HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-khk_Cyrl	9/23/2025	Halh Mongolian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,478,588	80,595,258	44,804,524 (55.59 %)	2.3B	13,456,001,552	22.06 GB

Top 10 domains

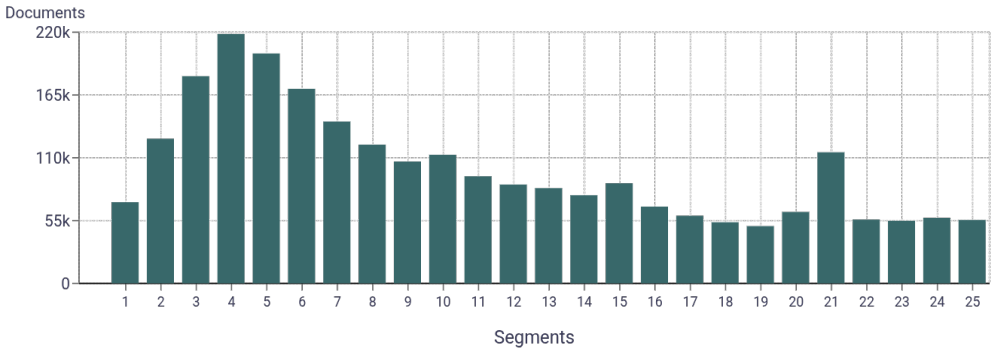
Domain	Docs	% of total
news.mn	80K	2.31%
montsame.mn	61K	1.74%
zindaa.mn	49K	1.41%
shuud.mn	39K	1.12%
eguur.mn	35K	0.99%
vip76.mn	34K	0.99%
blogspot.com	32K	0.93%
fact.mn	30K	0.87%
medee.mn	30K	0.85%
unuudur.mn	28K	0.82%

Top 10 TLDs

Domain	Docs	% of total
mn	1.8M	51.78%
com	353K	10.14%
pl	333K	9.57%
nl	149K	4.29%
gov.mn	147K	4.21%
de	93K	2.67%
eu	86K	2.47%
fr	81K	2.34%
be	73K	2.11%
org	53K	1.52%

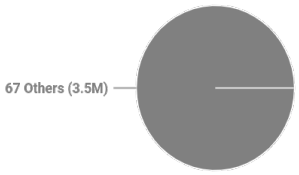
Documents size (in segments)

≤ 25 segments **72.26%** (2.5M documents)
> 25 segments **27.74%** (965K documents)



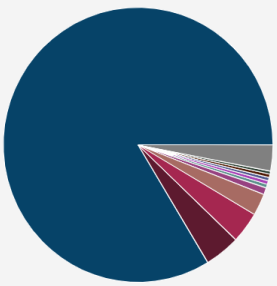
Document collections

CC = 88.87%
IA = 11.13%



Language Distribution

Number of segments in the Halh Mongolian corpus

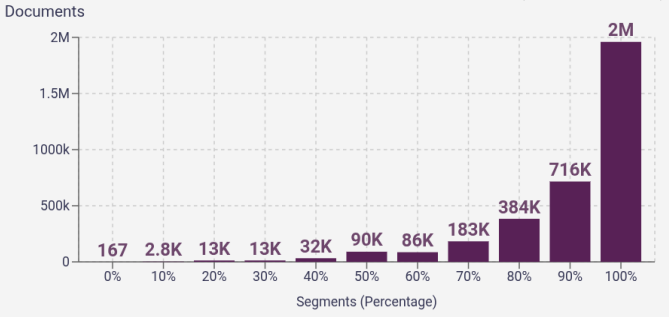


*Halh Mongolian identification might be inaccurate because it is not supported by FastSpell.

- Mongolian - 67M **(83.6%)**
- Italian - 3.4M **(4.2%)**
- English - 3M **(3.7%)**
- Russian - 2.1M **(2.6%)**
- Bulgarian - 652K **(0.8%)**
- Yakut - 348K **(0.4%)**
- Buriat - 337K **(0.4%)**
- Finnish - 317K **(0.4%)**
- Kyrgyz - 317K **(0.4%)**
- German - 286K **(0.4%)**
- 165 Others - 2.5M **(3.1%)**

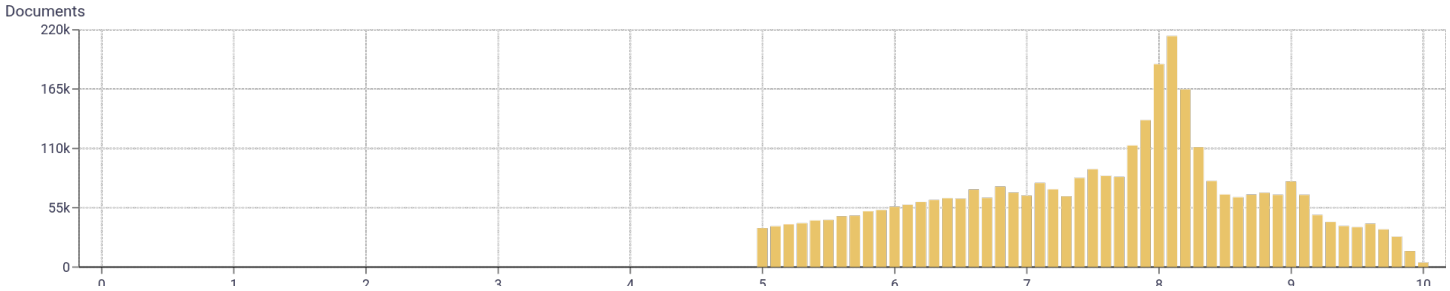
Percentage of segments in Halh Mongolian inside documents

segments < 50% - **1.74%** (61K documents)
segments ≥ 50% - **98.26%** (3.4M documents)



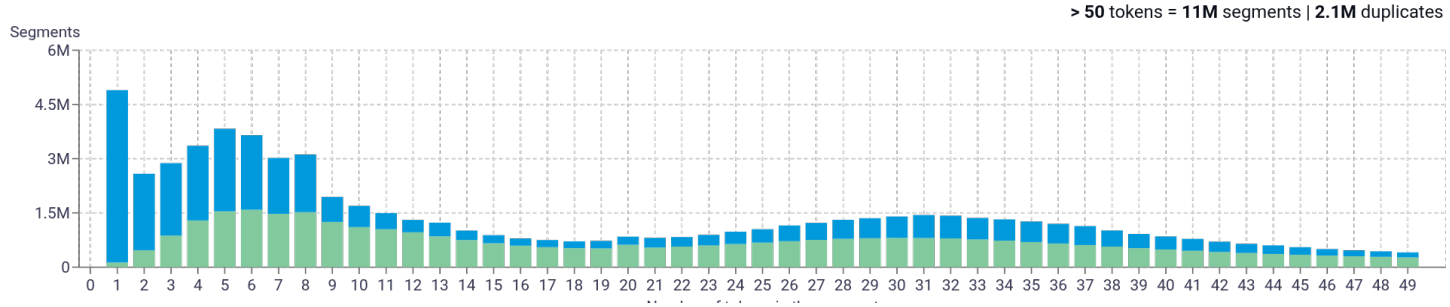
Distribution of documents by document score

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

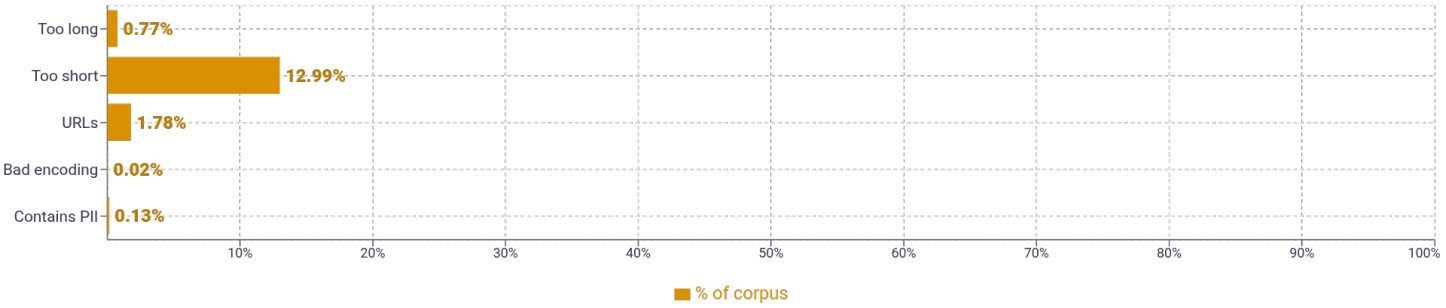


Segment length distribution by token

≤ 49 tokens = **69M** segments | **34M** duplicates
> 50 tokens = **11M** segments | **2.1M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	бутлуур | 19,613,046 · чулуу | 10,854,090 · машин | 9,429,080 · үнэ | 8,382,308 · бутлуурын | 7,988,951
2	тоног төхөөрөмж | 3,902,381 · чулуу бутлуур | 3,645,786 · уул уурхайн | 3,197,571 · хацарт бутлуур | 2,643,149 · чулуу бутлах | 1,780,469
3	уул уурхайн тоног | 738,663 · хоёр дахь гар | 681,581 · уурхайн тоног төхөөрөмж | 642,909 · чулуу бутлах машин | 508,078 · хийх элс машин | 452,984
4	уул уурхайн тоног төхөөрөмж | 526,680 · бутлуур нь шохойн чулуу | 283,175 · худалдах хоёр дахь гар | 274,029 · бутлах машин хийх элс | 230,553 · машин хийх элс машин | 224,633
5	бутлах машин хийх элс машин | 220,703 · чулуу бутлах машин хийх элс | 212,899 · машин хийх элс машин чулуу | 173,969 · хийх элс машин чулуу бутлах | 172,329 · бутлуур нь шохойн чулуу боржин | 151,097

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				