HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-ms.tsv	1/26/2025	English (en)	Malay (ms)

Volumes

Segments	SL tokens	SL characters	SL size
8,432,285	179M	933,359,950	893.65 MB

TL tokens	TL characters	TL size
174M	1,044,032,960	997.25 MB

Dataset top 10 domains

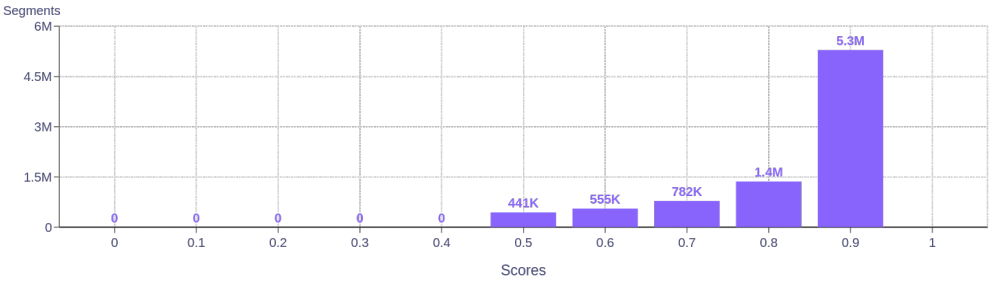
SL domain	Segments	TL domain	Segments
hotels.com	63.9%	hotels.com	23.1%
google.com	19.8%	wikipedia.org	8.4%
wikipedia.org	10.4%	google.com	7.5%
agoda.com	8.3%	agoda.com	5.9%
booking.com	7.1%	booking.com	3.7%
orangesmile.com	1.6%	blogspot.com	2.3%
lacroixx.com	1.5%	lacroixx.com	1.5%
itsmygame.org	1.4%	hotelscombined.my	1.3%
masterstudies.com	1.4%	itsmygame.org	1.1%
airwise.com	1.4%	airwise.com	1.1%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	182.2%	com	101.1%
org	19.5%	org	15.1%
net	5.3%	com.my	7.5%
com.my	3.1%	my	5.4%
gov.my	2.1%	net	4.1%
my	2.0%	gov.my	2.2%
co.uk	1.7%	ru	1.0%
ru	1.0%	info	0.7%
ca	1.0%	edu.my	0.6%
fm	0.8%	nu	0.5%

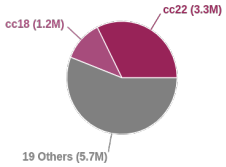
Translation likelihood

≥ 5 = 8.4M segments | **100.0%**
≥ 8 = 6.7M segments | **78.9%**
< 5 = 0 segments | **0.0%**



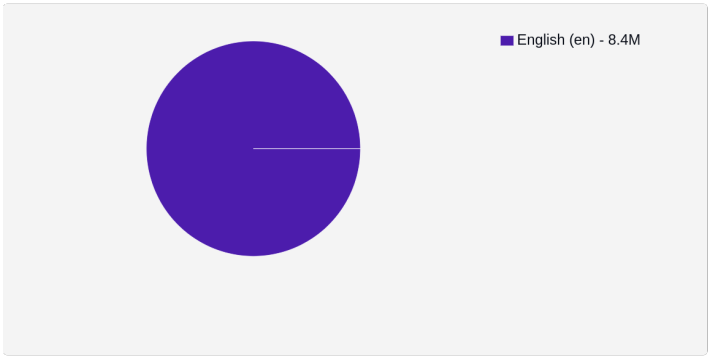
Collections

CC = 65.57%
IA = 34.43%

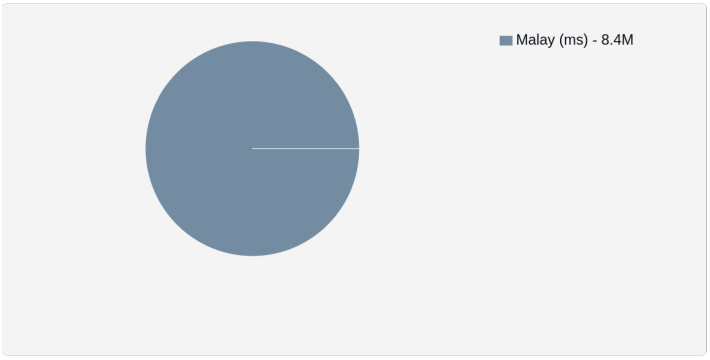


Language Distribution

Source

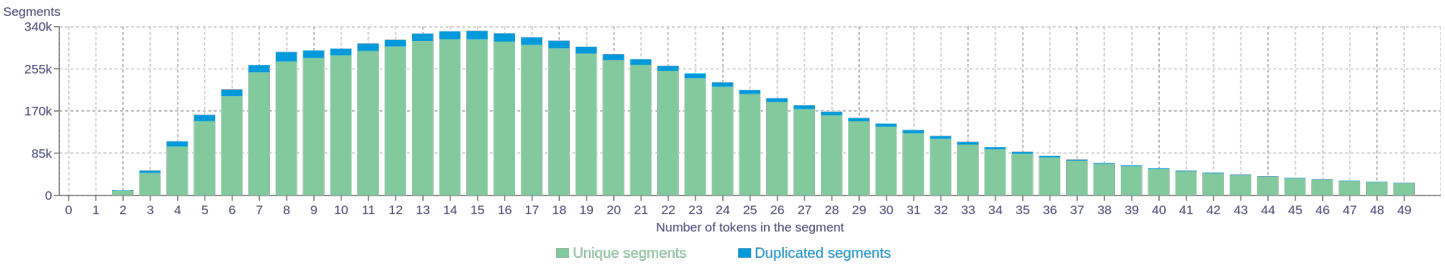


Target



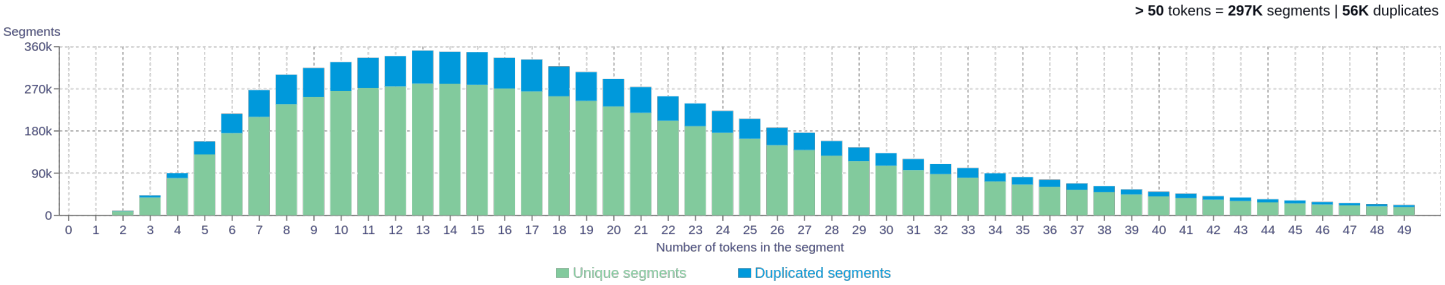
Source segment length distribution by token

<= 49 tokens = **7.7M** segments | **382K** duplicates
> 50 tokens = **334K** segments | **7.8K** duplicates

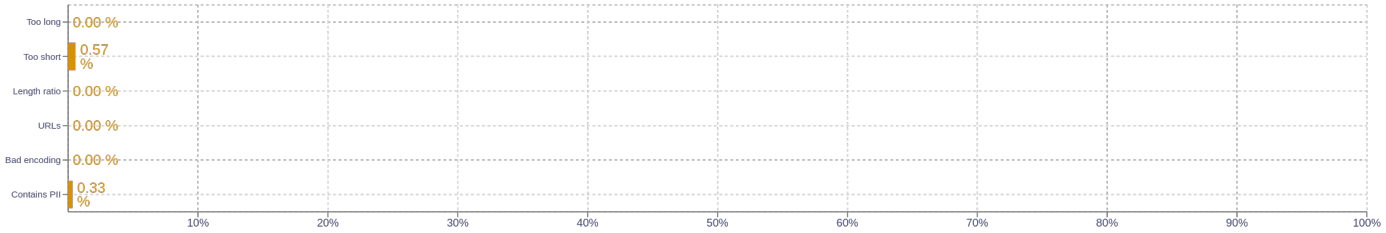


Target segment length distribution by token

<= 49 tokens = **6.5M** segments | **1.6M** duplicates
> 50 tokens = **297K** segments | **56K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 442938 · also | 375479 · use | 306657 · one | 285506 · may | 280405
2	personal data | 70848 · personal information | 61634 · public areas | 40530 · privacy policy | 36556 · kuala lumpur | 31854
3	wi-fi in public | 23435 · available to hotel | 18578 · terms and conditions | 18481 · choice for travelers | 16476 · like the game | 15652
4	hotels on a map | 27564 · wi-fi in public areas | 23412 · available to hotel guests | 18552 · wi-fi available to hotel | 18551 · wi-fi in all rooms | 16759
5	wi-fi available to hotel guests | 18551 · free wi-fi in all rooms | 16749 · people looked at this hotel | 12432 · hotel in the last hour | 12432 · cards and reserves the right | 9876

Target n-grams

Size	n-grams
1	anda | 2924614 · hotel | 681806 · permainan | 327120 · bilik | 295982 · laman | 254012
2	laman web | 188405 · peribadi anda | 87062 · data peribadi | 85495 · berjalan kaki | 50491 · sekiranya anda | 46497
3	data peribadi anda | 49798 · minit berjalan kaki | 39443 · laman web anda | 31257 · terma dan syarat | 19440 · permainan dalam talian | 19342
4	fi untuk tetamu hotel | 19033 · berkualiti tinggi bergerak telefon | 18167 · percuma dalam semua bilik | 10580 · kad kredit sebelum ketibaan | 9921 · kredit sebelum ketibaan anda | 9912
5	makan tengah hari dan makan | 14794 · hotel ini dalam masa sejam | 13163 · fi percuma dalam semua bilik | 10579 · kad kredit sebelum ketibaan anda | 9912 · menahan jumlah wang yang ditetapkan | 9638

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