HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
nob_Latn.jsonl.tsv	6/20/2025	Norwegian Bokmål (nb)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
27,053,789	675,891,305	279,861,405 (41.41 %)	24B	132,594,741,063	126.24 GB

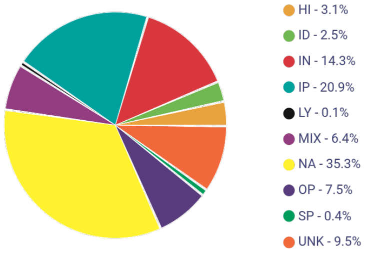
Top 10 domains

Domain	Docs	% of total
blogspot.com	1.2M	4.37%
blogg.no	1.1M	3.90%
blogspot.no	711K	2.63%
wikipedia.org	594K	2.20%
aftenposten.no	386K	1.43%
docplayer.me	376K	1.39%
dagbladet.no	330K	1.22%
wordpress.com	267K	0.99%
tripadvisor.com	261K	0.96%
nrk.no	222K	0.82%

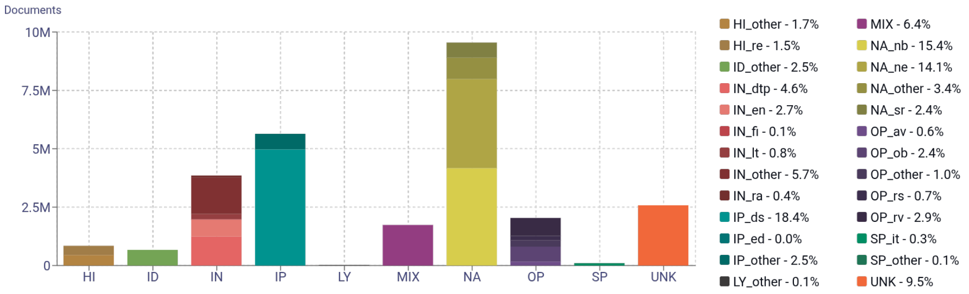
Top 10 TLDs

Domain	Docs	% of total
no	17M	64.44%
com	5.8M	21.34%
org	997K	3.68%
net	465K	1.72%
me	385K	1.42%
eu	385K	1.42%
kommune.no	189K	0.70%
info	179K	0.66%
ru	94K	0.35%
se	83K	0.31%

Register labels

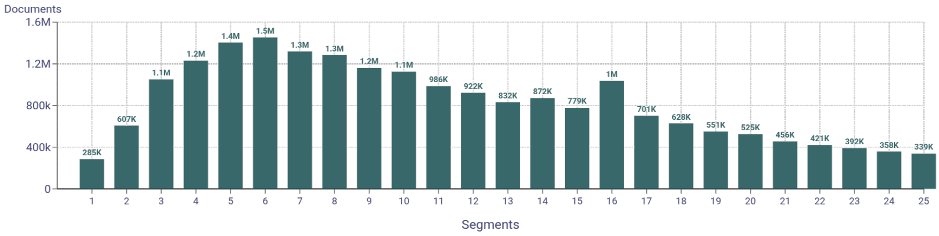


🤖 **MT**:7.5% | 2M Documents



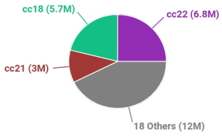
Documents size (in segments)

<= 25 segments **76.57%** (21M documents)
> 25 segments **23.43%** (6.3M documents)



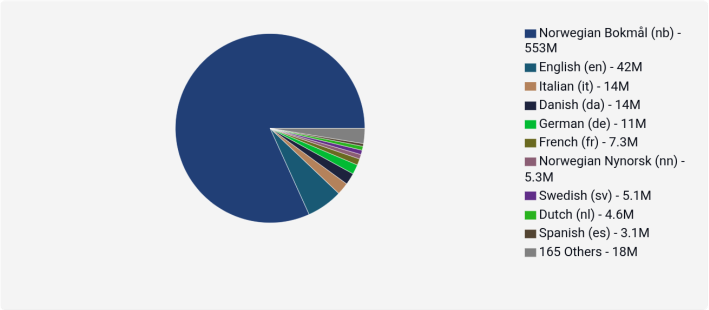
Documents by collection

CC = 69.54%
IA = 30.46%

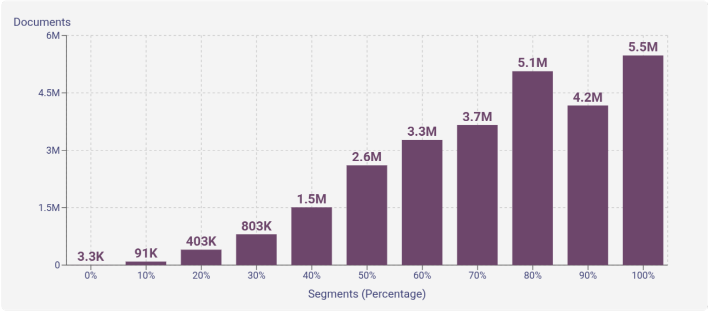


Language Distribution

Number of segments in the Norwegian Bokmål (nb) corpus

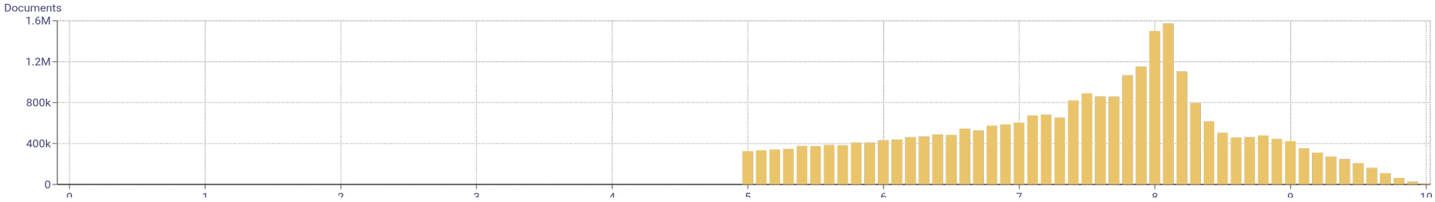


Percentage of segments in Norwegian Bokmål (nb) inside documents



Distribution of documents by document score

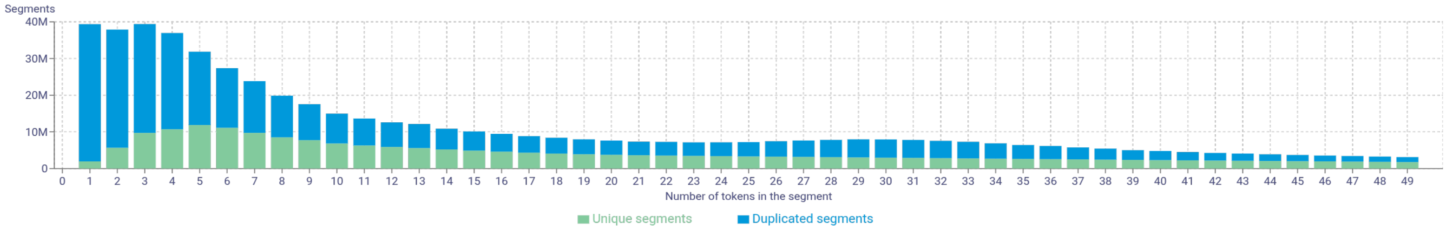
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (27M documents)



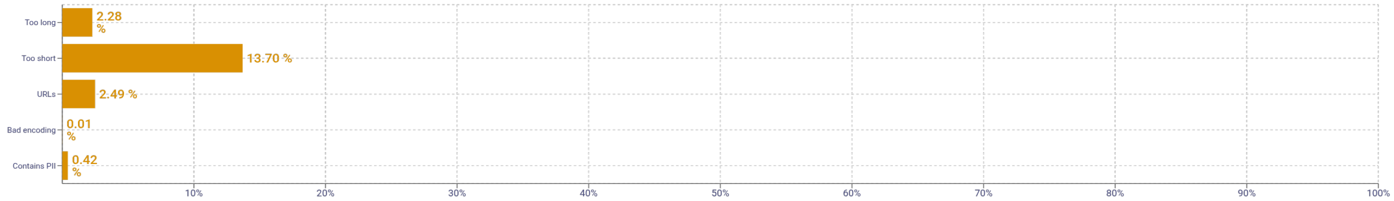
Segment length distribution by token

≤ **49** tokens = **209M** segments | **353M** duplicates

> **50** tokens = **114M** segments | **43M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	mer | 40660736 andre | 37698964 må | 33647299 år | 32674881 to | 30595607
2	blant annet | 5968702 les mer | 5636082 thai massasje | 2719029 of the | 2289240 online dating | 2016749
3	rett og slett | 1782276 først og fremst | 1376194 barn og unge | 1034371 universitetet i oslo | 599666 thai massasje oslo | 590871
4	skjult id med pseudonym | 496477 løpet av den siste | 451921 sett på dette hotellet | 436237 løpet av de siste | 192955 klagenemnda for offentlige anskaffelser | 183666
5	ønsker å knulle gift mann | 463610 løpet av den siste timen | 437528 mer mer mer mer mer | 176914 høgskolen i oslo og akershus | 120623 mandag tirsdag onsdag torsdag fredag | 116824

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