

# Multilingual Corpus with Lexical, Syntactic, and Semantic Features

Shunxin Yao  
Queensland University of Technology, Brisbane, Australia  
n12012416@qut.edu.au

2026-10-03    Version 1.0

## 1 Description

This dataset contains 1,860 Wikipedia articles across five languages:

- English (391)
- Chinese (329)
- Spanish (384)
- French (384)
- German (372)

Each document is accompanied by:

- Lexical features (TTR, MTLD, entropy, hapax ratio)
- Syntactic features (mean dependency distance, tree height, POS and dependency distributions)
- Semantic features (LaBSE document embeddings, 768 dimensions)

The dataset is designed for reproducible cross-linguistic comparison and can be reused for typological, computational, and quantitative linguistic research.

## 2 File Structure

```
Dataset_Upload/  
|-- data/  
|   |-- corpus_meta_v6.csv  
|   |-- corpus_text_v6.csv  
|   |-- features_vocab_clean.parquet  
|   |-- features_syntax_clean.parquet  
|   |-- features_embed.parquet  
|   |-- analysis_df.parquet  
|   |-- embed_matrix.rds  
|   +-- mixed_model_results.csv  
|-- code/  
|   |-- extract_features.py
```

```

|   |-- 01_prepare.R
|   |-- 02b_permanova_terms.R
|   |-- 03_cluster_mds.R
|   |-- 04_mantel.R
|   |-- 05_mixed_models.R
|   +-- 06_all_figures.R
|-- figures/
|   |-- fig_length_boxplot.png
|   |-- fig_distance_heatmap.png
|   |-- fig_pca.png
|   |-- fig_mds_language.png
|   |-- fig_dendrogram.png
|   |-- fig_silhouette.png
|   |-- fig_mantel_scatter.png
|   +-- fig_mixed_forest.png
|-- README.md
+-- data_dictionary.md

```

### 3 Data Sources

- **Wikipedia:** Article texts were retrieved via the MediaWiki API.
- **Wikidata:** 15 concepts were aligned across languages using Wikidata QIDs.
- **Keyword sampling:** 20 additional keywords per language were used to retrieve comparable articles.

All articles were truncated to the first 2,000 tokens to control for text length.

### 4 Methods

#### 4.1 Feature Extraction (Python)

- **Tokenization, POS tagging, dependency parsing:** Stanza
- **Semantic embeddings:** LaBSE (sentence-transformers/LaBSE)
- **Feature matrix:** exported as Parquet files

#### 4.2 Statistical Modeling (R)

- **PERMANOVA:** `vegan::adonis2`
- **PCA:** `FactoMineR::PCA`
- **Clustering:** `cluster::hclust`
- **Mantel test:** `vegan::mantel`
- **Mixed-effects models:** `lme4::lmer`

All analyses controlled for text length and sampling stage.

## 5 Usage

### 5.1 Load data in Python

```
import pandas as pd
meta = pd.read_csv("data/corpus_meta_v6.csv")
vocab = pd.read_parquet("data/features_vocab_clean.parquet")
syntax = pd.read_parquet("data/features_syntax_clean.parquet")
embed = pd.read_parquet("data/features_embed.parquet")
```

### 5.2 Load data in R

```
library(arrow)
df <- read_parquet("data/analysis_df.parquet")
E <- readRDS("data/embed_matrix.rds")
```

### 5.3 Reproduce the full pipeline

1. Run `code/extract_features.py` in Python.
2. Run R scripts in order: `01_prepare.R` to `06_all_figures.R`.
3. See `data_dictionary.md` for variable descriptions.

## 6 Data Dictionary

See `data_dictionary.md` for a full description of all variables.

## 7 License

- **Wikipedia article texts** (`corpus_text_v6.csv`): CC BY-SA 4.0
- **Derived features** (Parquet files, RDS, CSV): CC BY 4.0
- **Code** (Python and R scripts): MIT License

Please cite the dataset if you use it.

## 8 Citation

If you use this dataset, please cite:

Yao, S. (2026). Multilingual Corpus with Lexical, Syntactic, and Semantic Features (Version 1.0) [Data set].

## 9 Acknowledgments

The author thanks the Queensland University of Technology for computational resources.

## 10 Contact

For questions, please contact: `n12012416@qut.edu.au`