HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-yue_Hant	9/24/2025	Cantonese (yue)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
217,261	4,618,489	3,461,753 (74.95 %)	167M	272,601,981	661.32 MB

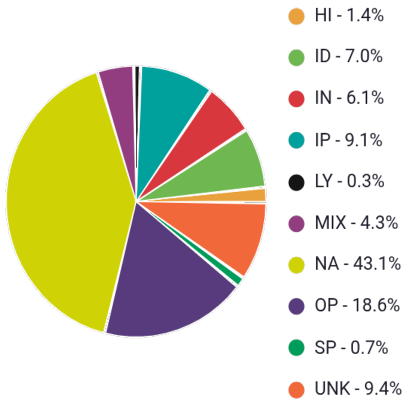
Top 10 domains

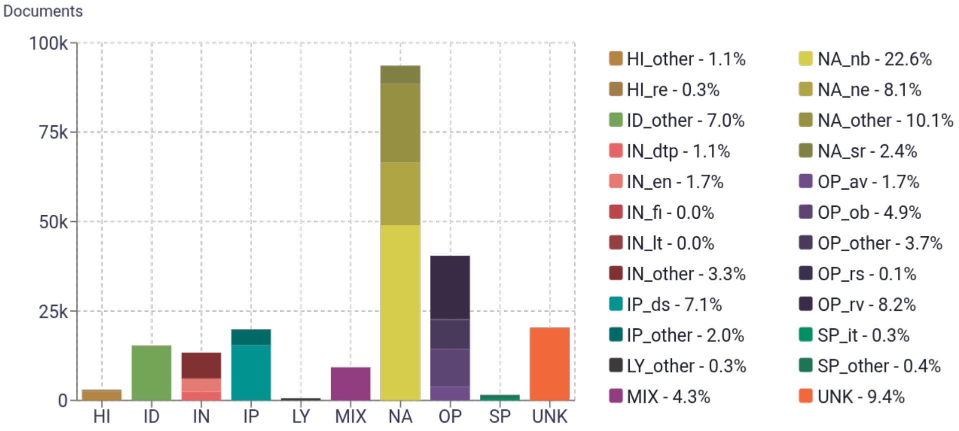
Domain	Docs	% of total
blogspot.hk	13K	6.04%
blogspot.com	11K	4.99%
openrice.com	10K	4.63%
on.cc	7.3K	3.38%
hotels.com	4.2K	1.92%
wikipedia.org	3.9K	1.78%
presslogic.com	3.7K	1.71%
fanpiece.com	3.6K	1.65%
yahoo.com	3.4K	1.57%
hkgolden.com	3.3K	1.53%

Top 10 TLDs

Domain	Docs	% of total
com	123K	56.83%
hk	32K	14.59%
com.hk	23K	10.73%
cc	7.7K	3.54%
org	5.9K	2.74%
net	5.8K	2.66%
tw	2.4K	1.09%
name	2.2K	1.02%
me	2.2K	0.99%
info	2K	0.92%

Register labels

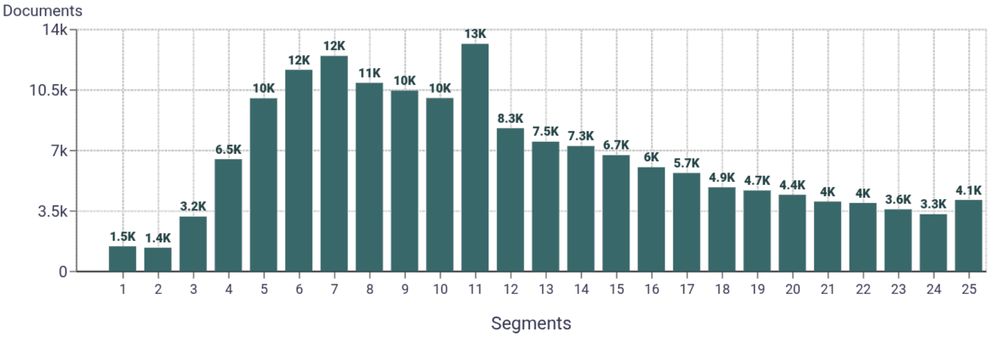




MT:0.2% | 399 Documents

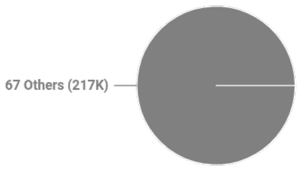
Documents size (in segments)

≤ 25 segments **76.37%** (166K documents)
> 25 segments **23.63%** (51K documents)



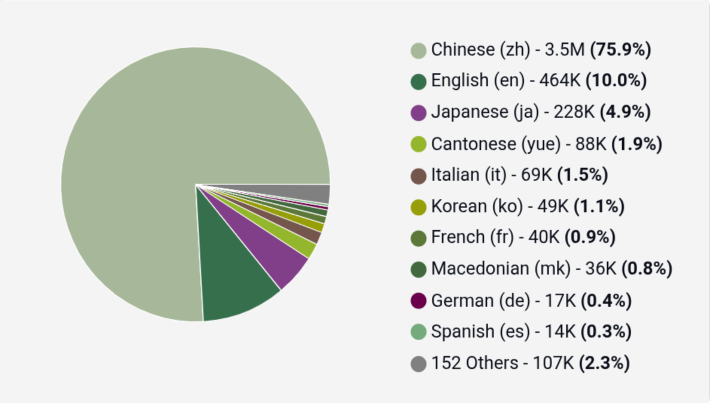
Document collections

CC = 81.66%
IA = 18.34%

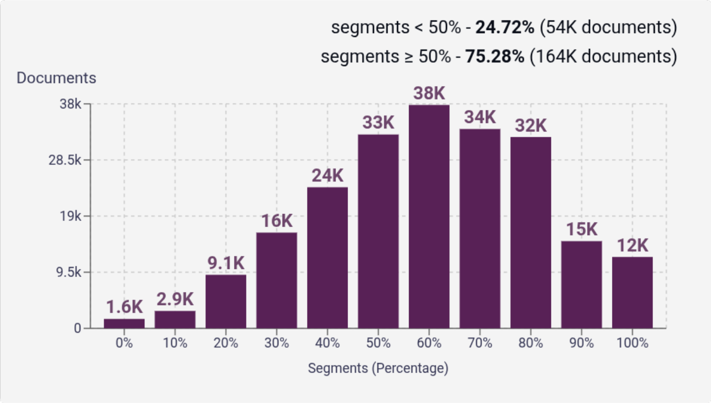


Language Distribution

Number of segments in the Cantonese (yue) corpus

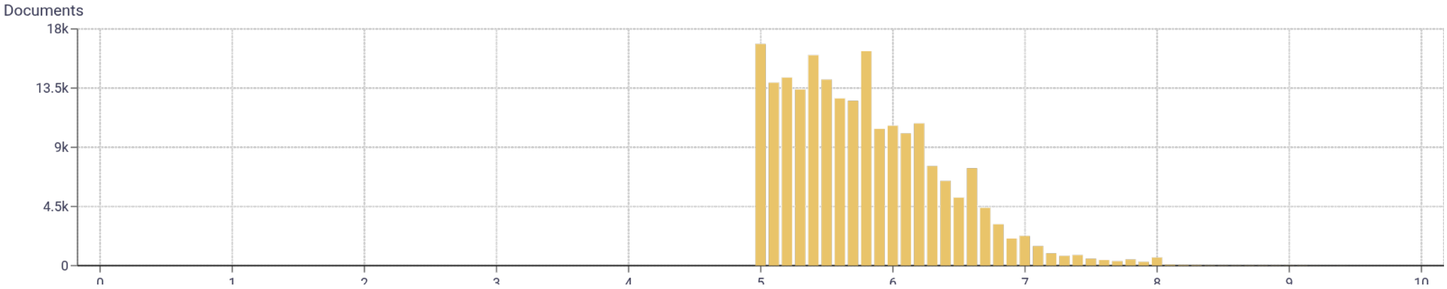


Percentage of segments in Cantonese (yue) inside documents



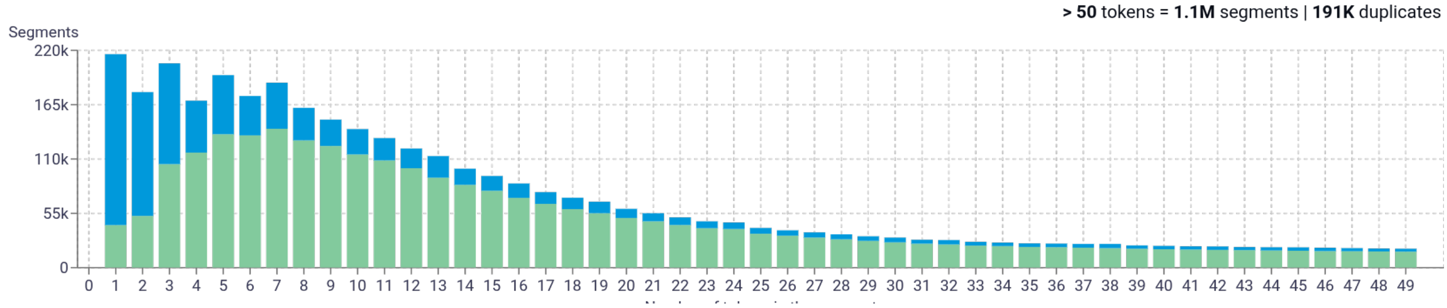
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (217K documents)

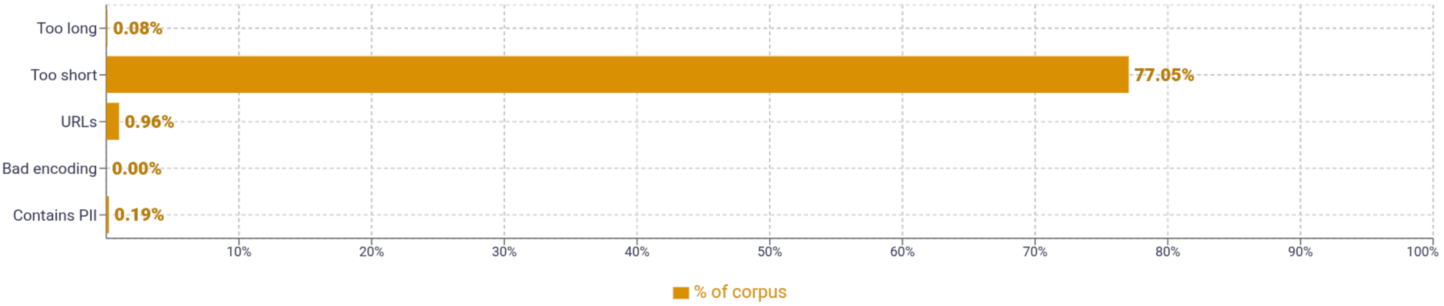


Segment length distribution by token

≤ 49 tokens = **3.6M** segments | **969K** duplicates
> 50 tokens = **1.1M** segments | **191K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	的 | 818,762 既 | 660,802 但 | 522,544 人 | 520,089 一 | 512,848
2	回覆 刪除 | 60,816 好多 人 | 44,637 認 為 | 42,171 出 嚟 | 41,312 一 間 | 36,002
3	忍 唔 住 | 10,221 必 買手 信 | 7,762 必 食 必 | 7,683 食 必 買 | 7,648 玩 得 開心 | 4,818
4	必 食 必 買 | 7,646 識 揀 一定 揀 | 3,859 讀者 有 任何 關於 | 3,846 任何 新奇 有趣 事物 | 3,846 大家 可以 透過 facebookinbox | 3,845
5	讀者 有 任何 關於 美食 | 3,846 歡迎 話 比 小編 們 | 3,843 直接 將 資料 傳 比 | 3,842 將 資料 傳 比 小編 | 3,841 買手 信 帶 返 香港 | 3,824

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				