

General overview

Corpus	Date	Language
hplt-v3-arz_Arab	10/3/2025	Egyptian Arabic (arz)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
94,125	1,122,735	877,179 (78.13 %)	21.87%	32M	175,229,253	302.43 MB

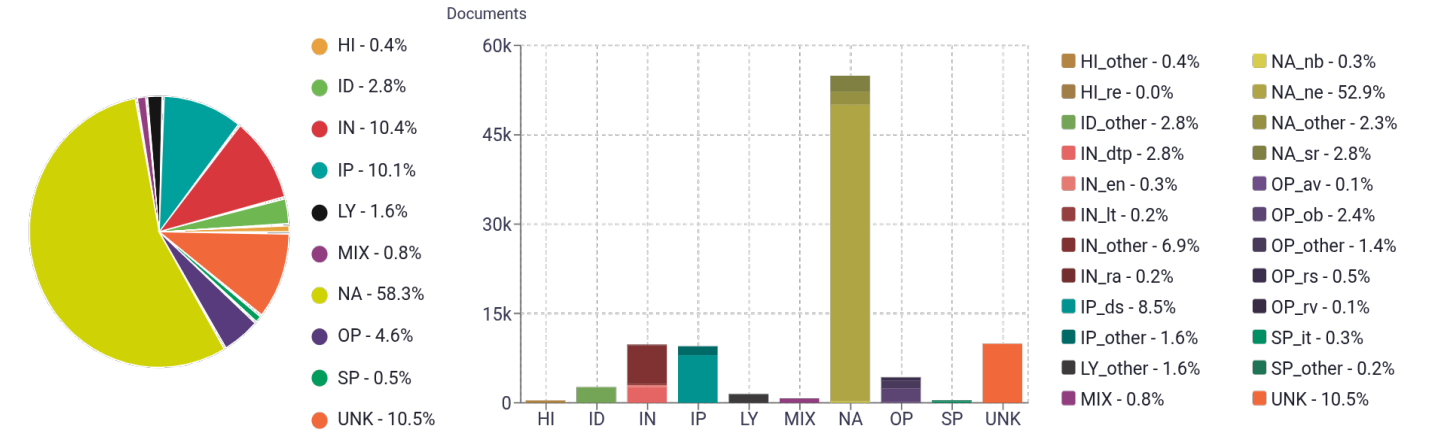
Top 10 domains

Domain	Docs	% of total
youm7.com	3.9K	4.14%
akhbarak.net	2K	2.10%
blogspot.com	1.4K	1.48%
jobs-arab.com	1.2K	1.32%
egybest.site	1.2K	1.31%
forbeseg.com	1.2K	1.28%
almaalnews.com	1.2K	1.27%
ahram.org.eg	1.2K	1.26%
egybest.pw	864	0.92%
kilma.net	822	0.87%

Top 10 TLDs

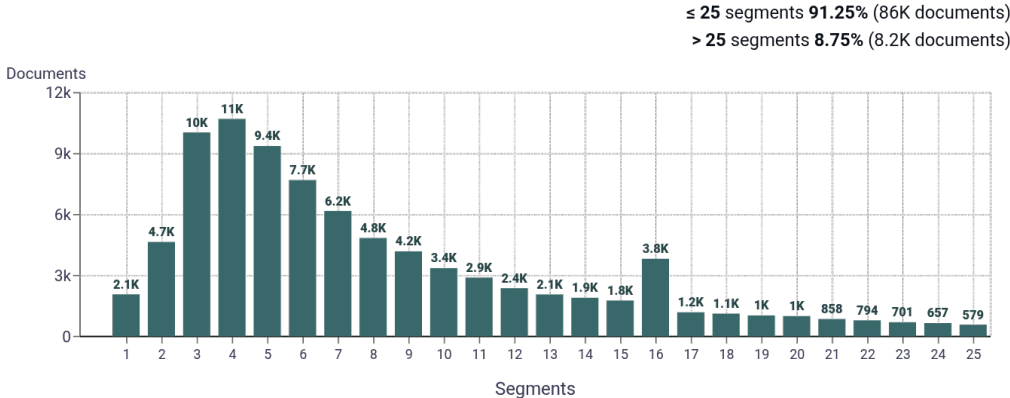
Domain	Docs	% of total
com	64K	67.46%
net	11K	11.27%
org	4.3K	4.61%
news	1.8K	1.88%
org.eg	1.3K	1.39%
site	1.3K	1.33%
edu.eg	1K	1.10%
pw	867	0.92%
com.eg	864	0.92%
co	747	0.79%

Register labels

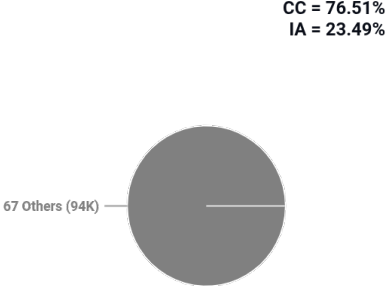


MT:2.2% | 2.1K Documents

Documents size (in segments) ⓘ

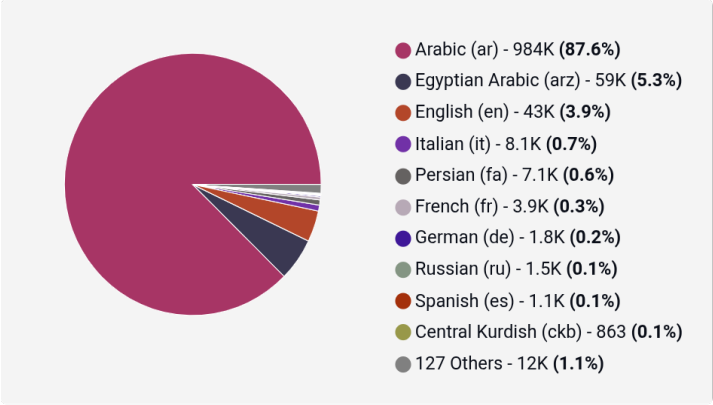


Document collections

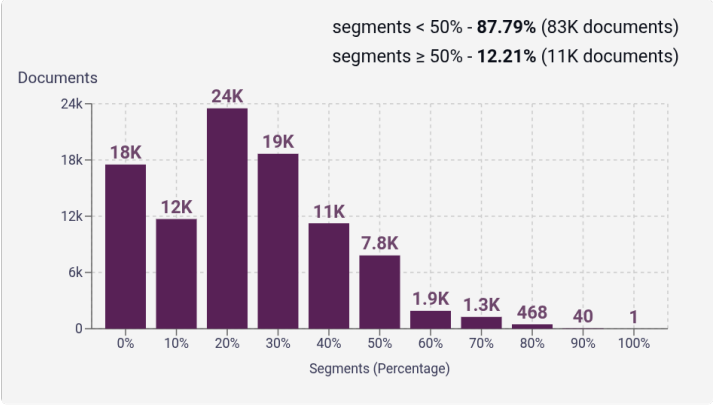


Language Distribution

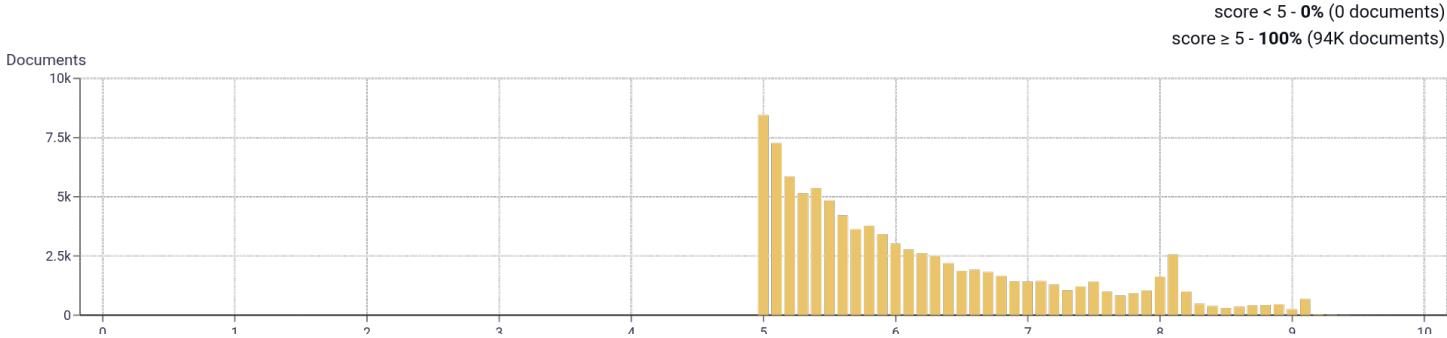
Number of segments in the Egyptian Arabic (arz) corpus



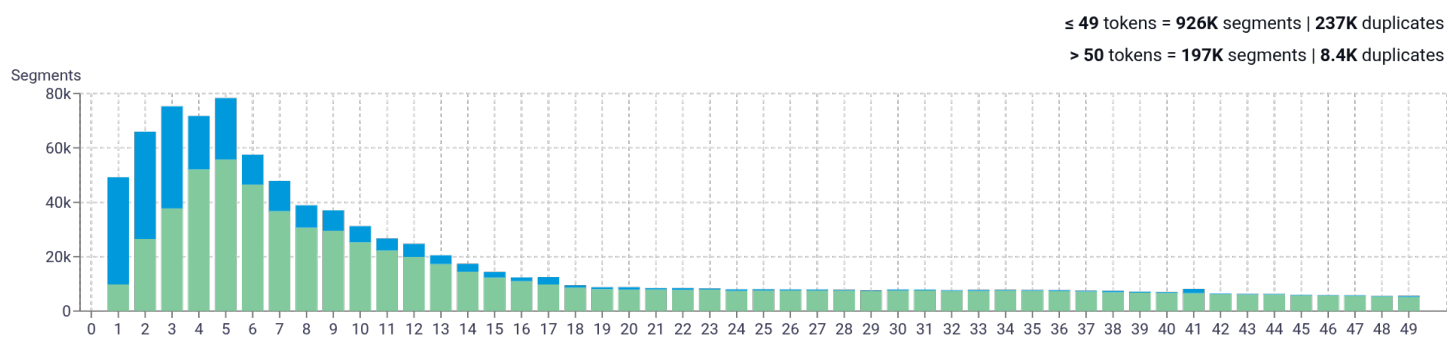
Percentage of segments in Egyptian Arabic (arz) inside documents



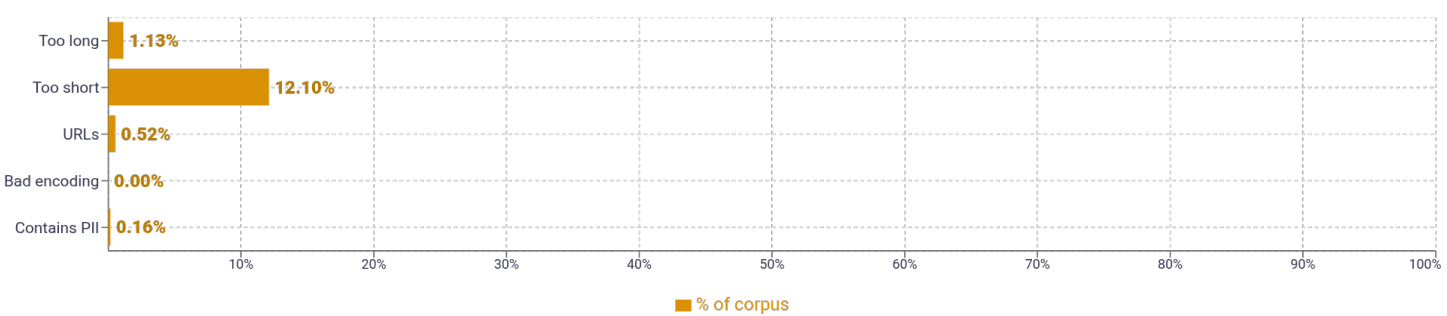
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	705,516 فى 133,234 مصر 125,441 التى 78,118 خلال 56,124 المصرية	
2	31,016 فى مصر 13,650 فى مجال 8,863 أكمل القراءة 8,099 الولايات المتحدة 7,741 الفوركس فى	
3	7,722 وسطاء الفوركس فى 3,456 الرئيس عبد الفتاح 3,350 عبد الفتاح السيسي 2,731 رئيس مجلس إدارة 2,662 وزارة التربية والتعليم	
4	2,518 الرئيس عبد الفتاح السيسي 2,057 أضافه تعليقك سوف يظهر 1,842 الولايات المتحدة الأمريكية المملكة المتحدة مصرفرنساالهندكندااليابانكوريا 1,161 توخى الحذر من اعلانات 1,161 الحذر من اعلانات النصب	
5	2,057 تم أضافه تعليقك سوف يظهر 2,057 تعليقك سوف يظهر بعد المراجعة 1,161 توخى الحذر من اعلانات النصب 1,161 برجاء توخى الحذر من اعلانات 1,161 الحذر من اعلانات النصب والاحتيال	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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				