HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-vi.tsv	1/28/2025	English (en)	Vietnamese (vi)

Volumes

Segments	SL tokens	SL characters	SL size
19,231,770	422M	2,161,036,870	2.02 GB

TL tokens	TL characters	TL size
565M	2,306,240,114	2.79 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
wikipedia.org	11.9%	wikipedia.org	9.8%
alibaba.com	11.3%	alibaba.com	9.0%
hotels.com	8.3%	hotels.com	3.9%
google.com	7.9%	google.com	3.7%
microsoft.com	2.4%	tripadvisor.com.vn	1.9%
wikihow.com	1.8%	wikihow.vn	1.8%
agoda.com	1.4%	microsoft.com	1.6%
biblegateway.com	1.4%	biblegateway.com	1.2%
booking.com	1.3%	agoda.com	1.2%
softoware.net	1.2%	softoware.net	1.1%

Dataset top 10 TLDs

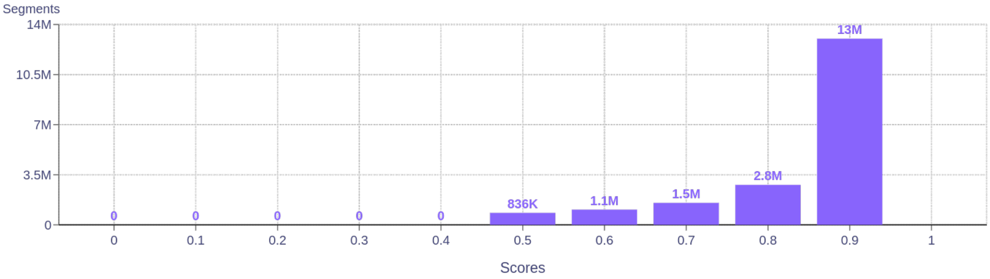
SL domain	Segments	TL domain	Segments
com	107.3%	com	71.4%
org	21.2%	org	17.1%
net	5.6%	vn	14.1%
vn	3.2%	com.vn	7.3%
com.vn	2.0%	net	5.5%
co.uk	1.8%	edu.vn	1.2%
edu.vn	1.0%	info	0.8%
ca	0.9%	io	0.7%
in	0.9%	gov	0.6%
com.au	0.8%	ru	0.5%

Translation likelihood

≥ 5 = 19M segments | **100.0%**

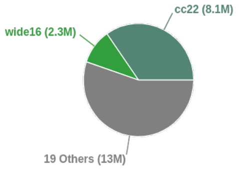
≥ 8 = 16M segments | **82.2%**

< 5 = 0 segments | **0.0%**



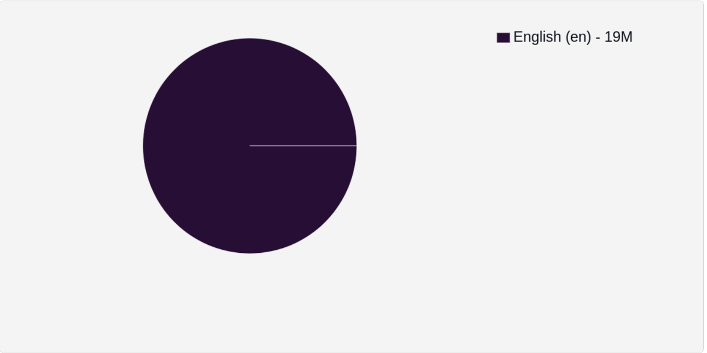
Collections

CC = 65.72%
IA = 34.28%

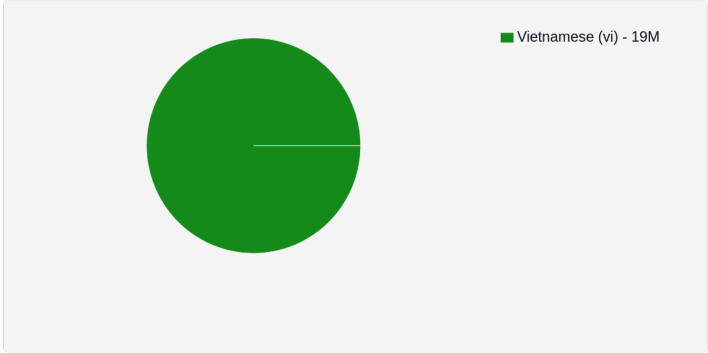


Language Distribution

Source



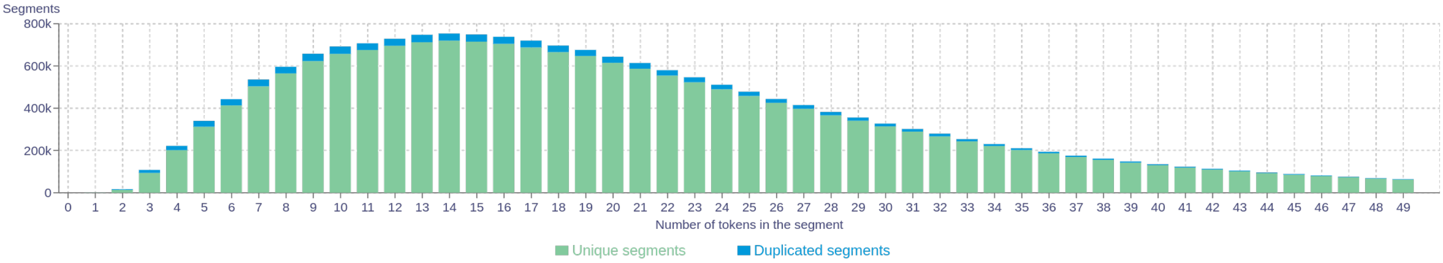
Target



Source segment length distribution by token

<= 49 tokens = **17M** segments | **850K** duplicates

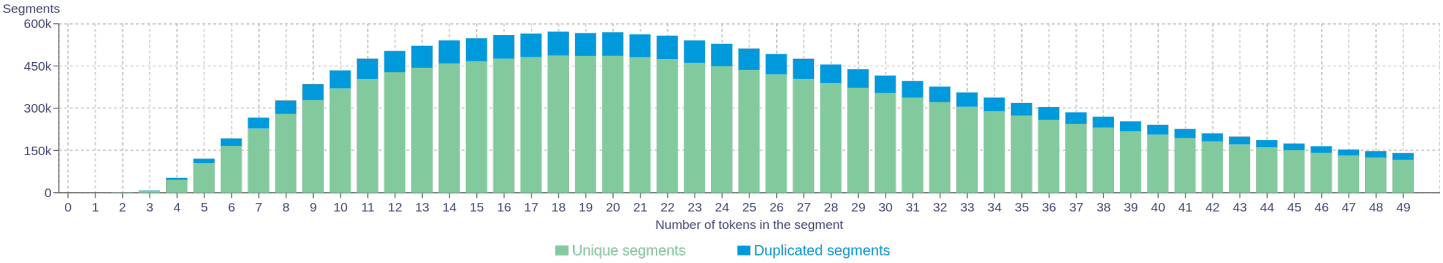
> 50 tokens = **905K** segments | **23K** duplicates



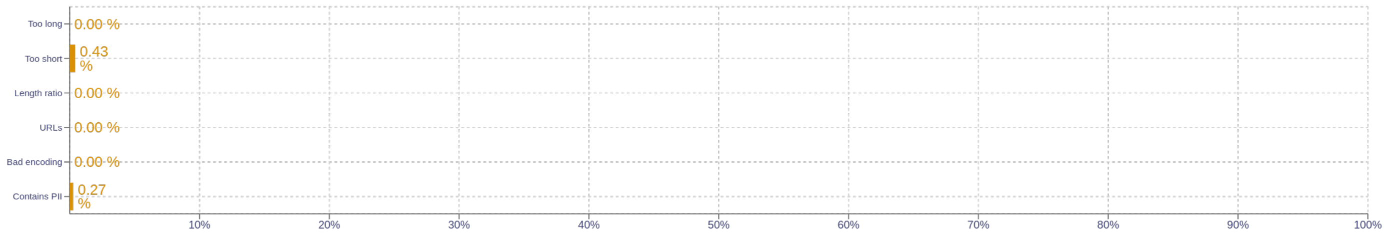
Target segment length distribution by token

<= 49 tokens = **14M** segments | **2.5M** duplicates

> 50 tokens = **2.3M** segments | **327K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 836632 one | 735762 new | 653186 use | 639160 time | 630212
2	united states | 81430 samsung galaxy | 56639 make sure | 55620 high quality | 52332 personal information | 47371
3	around the world | 31513 ho chi minh | 31123 hotel is within | 30464 chi minh city | 25614 call of duty | 20428
4	ho chi minh city | 25570 one of the best | 17567 within a 10-minute walk | 12346 located in the heart | 11710 music singers and bands | 10629
5	tripadvisor is proud to partner | 18060 streamed directly from their servers | 9916 player fm and our community | 9916 forget to rate this game | 9020 game with your best friends | 8870

Target n-grams

Size	n-grams
1	thể | 3419198 chúng | 2664929 dụng | 2157369 công | 2114699 hàng | 1613499
2	sử dụng | 1325102 cung cấp | 803639 sản phẩm | 602301 thông tin | 591538 dịch vụ | 584999
3	đức chúa trời | 109771 thể sử dụng | 95238 cách sử dụng | 74652 trường đại học | 66021 mối quan hệ | 57881
4	chúng tôi có thể | 99839 cung cấp dịch vụ | 56271 thông tin cá nhân | 53008 thể được sử dụng | 42015 liên hệ với chúng | 41932
5	tọa lạc tại khu vực | 27446 thành phố hồ chí minh | 22463 internet không dây miễn phí | 18839 đặt phòng tại khách sạn | 18289 tự hào khi được hợp | 18093

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