HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-he.tsv	1/27/2025	English (en)	Hebrew (he)

Volumes

Segments	SL tokens	SL characters	SL size
8,686,089	218M	1,125,131,684	1.05 GB

TL tokens	TL characters	TL size
198M	942,884,908	1.49 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
hotels.com	41.7%	hotels.com	16.2%
alibaba.com	20.8%	alibaba.com	14.8%
wikipedia.org	9.6%	wikipedia.org	8.6%
booking.com	6.5%	booking.com	3.4%
microsoft.com	3.2%	tripadvisor.co.il	2.3%
agoda.com	2.9%	microsoft.com	2.1%
kayak.com	2.0%	agoda.com	2.0%
booked.net	1.9%	kayak.com	1.8%
softoware.net	1.7%	sacred-texts.com	1.5%
google.com	1.6%	medwowglobal.com	1.5%

Dataset top 10 TLDs

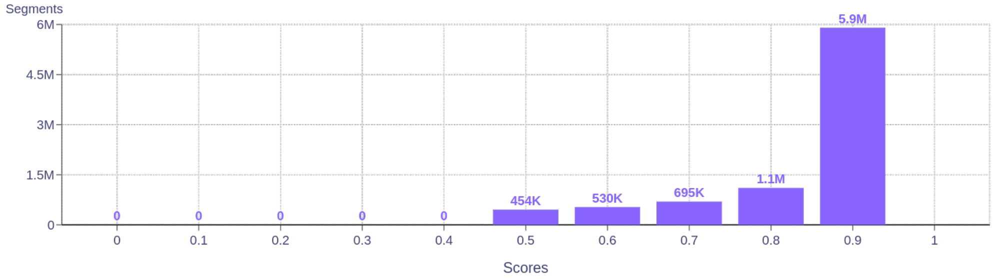
SL domain	Segments	TL domain	Segments
com	167.9%	com	100.2%
org	18.7%	co.il	15.2%
net	10.9%	org	15.0%
co.il	2.8%	net	6.0%
org.il	2.6%	org.il	3.8%
co.uk	2.6%	info	1.5%
info	1.6%	ac.il	0.9%
ca	1.1%	co.uk	0.6%
in	0.9%	com.br	0.4%
ac.il	0.8%	de	0.4%

Translation likelihood

≥ 5 = 8.7M segments | **100.0%**

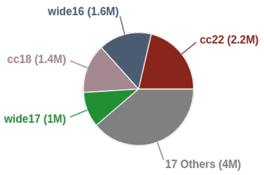
≥ 8 = 7M segments | **80.7%**

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



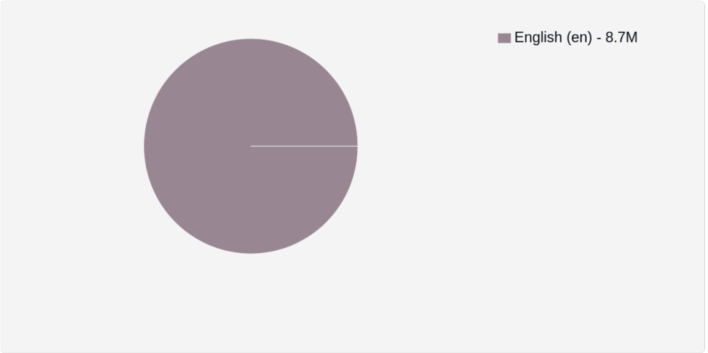
Collections

CC = 55.40%
IA = 44.60%

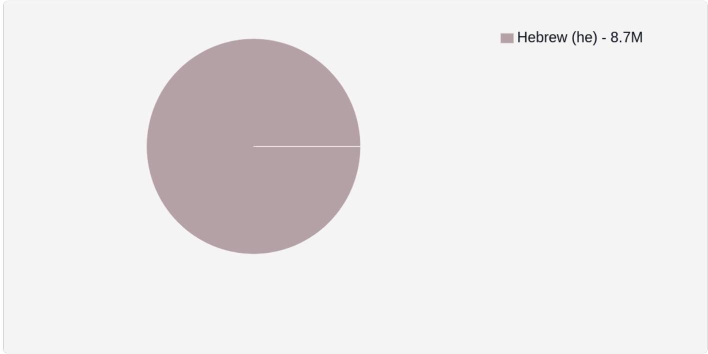


Language Distribution

Source



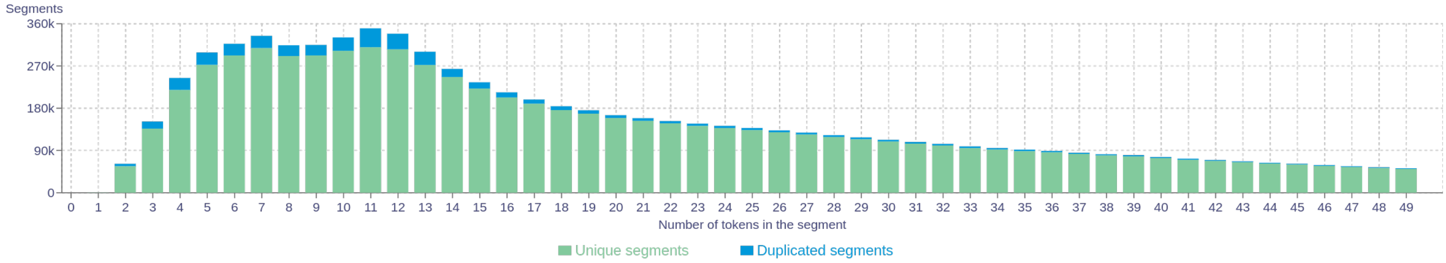
Target



Source segment length distribution by token

<= 49 tokens = **7.2M** segments | **455K** duplicates

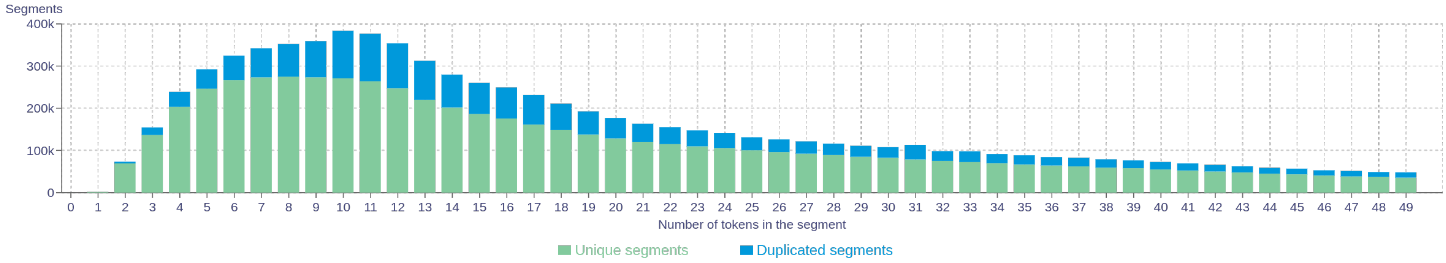
> 50 tokens = **1.1M** segments | **27K** duplicates



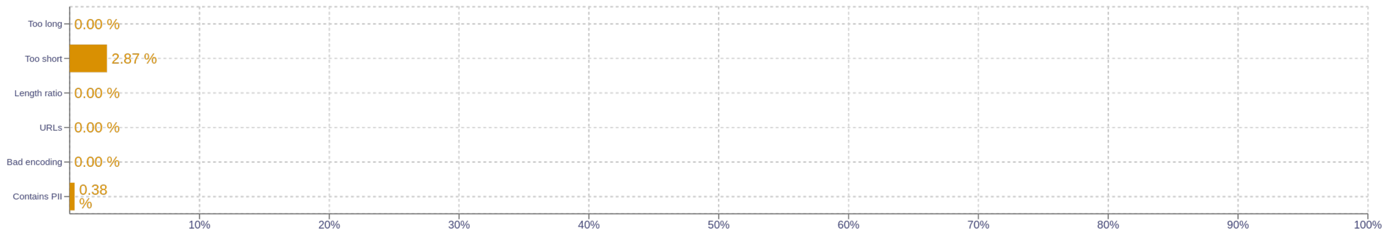
Target segment length distribution by token

<= 49 tokens = **5.9M** segments | **2M** duplicates

> 50 tokens = **799K** segments | **185K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 893141 map | 428369 use | 365449 also | 339767 km | 329413
2	show map | 255988 city center | 81301 united states | 79676 km away | 73783 special offers | 72682
3	find the perfect | 73047 proud to partner | 72471 tripadvisor is proud | 72448 reservations with confidence | 72443 discounts and special | 71273
4	find the perfect hotel | 71823 discounts and special offers | 71268 always with the best | 71238 best discounts and special | 71236 hotel for both holiday | 44546
5	tripadvisor is proud to partner | 72448 month to find the perfect | 71236 best discounts and special offers | 71236 always with the best discounts | 71236 travellers each month to find | 44546

Target n-grams

Size	n-grams
1	873373 | ב 643429 | כדי 633416 | המלון 555708 | עבור 430672 | מ
2	344614 | הצג מפה 120747 | דקות הליכה 115028 | ממרכז העיר 100037 | דקות נסיעה 97711 | נמצא במרחק
3	83767 | עריכת קוד מקור 73472 | ק"מ ממרכז העיר 72494 | גאה להיות שותף 72489 | כדי שתוכל להזמין 72488 | שתוכל להזמין בבטחה
4	72488 | שתוכל להזמין בבטחה מקומות 72488 | כדי שתוכל להזמין בבטחה tripadvisor 72488 | גאה להיות שותף 72273 | למצוא את המלון המושלם 72135 | עבור חופשות והן לנסיעות
5	72488 | כדי שתוכל להזמין בבטחה מקומות 72135 | עבור חופשות והן לנסיעות עסקיות 72135 | מסייעים למיליוני מטיילים בכל חודש 72135 | למיליוני מטיילים בכל חודש למצוא 72135 | חודש למצוא את המלון המושלם

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