HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ary_Arab	10/3/2025	Moroccan Arabic (ary)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
17,503	184,629	169,907 (92.03 %)	7.97%	6.5M	32,631,456	56.06 MB

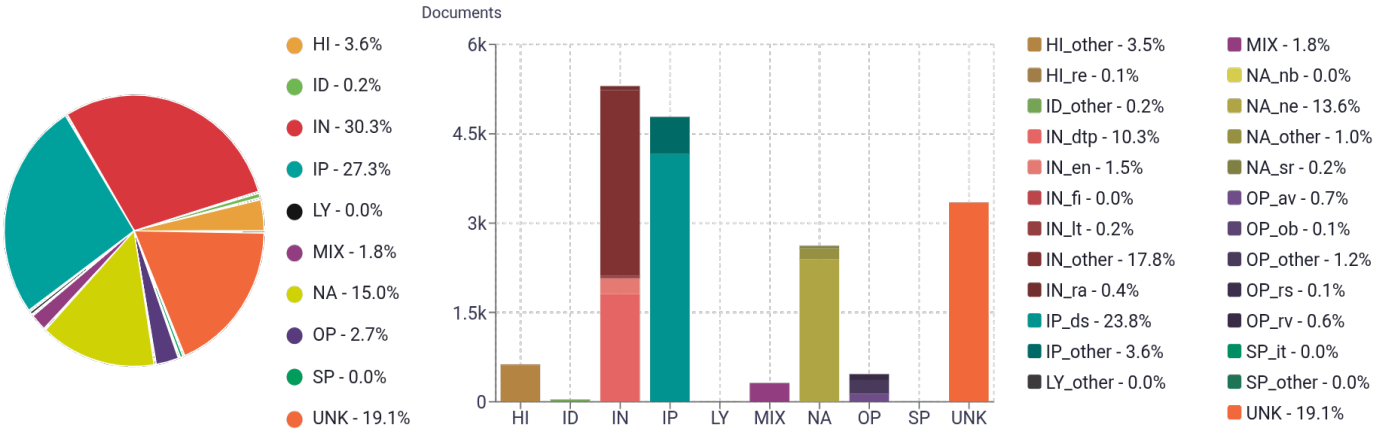
Top 10 domains

Domain	Docs	% of total
tafsir-7ulm.com	420	2.40%
venngage.com	115	0.66%
ktbbh.com	114	0.65%
orange.com	103	0.59%
wikipedia.org	101	0.58%
ktbm.net	89	0.51%
e3arabi.com	89	0.51%
oregonscienceol...	80	0.46%
aspose.com	80	0.46%
facts-news.org	78	0.45%

Top 10 TLDs

Domain	Docs	% of total
com	12K	69.20%
net	1.5K	8.32%
org	801	4.58%
ae	230	1.31%
ma	188	1.07%
news	176	1.01%
info	165	0.94%
me	115	0.66%
co	105	0.60%
sa	102	0.58%

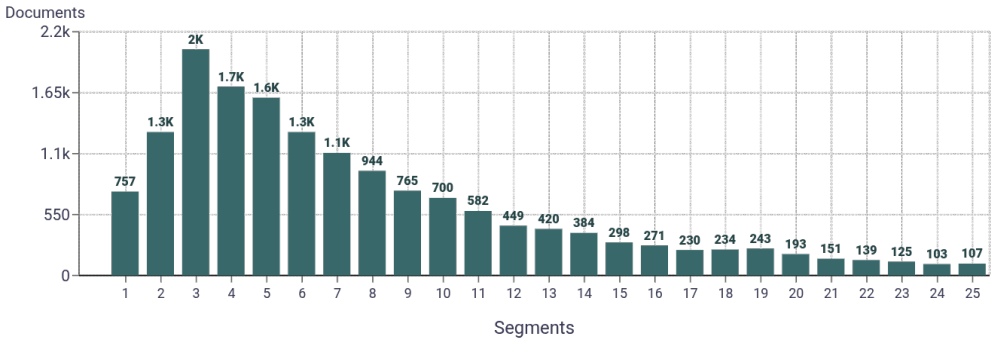
Register labels



MT:10.6% | 1.9K Documents

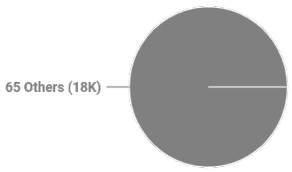
Documents size (in segments)

≤ 25 segments **92.2%** (16K documents)
> 25 segments **7.8%** (1.4K documents)



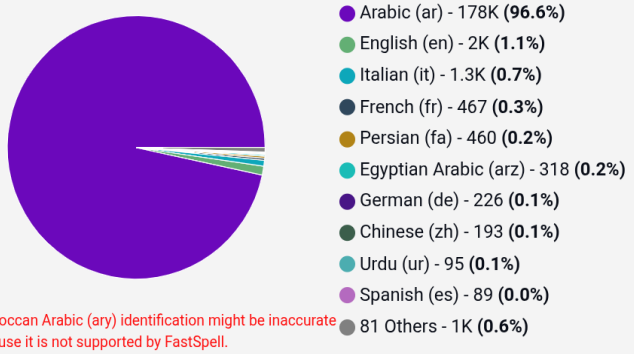
Document collections

CC = 98.50%
IA = 1.50%



Language Distribution

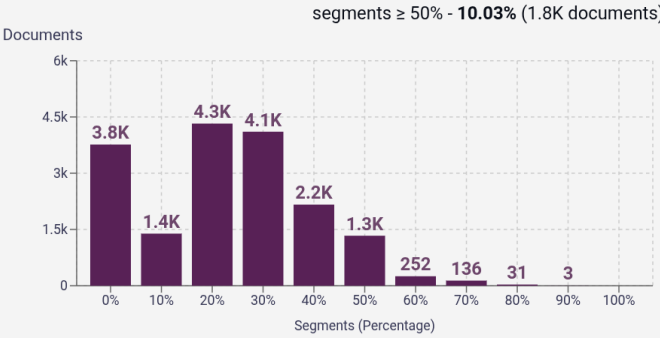
Number of segments in the Moroccan Arabic (ary) corpus



*Moroccan Arabic (ary) identification might be inaccurate because it is not supported by FastSpell.

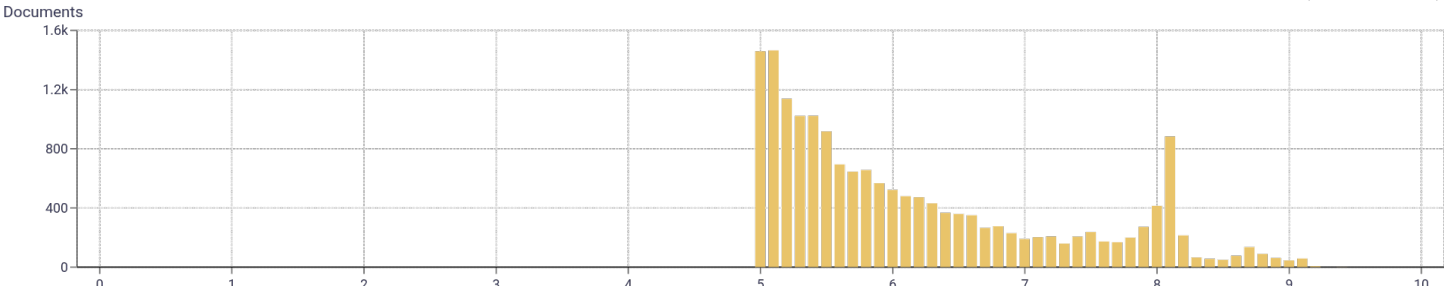
Percentage of segments in Moroccan Arabic (ary) inside documents

segments < 50% - **89.97%** (16K documents)
segments ≥ 50% - **10.03%** (1.8K documents)



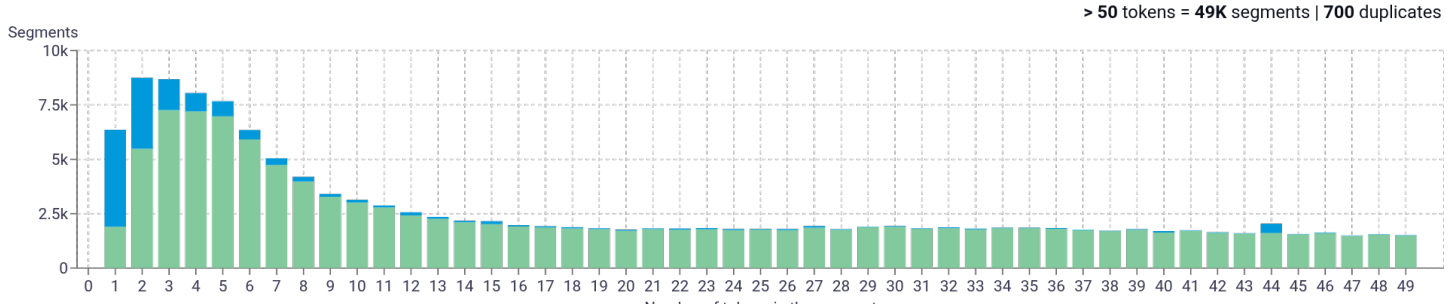
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (18K documents)

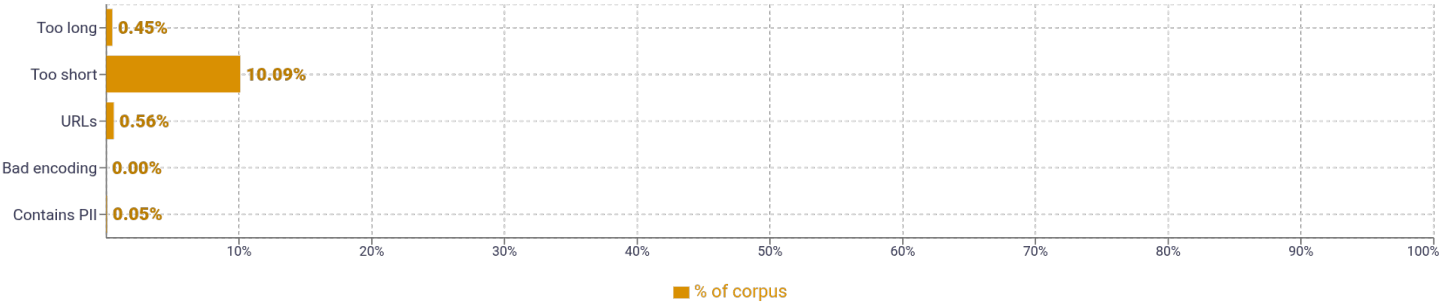


Segment length distribution by token

≤ **49** tokens = **136K** segments | **14K** duplicates
> **50** tokens = **49K** segments | **700** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	يمكن | 29,746 بشكل | 19,211 أيض | 18,966 يتم | 17,433 خلال | 14,402
2	سبيل المثال | 3,143 الذكاء الاصطناعي | 3,031 عبر الإنترنت | 2,947 بشكل عام | 2,266 بشكل كبير | 2,036
3	يمكن أن يكون | 1,990 يمكن أن تكون | 1,196 يمكن أن يؤدي | 914 يجب أن يكون | 878 عندما يتعلق الأمر | 748
4	رؤية الثعابين في الحلم | 421 والخصوم الذين قد يحاولون | 420 قتل الثعبان الأول يشير | 420 فتعني أن اقرأ المزيد | 420 الحلم إلى الأعداء والخصوم | 420
5	يشير إلى قدرتك على التغلب | 420 يرمز رؤية الثعابين في الحلم | 420 والخصوم الذين قد يحاولون إيذاءك | 420 الثعبان الأول يشير إلى قدرتك | 420 الثعابين في الحلم إلى الأعداء | 420

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				