HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
21,616,652	531M	2,788,991,995	2.61 GB

TL tokens	TL characters	TL size
479M	3,042,855,964	3.07 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
hotels.com	21.5%	alibaba.com	8.5%
alibaba.com	10.9%	hotels.com	8.2%
google.com	9.1%	google.com	3.3%
microsoft.com	4.0%	microsoft.com	3.0%
wikipedia.org	3.3%	wikipedia.org	2.7%
booking.com	3.2%	booking.com	1.7%
tumblr.com	1.6%	tumblr.com	1.2%
hostelworld.com	1.2%	tripadvisor.com.tr	1.2%
office.com	1.1%	hostelworld.com	1.0%
dhgate.com	1.1%	office.com	1.0%

Dataset top 10 TLDs

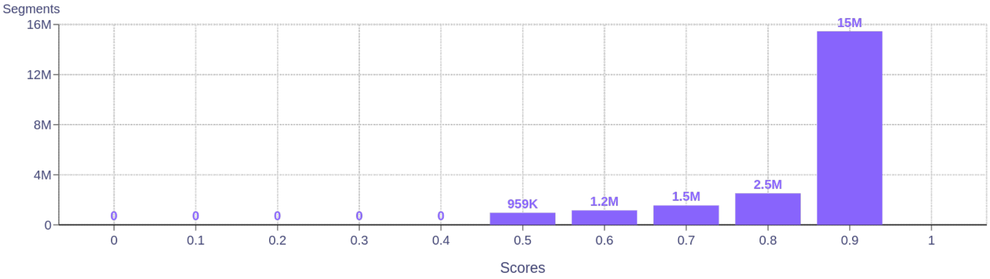
SL domain	Segments	TL domain	Segments
com	150.3%	com	97.5%
org	10.8%	com.tr	11.9%
net	7.1%	org	8.4%
co.uk	3.3%	net	6.0%
com.tr	2.2%	edu.tr	1.3%
de	1.3%	de	1.1%
edu.tr	1.3%	eu	0.8%
eu	1.1%	org.tr	0.7%
com.au	1.1%	info	0.5%
ie	0.9%	biz.tr	0.5%

Translation likelihood

≥ 5 = 22M segments | **100.0%**

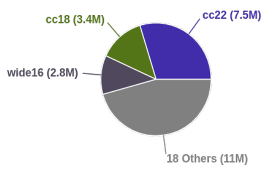
≥ 8 = 18M segments | **83.1%**

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



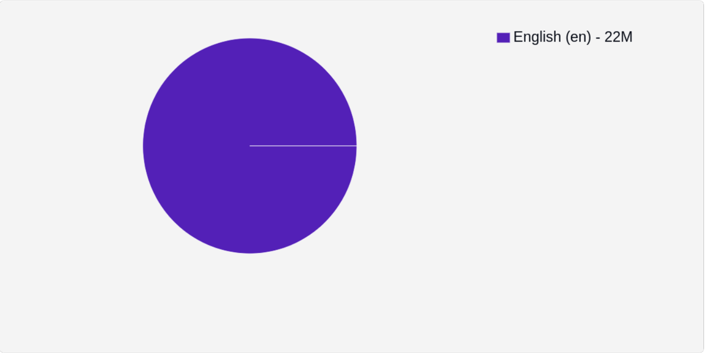
Collections

CC = 62.80%
IA = 37.20%

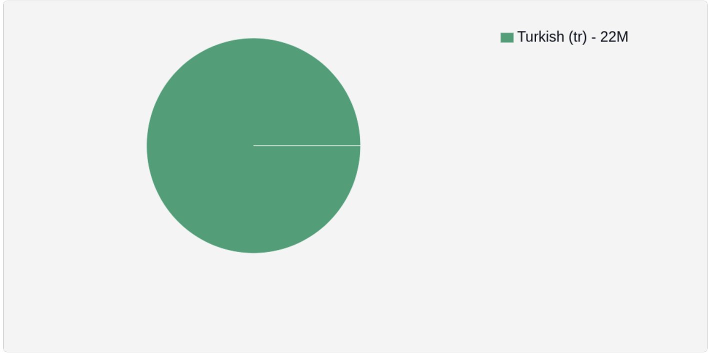


Language Distribution

Source



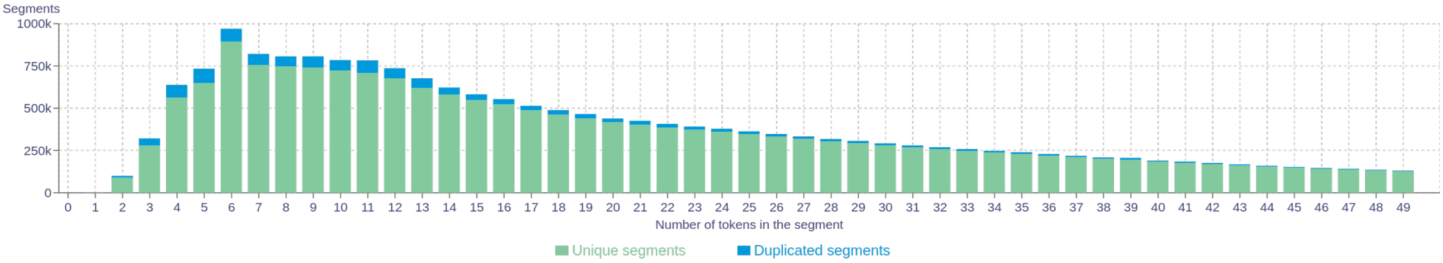
Target



Source segment length distribution by token

<= 49 tokens = **18M** segments | **1.2M** duplicates

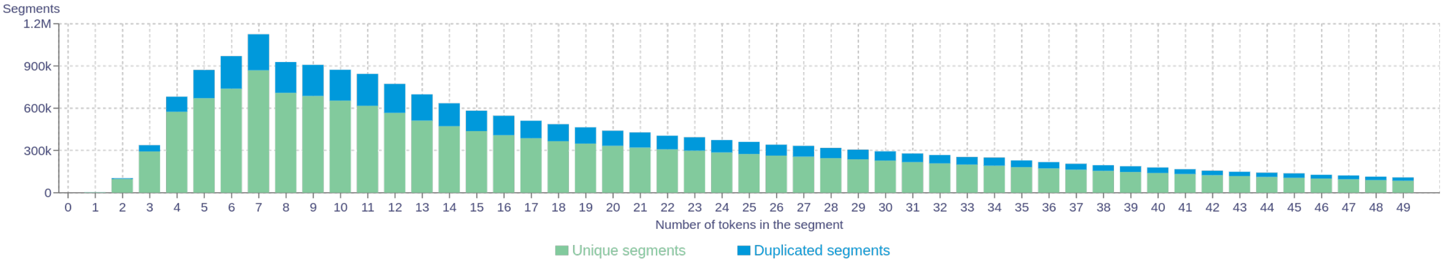
> 50 tokens = **2.5M** segments | **68K** duplicates



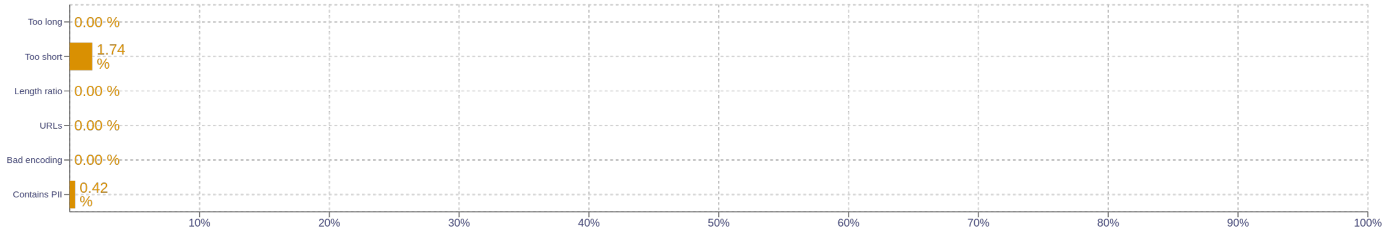
Target segment length distribution by token

<= 49 tokens = **15M** segments | **4.6M** duplicates

> 50 tokens = **1.8M** segments | **375K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 1072832 use | 943426 free | 828252 new | 825052 data | 775618
2	show map | 233851 personal data | 213628 free shipping | 167592 piece free | 117204 city center | 114337
3	proud to partner | 121313 reservations with confidence | 121229 tripadvisor is proud | 121227 piece free shipping | 116396 find the perfect | 107903
4	discounts and special offers | 106415 find the perfect hotel | 106412 always with the best | 106389 best discounts and special | 106361 vacation and business trips | 56579
5	tripadvisor is proud to partner | 121227 month to find the perfect | 106361 best discounts and special offers | 106361 always with the best discounts | 106360 travelers each month to find | 56579

Target n-grams

Size	n-grams
1	bir | 7046356 olarak | 1895068 olan | 977457 yeni | 863635 ücretsiz | 831400
2	herhangi bir | 409074 bir şekilde | 232105 yer alan | 213898 fazla bilgi | 193980 haritayı göster | 191672
3	kullanıcısından yeniden blogladı | 180189 parça ücretsiz sevkiyat | 147557 zaman en iyi | 123290 indirimler ve özel | 122089 yapmaktan gurur duyar | 122074
4	tatil hem de iş | 122064 iş ortaklığı yapmaktan gurur | 122058 zaman en iyi indirimler | 122057 seyahatleri için yardımcı oluyoruz | 122057 seyahat edene hem tatil | 122057
5	tatil hem de iş seyahatleri | 122057 milyonlarca seyahat edene hem tatil | 122057 iş seyahatleri için yardımcı oluyoruz | 122057 iş ortaklığı yapmaktan gurur duyar | 122057 iyi indirimler ve özel tekliflerle | 122057

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