To extract texts for the one hundred language sample, we use existing resources, e.g., Project Gutenberg,[2] Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights.[3] These resources cover around one half of our target languages. For the rest of the sample, we turn to sources of language documentation and description: manually collected translations, transcriptions, and grammatical annotations. Given that the available resources for languages greatly

[2] https://www.gutenberg.org/

[3] http://unicode.org/udhr/

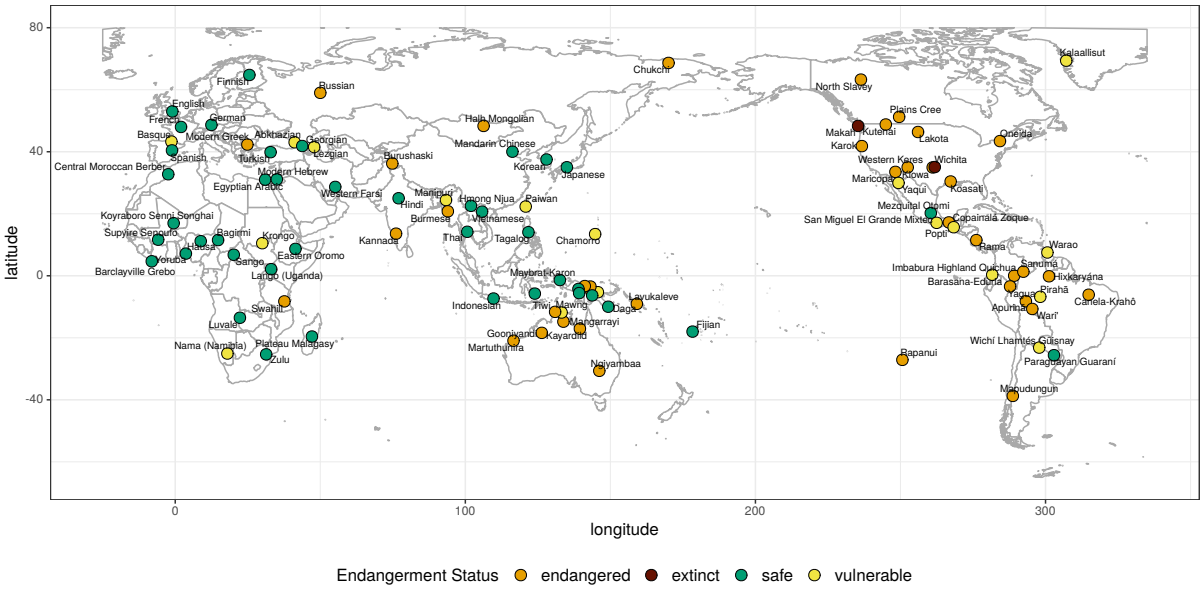

Figure 1: WALS 100 language sample with endangerment status. These languages stem from 68 language families according to WALS, and 61 top level families according to Glottolog.

vary in terms of size, we take several design decisions on how and which data to include.

2.3. Sampling texts from rich resources

For some languages, the number of available texts is very large. To keep the overall size of our data set easy to manage, 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. This is the size at which quantitative measures like unigram entropy reach stabilization (Bentz et al., 2017b; Bentz et al., 2017a). In addition to the size, we limit the number of text units to 100 per online source. Thus the total size of the data for all the available languages extracted from the same source cannot be greater than 5,000,000 tokens.

We implement sampling in web crawlers written to collect the data from the original web pages. For each language in each online resource, we perform the following:

1. Identify how many samples of 50,000 tokens can be drawn from the text.
2. Identify the potential starting points. These are typically at the beginning of a sentence, but they can be defined in terms of smaller or bigger units depending on the genre of the text.
3. Choose a random starting point for the current sample.
4. Store 50,000 tokens following the starting point:
   (a) If the end of the text is reached before the given size, store this piece and continue from the beginning of the text until the sample has 50,000 tokens, then store it.
5. Continue sampling from the remaining text.

The current approach for identifying starting points is relatively simple and might be improved in the future. For now, the program looks for the starting points which appear directly after blank lines (visual division of the text), or after short lines (which usually show the end of paragraphs). If there are no blank lines and no short lines, then the starting point is a random line, which begins right after a carriage return symbol. Sometimes, we look for specific punctuation marks before the carriage return symbol depending on a source or genre. For example, in the OpenSubtitles corpus, we prefer lines ending in a question mark as potential starting points.

To represent various genres, we divide all available resources into a number of categories and then aim to collect at least one text unit from each category. We agreed on six broader categories: *fiction*, *non-fiction*, *conversation*, *professional*, *technical*, and *grammar*. The first five categories are obtained by aggregating 23 genres identified empirically in corpus linguistics (Biber, 1991). We added the sixth category to accommodate the examples found in grammars. Table 1 shows the correspondence between broader categories, which we use to describe genres, the original fine-grained genres, and the mode (written or spoken). The last column contains a few examples of how available online resources can be classified with respect to the genre. Table 2 shows the current size of samples per genre.

Table 2: Summary statistics.

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.

2.4. Metadata

Without detailed metadata, comparative analyses of text samples are often not straightforwardly interpretable (Koplenig, 2017). Therefore, for each text unit we provide a metadata header inspired by the format in the Parallel Bible Corpus (Mayer and Cysouw, 2014). It consists of a two column tab-delimited list of metadata categories, including standardized information about the language, the text, its mode and genre, when it was collected, etc. Two features particularly relevant from a corpus linguistic point of view are the *modality* (or *mode*) and the *genre*. To each text we assign a mode (spoken or written),[4] and a *broad genre* as described above. All the metadata fields and a description of their values is given below:

- `language_name_wals`: language name in the WALS 100 language sample;
- `language_name_glotto`: language name in Glottolog 4.5;
- `iso639_3`: ISO 639-3 code as a unique language name identifier;
- `year_composed`: year in which the text was written or recorded;
- `year_published`: year in which the text was published;
- `mode`: spoken or written;
- `genre_broad`: broad genre (conversation, fiction, grammar, nonfiction, professional, technical);
- `genre_narrow`: narrow genre;
- `writing_system`: ISO 15924 four letter code identifying the script used in the text (e.g., Latin: Latn, Cyrillic: Cyrl);[5]
- `special_characters`: particular characters/diacritics introduced in a text;
- `short_description`: short description of the content of the text (e.g., an English title given to oral stories);
- `source`: URL (with date) for online texts; bibliographic reference for books, articles etc.
- `copyright_short`: some sources give specific short copyright phrases which are repeated here;
- `copyright_long`: full copyright statement as given by the source;
- `sample_type`: 'whole' (for the documents containing less or equal to 50K tokens) or 'part' (for the samples taken from a larger document);
- `comments`: further comments that are necessary for understanding the transcriptions of texts.

[4] A third mode *signed* is possible, but our collection does not currently include transliterated texts of sign languages.

[5] `https://en.wikipedia.org/wiki/ISO_15924`

3. Resource Development

Since each text in the TeDDi sample is potentially of a different format, e.g., free text, parallel text, annotated interlinear glossed text, we had to develop a pipeline to extract, transform, and load (ETL) the data into a syntactically and semantically interoperable format. In the following sections, we describe our ETL pipeline.

3.1. Text Input Formats

At a minimum, each text file entered into the TeDDi sample needs to contain: 1) a metadata header; 2) lines of text written in the respective language and script specified in the metadata header. Thus, each file is divided into a metadata header and a body of text. The metadata header includes all consecutive lines that are prefixed with a hash symbol. The body is comprised of the rest of the text in the file. For this body of text, there are currently ten different input formats in the TeDDi sample:

1. Universal Declaration of Human Rights (UDHR),
2. Manual/Transkribus transcription,[6]
3. Parallel Bible Corpus (PBC),
4. Manual transcription with translation,
5. Manual transcription with glossing,
6. Manual transcription with further annotation layers,
7. Open Subtitles,
8. Project Gutenberg,
9. Hand annotated bibles,
10. Paragraph based format.

[6] `https://readcoop.eu/transkribus/`

An illustration of the UDHR in Central Moroccan Berber with metadata header and text body is given in Figure 2. This corresponds to format number 1 above, in which there is no further annotation at all – just plain text.

```
# language_name_wals:   Berber (Middle Atlas)
# language_name_glotto: Central Moroccan Berber
# iso639_3: tzm
# year_composed:    1996 – 2009
# year_published:   NA
# mode: written
# genre_broad:  professional
# genre_narrow: official_documents
# writing_system:   Latn
# special_characters:   NA
# short_description:    Universal Declaration of Human Rights
# source:   https://unicode.org/udhr/d/udhr_tzm.txt (30/01/2019)
# copyright_short:  © 1996 – 2009 The Office of the High Commissioner for Human Rights. This plain text version prepared
# copyright_long:   NA
# sample_type:  whole
# comments: NA

TIṢERRIḤT TAGRAGHLANT IZERFAN N WEMDAN
    Yesbudan d yesse ran deg wegraw amatu seg use ti-s 217 A. (III) di 10 dujember 1948

    ANAKCHAM
    Imi asmussen n lḥwerma i ttalasen akkw yâggalen n twachult talsawt d yizerfan n sen yemsawan, d nitni i d llsas n tle
    Imi kra n widn nesmussun ara izerfan n wemdan d widn iheqqren s lkhwedm n lewhuch yesserfayen tamsakwit n talsa akkw
    Imi tebbwi – dd nnig kra yellan ad ttuhudden izerfan n wemdan s nnidam azerfan i wakkn ur yettuḥerrs ara wemdan di ta
    Imi yessefk ad tennerni tegmi d wassaghen n tmidwa d lemḥibba gar yeghlanen.
    Imi di lqanun izerfan n wemdan imudan n ledjnas yedduklen berrḥen i tikkelt tajdidt laman n sen deg zerfan ilsasiyen
    Imi, timura yetteskin di ledjnas yedduklen lezment iman n sent ad myallent (ad mmâwanent) i wakken ad demnent nitenti
    Imi afham amechruk n tlella d yizerfan agi degs lqima tameqwrant i wakken ad uqment leqder tmura leqder i lâhd n sent
    Tajmâit tamatawt
    Tberreḥ belli tiserriḥt agi tagraghlant izerfan n wemdan d anawad amechruk igher ssaramen imudan d yeghlanen akken ma
```

Figure 2: Example of UDHR format with metadata header and plain text.

Another example, in this case of format number 6, is given in Figure 3. Here, several layers of annotation (phonological, segmentation, glossing, etc.) are available. The availability of annotation layers depends on the source the text is taken from.

```
<line_1>        Dankiya kunawunaya barjijarranth !
<phonological>  ŋankia kunaunaja paɻcicaraɳʈa
<segmentation>  ŋan+ki-a kuna+kuna+ki-a paɻci-c+ŋara-iɳʈa-ø
<morphomic>     this-μloc-t ‹child_NL-child_NL›-μloc-t ‹fall-j›-μ̃cons-μobl-t
<glossing>      this-CMP ‹child›-CMP ‹fall›-PST-SEJ
<translation>   ‘ This child has been born ! ’
<comment>       This is example 1.3 in Round (2013), p. 10
```

Figure 3: Example of a sentence with multiple annotation layers extracted from a Kayardild grammar (Round, 2012).

3.2. Data Transformation

Our data extraction and aggregation pipeline accepts these formats as input and outputs a unified relational database. Our pipeline follows the ETL paradigm and parses the different text corpus input formats and corpus-specific annotation schemes, and then brings them together into a structurally and conceptually interoperable database. This process is illustrated in Figure 4.

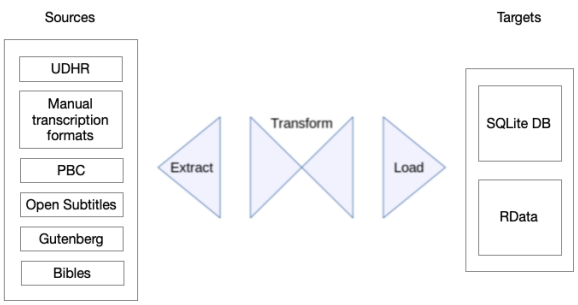

Figure 4: TeDDi sample database aggregation pipeline.

We have written the ETL pipeline in Python (Van Rossum and Drake, 2009) using SQLAlchemy (Bayer, 2012), an object-relational mapper with which we load the input parsed corpus data into a relational database model. Our current output formats are a SQLite database (Hipp, 2020), an R data object, and the Cross-Linguistic Data Format (CLDF; Forkel et al. (2018)). We describe each of these output formats in turn.

Our relational database schema is comprised of four tables, as illustrated in Figure 5. The main table is the LANGUAGE table, which is related in a cascading one-to-many relationship with the CORPUS, FILE, and LINE tables.

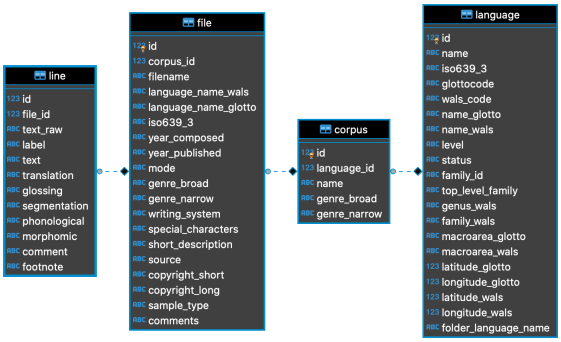

Figure 5: TeDDi relational database schema.

The LANGUAGE table contains metadata about our language sample, including for example: ISO 639-3, Glottolog, and WALS language codes; genealogical and geographic information about each language from Glottolog and WALS; and information about where the input files for each language reside. The CORPUS table provides information about the file folder structure, which divides the files into genres, as described above. The FILE table includes information about each file in the TeDDi sample, including all metadata in each file's metadata header.

The body contains the textual information for each file. Our ETL pipeline identifies each input format and calls an input format specific parsing routine to extract the text and annotations. These data are then stored in the LINE table, in which each row in the table represents a line in the input text and annotation information about that line is given in separate columns. An illustration is given in Figure 6.

We chose SQLite as our database format because it is public domain, requires no special setup or configuration, and users of our GitHub repository can easily generate the database themselves – with all or some subset of the corpora. Moreover, because we use an object-relational mapping process in our ETL pipeline, users can also easily load the data into different database management systems, so they are not limited to SQLite.

Figure 6: Example of TeDDi sample database line table.

id	file_id	text_raw	label	text	translation	glossing	segmentation	phonological	morphomic	comment
79477	19	ತನ್ನ ಕರ್ತೃತ್ವದಲ್ಲಿಯಾದ ಯಾವುದೇ ವೈಜ್ಞಾನಿಕ, ಸಾಹಿ...	NULL	ತನ್ನ ಕರ್ತೃತ್ವದಲ್ಲಿಯಾದ ಯಾವುದೇ ವೈಜ್ಞಾನಿಕ, ಸಾಹಿ...	NULL	NULL	NULL	NULL	NULL	NULL
79478	19	ನಿಬಂಧನೆ ೨೮.	NULL	ನಿಬಂಧನೆ ೨೮.	NULL	NULL	NULL	NULL	NULL	NULL
79479	19	ಈ ಪ್ರಕಟನೆಯಲ್ಲಿ ಸೂಚಿತವಾದ ಹಕ್ಕುಗಳನ್ನು ಸ್ವಾತಂ...	NULL	ಈ ಪ್ರಕಟನೆಯಲ್ಲಿ ಸೂಚಿತವಾದ ಹಕ್ಕುಗಳನ್ನು ಸ್ವಾತಂ...	NULL	NULL	NULL	NULL	NULL	NULL
79480	19	ನಿಬಂಧನೆ ೨೯.	NULL	ನಿಬಂಧನೆ ೨೯.	NULL	NULL	NULL	NULL	NULL	NULL
79481	19	ಕೇವಲ ಯಾವ ಕರ್ತವ್ಯಗಳಿಂದ ತನ್ನ ವ್ಯಕ್ತಿತ್ವದ ನಿರಾ...	NULL	ಕೇವಲ ಯಾವ ಕರ್ತವ್ಯಗಳಿಂದ ತನ್ನ ವ್ಯಕ್ತಿತ್ವದ ನಿರಾ...	NULL	NULL	NULL	NULL	NULL	NULL
79482	19	ಪ್ರಜಾಪ್ರಭುತ್ವ ಸಮಾಜದಲ್ಲಿ ಕೇವಲ ಇತರರ ಹಕ್ಕು ಸ್ವಾ...	NULL	ಪ್ರಜಾಪ್ರಭುತ್ವ ಸಮಾಜದಲ್ಲಿ ಕೇವಲ ಇತರರ ಹಕ್ಕು ಸ್ವಾ...	NULL	NULL	NULL	NULL	NULL	NULL
79483	19	ಸಂಯುಕ್ತ ರಾಷ್ಟ್ರಗಳ ಉದ್ದೇಶಗಳಿಗೂ ತತ್ವಗಳಿಗೂ ವಿ...	NULL	ಸಂಯುಕ್ತ ರಾಷ್ಟ್ರಗಳ ಉದ್ದೇಶಗಳಿಗೂ ತತ್ವಗಳಿಗೂ ವಿ...	NULL	NULL	NULL	NULL	NULL	NULL
79484	19	ನಿಬಂಧನೆ ೩೦.	NULL	ನಿಬಂಧನೆ ೩೦.	NULL	NULL	NULL	NULL	NULL	NULL
79485	19	ಯಾವ ಪ್ರಾಂತಕ್ಕಾಗಿಯೋ ಸಂಘಕ್ಕಾಗಿಯೋ ವ್ಯಕ್ತಿಗಾ...	NULL	ಯಾವ ಪ್ರಾಂತಕ್ಕಾಗಿಯೋ ಸಂಘಕ್ಕಾಗಿಯೋ ವ್ಯಕ್ತಿಗಾ...	NULL	NULL	NULL	NULL	NULL	NULL
79486	20	<line_1> Dankiya kunawunaya barjijarranth !	<line_1>	Dankiya kunawunaya barjijarranth !	‘ This child has…	this-CMP ‹c…	ṭan+ki-a kuna+kun…	ṭankia kunaunaja…	this-μloc-t ‹child…	This is exam…
79487	20	<line_2> Ngada mungurru , makuntha yala…	<line_2>	Ngada mungurru , makuntha yalawujarrant…	‘ I know that th…	1sg know w…	ŋaṭ-ta muŋuru-a m…	NULL	1sg-T know-T…	This is exam…
79488	20	<line_3> Dathina kunawuna jungarr .	<line_3>	Dathina kunawuna jungarr .	‘ That child is b…	that ‹child› big	ṭaṭina kuna+kuna-…	NULL	that.T ‹child_N…	This is exam…
79489	20	<line_4> Darrathiwu ngakulda wuranku diy…	<line_4>	Darrathiwu ngakulda wuranku diyaju .	‘ We’ll eat the f…	hot-FUT 1-2…	ṭaraṭi+kuu-ø ŋa-ku…	NULL	hot-μPROP-T 1…	This is exam…
79490	20	<line_5> Ngada kurrinngarrba wuranngarr…	<line_5>	Ngada kurrinngarrba wuranngarrb , ngumb…	‘ Having seen t…	1sg ‹see-‹A…	ŋaṭ-ta kuri-c-n-ŋar…	NULL	1sg-T ‹see-J-‹…	This is exam…
79491	21	40001001 Nēs ge ǂkhanis Jesub Xristu…	40001001	Nēs ge ǂkhanis Jesub Xristub !haos disa ,…	NULL	NULL	NULL	NULL	NULL	NULL
79492	21	40001002 Abrahammi ge lsaka ge hō ,…	40001002	Abrahammi ge lsaka ge hō , tsî lsaki ge Ja…	NULL	NULL	NULL	NULL	NULL	NULL
79493	21	40001003 Tsî Judab ge Peresi tsî Sera…	40001003	Tsî Judab ge Peresi tsî Serab tsîkha Tamar…	NULL	NULL	NULL	NULL	NULL	NULL
79494	21	40001004 tsî Arammi ge Aminadaba ge…	40001004	tsî Arammi ge Aminadaba ge hō , tsî Amina…	NULL	NULL	NULL	NULL	NULL	NULL

3.3. Data Availability

The input corpora and the source code for processing them is available in the TeDDi sample GitHub repository.[7] Currently, the SQLite version of the database is 2.6 GB in size. Therefore we also provide more lightweight versions of the database tables as CSV files and as a serialized R data object ($\sim$ 700MB) for users who prefer to interact with the parsed text corpora with programs such as R (R Core Team, 2021), a free statistical programming language and software environment popular among data scientists. The data formats are available online.[8]

Lastly, we export the data into CLDF. CLDF is built on the W3C's Model for Tabular Data and Metadata on the Web[9] and the Metadata Vocabulary for Tabular Data.[10] The CLDF model is ideally suited for sharing Unicode-compliant CSV text files that are made ontologically aware via a strict linguistics ontology encoded in JSON-DL. Some advantages of this infrastructure include: useful delineation of data and tools, standardized tabular data on the web, pipeline style data transformation procedures, and standards for describing analysis workflows, e.g., the Common Workflow Language.[11] The TeDDi sample in CLDF is made available through a separate Github repository.[12]

3.4. Data Summary Statistics

The current version of the TeDDi sample contains more than 20K texts from 89 different languages, stemming from 58 language families (according to WALS), and written and encoded in 16 different scripts. Table 2 (above) gives some summary statistics split by genre. Figure 7 visualizes the currently represented languages on a world map. Figure 8 gives the coverage (in percent of 192 WALS features) for the languages for which text material is currently available in the TeDDi sample.

3.5. Copyright

We publish the overall collection of texts and database infrastructure under a CC BY-NC-SA 4.0 license.[13] Note that in some cases, the original texts have more (or less) restrictive licenses. The particular copyright is given in the metadata header for each text and it should be adhered to in further use cases. Also, we are aware that our collection includes texts of minority languages, which, in some cases, might be considered legacy materials with unclear copyright conditions. Therefore, we follow the so-called "takedown principle", i.e., we can remove such material if contacted by people aggrieved by it (Seyfeddinipur et al., 2019, p. 554).

4. Added Value and Use Cases

The main purpose of the TeDDi sample is to approximate the diversity of languages across the world by means of text samples. The availability of this dataset means that we can extract linguistic features from text directly and automatically, and compare languages on these grounds. In this sense, we aim to complement the existing knowledge about the structure of languages, which mostly consists of high-level feature-value pairs stored in linguistic databases.

A downside of the current dataset is that – in many cases – it does not provide rich linguistic annotations. However, as an upside, it provides rich metadata for each text, and it samples from diverse languages, genres, and scripts. This lends itself, for instance, to quantitative linguistics analyses of laws of language (Piantadosi et al., 2011; Bentz and Ferrer-i-Cancho, 2016; Levshina and Moran, 2021). In this case, the relevant features can be extracted without further linguistic annotation. These include, but are not limited to, information-theoretic measures such as n-gram entropy, or various indicators of morphological complexity such as the mean word length or subword recur-

[7] https://github.com/MorphDiv/TeDDi_sample

[8] https://drive.switch.ch/index.php/s/MJv7xFkzqlzFn0y

[9] https://www.w3.org/TR/tabular-data-model/#standard-file-metadata

[10] https://www.w3.org/TR/tabular-metadata/

[11] https://www.commonwl.org

[12] https://github.com/cldf-datasets/TeDDi_sample

[13] https://creativecommons.org/licenses/by-nc-sa/4.0/

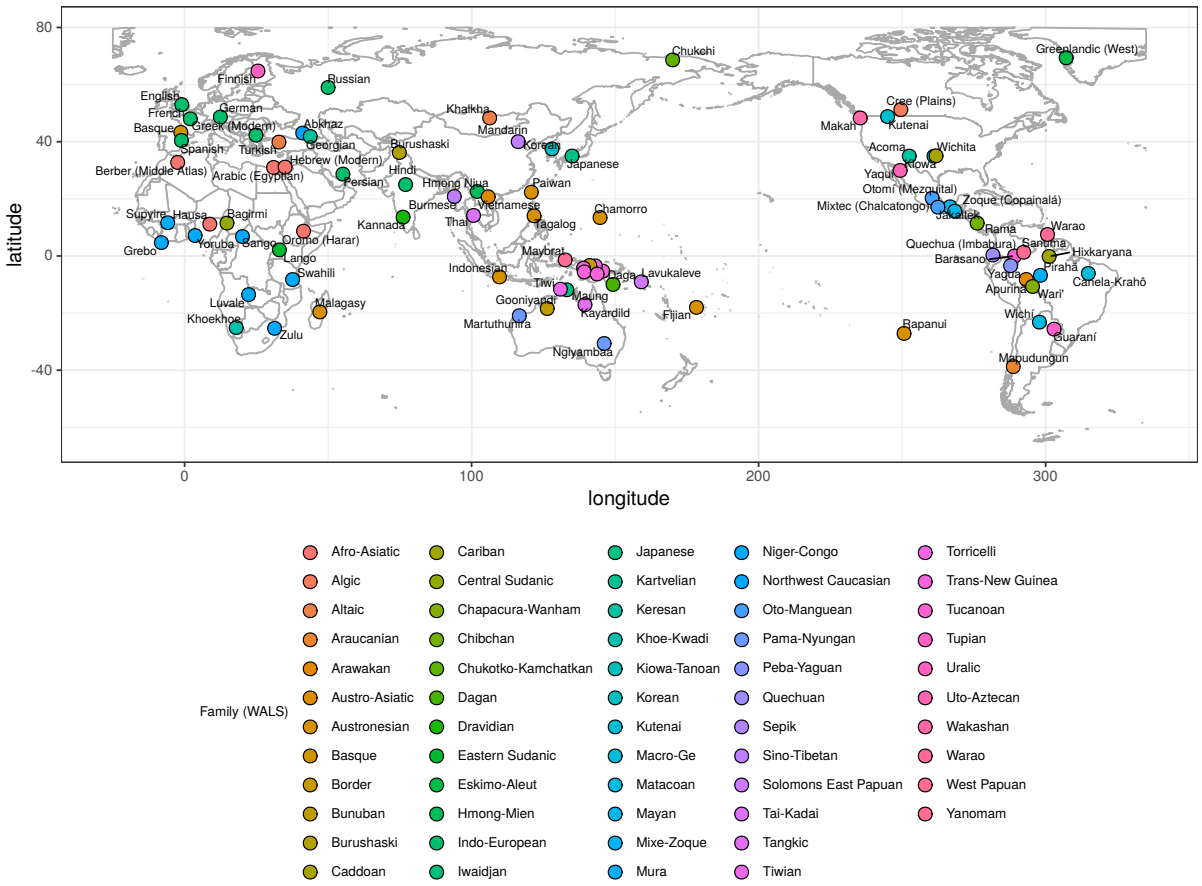

Figure 7: Languages included in the current version of the corpus. These are 89 languages according to ISO 639-3 codes from 58 language families according to WALS (colors).

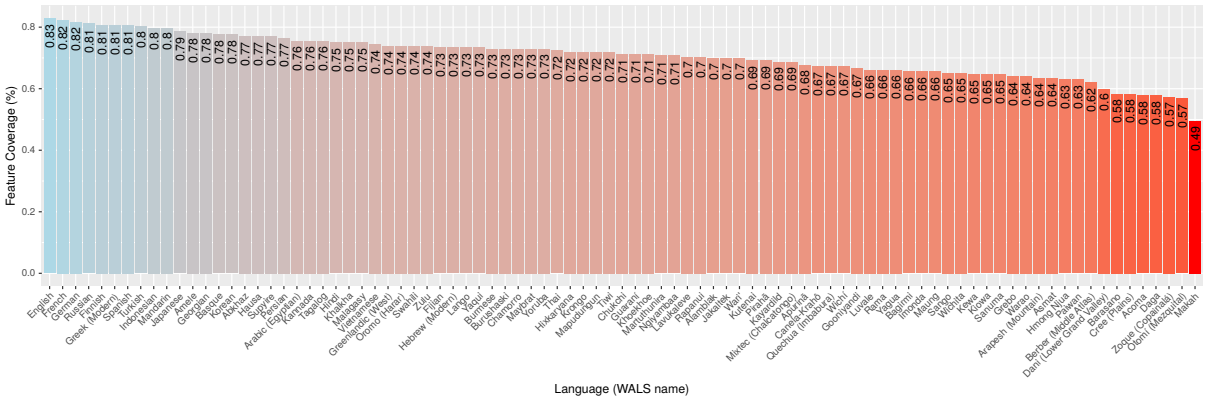

Figure 8: Coverage in percent of overall 192 WALS features for the 89 languages currently included in the TeDDi sample.

rence patterns (Gutierrez-Vasques and Mijangos, 2019; Gutierrez-Vasques et al., 2021). Such features can then be directly compared to higher-level typological features derived from chapters of the WALS (Bentz et al., 2016). This is here facilitated by the relatively high coverage of WALS features (49% to 83%) for the languages of the TeDDi sample (see Figure 8).

With languages properly sampled and described in terms of objectively comparable features, we can address various questions regarding the distribution of linguistic types. The match between the languages in the TeDDi sample and the WALS 100 language sample also enables token-based typological analyses (Levshina, 2019). We can quantify and visualise linguis-

tic diversity (e.g., what language types are most common?) or model potential relationships between linguistic types and various conditions of language use (e.g., what kind of languages are spoken where?). In addition to the known text features, text samples can be used by researchers to come up with novel linguistic features and measures that can improve our knowledge of linguistic diversity.

We underline that our collection of text samples in diverse languages enables a better cooperation between linguistics and NLP. Although most of the texts in our sample are rather short, they can still serve as test cases for assessing cross-linguistic generalization of multilingual models. This is especially true for the tasks that require only raw text (text segmentation, language modelling, machine translation, automatic language identification). For other tasks, additional annotation would be required, but this should be facilitated by the fact that the texts are easily accessible and ready to be put through NLP pipelines or imported into specialised annotation software.

5. Acknowledgements

SM, CB, XGV, OS, TS were funded by the Swiss National Science Foundation (SNSF; grant number 176305). SM was funded by the SNSF (Grant No. PCEFP1_186841). We thank Zifan Jiang for the CLDF conversion and Robert Forkel and Sebastian Bank for helpful feedback.

6. Bibliographical References

Bakker, D. (2011). Language Sampling. In J. J. Song, editor, *Handbook of Linguistic Typology*. Oxford University Press, Oxford, UK.

Bayer, M. (2012). Sqlalchemy. In Amy Brown et al., editors, *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. aosabook.org.

Bentz, C. and Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen.

Bentz, C., Soldatova, T., Koplenig, A., and Samardžić, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora.

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i-Cancho, R. (2017a). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6).

Bentz, C., Alikaniotis, D., Samardžić, T., and Buttery, P. (2017b). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, 24(2-3):128–162.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M. (2013). Introduction. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Gutierrez-Vasques, X. and Mijangos, V. (2019). Productivity and predictability for measuring morphological complexity. *Entropy*, 22(1):48.

Gutierrez-Vasques, X., Bentz, C., Sozinova, O., and Samardzic, T. (2021). From characters to words: the turning point of bpe merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). *Glottolog 4.4*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–135.

Hipp, R. D. (2020). *SQLite*. https://www.sqlite.org.

Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the google ngram data sets—reconstructing the composition of the german corpus in times of wwii. *Digital Scholarship in the Humanities*, 32(1):169–188.

Levshina, N. and Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, 7(s3).

Levshina, N. (2019). Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings from LREC 2016*, pages 923–929. European Language Resources Association.

Majid, A. and Levinson, S. C. (2010). Weird languages have misled us, too. *Behavioral and Brain Sciences*, 33(2-3):103.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3158–3163.

Moran, S. (2012). *Phonetics Information Base and Lexicon*. Ph.D. thesis, University of Washington.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communi-

cation. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

R Core Team, (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Round, E. (2012). *Kayardild morphology and syntax*. Oxford University Press.

Seyfeddinipur, M., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B., Cruz, H., Drude, S., Epps, P., Ferreira, V., Galucio, A., et al. (2019). Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation*, 13:545–563.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.