HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
17,505,366	508M	2,711,238,177	2.53 GB

TL tokens	TL characters	TL size
497M	2,506,340,590	4.15 GB

Dataset top 10 domains

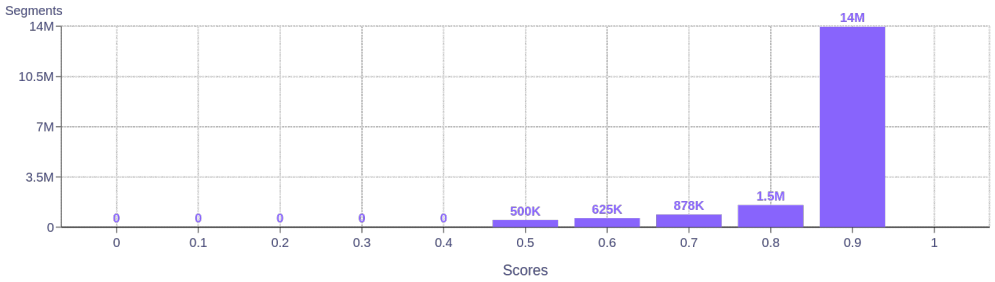
SL domain	Segments	TL domain	Segments
hotels.com	9.0%	hotels.com	4.4%
wikipedia.org	3.5%	wikipedia.org	2.8%
alibaba.com	3.3%	alibaba.com	2.8%
microsoft.com	2.0%	microsoft.com	1.5%
booking.com	1.6%	ohchr.org	1.2%
ohchr.org	1.3%	booking.com	1.2%
office.com	0.9%	office.com	0.8%
airwise.com	0.9%	airwise.com	0.8%
wikihow.com	0.7%	ciwanekurd.net	0.7%
orangesmile.com	0.7%	wikihow.com	0.7%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	101.3%	com	80.5%
org	21.0%	org	18.6%
net	7.5%	net	7.2%
ae	2.0%	ae	3.4%
co.uk	1.6%	com.eg	1.4%
info	0.8%	sa	1.0%
edu	0.8%	ma	0.8%
de	0.7%	info	0.8%
ru	0.6%	de	0.7%
ca	0.6%	ru	0.6%

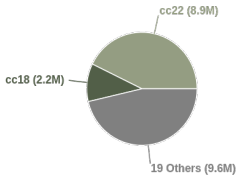
Translation likelihood

≥ 5 = 18M segments | **100.0%**
≥ 8 = 16M segments | **88.6%**
< 5 = 0 segments | **0.0%**



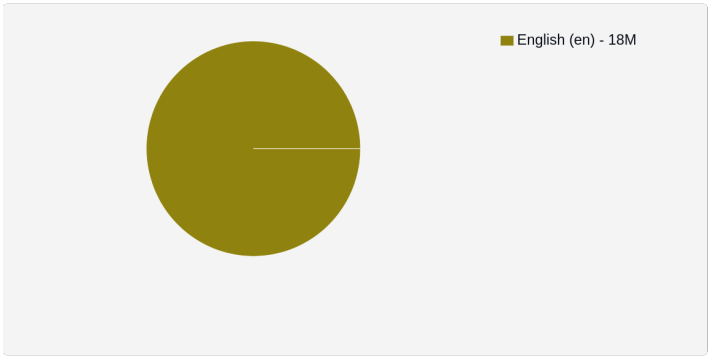
Collections

CC = 71.65%
IA = 28.35%

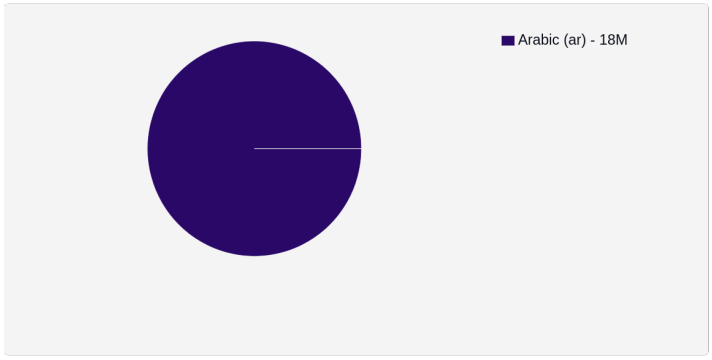


Language Distribution

Source

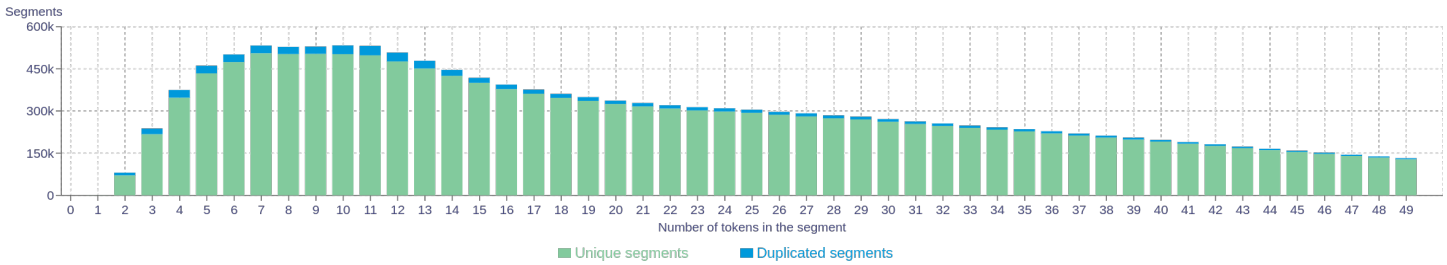


Target



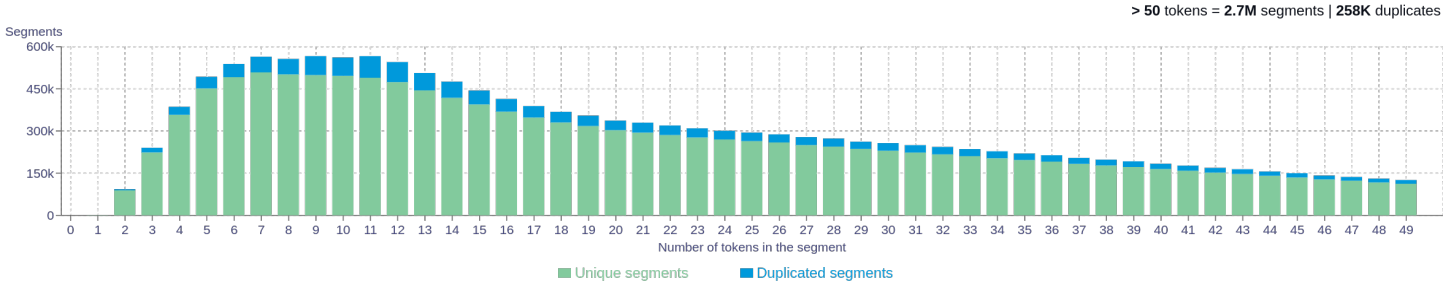
Source segment length distribution by token

<= 49 tokens = **14M** segments | **652K** duplicates
> 50 tokens = **2.8M** segments | **72K** duplicates

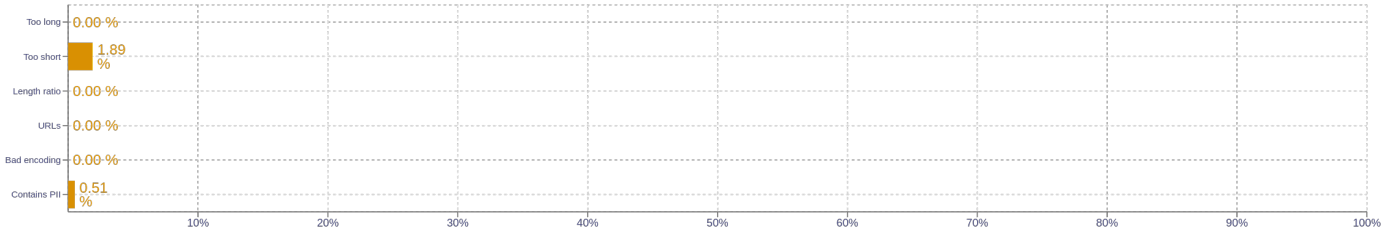


Target segment length distribution by token

<= 49 tokens = **13M** segments | **1.6M** duplicates
> 50 tokens = **2.7M** segments | **258K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 855982 · one | 788031 · use | 729131 · new | 686901 · information | 680604
2	show map | 216877 · human rights | 204723 · united states | 143294 · personal data | 94037 · united nations | 91362
3	around the world | 48639 · see on map | 48424 · proud to partner | 41906 · tripadvisor is proud | 41755 · reservations with confidence | 41754
4	discounts and special offers | 39776 · find the perfect hotel | 39712 · always with the best | 39697 · best discounts and special | 39694 · km from city centre | 28378
5	tripadvisor is proud to partner | 41754 · best discounts and special offers | 39694 · month to find the perfect | 39693 · always with the best discounts | 39693 · travellers each month to find | 20154

Target n-grams

Size	n-grams
1	يمكن | 981214 · خلال | 930689 · يتم | 681969 · عام | 676644 · يمكنك | 667373
2	الولايات المتحدة | 223972 · اعرض الخريطة | 187615 · سبيل المثال | 146184 · الأمم المتحدة | 127063 · عبر الإنترنت | 112328
3	الولايات المتحدة الأمريكية | 65445 · الإمارات العربية المتحدة | 64262 · يجب أن يكون | 46755 · المملكة العربية السعودية | 43982 · يمكن أن يكون | 42916
4	بحيث يمكنك القيام بحجوزاتك | 42026 · عمل على حد سواء | 41940 · العثور على الفندق المثالي | 41940 · أفضل الخصومات والعروض الخاصة | 41935 · وذلك مع القيام دائم | 41934
5	والقيام برحلة عمل على حد | 41934 · لقضاء عطلة والقيام برحلة عمل | 41934 · شهر في العثور على الفندق | 41934 · برحلة عمل على حد سواء | 41934 · بتقديم أفضل الخصومات والعروض الخاصة | 41934

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