HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-th.tsv	1/24/2025	English (en)	Thai (th)

Volumes

Segments	SL tokens	SL characters	SL size
4,088,354	117M	606,353,124	580.64 MB

TL tokens	TL characters	TL size
128M	606,017,378	1.5 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
google.com	12.0%	tripadvisor.com	5.3%
hotels.com	6.6%	google.com	5.0%
microsoft.com	4.9%	hotels.com	3.5%
tripadvisor.com	4.7%	expedia.co.th	3.4%
wikipedia.org	3.2%	microsoft.com	3.4%
agoda.com	1.8%	wikipedia.org	2.8%
hotelscombined.com	1.7%	hotelscombined.co.th	2.7%
wikihow.com	1.7%	wikihow.com	1.5%
apple.com	1.4%	agoda.com	1.5%
biblegateway.com	1.4%	biblegateway.com	1.2%

Dataset top 10 TLDs

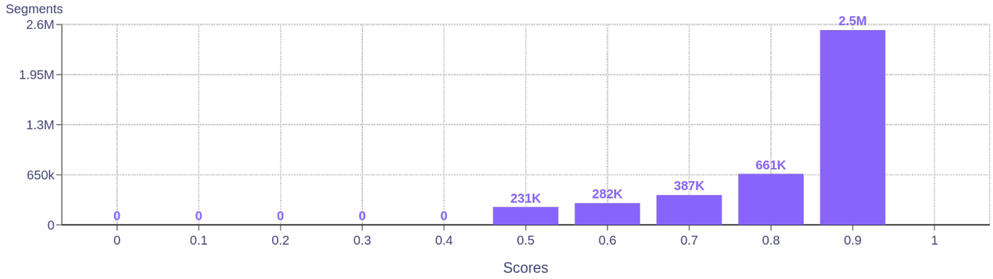
SL domain	Segments	TL domain	Segments
com	122.5%	com	87.9%
org	9.4%	co.th	11.6%
net	5.4%	org	7.7%
co.uk	3.0%	net	4.6%
co.th	2.8%	co	0.7%
com.sg	1.5%	info	0.6%
ca	1.5%	in.th	0.5%
com.au	1.4%	or.th	0.5%
in	1.2%	asia	0.5%
ie	1.0%	io	0.4%

Translation likelihood

≥ 5 = 4.1M segments | **100.0%**

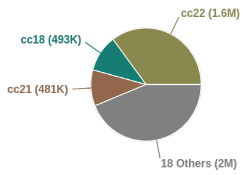
≥ 8 = 3.2M segments | **78.0%**

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



Collections

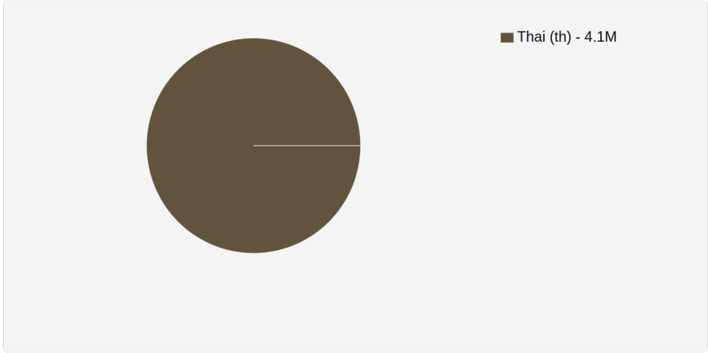
CC = 67.45%
IA = 32.55%



Language Distribution

Source

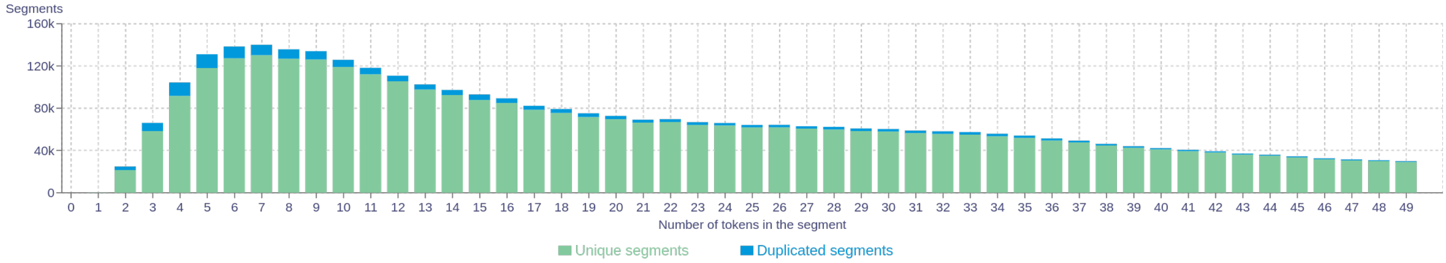
Target



Source segment length distribution by token

<= 49 tokens = **3.2M** segments | **177K** duplicates

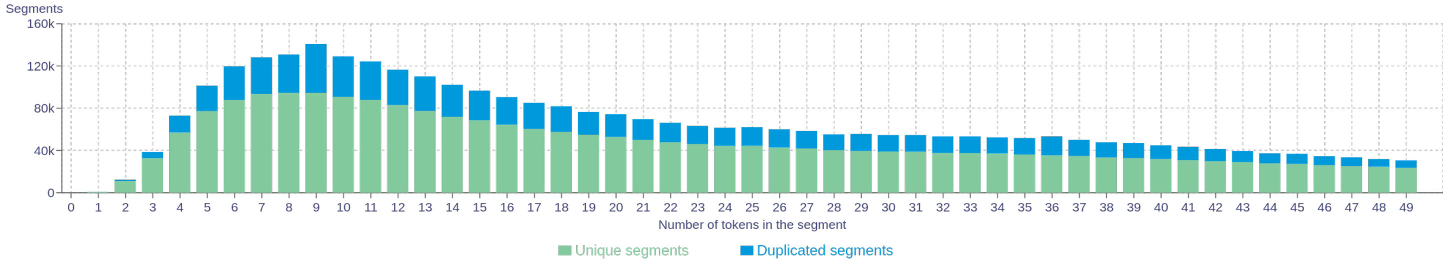
> 50 tokens = **693K** segments | **15K** duplicates



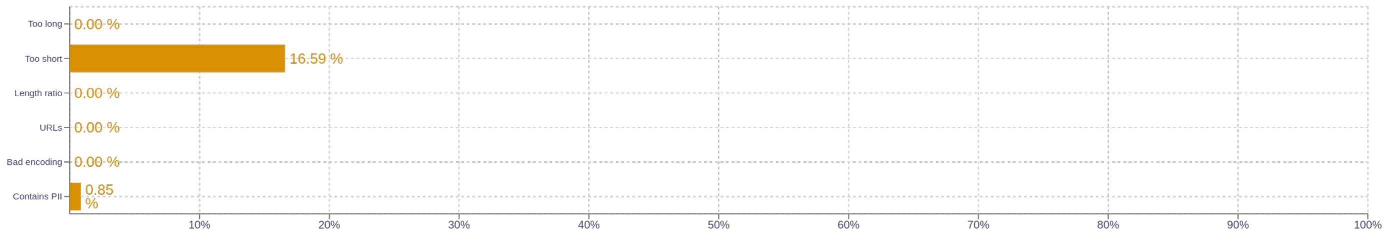
Target segment length distribution by token

<= 49 tokens = **2.4M** segments | **922K** duplicates

> 50 tokens = **812K** segments | **210K** duplicates



Segment pair noise distribution



Source n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>hotel | 457544
information | 269098
use | 253791
best | 222406
business | 218176</td></tr>
<tr><td>2</td><td>special offers | 133655
perfect hotel | 132805
business trips | 132734
help millions | 132682
best discounts | 132682</td></tr>
<tr><td>3</td><td>proud to partner | 133944
tripadvisor is proud | 133914
reservations with confidence | 133902
find the perfect | 132920
discounts and special | 132728</td></tr>
<tr><td>4</td><td>discounts and special offers | 132723
find the perfect hotel | 132682
always with the best | 132670
best discounts and special | 132668
vacation and business trips | 84435</td></tr>
<tr><td>5</td><td>tripadvisor is proud to partner | 133914
month to find the perfect | 132668
best discounts and special offers | 132668
always with the best discounts | 132668
travelers each month to find | 84435</td></tr>
</table>

Target n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>คุณ | 1654317
สามารถ | 771254
ใช้ | 594674
สำหรับ | 591427
ข้อมูล | 533194</td></tr>
<tr><td>2</td><td>คุณ สามารถ | 188912
ค้นหา โรงแรม | 152206
จอง ห้องพัก | 136329
ข้อมูล ส่วนบุคคล | 134876
นับ ล้าน | 134836</td></tr>
<tr><td>3</td><td>คุณ จึง สามารถ | 136026
โรงแรม ที่ เหมาะสำหรับ | 134436
ข้อเสนอ สุด พิเศษ | 134078
สามารถ จอง ห้องพัก | 134009
ส่วนลด และ ข้อเสนอ | 133880</td></tr>
<tr><td>4</td><td>คุณ จึง สามารถ จอง | 133956
ค้นหา โรงแรม ที่ เหมาะสำหรับ | 133778
โรงแรม ที่ เหมาะสำหรับ ทริป | 133770
แต่ละ เดือน เรา ช่วย | 133770
เหมาะสำหรับ ทริป วันหยุด พักผ่อน | 133770</td></tr>
<tr><td>5</td><td>คุณ จึง สามารถ จอง ห้องพัก | 133956
ส่วนลด และ ข้อเสนอ สุด พิเศษ | 133771
โรงแรม ที่ เหมาะสำหรับ ทริป วันหยุด | 133770
เดือน เรา ช่วย ให้ นักท่องเที่ยว | 133770
ล้าน ค้นหา โรงแรม ที่ เหมาะสำหรับ | 133770</td></tr>
</table>

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