HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
18,894,019	332M	1,727,670,171	1.61 GB

TL tokens	TL characters	TL size
570M	1,241,550,816	2.82 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
google.com	16.7%	hotels.com	8.5%
hotels.com	16.5%	google.com	6.8%
venere.com	6.7%	venere.com	6.5%
microsoft.com	5.5%	microsoft.com	4.4%
alibaba.com	4.6%	alibaba.com	4.3%
cisco.com	4.3%	cisco.com	3.9%
wikipedia.org	4.0%	wikipedia.org	3.3%
made-in-china.com	3.2%	tumblr.com	2.9%
tumblr.com	3.1%	made-in-china.com	2.3%
apple.com	2.7%	amazon.com	1.7%

Dataset top 10 TLDs

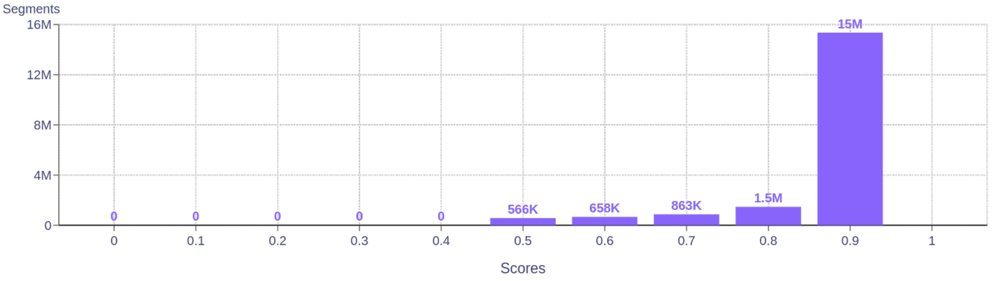
SL domain	Segments	TL domain	Segments
com	160.9%	com	115.5%
org	13.1%	jp	11.6%
net	6.4%	org	9.4%
jp	1.7%	net	4.7%
co.uk	1.4%	co.jp	3.4%
co.jp	0.8%	info	0.8%
info	0.7%	io	0.6%
io	0.7%	ac.jp	0.4%
com.au	0.6%	us	0.3%
us	0.6%	co	0.3%

Translation likelihood

≥ 5 = 19M segments | **100.0%**

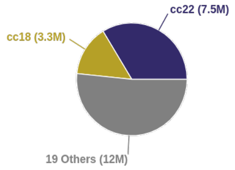
≥ 8 = 17M segments | **89.0%**

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



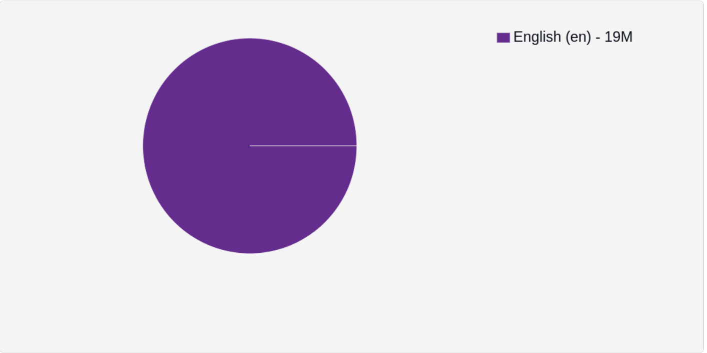
Collections

CC = 63.80%
IA = 36.20%

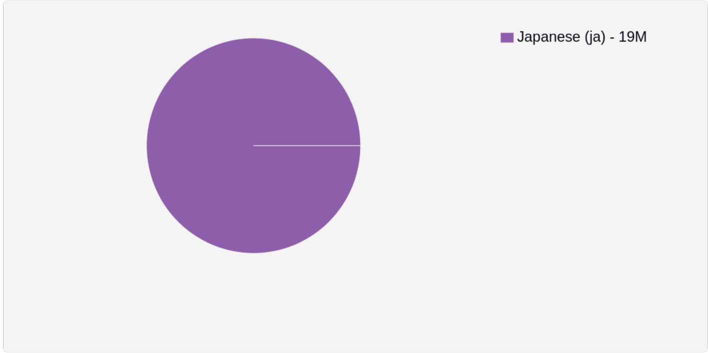


Language Distribution

Source



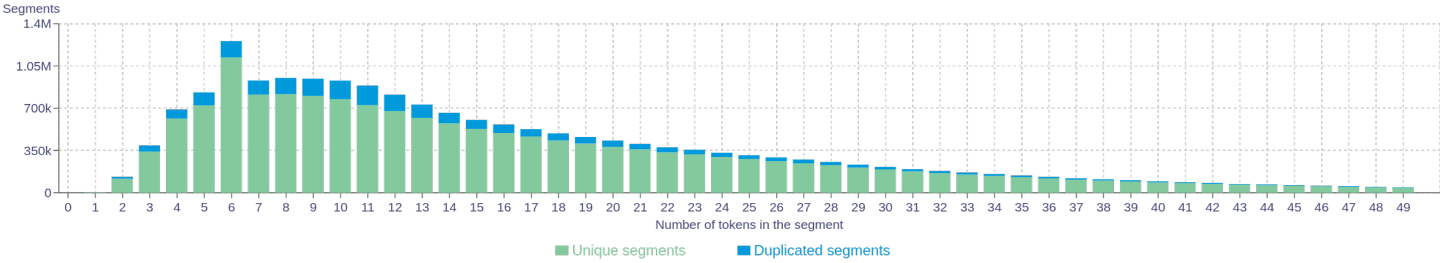
Target



Source segment length distribution by token

<= 49 tokens = **16M** segments | **2.3M** duplicates

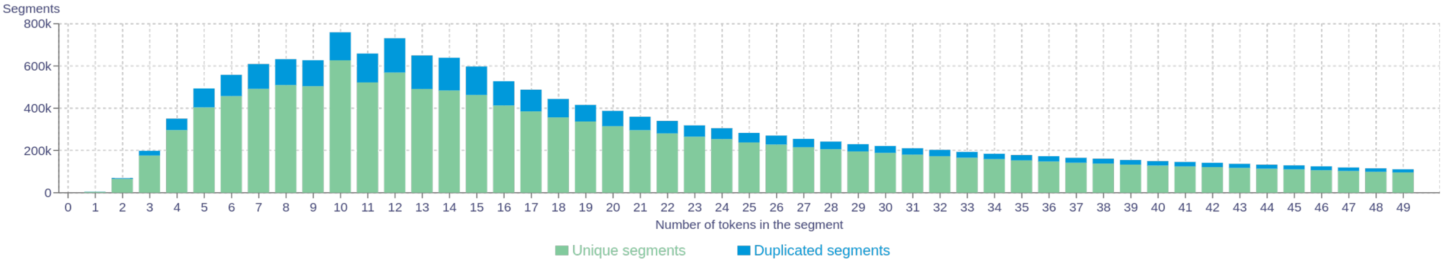
> 50 tokens = **644K** segments | **51K** duplicates



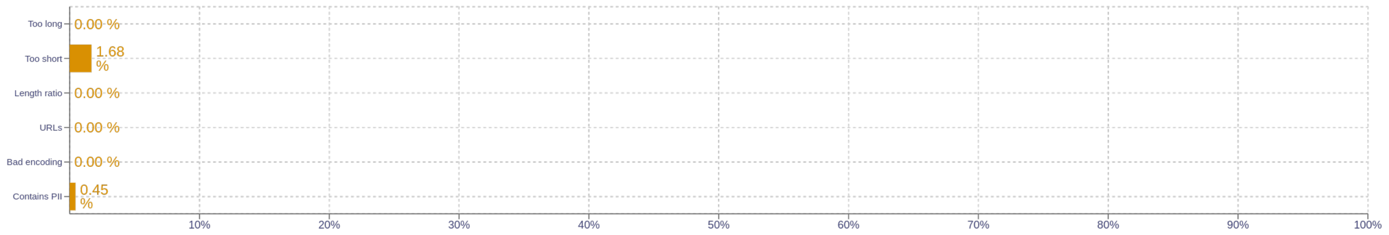
Target segment length distribution by token

<= 49 tokens = **13M** segments | **2.8M** duplicates

> 50 tokens = **3.3M** segments | **389K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	use | 737033 map | 713021 data | 686118 new | 630348 hotel | 580702
2	show map | 412451 open map | 117097 united states | 93408 ip address | 60086 make sure | 50891
3	see on map | 65421 hotel is within | 32519 create a new | 26217 proud to partner | 21670 tripadvisor is proud | 21648
4	one of the following | 17961 km from city centre | 13404 within a 10-minute walk | 13243 located in the heart | 12324 within a 15-minute walk | 8882
5	tripadvisor is proud to partner | 21648 hotel is within a 10-minute | 8280 driving directions to the hotel | 7270 least street address and city | 7268 proud to partner with booking.com | 6492

Target n-grams

Size	n-grams
1	ます | 9995051 て | 9674375 で | 7739973 する | 6885152 た | 4919460
2	さ れ | 3124465 て い | 1910674 でき ます | 1567741 い ます | 1503484 て いる | 1478316
3	て い ます | 1472192 さ れ て | 936314 さ れ た | 904483 さ れ ます | 771961 こと が でき | 711988
4	こと が でき ます | 622983 さ れ て いる | 438997 さ れ て い | 411873 する こと が でき | 400873 れ て い ます | 390381
5	する こと が でき ます | 355866 リブ ログ し まし た | 344854 さ れ て い ます | 279550 する 必要 が あり ます | 218653 提供 し て い ます | 104456

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