HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-smo_Latn	9/18/2025	Samoan (sm)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
161,099	3,297,842	2,673,531 (81.07 %)	137M	580,542,204	564.64 MB

Top 10 domains

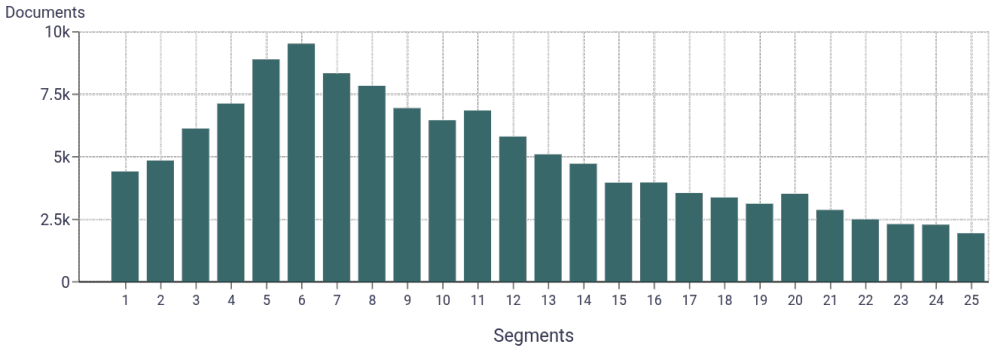
Domain	Docs	% of total
samoatimes.co.nz	9.1K	5.67%
radiosamoa.co.nz	6.8K	4.21%
samoanews.com	6.3K	3.92%
martech.zone	5K	3.09%
eturbonews.com	4.9K	3.03%
jw.org	3.9K	2.43%
actualidadiphon...	3K	1.88%
radiopolynesias...	1.9K	1.20%
actualidadgadge...	1.9K	1.17%
samoaobserver.ws	1.8K	1.13%

Top 10 TLDs

Domain	Docs	% of total
com	115K	71.09%
co.nz	16K	9.94%
org	9.3K	5.78%
zone	5K	3.09%
net	2.6K	1.60%
ws	2.2K	1.34%
ru	888	0.55%
pt	812	0.50%
es	681	0.42%
fr	640	0.40%

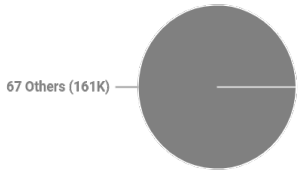
Documents size (in segments)

≤ 25 segments **78.5%** (126K documents)
> 25 segments **21.5%** (35K documents)



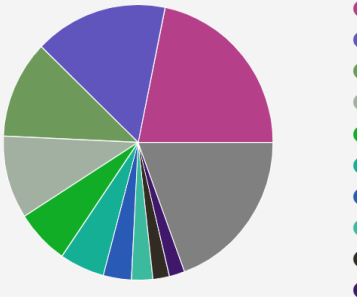
Document collections

CC = 97.56%
IA = 2.44%



Language Distribution

Number of segments in the Samoan (sm) corpus

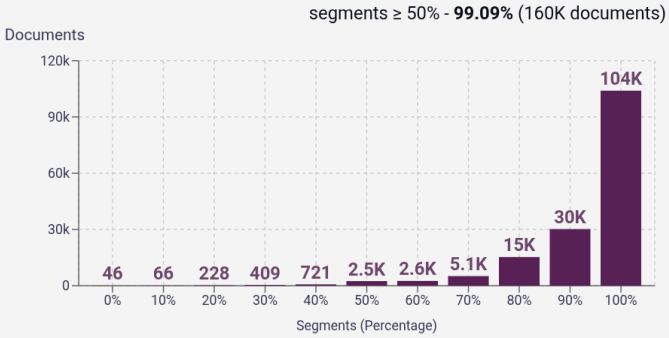


- French (fr) - 719K **(21.8%)**
- English (en) - 527K **(16.0%)**
- Portuguese (pt) - 380K **(11.5%)**
- Italian (it) - 322K **(9.8%)**
- Spanish (es) - 212K **(6.4%)**
- Romanian (ro) - 178K **(5.4%)**
- Breton (br) - 111K **(3.4%)**
- Ido (io) - 84K **(2.5%)**
- Finnish (fi) - 65K **(2.0%)**
- Welsh (cy) - 62K **(1.9%)**
- 160 Others - 640K **(19.4%)**

*Samoan (sm) identification might be inaccurate because it is not supported by FastSpell

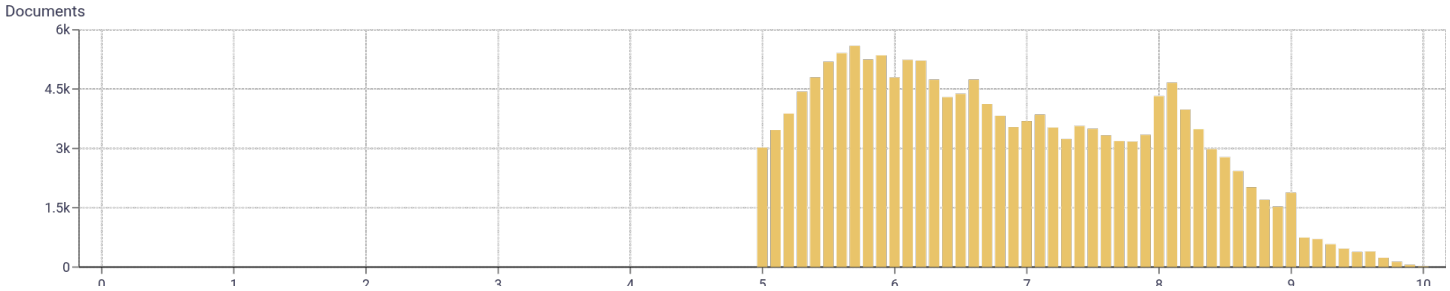
Percentage of segments in Samoan (sm) inside documents

segments < 50% - **0.91%** (1.5K documents)
segments ≥ 50% - **99.09%** (160K documents)



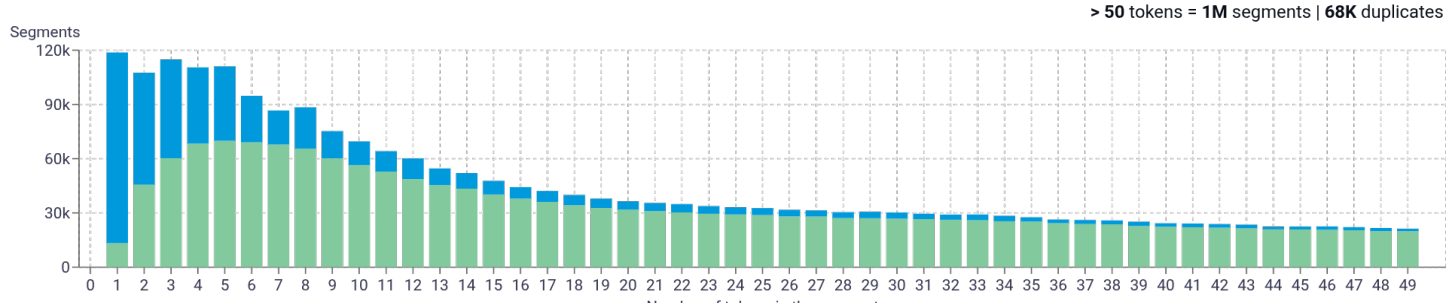
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (161K documents)

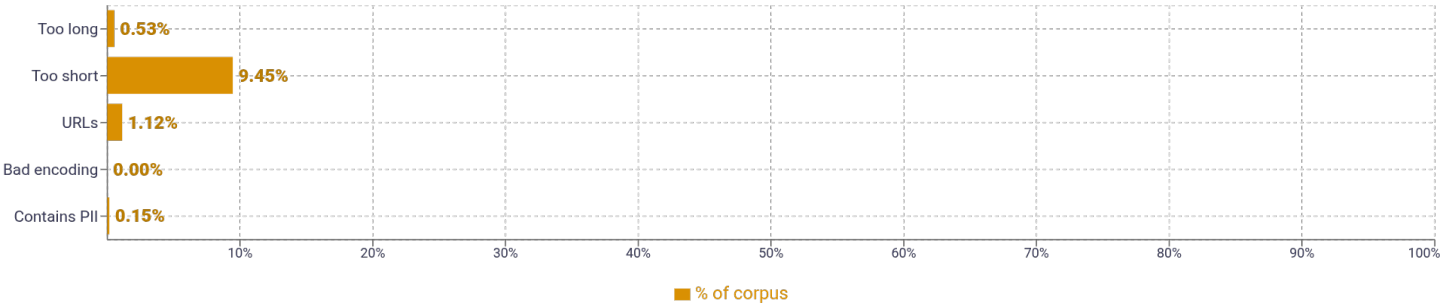


Segment length distribution by token

≤ 49 tokens = **2.3M** segments | **558K** duplicates
> 50 tokens = **1M** segments | **68K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	te | 844,413 · mafai | 683,817 · pe | 453,172 · lelei | 445,456 · matou | 444,782
2	ou te | 114,244 · matou te | 111,823 · te oe | 89,080 · nai lo | 71,845 · la matou | 52,733
3	afai e te | 45,637 · mafai ona tatou | 21,554 · pito i luga | 21,488 · luga ole laiga | 20,275 · alualu i luma | 16,490
4	fesoasoani ia te oe | 14,548 · afai e te mana'o | 8,393 · luga o le initaneti | 7,766 · luga o le upega | 6,808 · mafai ona e faia | 6,551
5	tu'uina atu ia te oe | 5,317 · ta'u atu ia te oe | 4,025 · pe a fai e te | 3,410 · luga o le upega tafa'ilagi | 3,213 · fa'aali atu ia te oe | 2,963

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				