

General overview

Corpus	Date	Language
hplt-v3-zgh_Tfng	10/3/2025	Standard Moroccan Tamazight (zgh)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
3,488	34,984	30,507 (87.20 %)	12.80%	1.4M	6,522,944	15.71 MB

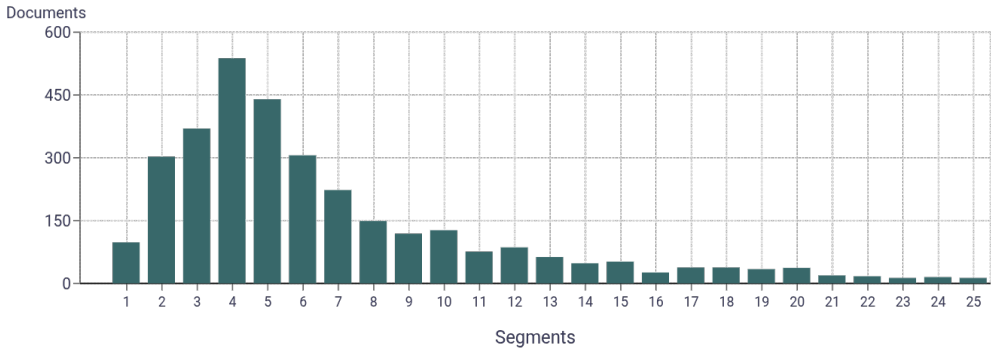
Top 10 domains

Domain	Docs	% of total
aps.dz	644	18.46%
mapamazighe.ma	611	17.52%
wikipedia.org	370	10.61%
wikimedia.org	346	9.92%
cg.gov.ma	286	8.20%
mapnews.ma	195	5.59%
amadalamazigh.p...	129	3.70%
haca.ma	121	3.47%
mjcc.gov.ma	78	2.24%
minculture.gov.ma	72	2.06%

Top 10 TLDs

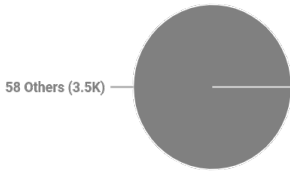
Domain	Docs	% of total
ma	1.3K	36.87%
org	749	21.47%
dz	651	18.66%
gov.ma	462	13.25%
press.ma	129	3.70%
com	126	3.61%
press	40	1.15%
net	26	0.75%
de	9	0.26%
edu.ly	2	0.06%

Documents size (in segments) ⓘ



≤ 25 segments **93.12%** (3.2K documents)
> 25 segments **6.88%** (240 documents)

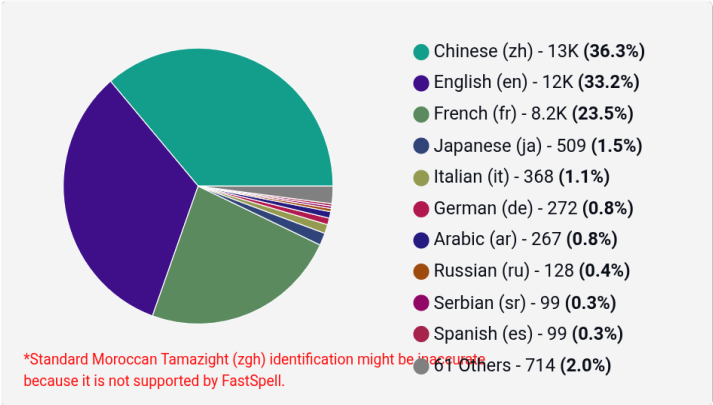
Document collections



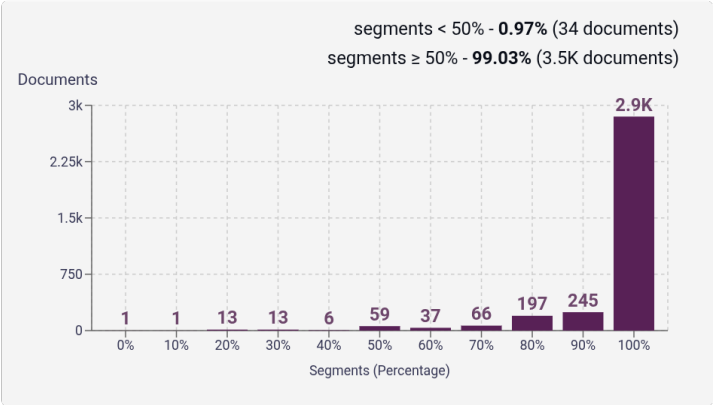
CC = **96.82%**
IA = **3.18%**

Language Distribution

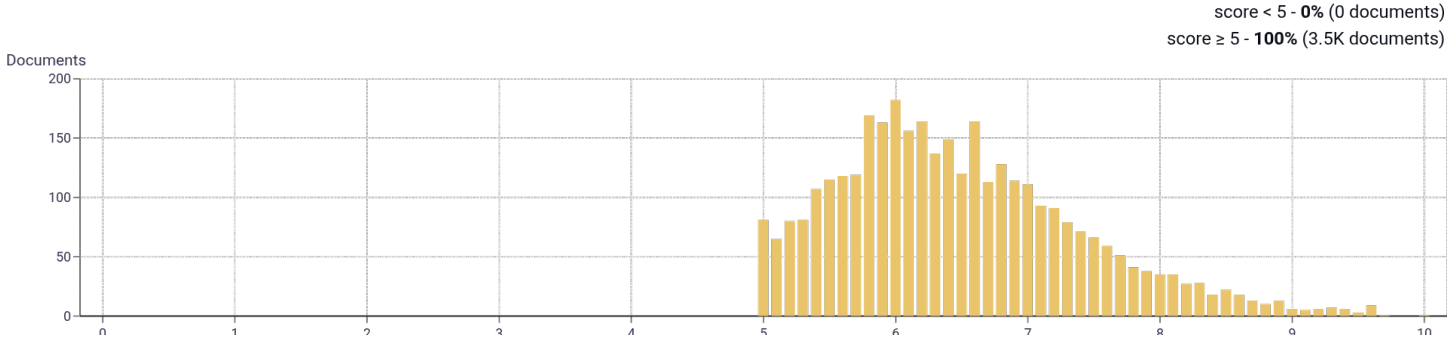
Number of segments in the Standard Moroccan Tamazight (zgh) corpus



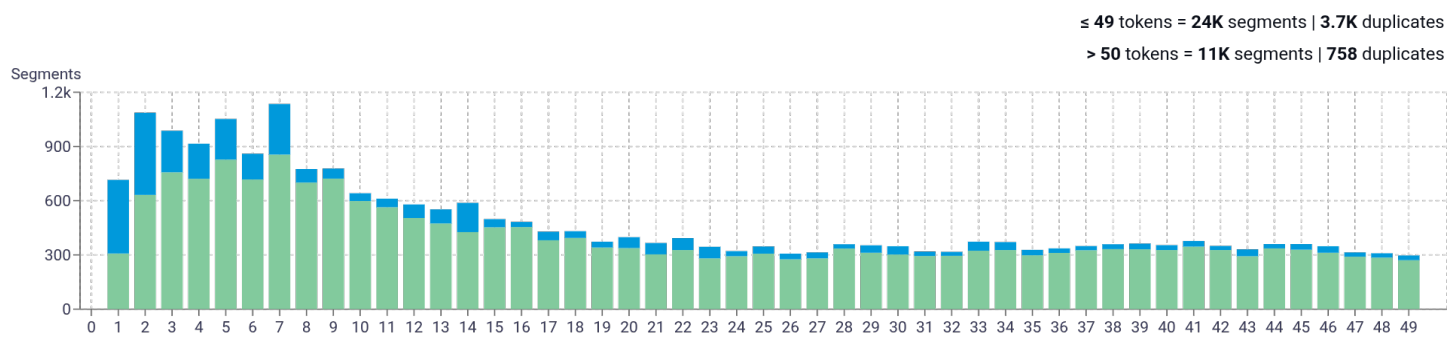
Percentage of segments in Standard Moroccan Tamazight (zgh) inside documents



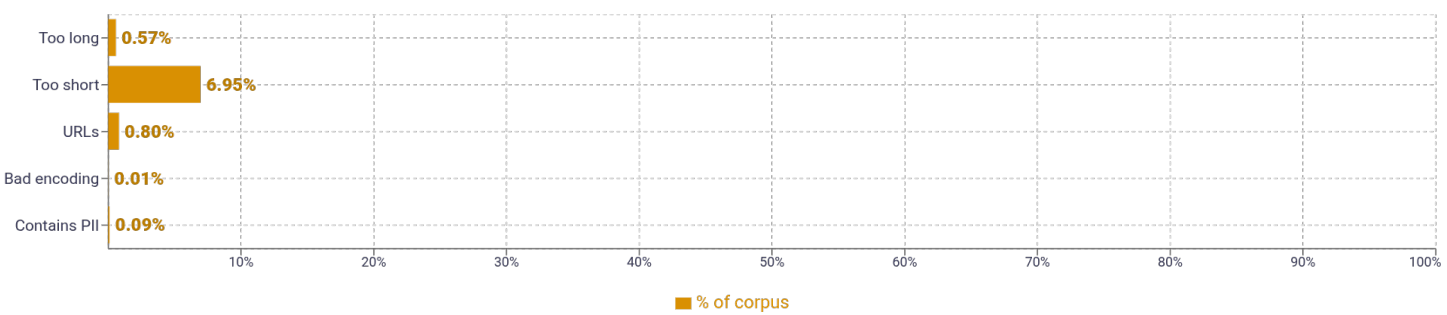
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>ΠΟ 9,174</div> <div>ΠΙΣ 6,091</div> <div>ΧΗ 6,033</div> <div>Πο 6,028</div> <div>οο 5,377</div>	
2	<div>Πο/ΛΣΟ οΧΙΠΙΣΛ 1,032</div> <div>ΣΕΥοΠοΕ οΘΙΣΨQ 848</div> <div>Πο/ΚΣΙ οΥ 765</div> <div>ΣΧΟοΠ οΕο++οΥ 708</div> <div>ΠΣΟΘ ΘΕΣΘ 648</div>	
3	<div>ΘοΘ Ι Πο/ΛΣΟ 1,467</div> <div>οΕο++οΥ Ι ΣΕΥοΠοΕ 784</div> <div>ΣΨΟΗοΙ Ι ΣΗΧοΙ 683</div> <div>ΣΙΘΟΣΧΗ Ι +ΘοΕ+ 641</div> <div>ΥΨΗ Πο/ΚΣΙ οΥ 519</div>	
4	<div>ΘοΘ Ι Πο/ΛΣΟ οΧΙΠΙΣΛ 1,010</div> <div>οΕο++οΥ Ι ΣΕΥοΠοΕ οΘΙΣΨQ 717</div> <div>ΣΧΟοΠ οΕο++οΥ Ι ΣΕΥοΠοΕ 630</div> <div>οΧΙΠΙΣΛ ΕΣ/ΕΕΛ ΠΣΟΘ ΘΕΣΘ 331</div> <div>Πο/ΛΣΟ οΧΙΠΙΣΛ ΕΣ/ΕΕΛ ΠΣΟΘ 328</div>	
5	<div>ΣΧΟοΠ οΕο++οΥ Ι ΣΕΥοΠοΕ οΘΙΣΨQ 578</div> <div>ΘοΘ Ι Πο/ΛΣΟ οΧΙΠΙΣΛ ΕΣ/ΕΕΛ 338</div> <div>Πο/ΛΣΟ οΧΙΠΙΣΛ ΕΣ/ΕΕΛ ΠΣΟΘ ΘΕΣΘ 309</div> <div>ΣΟΣΙοΧ οΧΙΠΙοΙ Ι +ΣΟΘΙο +οΕοΨΣΥ+ 225</div> <div>οοΕΣΟ Ι ΣΨΟΗοΙ Ι ΣΗΧοΙ 205</div>	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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	ntp
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				