

General overview

Corpus	Date	Language
hplt-v3-nno_Latn	9/18/2025	Norwegian Nynorsk (nn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,509,659	31,904,198	20,530,211 (64.35 %)	909M	4,894,995,469	4.67 GB

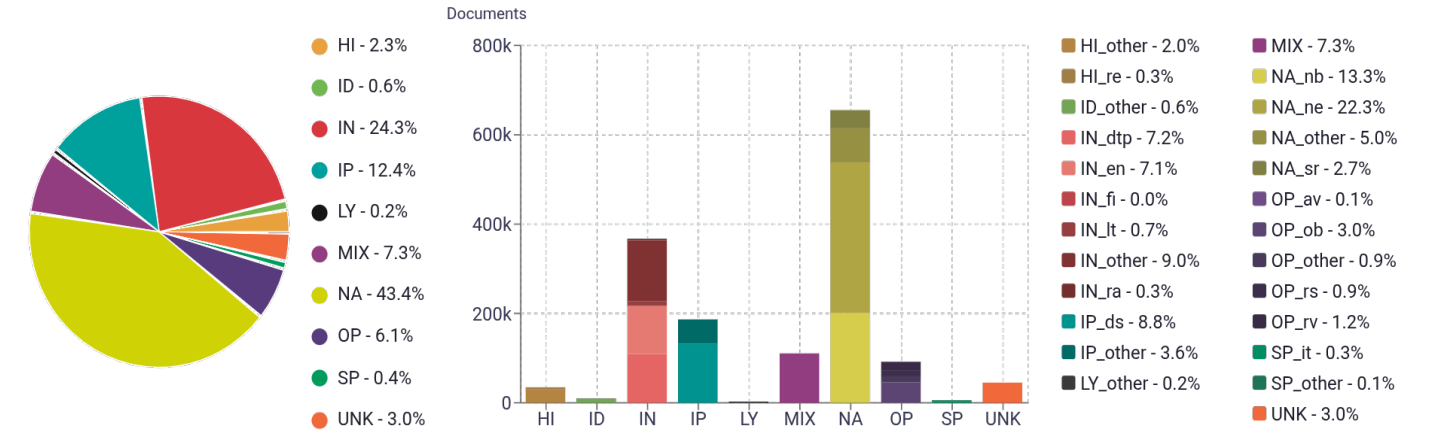
Top 10 domains

Domain	Docs	% of total
wikipedia.org	91K	6.05%
blogspot.com	77K	5.10%
nrk.no	66K	4.40%
docplayer.me	26K	1.74%
framtida.no	24K	1.57%
wordpress.com	21K	1.42%
blogg.no	17K	1.13%
midtsiden.no	17K	1.11%
blogspot.no	17K	1.11%
ndia.no	16K	1.05%

Top 10 TLDs

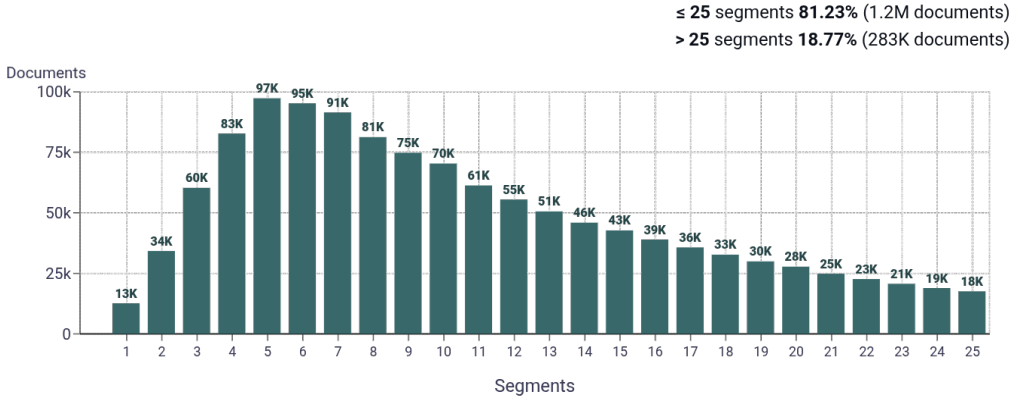
Domain	Docs	% of total
no	1M	69.07%
com	188K	12.47%
org	113K	7.47%
kommune.no	60K	3.99%
me	26K	1.75%
net	21K	1.37%
info	12K	0.80%
vgs.no	7.8K	0.52%
biz	4.7K	0.31%
eu	4.4K	0.29%

Register labels

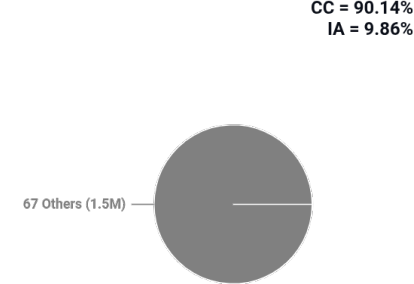


MT:0.9% | 14K Documents

Documents size (in segments)

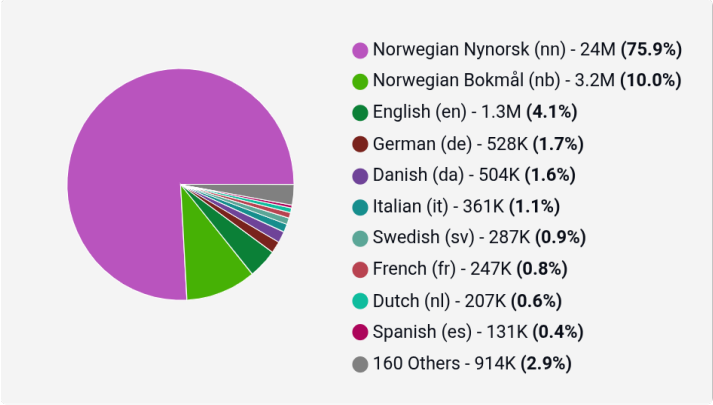


Document collections

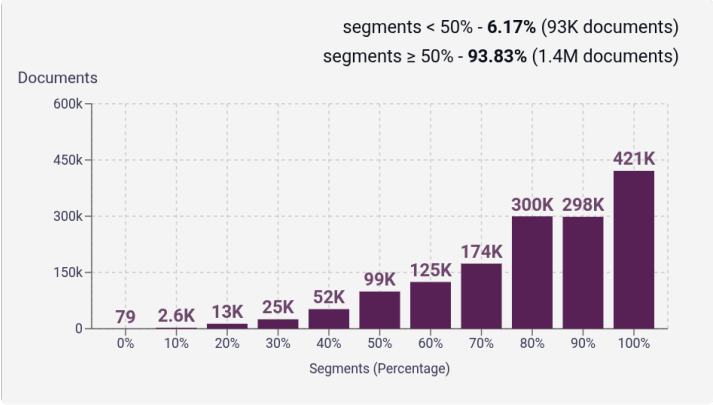


Language Distribution

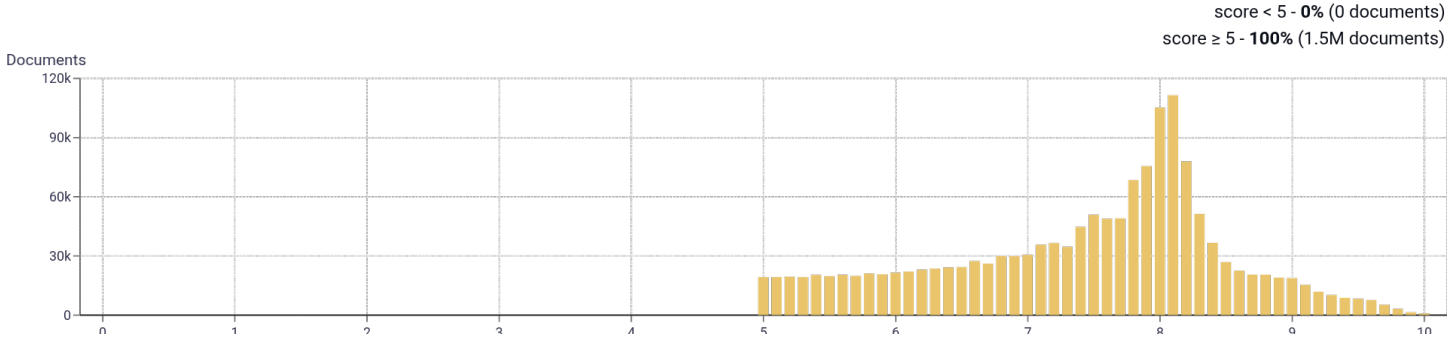
Number of segments in the Norwegian Nynorsk (nn) corpus



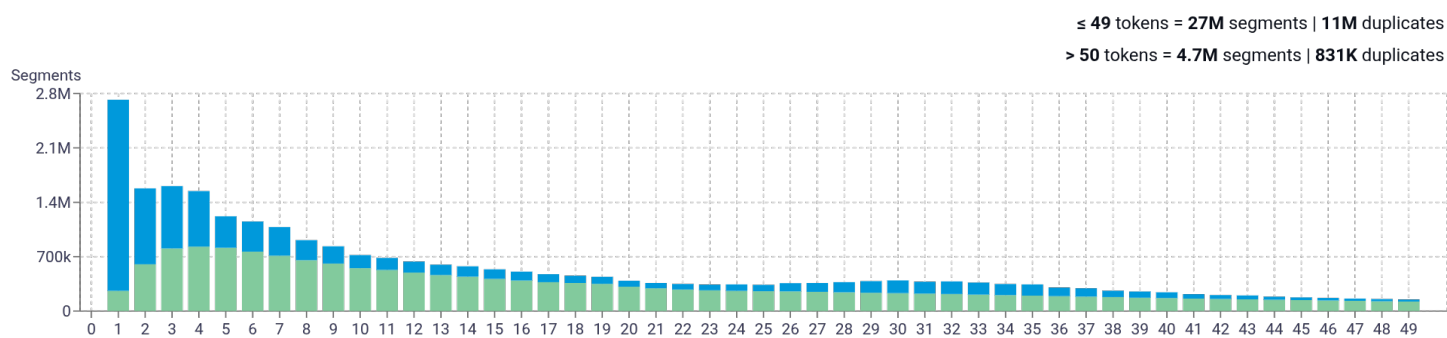
Percentage of segments in Norwegian Nynorsk (nn) inside documents



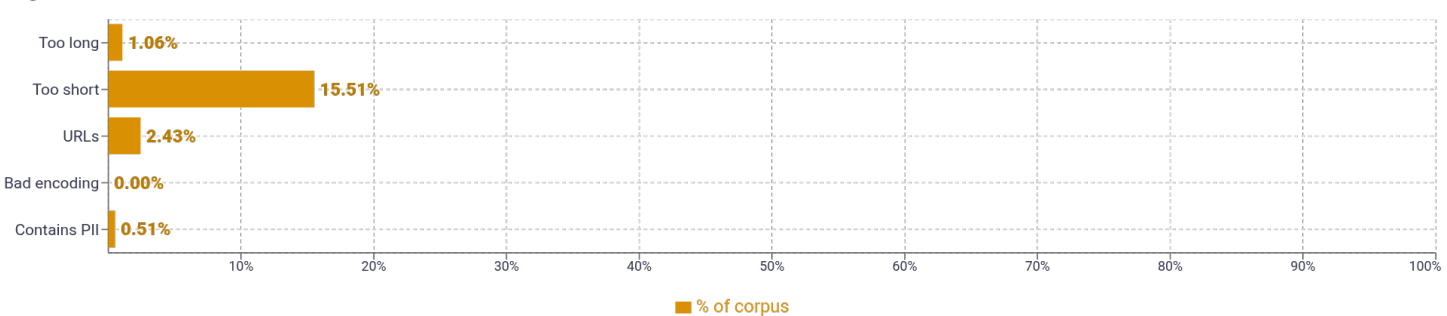
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	år   1,517,506    få   1,252,626    detaljer   1,109,032    to   1,104,869    seier   1,019,690	
2	endre wikiteksten   170,526    blant anna   84,378    millionar kroner   82,402    ta kontakt   76,591    n s   71,066	
3	sogn og fjordane   243,383    møre og romsdal   164,310    barn og unge   63,731    nor og døydde   50,923    rett og slett   47,604	
4	legg inn en kommentar   41,172    møre og romsdal fylkeskommune   28,533    helse vest rhf dato   22,316    m ug b s   17,875    k ug b d   17,537	
5	wikipedia på engelsk oppgav desse   12,999    teneste for publisering av kjelder   11,641    publisering av kjelder på internett   11,175 fylkesleksikon for sogn og fjordane   10,471    k ug b d datter   9,516	

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 of 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				