

General overview

Corpus	Date	Language
hplt-v3-aeb_Arab	10/3/2025	Tunisian Arabic (aeb)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
177	2,462	1,993 (80.95 %)	19.05%	33K	175,248	306.79 KB

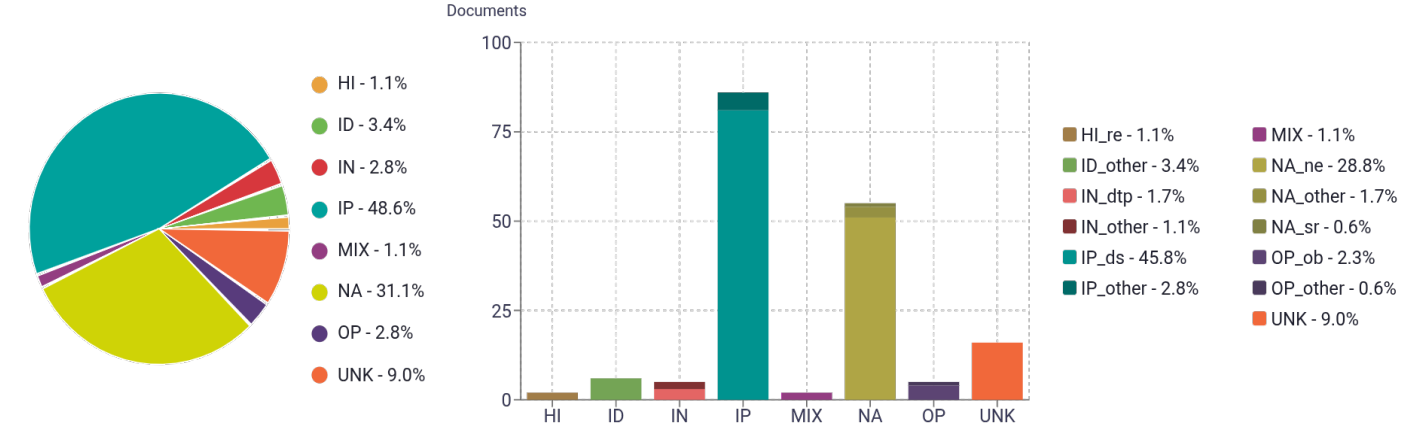
Top 10 domains

Domain	Docs	% of total
q3sk.online	18	10.17%
esheek.cam	16	9.04%
shahline.net	9	5.08%
blogspot.com	4	2.26%
tunmix.com	3	1.69%
tunisiaface.net	3	1.69%
tadwinet.net	3	1.69%
journaltunisie.net	3	1.69%
elriyadh.news	3	1.69%
aratk.com	3	1.69%

Top 10 TLDs

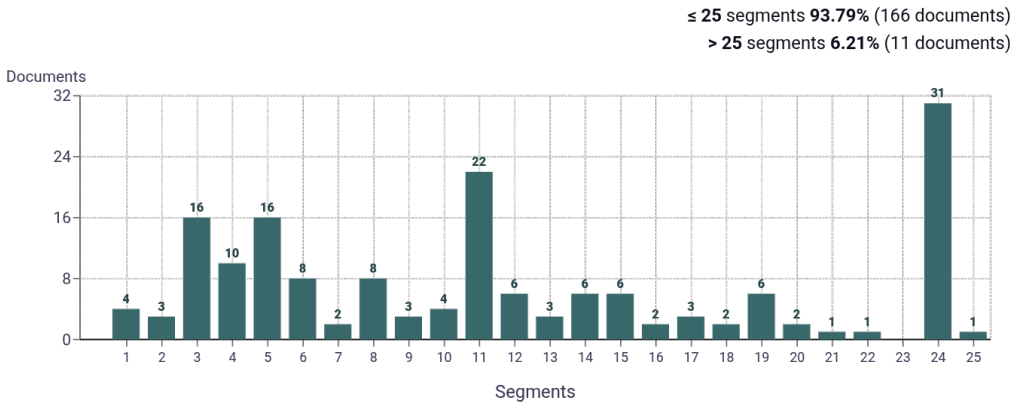
Domain	Docs	% of total
com	72	40.68%
net	35	19.77%
online	19	10.73%
cam	16	9.04%
tn	9	5.08%
news	3	1.69%
info	3	1.69%
org	2	1.13%
me	2	1.13%
live	2	1.13%

Register labels

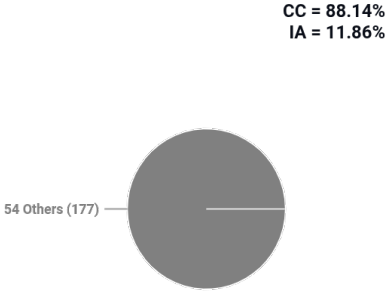


MT:1.1% | 2 Documents

Documents size (in segments) ⓘ

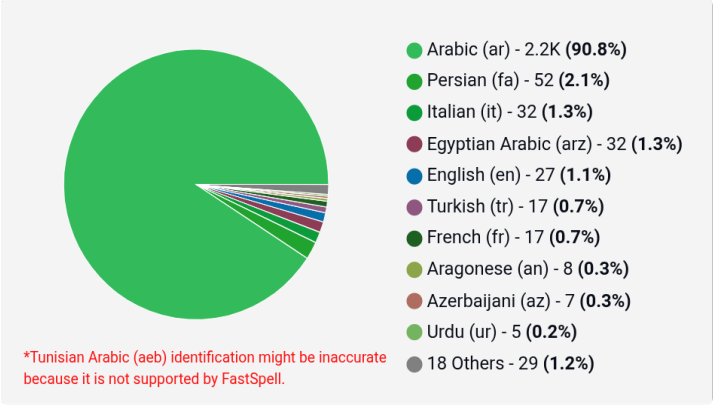


Document collections

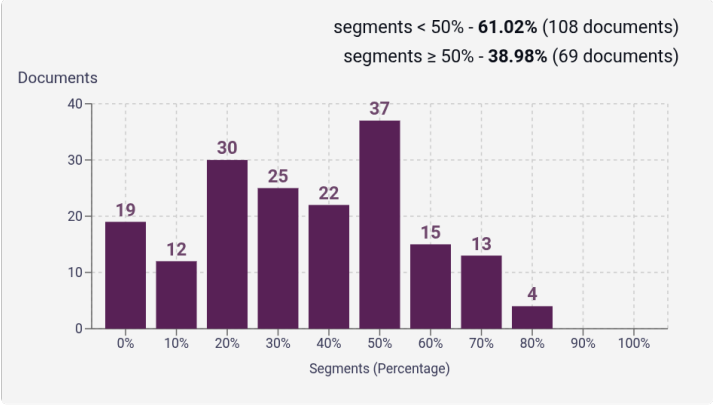


Language Distribution

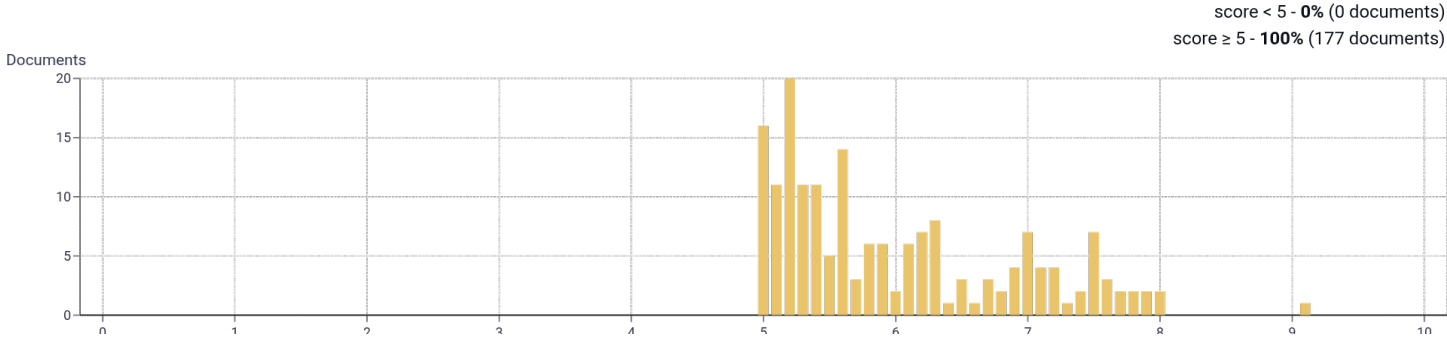
Number of segments in the Tunisian Arabic (aeb) corpus



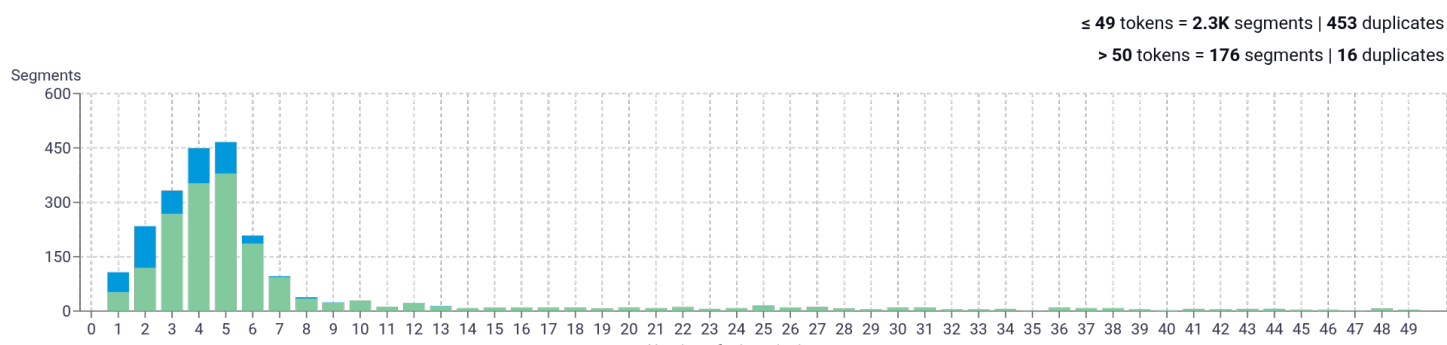
Percentage of segments in Tunisian Arabic (aeb) inside documents



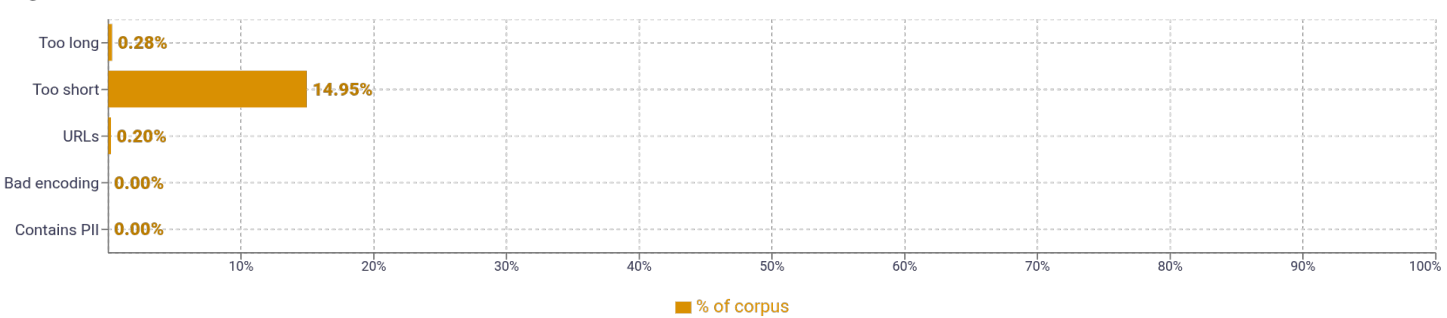
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	مسلسل 859 الحلقة 641 صيانة 358 بوش 339 قناة 178	
2	صيانة كيربازي 116 صيانة بوش 87 صيانة سيلتال 66 ورقة ورقة 65 رقم صيانة 59	
3	ورقة ورقة ورقة 63 قناة الحوار التونسي 40 موقع قصة عشق 35 رقم صيانة بوش 32 مشاهدة وتحميل مسلسل 29	
4	ورقة ورقة ورقة ورقة 61 وتدور قصة المسلسل حول 27 صيانة بوش بكفر الشيخ 23 رقم صيانة بوش بكفر 23 رقم صيانة نلاجات بوش 20	
5	ورقة ورقة ورقة ورقة ورقة 59 رقم صيانة بوش بكفر الشيخ 23 وعلى موقع قصة عشق بالترجمة 16 موقع قصة عشق بالترجمة العربية 16 مشاهدة مسلسل زهرة النالوت الحلقة 15	

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	dt
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				