HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
isl_Latn.jsonl.tsv	9/19/2024	Icelandic (is)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,840,735	69,643,257	28,868,018 (41.45 %)	1.7B	9,526,444,446	9.8 GB

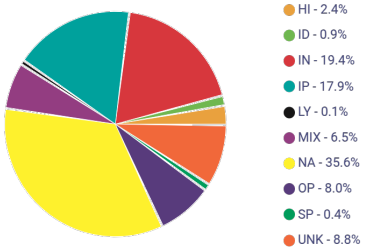
Top 10 domains

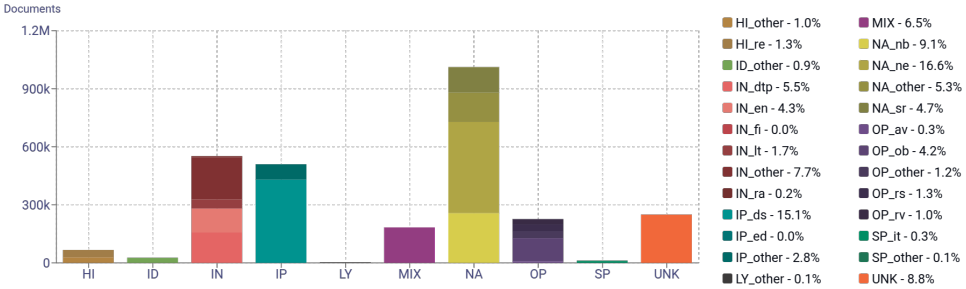
Domain	Docs	% of total
hotels.com	131K	4.62%
wikipedia.org	118K	4.14%
visir.is	71K	2.50%
mbl.is	70K	2.46%
blog.is	61K	2.14%
blogspot.com	53K	1.87%
althingi.is	32K	1.13%
ruv.is	27K	0.97%
dv.is	26K	0.93%
skessuhorn.is	26K	0.91%

Top 10 TLDs

Domain	Docs	% of total
is	2.1M	72.84%
com	478K	16.81%
org	164K	5.77%
net	56K	1.98%
eu	9.4K	0.33%
info	8.2K	0.29%
dk	3.7K	0.13%
no	3.4K	0.12%
blog	2.9K	0.10%
co.uk	2.9K	0.10%

Register labels

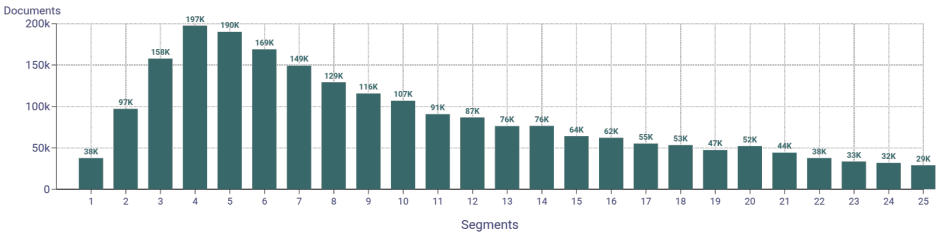




MT:8.9% | 254K Documents

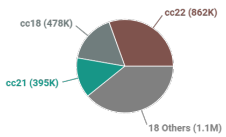
Documents size (in segments)

<= 25 segments **77.07%** (2.2M documents)
> 25 segments **22.93%** (651K documents)



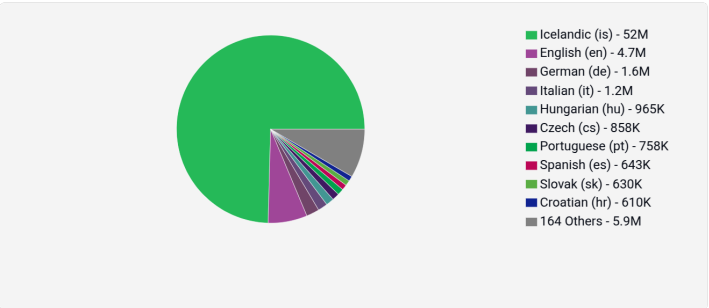
Documents by collection

CC = 72.16%
IA = 27.84%

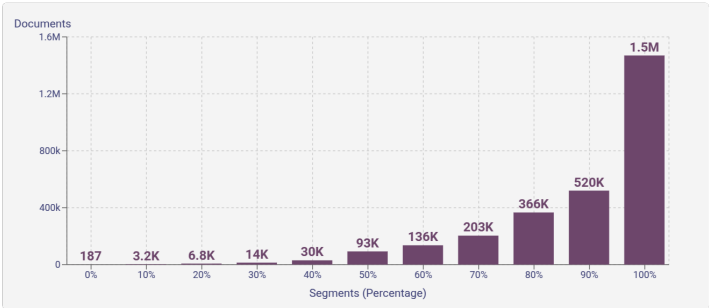


Language Distribution

Number of segments in the Icelandic (is) corpus

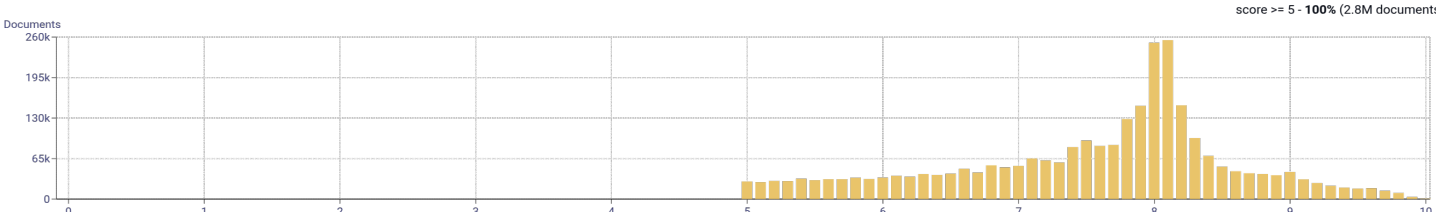


Percentage of segments in Icelandic (is) inside documents



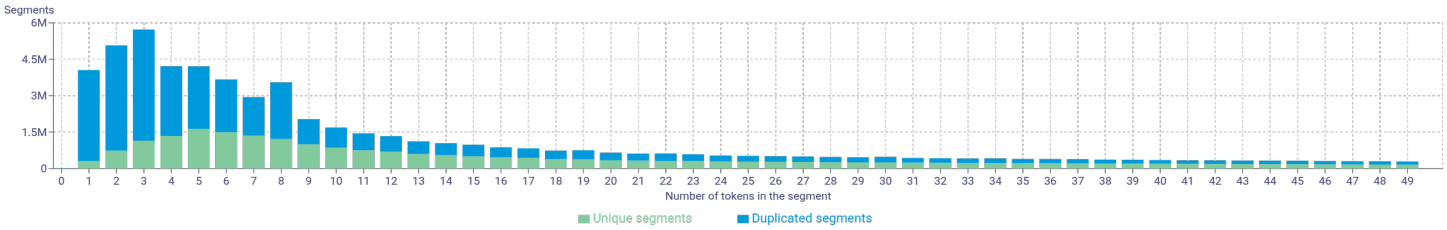
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.8M documents)

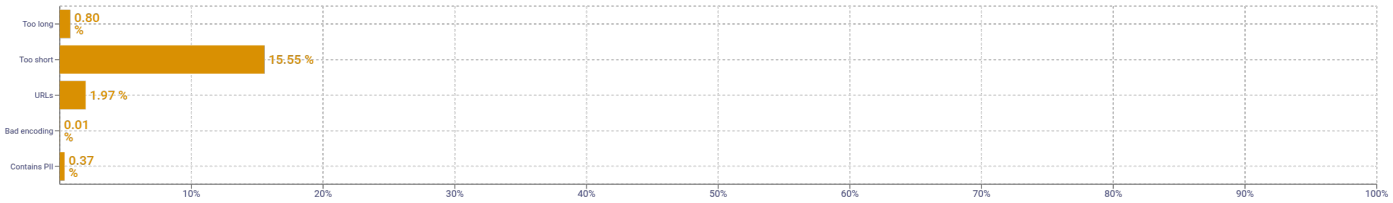


Segment length distribution by token

≤ 49 tokens = **23M** segments | **36M** duplicates
> 50 tokens = **11M** segments | **4.7M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	var | 9361168 · hafa | 4308729 · verið | 3803958 · sé | 2701146 · vera | 2588016
2	síðustu klukkustund | 691945 · einstaklingar skoðuðu | 691844 · hafi verið | 432232 · hafa verið | 420583 · lesa meira | 199970
3	hótel á síðustu | 691845 · skoðuðu þetta hótel | 691842 · hér á landi | 242118 · gr. laga nr. | 98705 · koma í veg | 98421
4	hótel á síðustu klukkustund | 691842 · einstaklingar skoðuðu þetta hótel | 691842 · gestur hefur gefið umsögn | 74119 · hótel og önnur gisting | 39004 · geta gjöld verið breytileg | 32854
5	skoðuðu þetta hótel á síðustu | 691842 · gefið er upp í bókunarstaðfestingunni | 35676 · upplýst okkur um eru innifalin | 32763 · vegar geta gjöld verið breytileg | 32761 · gjöld verið breytileg og farið | 32761

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