HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-isl_Latn	9/18/2025	Icelandic (is)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,295,927	93,441,531	56,799,911 (60.79 %)	2.7B	15,590,111,959	15.99 GB

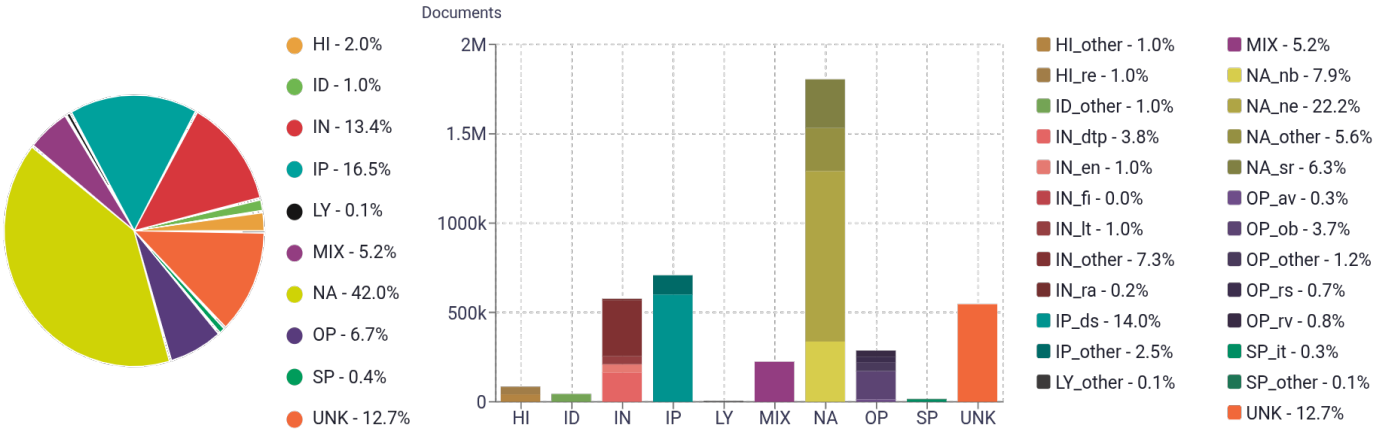
Top 10 domains

Domain	Docs	% of total
mbl.is	196K	4.57%
visir.is	170K	3.95%
hotels.com	95K	2.20%
blogspot.com	83K	1.94%
blog.is	81K	1.88%
dv.is	77K	1.80%
ruv.is	61K	1.42%
frettabladid.is	53K	1.24%
althingi.is	46K	1.08%
visindavefur.is	41K	0.95%

Top 10 TLDs

Domain	Docs	% of total
is	3.2M	73.64%
com	632K	14.70%
net	110K	2.57%
eu	102K	2.37%
org	90K	2.10%
dk	69K	1.60%
info	12K	0.29%
pt	8.4K	0.19%
co.uk	7.9K	0.18%
zone	7.3K	0.17%

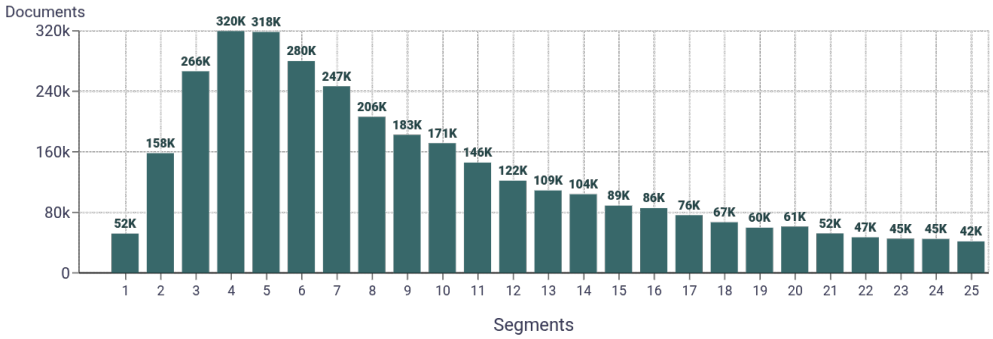
Register labels



MT:10.6% | 456K Documents

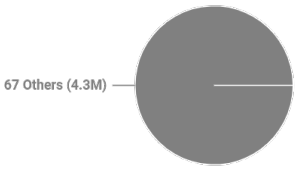
Documents size (in segments)

≤ 25 segments **78.03%** (3.4M documents)
> 25 segments **21.97%** (944K documents)



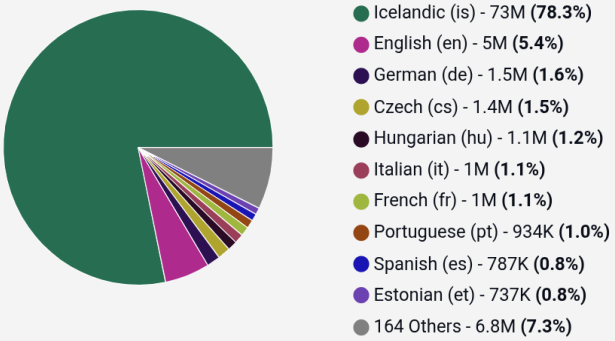
Document collections

CC = 90.90%
IA = 9.10%

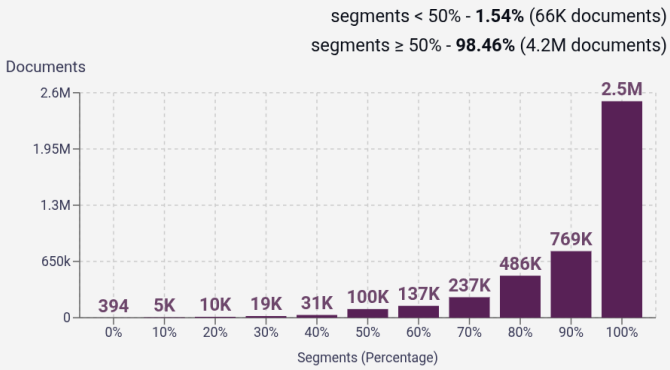


Language Distribution

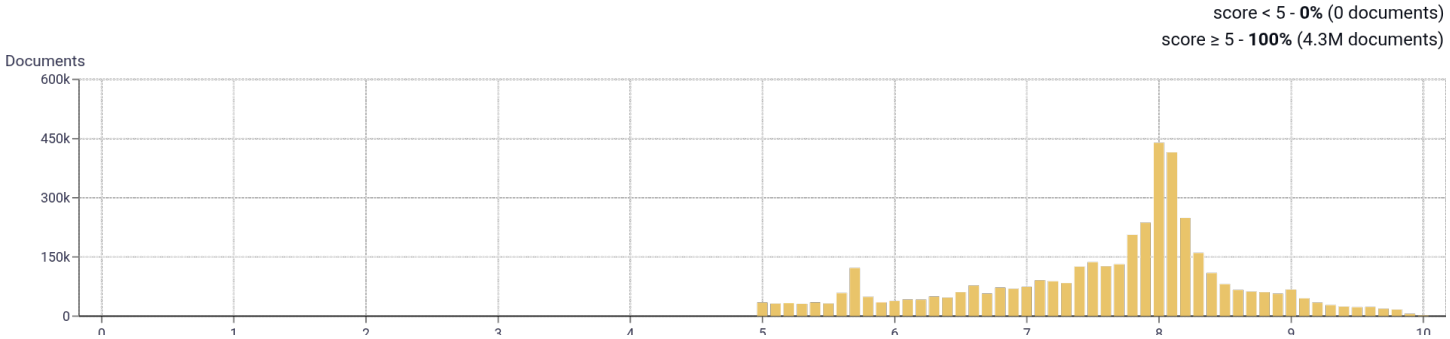
Number of segments in the Icelandic (is) corpus



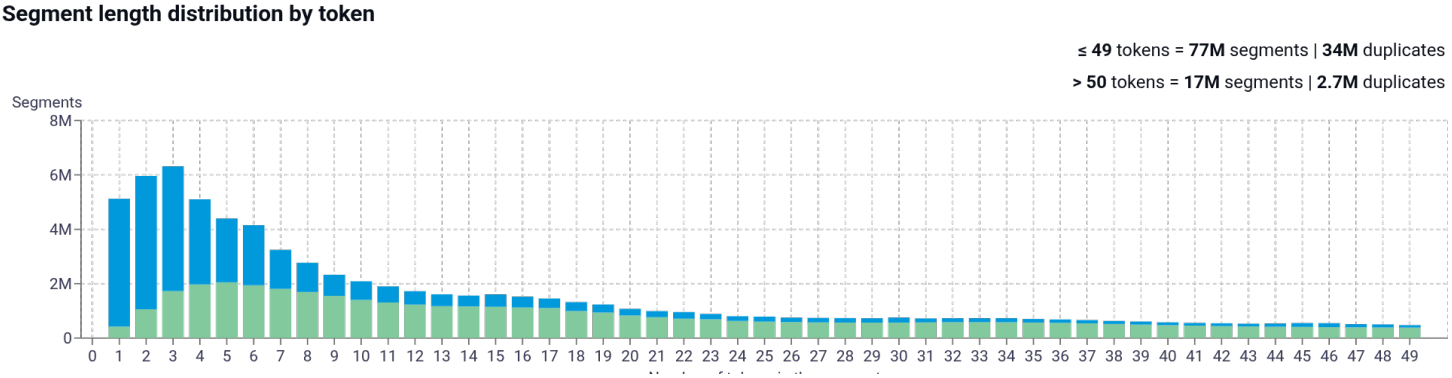
Percentage of segments in Icelandic (is) inside documents



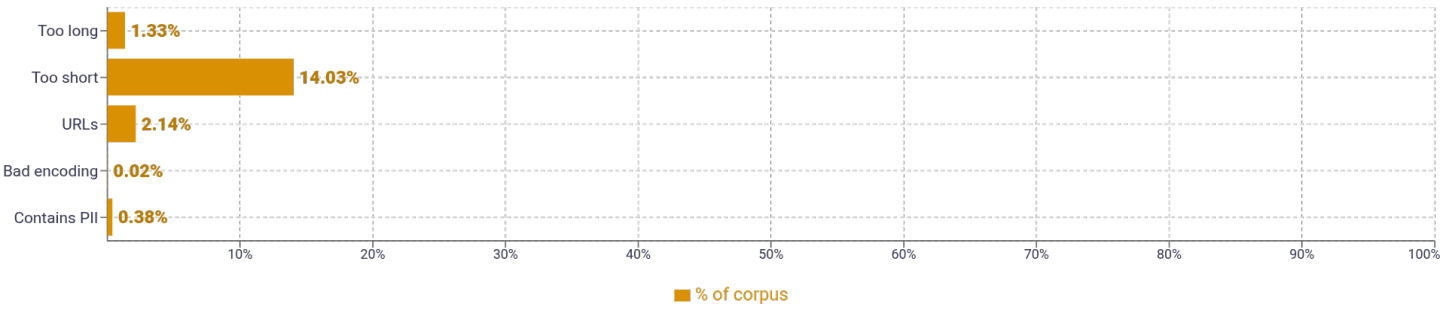
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	var | 11,769,650 · ókeypis | 10,564,455 · kynlíf | 6,830,425 · vændiskonur | 6,712,606 · klám | 6,347,756
2	erótískt nudd | 1,761,835 · ókeypis stefnumótasíða | 811,294 · ókeypis kynlíf | 747,299 · ókeypis klám | 727,139 · hafi verið | 703,708
3	kona að leita | 340,745 · hér á landi | 315,600 · leita að manni | 225,346 · konur sem leita | 206,388 · gr. laga nr. | 169,810
4	jerez de la frontera | 59,895 · hótel á síðustu klukkustund | 57,254 · einstaklingar skoðuðu þetta hótel | 57,254 · santa cruz de tenerife | 52,776 · las palmas de gran | 51,586
5	kona að leita að manni | 108,171 · konur sem leita að körlum | 71,400 · skoðuðu þetta hótel á síðustu | 57,254 · las palmas de gran canaria | 50,568 · gefið er upp í bókunarstaðfestingunni | 46,418

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				