HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-nob_Latn	9/18/2025	Norwegian Bokmål (nb)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
36,487,123	888,765,265	544,111,673 (61.22 %)	31B	171,279,234,150	163.04 GB

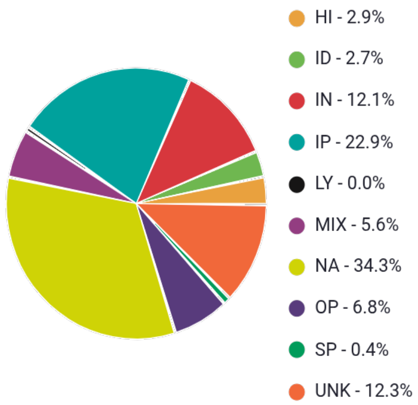
Top 10 domains

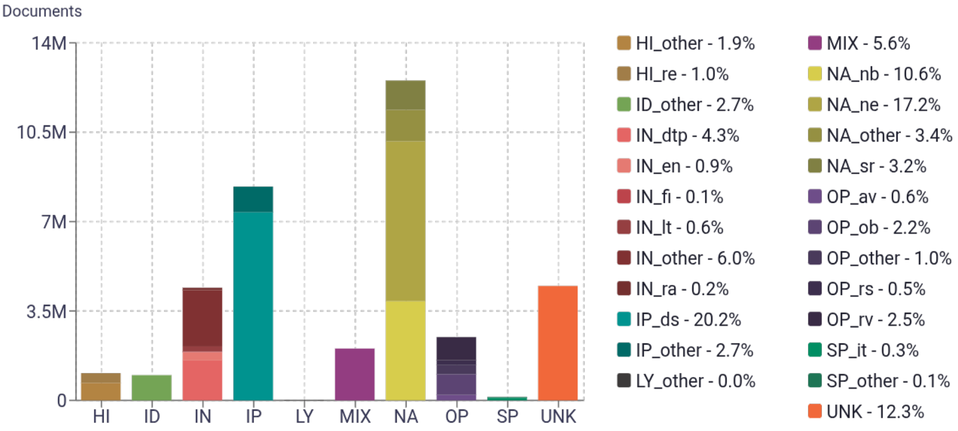
Domain	Docs	% of total
blogspot.com	1.2M	3.35%
blogg.no	990K	2.71%
dagbladet.no	620K	1.70%
nrk.no	378K	1.04%
tripadvisor.com	375K	1.03%
docplayer.me	346K	0.95%
aftenposten.no	337K	0.92%
nettavisen.no	321K	0.88%
wordpress.com	313K	0.86%
tv2.no	281K	0.77%

Top 10 TLDs

Domain	Docs	% of total
no	23M	63.88%
com	8M	21.99%
org	729K	2.00%
eu	709K	1.94%
net	550K	1.51%
me	362K	0.99%
kommune.no	271K	0.74%
ru	262K	0.72%
info	261K	0.72%
dk	190K	0.52%

Register labels

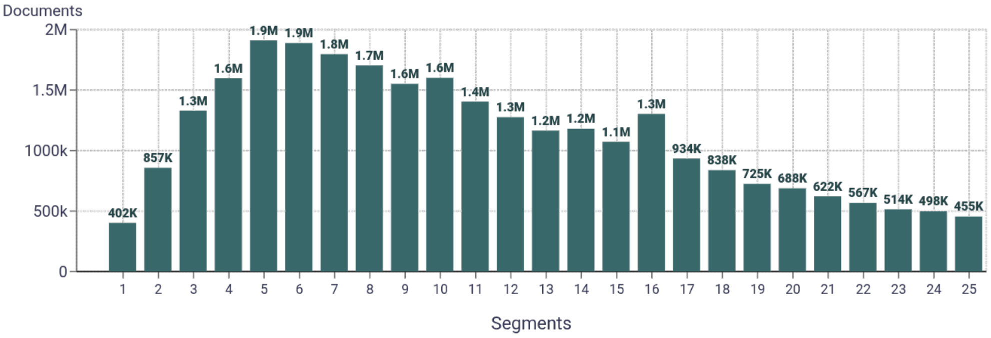




MT:10.1% | 3.7M Documents

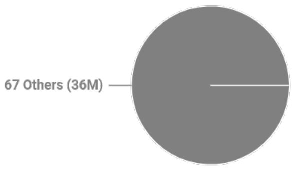
Documents size (in segments)

≤ 25 segments **76.42%** (28M documents)
> 25 segments **23.58%** (8.6M documents)



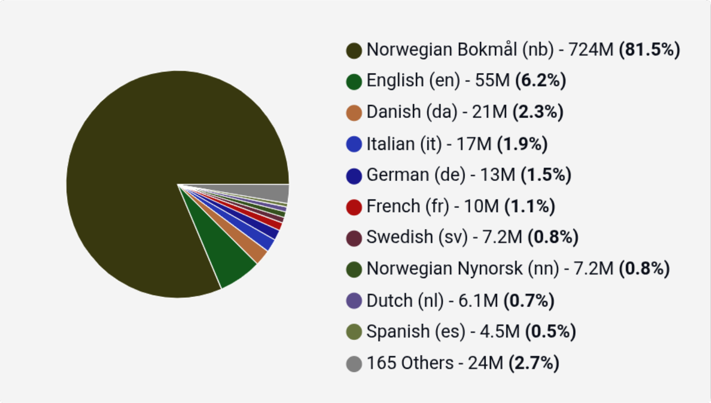
Document collections

CC = 89.22%
IA = 10.78%



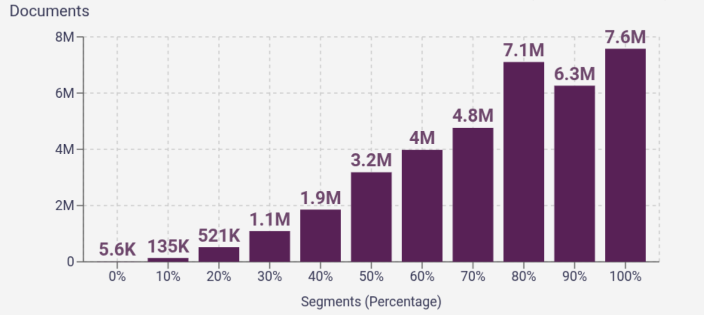
Language Distribution

Number of segments in the Norwegian Bokmål (nb) corpus



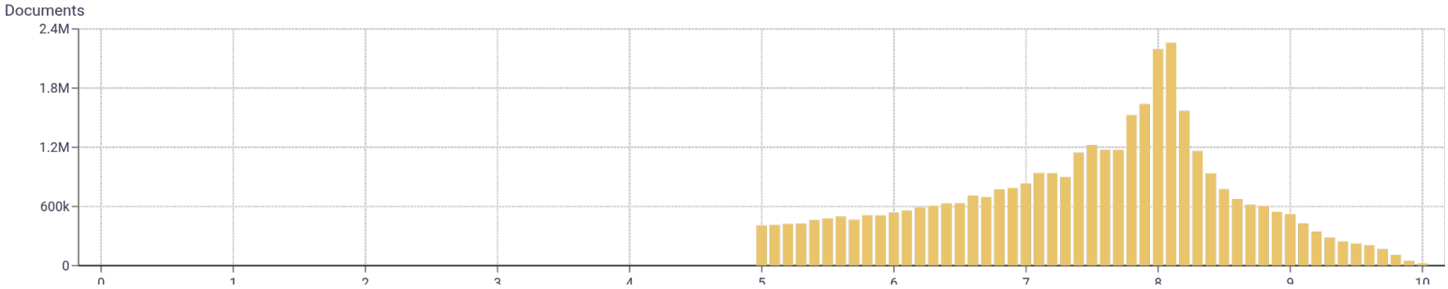
Percentage of segments in Norwegian Bokmål (nb) inside documents

segments < 50% - **9.90%** (3.6M documents)
segments ≥ 50% - **90.10%** (33M documents)



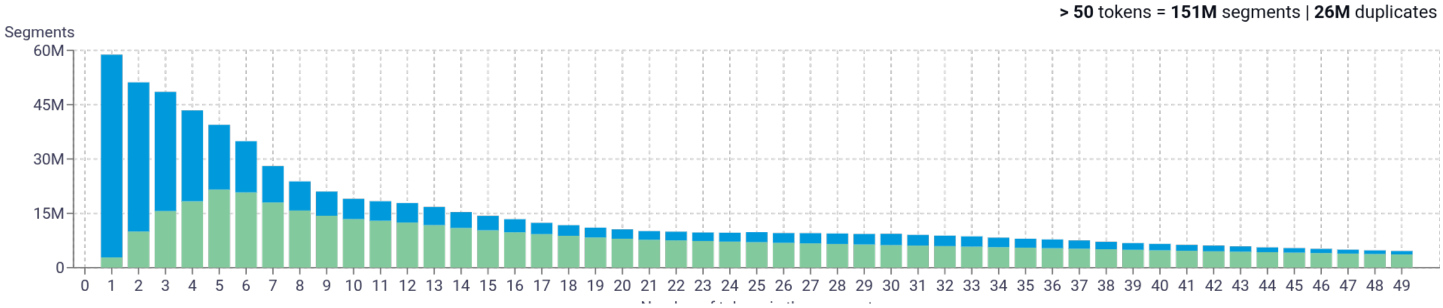
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (36M documents)

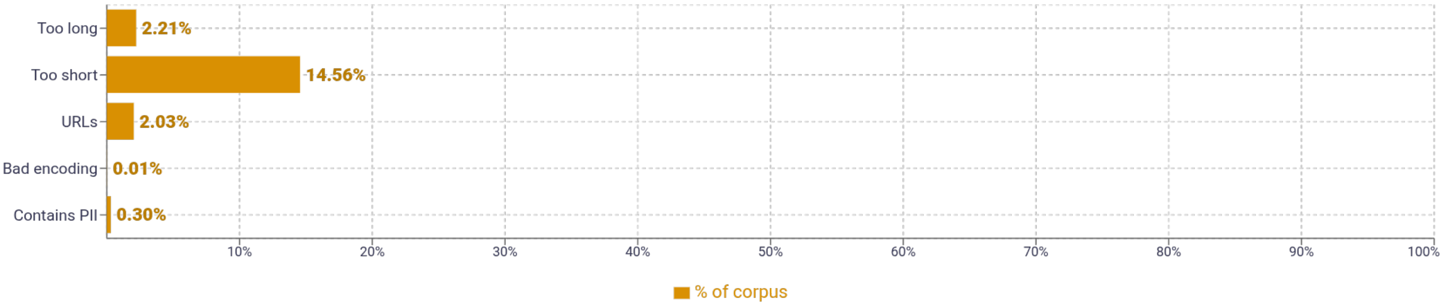


Segment length distribution by token

≤ 49 tokens = **737M** segments | **320M** duplicates
> 50 tokens = **151M** segments | **26M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	sex | 67,354,956 dating | 56,393,037 gratis | 51,036,857 mer | 51,026,305 andre | 47,523,020
2	blant annet | 7,153,802 dating nettsteder | 6,487,504 les mer | 6,333,557 thai massasje | 6,054,151 online dating | 4,899,230
3	rett og slett | 2,110,951 først og fremst | 1,660,008 thai massasje oslo | 1,282,473 barn og unge | 1,141,124 ønsker å knulle | 721,661
4	legg inn en kommentar | 714,405 ønsker å knulle gift | 686,168 skjult id med pseudonym | 356,901 massasje med happy ending | 290,477 løpet av den siste | 287,658
5	ønsker å knulle gift mann | 684,246 løpet av den siste timen | 267,237 logget inn for å kommentere | 226,041 bryr oss om ditt personvern | 205,993 logge inn på alle våre | 188,091

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				