HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-ko.tsv	1/29/2025	English (en)	Korean (ko)

Volumes

Segments	SL tokens	SL characters	SL size
18,393,859	417M	2,186,555,821	2.04 GB

TL tokens	TL characters	TL size
545M	1,302,556,372	2.59 GB

Dataset top 10 domains

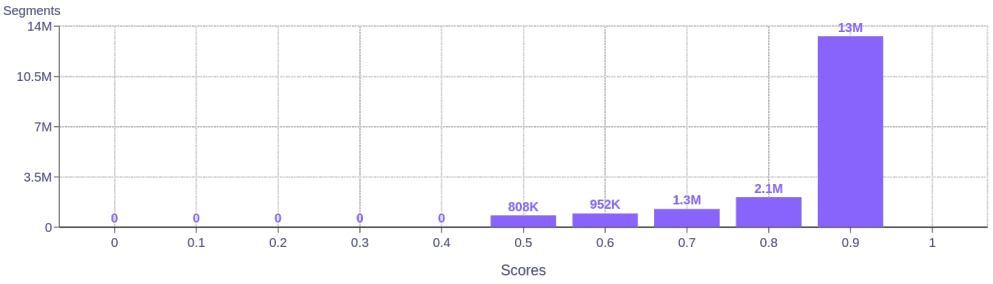
SL domain	Segments	TL domain	Segments
hotels.com	26.4%	hotels.com	11.3%
google.com	20.7%	google.com	7.3%
microsoft.com	8.6%	microsoft.com	6.8%
alibaba.com	6.9%	alibaba.com	5.7%
made-in-china.com	5.6%	made-in-china.com	3.3%
wikipedia.org	2.8%	expedia.co.kr	3.3%
iherb.com	2.3%	wikipedia.org	2.3%
apple.com	2.2%	destinia.kr	1.9%
booking.com	2.2%	hostelworld.com	1.8%
adobe.com	2.0%	amazon.com	1.7%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	184.1%	com	113.5%
org	10.7%	co.kr	9.3%
net	8.8%	org	8.2%
co.uk	2.3%	kr	5.8%
ca	1.0%	net	4.7%
com.au	0.9%	io	0.7%
io	0.8%	jp	0.7%
info	0.7%	info	0.5%
jp	0.7%	ru	0.4%
co.kr	0.7%	or.kr	0.4%

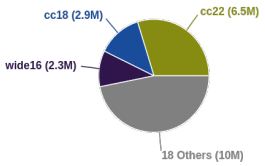
Translation likelihood

≥ 5 = 18M segments | **100.0%**
≥ 8 = 15M segments | **83.6%**
< 5 = 0 segments | **0.0%**



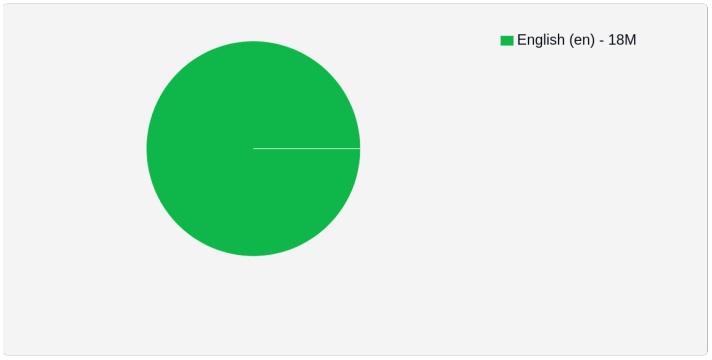
Collections

CC = 64.57%
IA = 35.43%

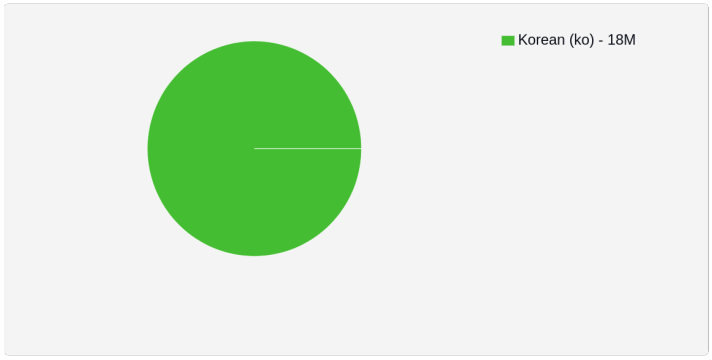


Language Distribution

Source

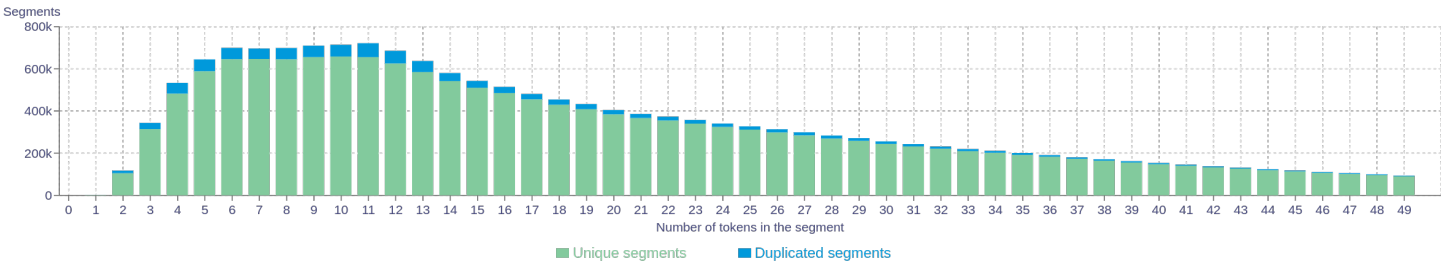


Target



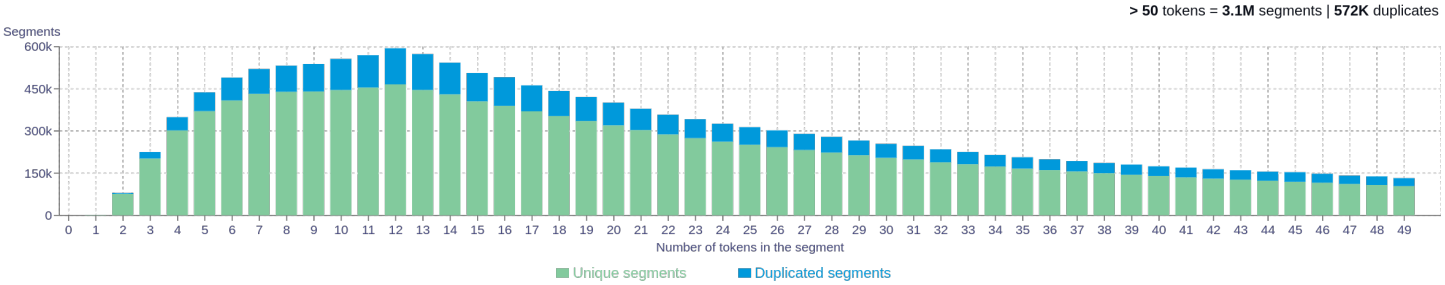
Source segment length distribution by token

<= 49 tokens = **16M** segments | **1.1M** duplicates
> 50 tokens = **1.5M** segments | **46K** duplicates

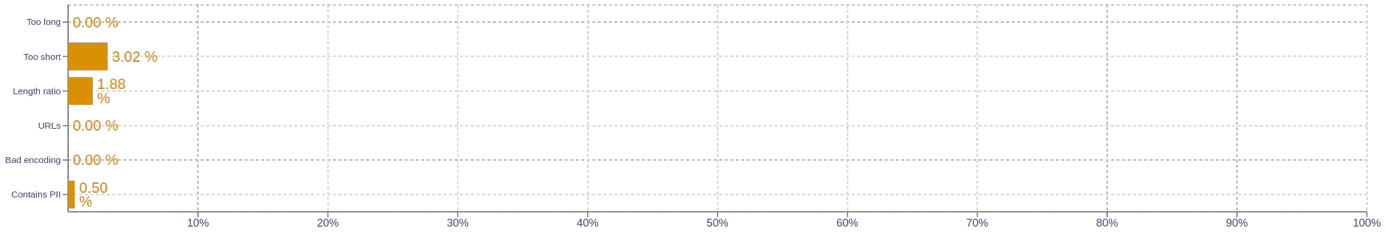


Target segment length distribution by token

<= 49 tokens = **12M** segments | **2.9M** duplicates
> 50 tokens = **3.1M** segments | **572K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 1112999 · use | 952328 · information | 818297 · data | 800193 · new | 671496
2	show map | 287810 · personal information | 153012 · city centre | 116687 · united states | 110574 · open map | 98846
3	see on map | 80873 · hotel is within | 69438 · km from city | 52862 · within a 10-minute | 41398 · terms and conditions | 30823
4	km from city centre | 52634 · within a 10-minute walk | 37006 · within a 15-minute walk | 24525 · located in the heart | 23264 · one of the following | 17078
5	hotel is within a 10-minute | 19874 · tripadvisor is proud to partner | 18991 · best discounts and special offers | 15278 · month to find the perfect | 15277 · always with the best discounts | 15277

Target n-grams

Size	n-grams
1	는 | 12713260 · 있 | 8040834 · 은 | 6878622 · 한 | 5333647 · 수 | 4407794
2	수 있 | 3776323 · 할 수 | 2513509 · 있 는 | 1608797 · 지 않 | 1055299 · 고 있 | 1025621
3	할 수 있 | 2220808 · 수 있 는 | 627516 · 되 어 있 | 381792 · 지도 보 기 | 304644 · 실 수 있 | 297758
4	할 수 있 는 | 421761 · 사용 할 수 있 | 236877 · 위치 하 고 있 | 148140 · 자세 한 내용 은 | 145468 · 할 수 있 도록 | 140314
5	이용 하 실 수 있 | 69013 · 대한 자세 한 내용 은 | 61968 · 사용 할 수 있 는 | 56085 · 이내 의 거리 에 있 | 44012 · 분 도보 거리 에 있 | 42868

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