HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-zgh_Tfng	9/18/2025	Standard Moroccan Tamazight (zgh)

Volumes

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

Top 10 domains

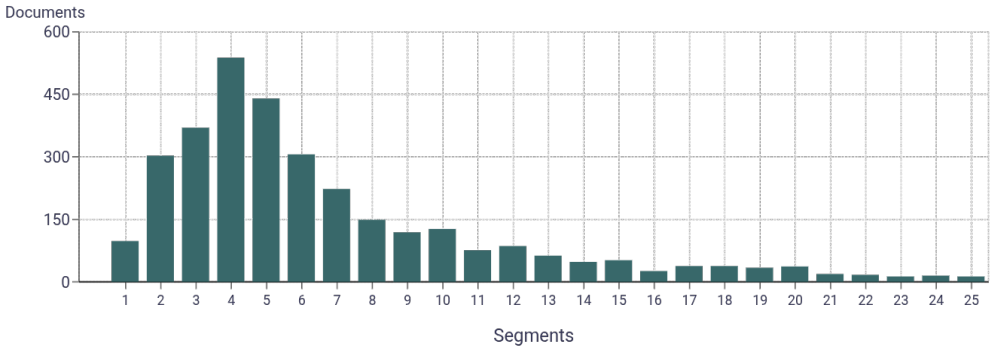
Domain	Docs	% of total
aps.dz	644	18.46%
mapamazighe.ma	611	17.52%
wikipedia.org	370	10.61%
wikimedia.org	346	9.92%
cg.gov.ma	286	8.20%
mapnews.ma	195	5.59%
amadalamazigh.p...	129	3.70%
haca.ma	121	3.47%
mjcc.gov.ma	78	2.24%
minculture.gov.ma	72	2.06%

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%

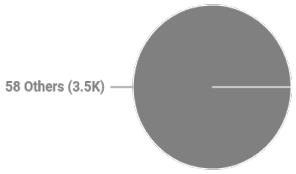
Documents size (in segments)

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



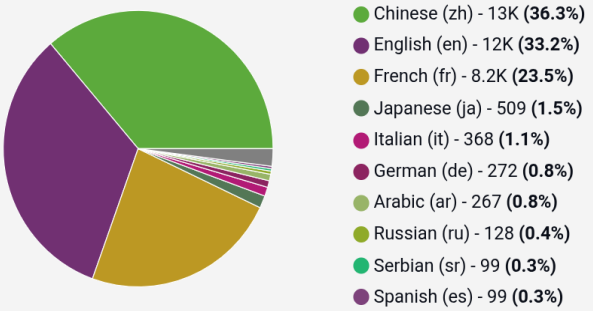
Document collections

CC = 96.82%
IA = 3.18%



Language Distribution

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

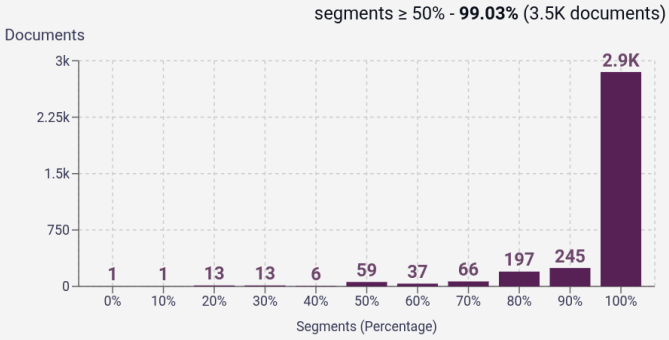


- Chinese (zh) - 13K **(36.3%)**
- English (en) - 12K **(33.2%)**
- French (fr) - 8.2K **(23.5%)**
- Japanese (ja) - 509 **(1.5%)**
- Italian (it) - 368 **(1.1%)**
- German (de) - 272 **(0.8%)**
- Arabic (ar) - 267 **(0.8%)**
- Russian (ru) - 128 **(0.4%)**
- Serbian (sr) - 99 **(0.3%)**
- Spanish (es) - 99 **(0.3%)**
- 61 Others - 714 **(2.0%)**

*Standard Moroccan Tamazight (zgh) identification might be inaccurate as it is not supported by FastSpell

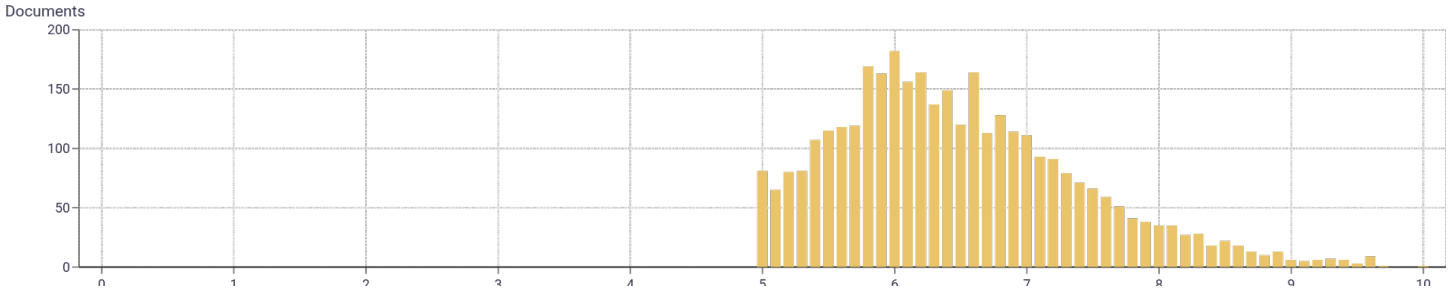
Percentage of segments in Standard Moroccan Tamazight (zgh) inside documents

segments < 50% - **0.97%** (34 documents)
segments ≥ 50% - **99.03%** (3.5K documents)



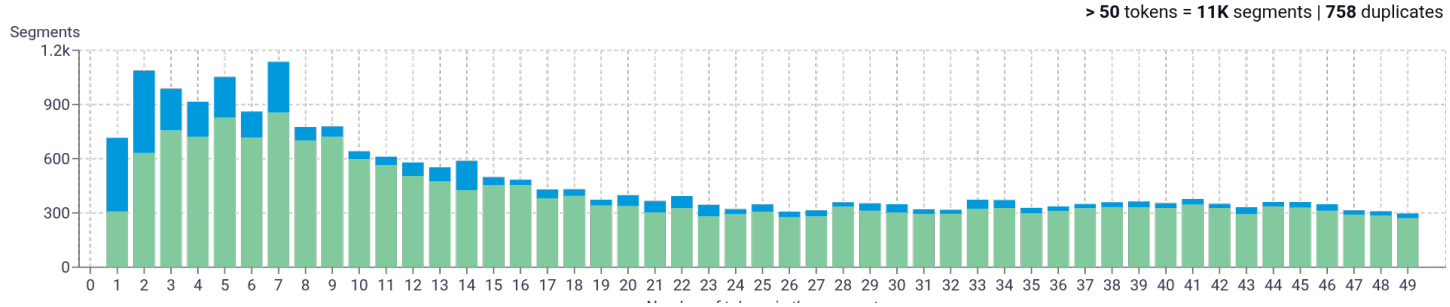
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (3.5K documents)

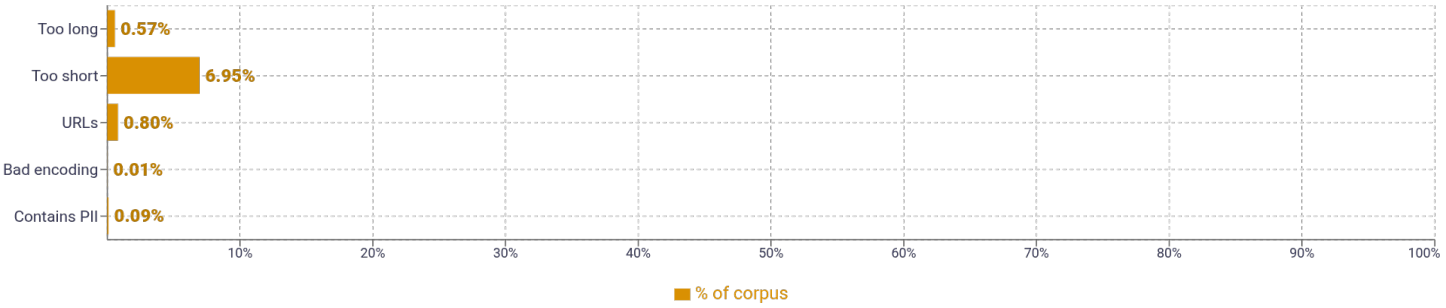


Segment length distribution by token

≤ 49 tokens = **24K** segments | **3.7K** duplicates
> 50 tokens = **11K** segments | **758** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ᐃ | 148,854 · ᐱ | 66,345 · ᕆ | 32,325 · ᓂᐱ | 24,124 · ᑐ | 22,711
2	ᓂᑐᑐ ᐃ | 2,375 · ᐅᓂᐅ ᐃ | 2,192 · ᓂᔾ ᐱ | 1,879 · ᐃ ᐸᐃᐅᓂᐦ | 1,768 · ᐃ ᒪᓂᐱᐱᔨᑐ | 1,551
3	ᐅᓂᐅ ᐃ ᒪᓂᐱᐱᔨᑐ | 1,467 · ᐃ ᒪᓂᐱᐱᔨᑐ ᓂᕿᐃᐃᔪᐱ | 1,028 · ᐃ ᔮᑕᔾᓂᒪᓂᕮ ᓂᑐᐃᔾᕐᑫ | 839 · ᓂᒐᓲᓲᓂᔾ ᐃ ᔮᑕᔾᓂᒪᓂᕮ | 784 · ᔪᔭᑐᐦᓂᒪ ᐃ ᔮᐦᕆᑐ | 683
4	ᐅᓂᐅ ᐃ ᒪᓂᐱᐱᔨᑐ ᓂᕿᐃᐃᔪᐱ | 1,010 · ᓂᒐᓲᓲᓂᔾ ᐃ ᔮᑕᔾᓂᒪᓂᕮ ᓂᑐᐃᔾᕐᑫ | 717 · ᔮᕿᑐᓂᒪ ᓂᒐᓲᓲᓂᔾ ᐃ ᔮᑕᔾᓂᒪᓂᕮ | 630 · ᐃ ᔪᔭᑐᐦᓂᒪ ᐃ ᔮᐦᕆᑐ | 548 · ᖁᔮᐦ ᒪᓂᖅᖅᔮᐃ ᓂᔾ ᐱ | 499
5	ᔮᕿᑐᓂᒪ ᓂᒐᓲᓲᓂᔾ ᐃ ᔮᑕᔾᓂᒪᓂᕮ ᓂᑐᐃᔾᕐᑫ | 578 · ᐃ ᔮᕿᑐᓂᒪ ᓂᒐᓲᓲᓂᔾ ᐃ ᔮᑕᔾᓂᒪᓂᕮ | 347 · ᐅᓂᐅ ᐃ ᒪᓂᐱᐱᔨᑐ ᓂᕿᐃᐃᔪᐱ ᑕᔮᐸᑕᑕᐱ | 338 · ᐃ ᒪᓂᐱᐱᔨᑐ ᓂᕿᐃᐃᔪᐱ ᑕᔮᐸᑕᑕᐱ ᒪᐃᑐᑐ | 325 · ᒪᓂᐱᐱᔨᑐ ᓂᕿᐃᐃᔪᐱ ᑕᔮᐸᑕᑕᐱ ᒪᐃᑐᑐ ᑐᕮᐃᑐ | 309

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				