

Data Dictionary

Shunxin Yao

2026-10-03

This document describes all variables in the dataset.

1 corpus_meta_v6.csv

Metadata for each Wikipedia article.

Variable	Type	Description
doc_id	string	Unique document identifier
language	string	Language name
lang_code	string	ISO 639-1 code
stage	string	Sampling stage (wikidata or keyword)
concept	string	Concept label for Wikidata-aligned articles
query	string	Search query used to retrieve the article
title	string	Wikipedia article title
wikidata_qid	string	Wikidata QID for aligned concepts
pageid	integer	Wikipedia page ID
revid	integer	Wikipedia revision ID
n_chars	integer	Number of characters
n_sections	integer	Number of sections

2 corpus_text_v6.csv

Full article text.

Variable	Type	Description
doc_id	string	Unique document identifier
text	string	Full article text

3 features_vocab_clean.parquet

Lexical features per document.

Variable	Type	Description
doc_id	string	Unique document identifier
n_tokens	integer	Number of tokens after truncation to 2000
n_types	integer	Number of unique tokens
ttr	float	Type-Token Ratio (n_types / n_tokens)
mtld	float	Measure of Textual Lexical Diversity

Variable	Type	Description
entropy	float	Shannon entropy of token distribution
hapax_ratio	float	Proportion of tokens occurring once

4 features_syntax_clean.parquet

Syntactic features per document.

Variable	Type	Description
doc_id	string	Unique document identifier
mean_dep_distance	float	Mean dependency distance
mean_tree_height	float	Mean dependency tree height
n_sentences	integer	Number of sentences
mean_sent_len	float	Mean sentence length in tokens
pos_*	float	Proportion of each Universal POS tag
dep_*	float	Proportion of each Universal dependency relation

5 features_embed.parquet

LaBSE document embeddings (768 dimensions).

Variable	Type	Description
doc_id	string	Unique document identifier
e0--e767	float	LaBSE embedding dimensions

6 analysis_df.parquet

Merged analysis table (metadata + lexical + syntactic features).

Variable	Type	Description
doc_id	string	Unique document identifier
language	string	Language name
stage	string	Sampling stage
n_chars	integer	Number of characters
ttr	float	Type-Token Ratio
mtld	float	MTLD
entropy	float	Shannon entropy
hapax_ratio	float	Hapax ratio
mean_dep_distance	float	Mean dependency distance
mean_tree_height	float	Mean tree height
log_chars	float	Natural log of n_chars

7 `embed_matrix.rds`

Document-level LaBSE embedding matrix (RDS format). Rows = documents, columns = 768 embedding dimensions.

8 `mixed_model_results.csv`

Fixed effects from mixed-effects models.

Variable	Type	Description
<code>feature</code>	string	Dependent variable name
<code>effect</code>	string	Effect type (fixed)
<code>term</code>	string	Predictor term
<code>estimate</code>	float	Model estimate
<code>std.error</code>	float	Standard error
<code>statistic</code>	float	t-statistic
<code>df</code>	float	Degrees of freedom
<code>p.value</code>	float	p-value
<code>conf.low</code>	float	Lower 95% confidence interval
<code>conf.high</code>	float	Upper 95% confidence interval

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