HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
khk_Cyrl.jsonl.tsv	9/26/2024	Mongolian (khk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,120,983	53,467,285	24,830,052 (46.44 %)	1.6B	9,275,953,976	11.45 GB

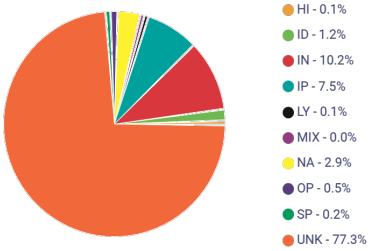
Top 10 domains

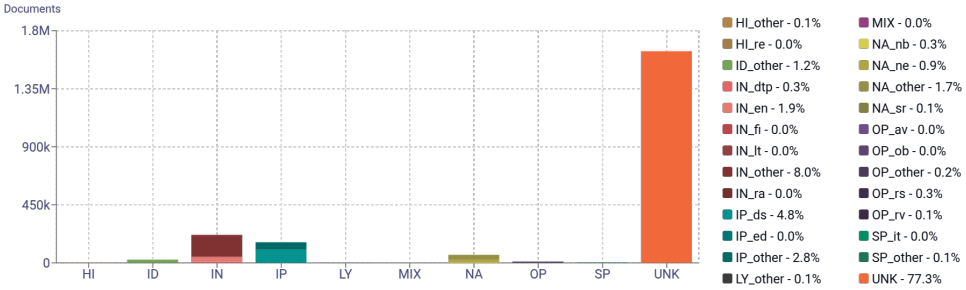
Domain	Docs	% of total
wikipedia.org	55K	2.61%
miss.mn	44K	2.06%
fact.mn	37K	1.73%
blogspot.com	36K	1.70%
olloo.mn	31K	1.48%
zindaa.mn	22K	1.02%
vip76.mn	21K	1.00%
shuud.mn	19K	0.92%
montsame.mn	19K	0.88%
ruvr.ru	18K	0.86%

Top 10 TLDs

Domain	Docs	% of total
mn	1.1M	49.73%
com	227K	10.71%
pl	166K	7.81%
nl	110K	5.18%
gov.mn	86K	4.05%
org	83K	3.93%
de	54K	2.55%
be	53K	2.51%
fr	45K	2.11%
es	33K	1.58%

Register labels

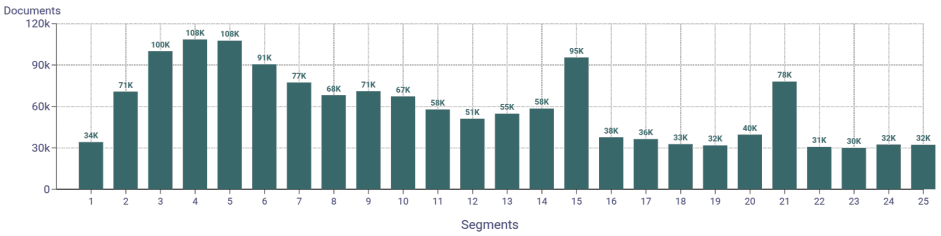




MT:32.1% | 680K Documents

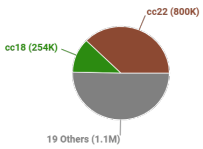
Documents size (in segments)

<= 25 segments **70.43%** (1.5M documents)
> 25 segments **29.57%** (627K documents)



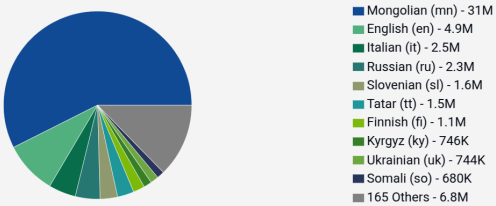
Documents by collection

CC = 64.79%
IA = 35.21%



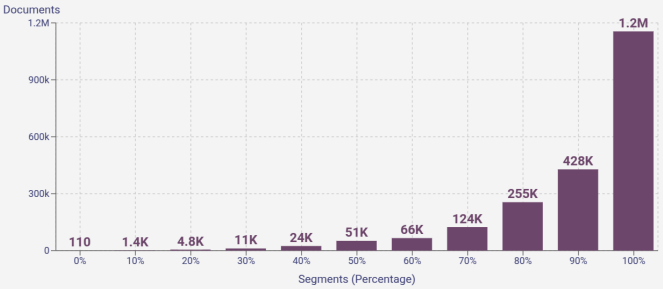
Language Distribution

Number of segments in the Mongolian (khk) corpus



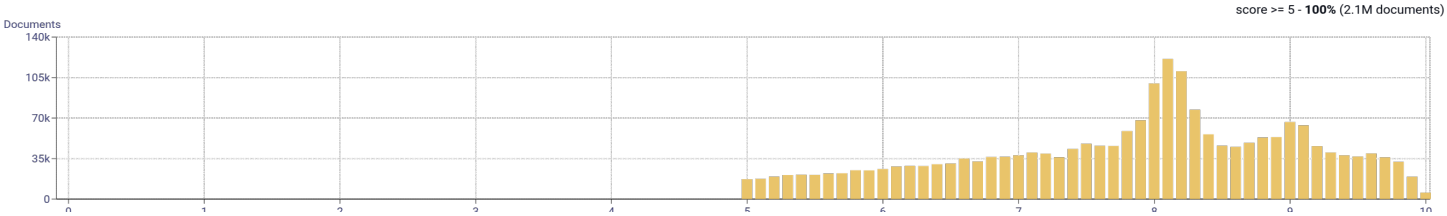
*Mongolian (khk) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Mongolian (khk) inside documents



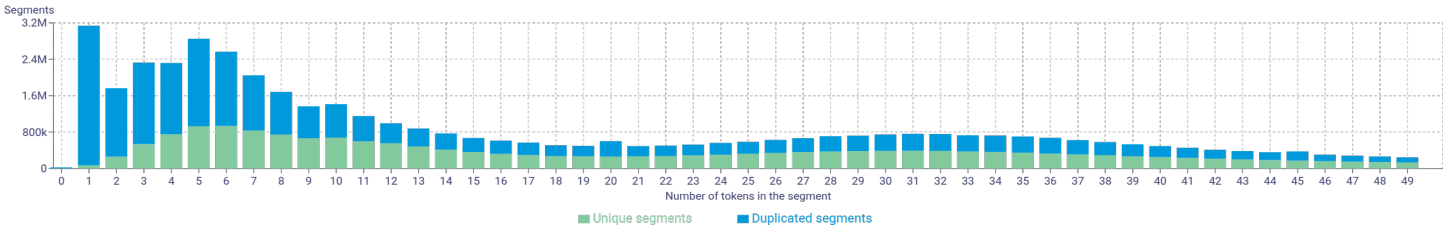
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.1M documents)

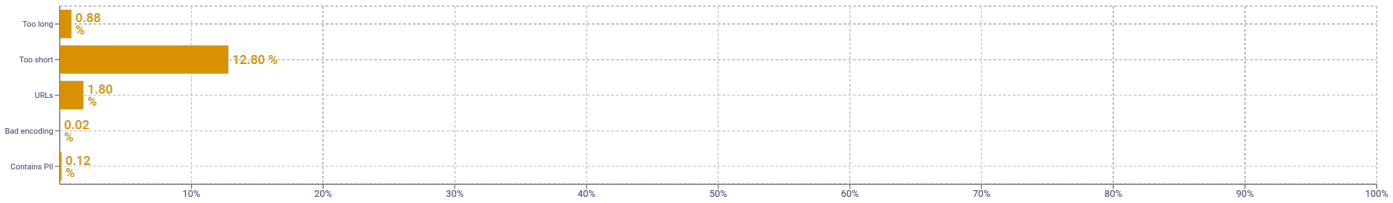


Segment length distribution by token

≤ 49 tokens = **18M** segments | **26M** duplicates
> 50 tokens = **9M** segments | **2.6M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	hb | 23447199 · бүтлүүр | 17520513 · чулуу | 9467368 · байна | 7938528 · машин | 7280864
2	чулуу бүтлүүр | 3344143 · тохор төхөөрөмж | 3061392 · хацарт бүтлүүр | 2484266 · үүл үүрхайн | 2371590 · бүтлүүр hb | 1877872
3	хоёр дахь гар | 664007 · үүл үүрхайн тохор | 578192 · тph чулуу бүтлүүр | 505749 · үүрхайн тохор төхөөрөмж | 487415 · чулуу бүтлах машин | 465725
4	үүл үүрхайн тохор төхөөрөмж | 406797 · худалдах хоёр дахь гар | 299706 · хоёр дахь гар hb | 288490 · бүтлүүр hb шохойн чулуу | 236101 · бүтлах машин хийх элс | 203175
5	худалдах хоёр дахь гар hb | 277487 · бүтлах машин хийх элс машин | 195862 · чулуу бүтлах машин хийх элс | 190785 · хуй хуй хуй хуй хуй | 167870 · хацарт бүтлүүр ре цүврал хацарт | 163694

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av