HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-mri_Latn	9/18/2025	Māori (mi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
203,007	4,122,704	3,172,670 (76.96 %)	160M	681,466,248	659.62 MB

Top 10 domains

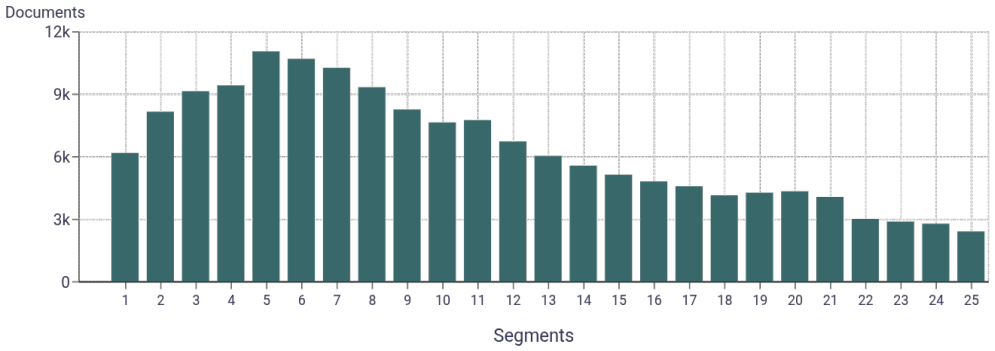
Domain	Docs	% of total
maoritelevision...	8.7K	4.27%
jw.org	4.2K	2.07%
martech.zone	3.8K	1.86%
victoria.ac.nz	3.6K	1.78%
eturbonews.com	3.5K	1.73%
teara.govt.nz	3K	1.47%
pricenedorogo.news	2.8K	1.37%
wondershare.com	2.5K	1.22%
teaomaori.news	2.2K	1.07%
maoridictionary...	2K	0.98%

Top 10 TLDs

Domain	Docs	% of total
com	136K	67.12%
org	14K	6.77%
govt.nz	6.4K	3.13%
news	6.1K	2.99%
co.nz	5K	2.45%
ac.nz	4.4K	2.16%
zone	3.8K	1.86%
net	2.9K	1.41%
org.nz	2.6K	1.30%
com.br	1.5K	0.74%

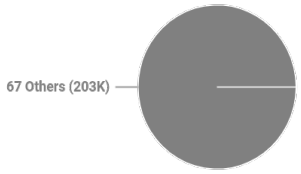
Documents size (in segments)

≤ 25 segments **78.26%** (159K documents)
> 25 segments **21.74%** (44K documents)



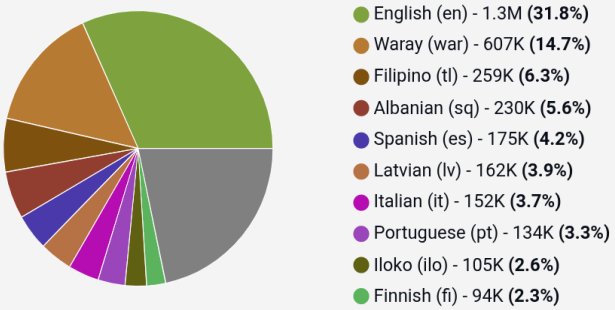
Document collections

CC = 93.73%
IA = 6.27%



Language Distribution

Number of segments in the Māori (mi) corpus

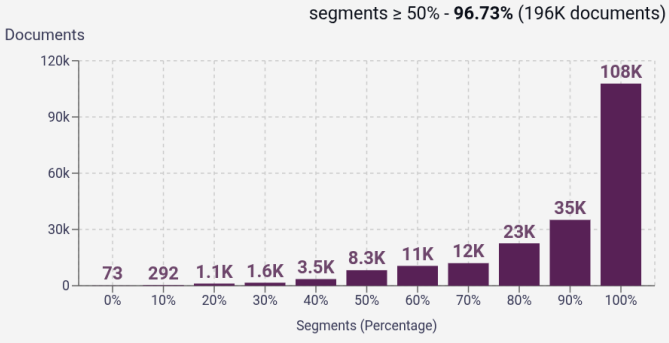


- English (en) - 1.3M **(31.8%)**
- Waray (war) - 607K **(14.7%)**
- Filipino (tl) - 259K **(6.3%)**
- Albanian (sq) - 230K **(5.6%)**
- Spanish (es) - 175K **(4.2%)**
- Latvian (lv) - 162K **(3.9%)**
- Italian (it) - 152K **(3.7%)**
- Portuguese (pt) - 134K **(3.3%)**
- Iloko (ilo) - 105K **(2.6%)**
- Finnish (fi) - 94K **(2.3%)**
- 163 Others - 895K **(21.7%)**

*Māori (mi) identification might be inaccurate because it is not supported by FastSpell

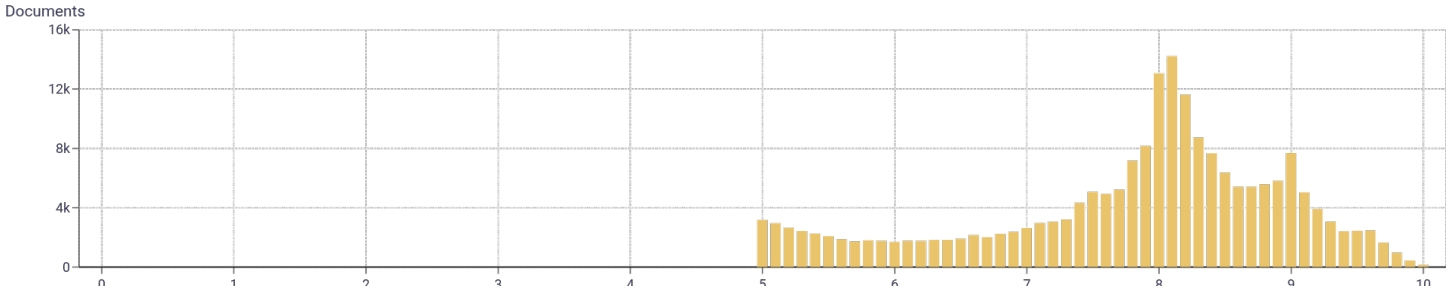
Percentage of segments in Māori (mi) inside documents

segments < 50% - **3.27%** (6.6K documents)
segments ≥ 50% - **96.73%** (196K documents)



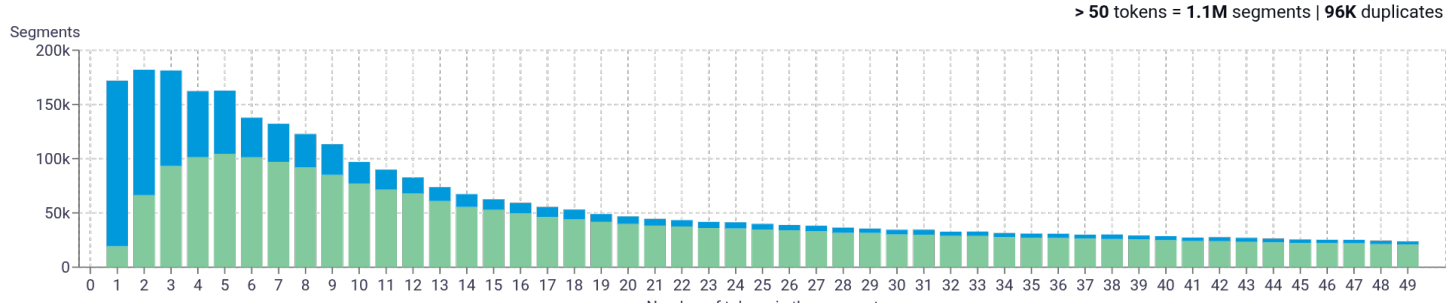
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (203K documents)

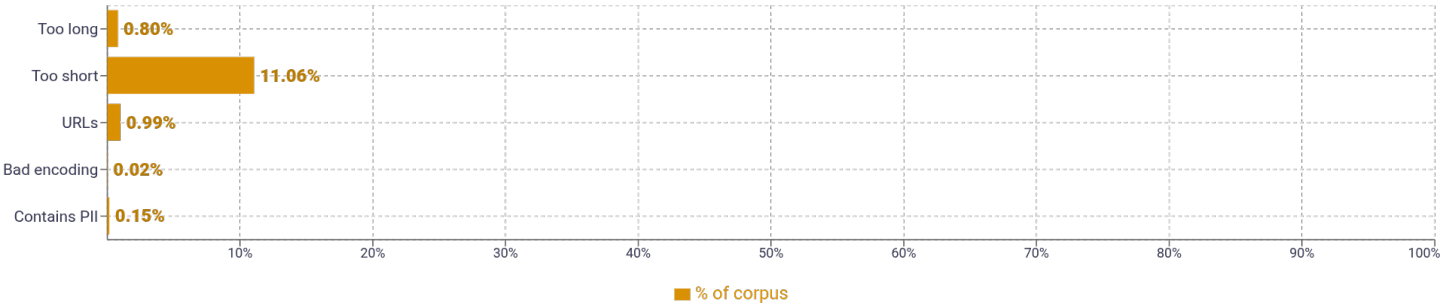


Segment length distribution by token

≤ 49 tokens = **3.1M** segments | **856K** duplicates
> 50 tokens = **1.1M** segments | **96K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	he | 2,001,428 · mo | 931,114 · mea | 888,274 · roto | 809,698 · ngā | 643,878
2	he mea | 137,813 · he aha | 100,921 · he pai | 93,518 · he maha | 76,018 · he tino | 60,467
3	he mea nui | 30,046 · kaore e taea | 21,699 · roto i to | 20,667 · roto i tenei | 19,511 · mo te wa | 18,680
4	noa i te ao | 9,891 · mo te wa roa | 9,878 · roto i te ao | 8,356 · tuatahi ki te korero | 7,627 · roto i te whare | 7,247
5	hei tuatahi ki te korero | 7,599 · puta noa i te ao | 6,857 · taea e koe te whakamahi | 6,163 · whakapiri mai ki a maatau | 4,535 · huri noa i te ao | 2,954

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Volumes - Tokens

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Type-Token Ratio

Lexical variety computed as *number or types (uniques)/number of tokens*, after removing punctuation (https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (https://github.com/mbanon/fastspell).

Distribution of segments by fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by average fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

Segment length distribution by token

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Segment noise distribution

Obtained with Bicleaner Hardrules (https://github.com/bitextor/bicleaner-hardrules/).

Frequent n-grams

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md, after removing n-grams starting or ending in a stopword. Stopwords from https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				