HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-arb_Arab	10/3/2025	Standard Arabic

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
50,071,127	755,927,954	564,065,401 (74.62 %)	25.38%	28B	146,475,244,972	246 GB

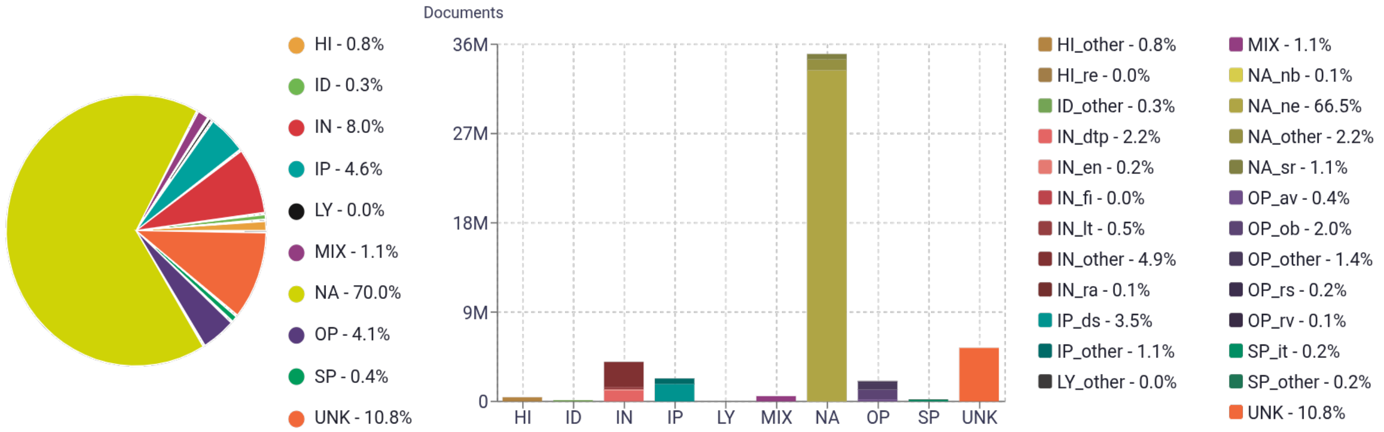
Top 10 domains

Domain	Docs	% of total
aljazeera.net	406K	0.81%
alittihad.ae	360K	0.72%
rt.com	246K	0.49%
netlify.app	241K	0.48%
web.app	227K	0.45%
blogspot.com	196K	0.39%
albayan.ae	194K	0.39%
alaraby.co.uk	168K	0.34%
sputniknews.com	167K	0.33%
alarabiya.net	163K	0.33%

Top 10 TLDs

Domain	Docs	% of total
com	28M	55.39%
net	7.4M	14.74%
org	2.7M	5.30%
ae	1.1M	2.17%
ma	903K	1.80%
ps	664K	1.33%
news	655K	1.31%
info	651K	1.30%
app	479K	0.96%
tv	421K	0.84%

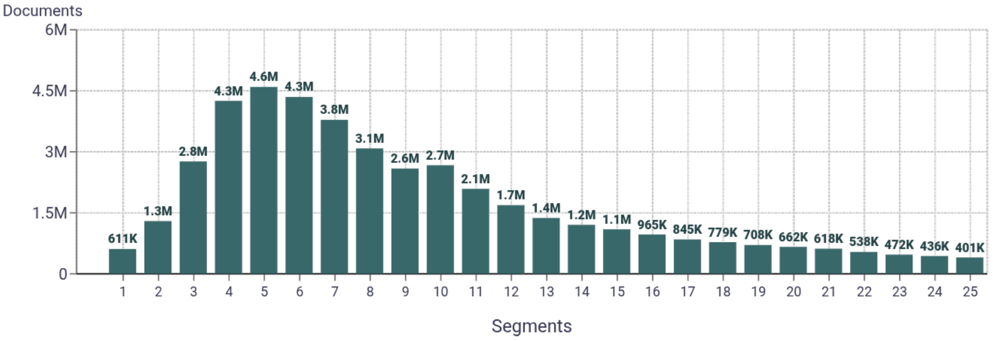
Register labels



MT:5.7% | 2.9M Documents

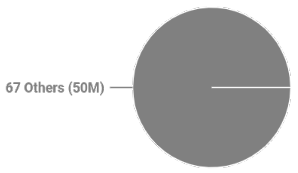
Documents size (in segments)

≤ **25** segments **87.59%** (44M documents)
> 25 segments **12.41%** (6.2M documents)



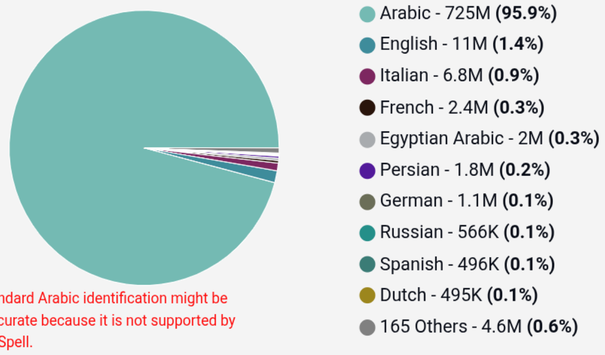
Document collections

CC = 89.55%
IA = 10.45%



Language Distribution

Number of segments in the Standard Arabic corpus

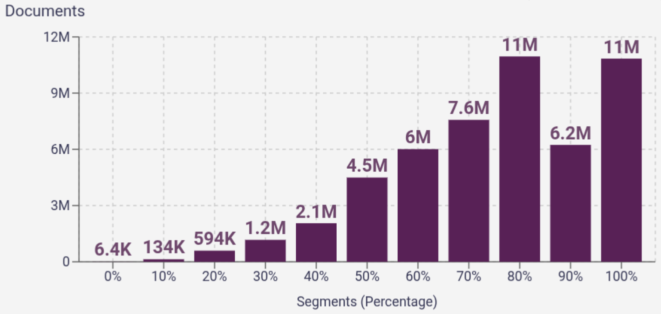


*Standard Arabic identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Standard Arabic inside documents

segments < 50% - **7.91%** (4M documents)

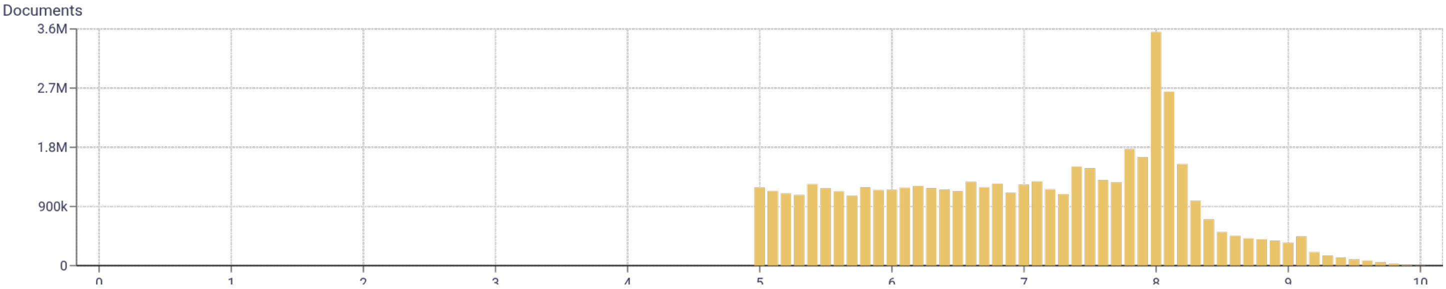
segments ≥ 50% - **92.09%** (46M documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

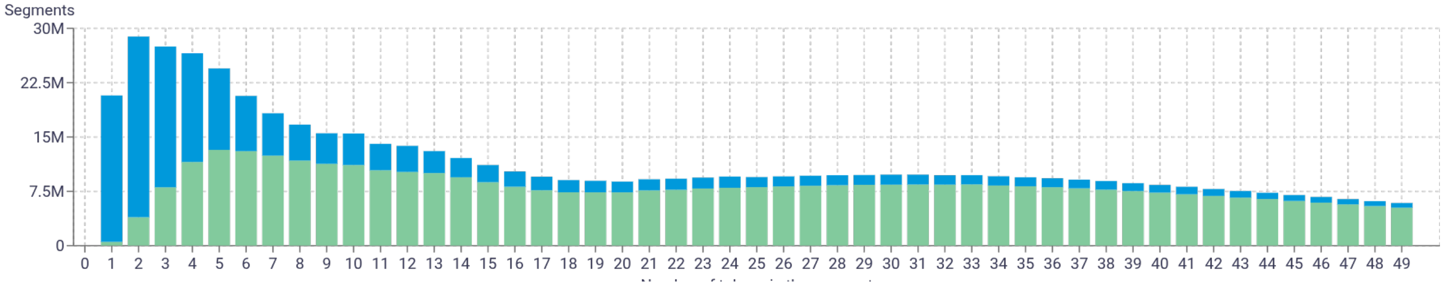
score ≥ 5 - **100%** (50M documents)



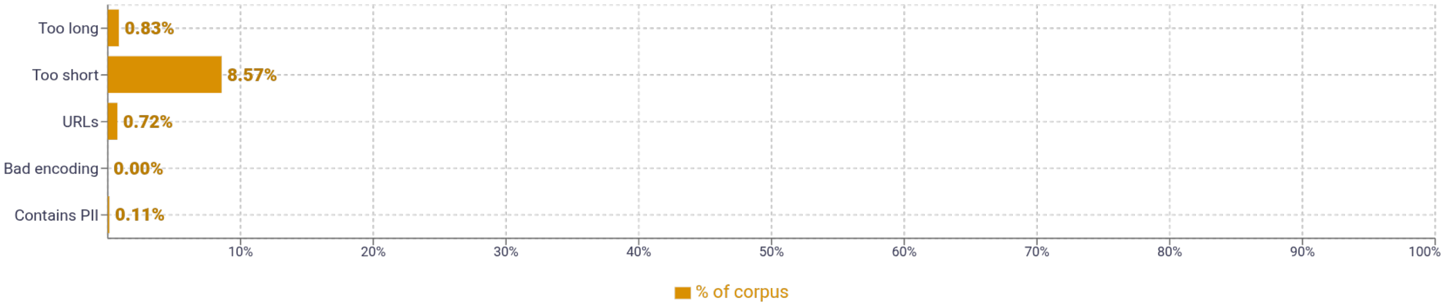
Segment length distribution by token

≤ 49 tokens = **582M** segments | **176M** duplicates

> 50 tokens = **174M** segments | **16M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	64,431,878 | خلال 36,834,705 | اليوم 33,624,991 | يمكن 31,858,950 | عام 31,534,283 | العام
2	11,770,444 | الولايات المتحدة 7,491,096 | الخيارات الثنائية 5,819,230 | أكمل القراءة 5,179,880 | الأمم المتحدة 4,888,034 | عبر الإنترنت
3	3,845,615 | المملكة العربية السعودية 1,944,553 | تداول العملات الأجنبية 1,857,390 | الإمارات العربية المتحدة 1,724,755 | الولايات المتحدة الأمريكية 1,667,447 | تداول الخيارات الثنائية
4	678,105 | دولة الإمارات العربية المتحدة 479,909 | صلى الله عليه وسلم 421,279 | بن زايد آل نهيان 417,975 | الأمين العام للأمم المتحدة 373,271 | الرئيس عبد الفتاح السيسي
5	308,125 | محمد بن راشد آل مكتوم 291,635 | الثنائية هي نوع من الأدوات 290,403 | الأدوات المالية التي تسمح للمستثمر 290,062 | المالية التي تسمح للمستثمر تحقيق 289,340 | تسمح للمستثمر تحقيق مكاسب مالية

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				