HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-taq_Tfng	9/18/2025	Tamasheq

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5	171	171 (100.00 %)	5K	26,633	65.85 KB

Top 10 domains

Domain	Docs	% of total
maghreb-dirassa...	2	40.00%
translatewiki.net	1	20.00%
fmsn.gov.ma	1	20.00%
cooplus.org	1	20.00%

Top 10 TLDs

Domain	Docs	% of total
com	2	40.00%
org	1	20.00%
net	1	20.00%
gov.ma	1	20.00%

Documents size (in segments) ⓘ
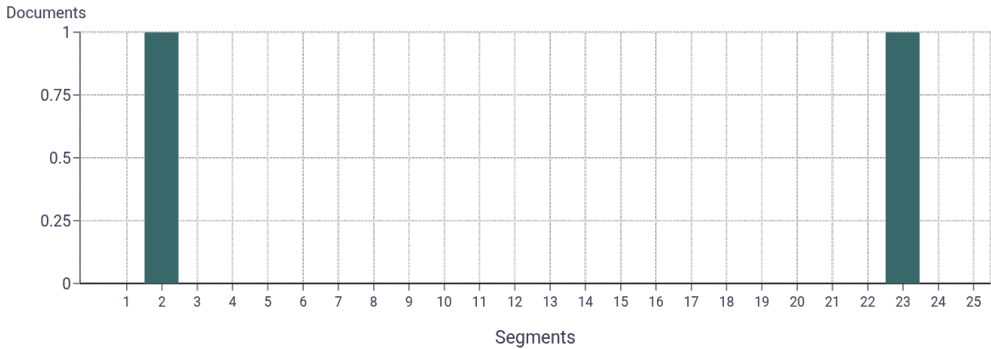
≤ 25 segments **40%** (2 documents)
> 25 segments **60%** (3 documents)

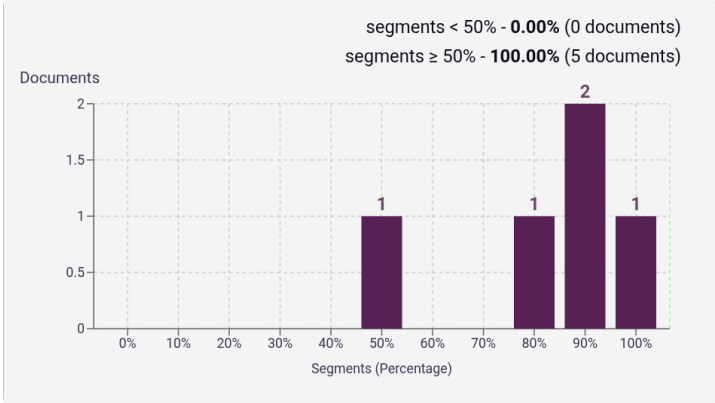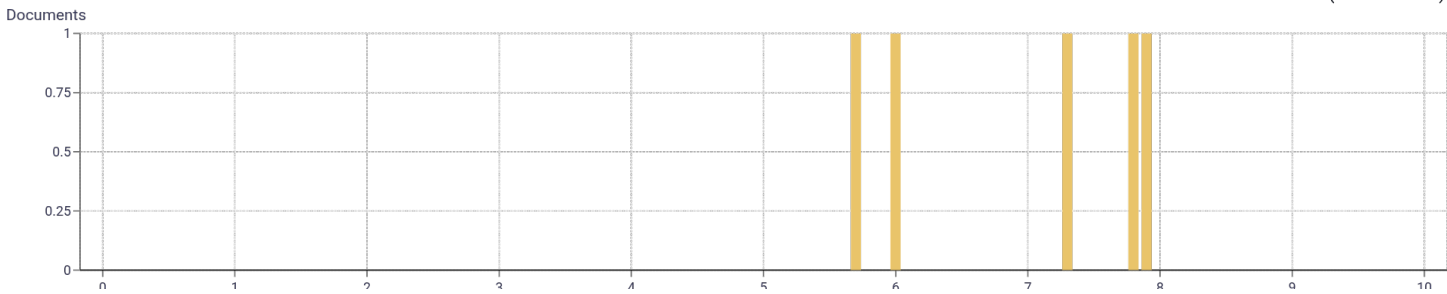## Document collections

CC = 100.00%
IA = 0.00%

- CC-MAIN-2021-10 (2)
- C-MAIN-2024-51 (1)
- CC-MAIN-2019
- CC-MAIN-2021-17 (1)
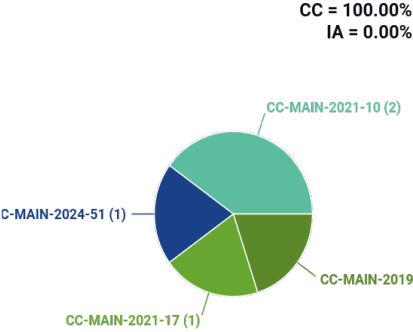
Language Distribution

Number of segments in the Tamasheq corpus

- Chinese - 105 **(61.4%)**
- French - 18 **(10.5%)**
- Japanese - 15 **(8.8%)**
- English - 14 **(8.2%)**
- Arabic - 14 **(8.2%)**
- Turkish - 1 **(0.6%)**
- Russian - 1 **(0.6%)**
- Portuguese - 1 **(0.6%)**
- Korean - 1 **(0.6%)**
- German - 1 **(0.6%)**

*Tamasheq identification might be inaccurate because it is not supported by FastSpell.
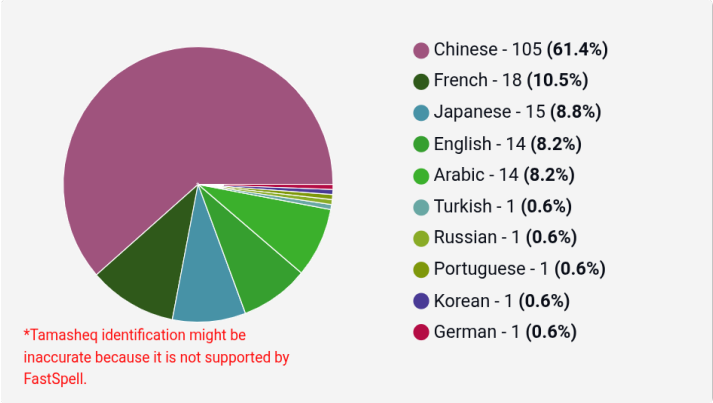
Percentage of segments in Tamasheq inside documents

segments < 50% - **0.00%** (0 documents)
segments ≥ 50% - **100.00%** (5 documents)

Distribution of documents by document score

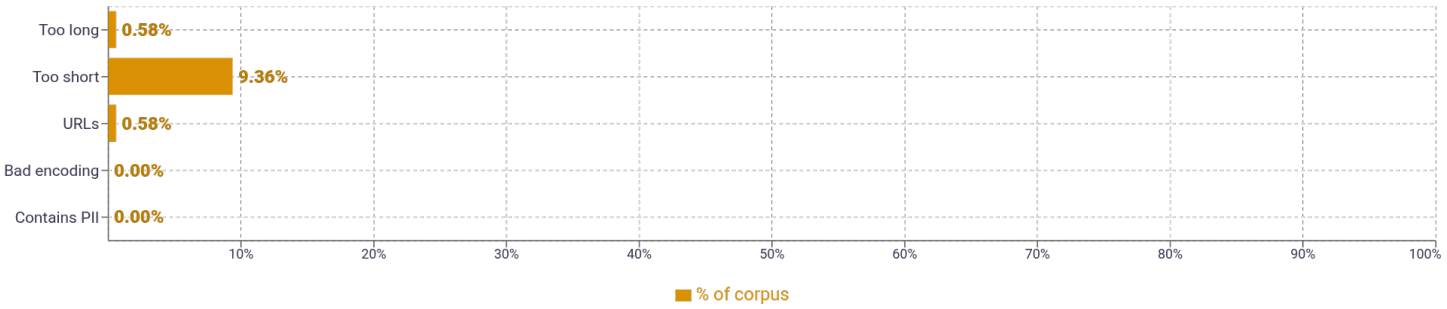score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (5 documents)

Segment length distribution by token

≤ 49 tokens = **142** segments | **0** duplicates
> 50 tokens = **29** segments | **0** duplicates

Segments

Segment noise distribution

Too long	**0.58%**
Too short	**9.36%**
URLs	**0.58%**
Bad encoding	**0.00%**
Contains PII	**0.00%**

■ % of corpus
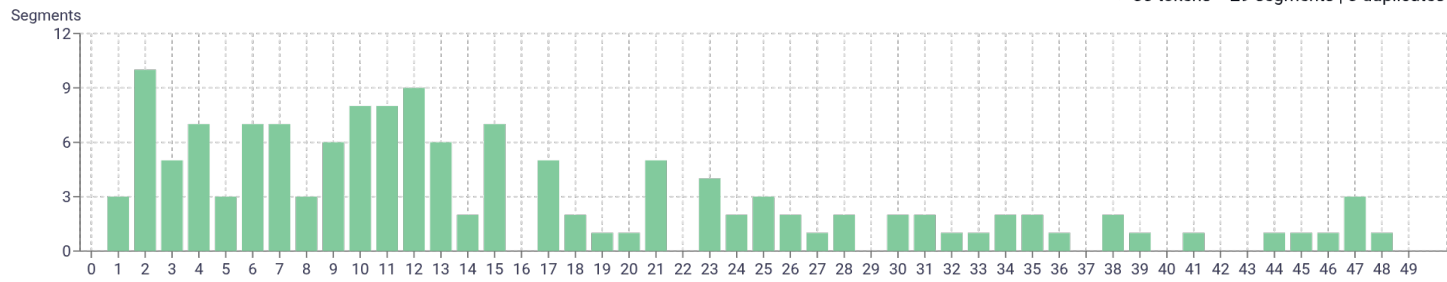
Frequent n-grams

SIZE	N-GRAMS	
1	+Φ⧧ | 157 +⧧ | 94 ₀I∧ | 74 ₀ | 74 ⧧Ħ | 71	⧉
2	⌐₀ϟ Θ⧧ | 22 ΣΘ ₀∧∧Ο⧧ΟΟ | 17 ⧧Ħ +Φ⧧ | 16 ΣΘ +Φ⧧ | 12 +⧧ +Φ⧧ | 11	⧉
3	Σ+ ⌐₀ϟ Θ⧧ | 5 Θ⧧ΟΟ⧧I₀ИИϟ Σ∧⧧I+ΣĦΣ₀ΘИ⧧ ΣIĦ⧧Ο⌐₀+Σ⧧I | 5 ⊔Σ∧И I⧧+ Θ⧧ | 4 ⊔Σ∧И Φ⧧Иϟ ϟ₀⧧ | 4 +Φ⧧ ₀ĦĦ⧧⌐+⧧∧ ⧧Ο⧧Ο | 4	⧉
4	⊔Σ∧И Φ⧧Иϟ ϟ₀⧧ +₀ | 3 Θ⧧Θ⧧⧧I₀ ⧧Ο ⧧+Φ⧧Ο ⌐⧧⌐Θ⧧И₀ΘΟϟ | 3 ⧧Ħ Θ⧧ΟΟ⧧I₀ИИϟ Σ∧⧧I+ΣĦΣ₀ΘИ⧧ ΣIĦ⧧Ο⌐₀+Σ⧧I | 3 Σ+ ⌐₀ϟ Θ⧧ Θ⧧ΟΟΣΘИ⧧ | 3 ∧⧧ И₀ ΘûΟ⧧+é I₀+Σ₀I₀И⧧ | 3	⧉
5	Θ⧧Θ⧧⧧I₀ ⧧Ο ⧧+Φ⧧Ο ⌐⧧⌐Θ⧧И₀ΘΟϟ Ο⧧⌐Ζ⧧Θ₀+ | 3 ⊔Σ∧И I⧧+ Θ⧧ ₀ΘИ⧧ +⧧ | 2 ∆⧧∆∆⧧Ο ∧⧧ И₀ ΘûΟ⧧+é I₀+Σ₀I₀И⧧ | 2 +⧧ Iϟ+ΣĦϟ +Φ⧧ ₀ĦĦ⧧⌐+⧧∧ ⧧Ο⧧Ο | 2 ⌐⧧ИИ⧧⌐+⧧∧ +Φ₀I ΣΟ +ϟΘΣ⌐₀ИИϟ ⌐⧧ИИ⧧⌐+⧧∧ | 2	⧉

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				