

General overview

Corpus	Date	Language
hplt-v3-mri_Latn	9/18/2025	Maori

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

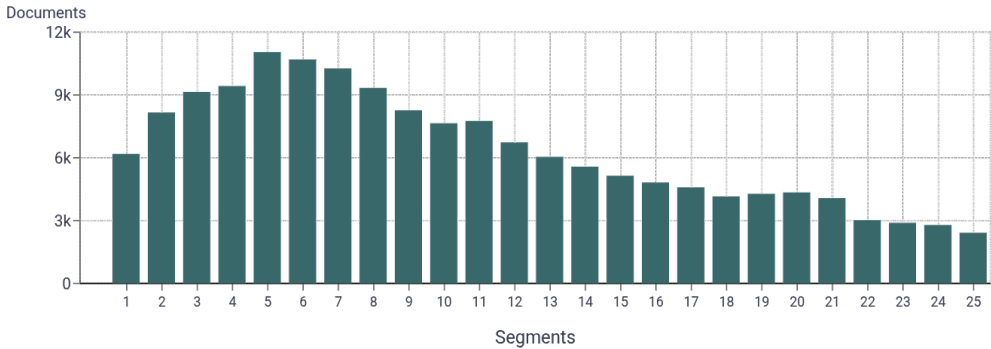
Top 10 domains

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%
pricedorogo.news	2.8K	1.37%
wondershare.com	2.5K	1.22%
teaomaori.news	2.2K	1.07%
maoridictionary...	2K	0.98%

Top 10 TLDs

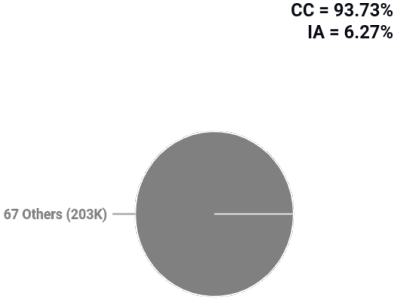
Domain	Docs	% of total
com	136K	67.12%
org	14K	6.77%
govt.nz	6.4K	3.13%
news	6.1K	2.99%
co.nz	5K	2.45%
ac.nz	4.4K	2.16%
zone	3.8K	1.86%
net	2.9K	1.41%
org.nz	2.6K	1.30%
com.br	1.5K	0.74%

Documents size (in segments) ⓘ



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

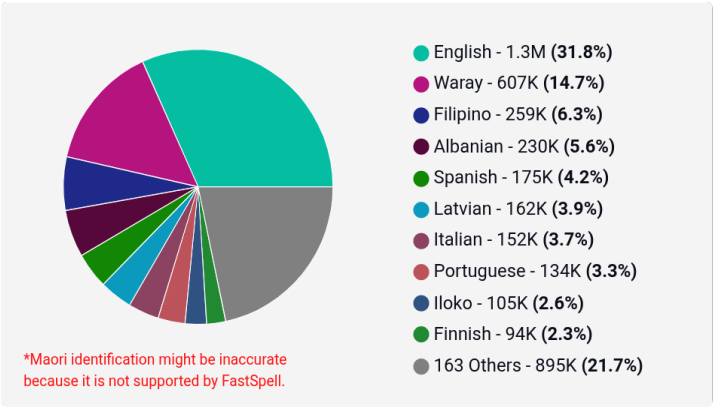
Document collections



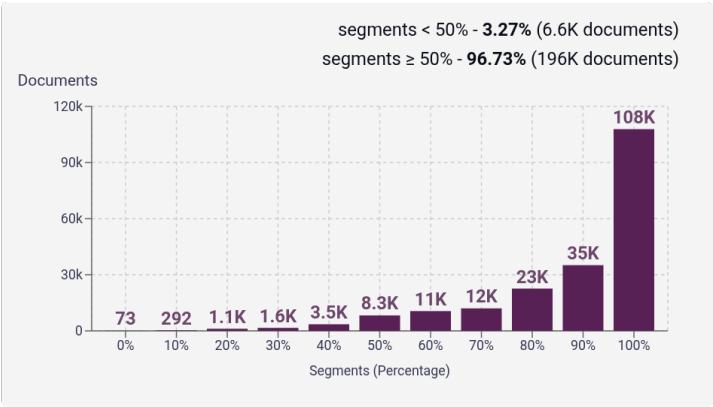
CC = 93.73%
IA = 6.27%

Language Distribution

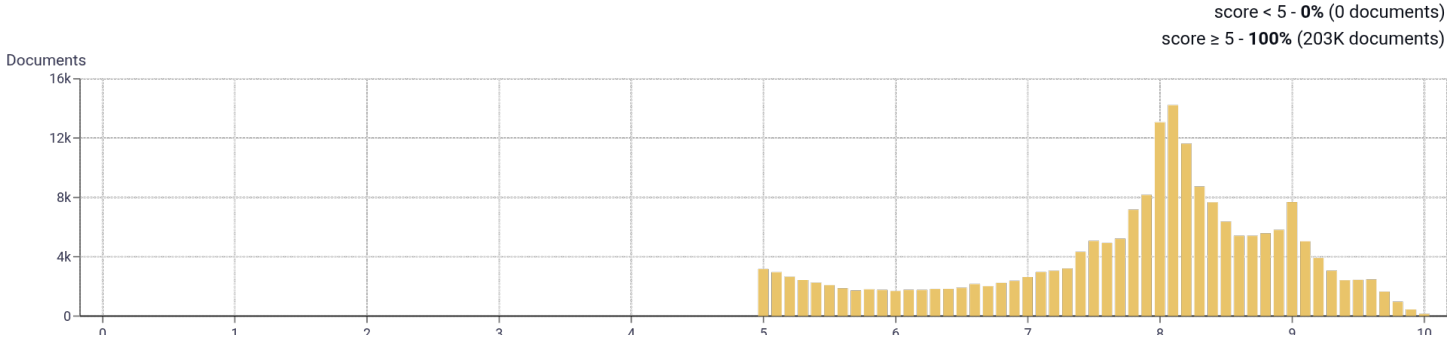
Number of segments in the Maori corpus



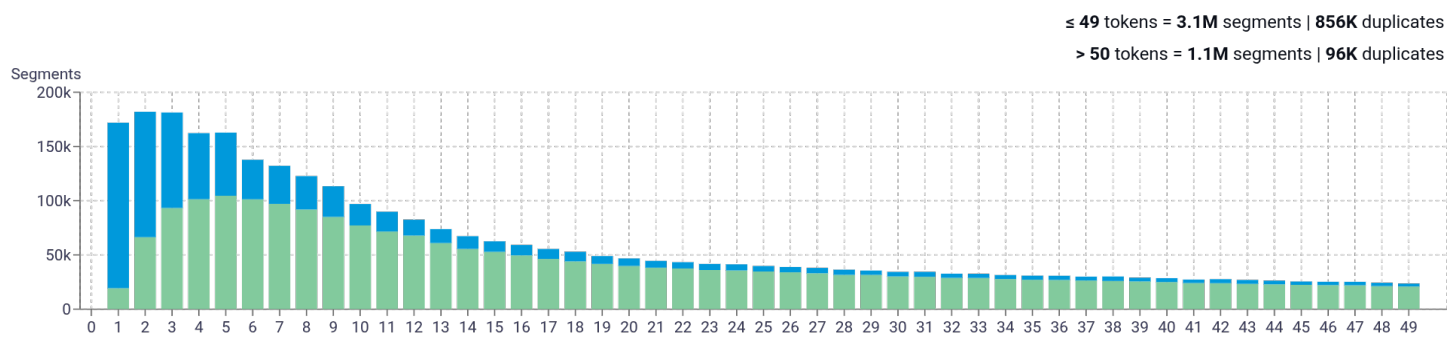
Percentage of segments in Maori inside documents



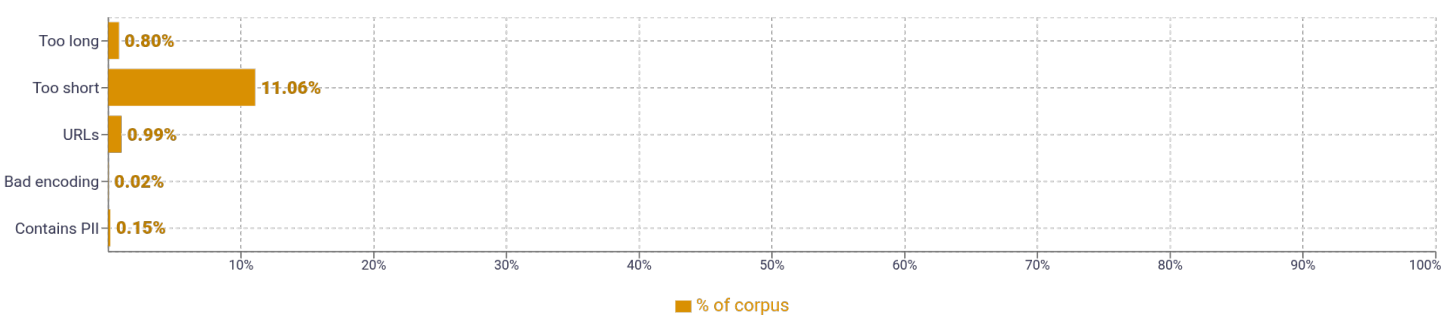
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				