

General overview

Corpus	Date	Language
hplt-v3-dan_Latn	9/18/2025	Danish (da)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
52,498,797	1,334,528,804	768,084,213 (57.55 %)	38B	207,000,581,242	197.48 GB

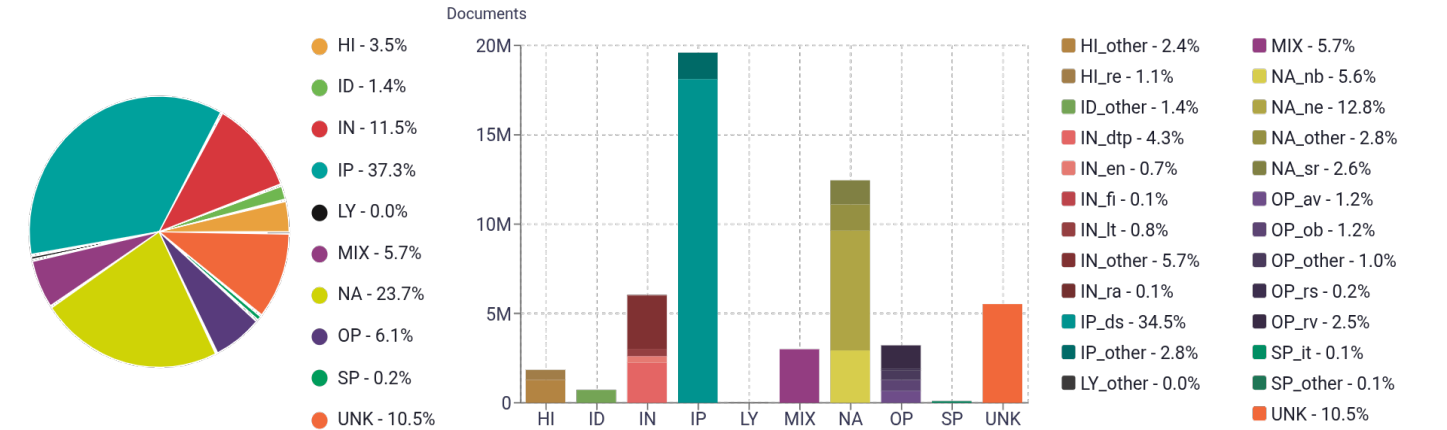
Top 10 domains

Domain	Docs	% of total
docplayer.dk	730K	1.39%
blogspot.com	601K	1.14%
billedeverden.com	448K	0.85%
avisen.dk	273K	0.52%
ekstrabladet.dk	260K	0.50%
nordjyske.dk	241K	0.46%
tripadvisor.dk	236K	0.45%
politiken.dk	213K	0.41%
wordpress.com	188K	0.36%
gucca.dk	174K	0.33%

Top 10 TLDs

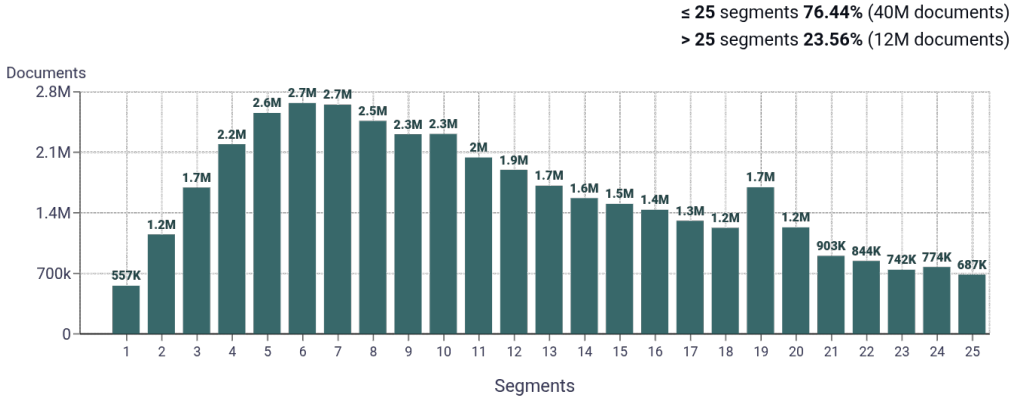
Domain	Docs	% of total
dk	40M	75.35%
com	7.9M	14.98%
eu	1.1M	2.06%
org	670K	1.28%
net	529K	1.01%
nu	353K	0.67%
se	337K	0.64%
info	274K	0.52%
ru	132K	0.25%
co	100K	0.19%

Register labels

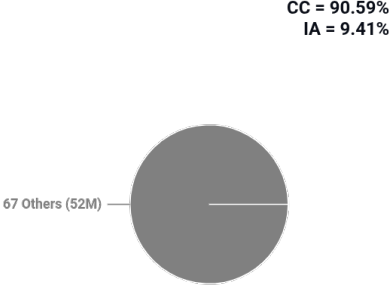


MT:8.9% | 4.7M Documents

Documents size (in segments) ⓘ

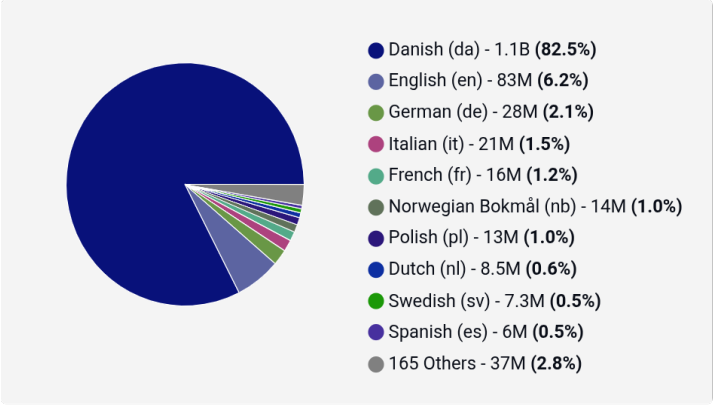


Document collections

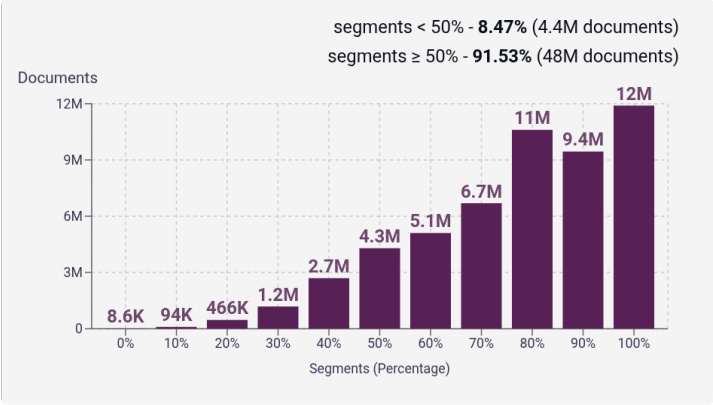


Language Distribution

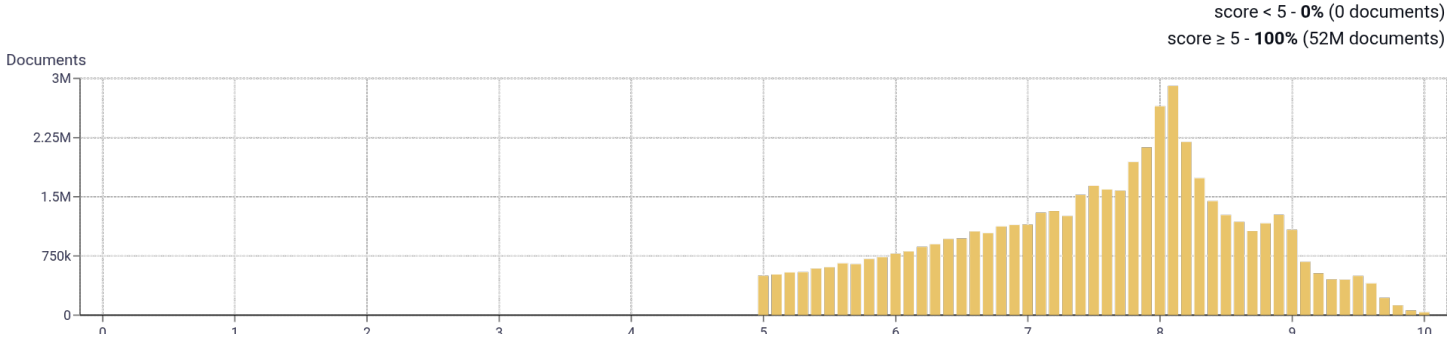
Number of segments in the Danish (da) corpus



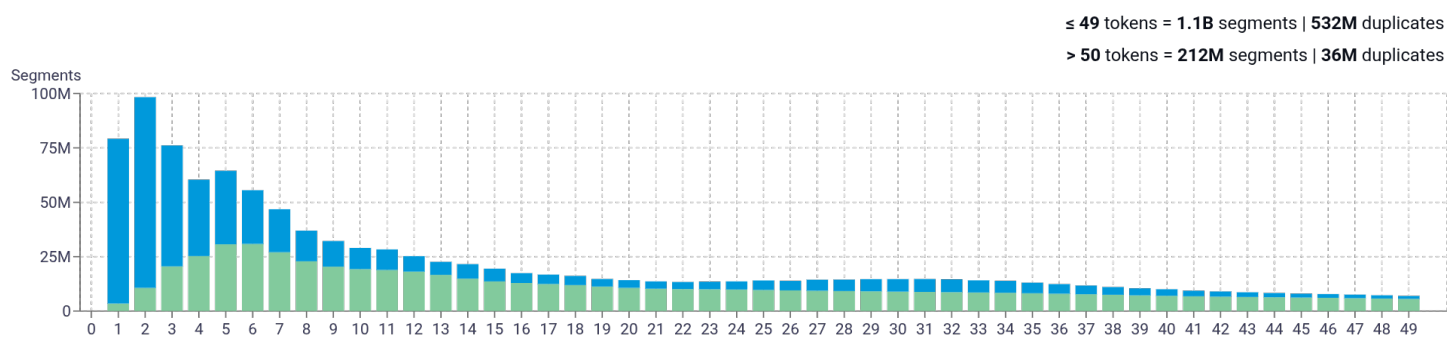
Percentage of segments in Danish (da) inside documents



Distribution of documents by document score

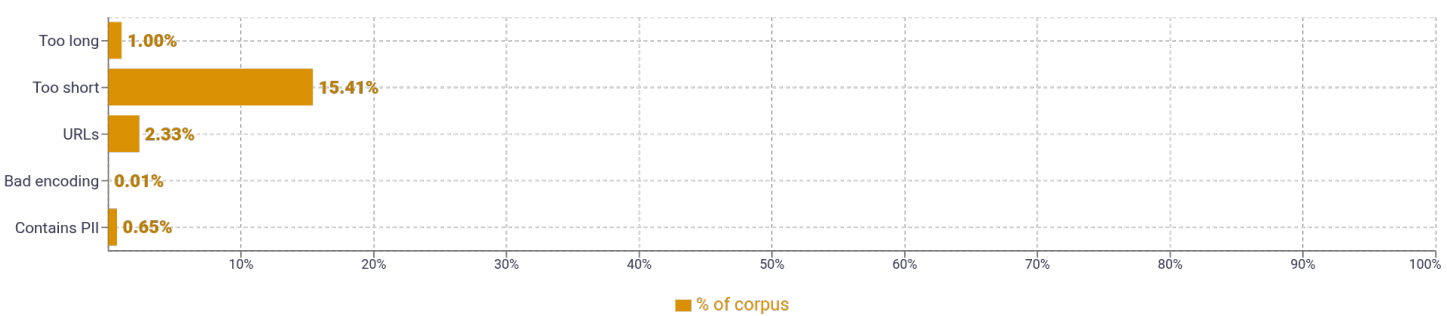


Segment length distribution by token



≤ 49 tokens = 1.1B segments | 532M duplicates
> 50 tokens = 212M segments | 36M duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>kan 264,430,242</div> <div>så 136,872,445</div> <div>massage 127,530,216</div> <div>ved 103,640,013</div> <div>mere 90,843,835</div>	
2	<div>thai massage 43,911,500</div> <div>læs mere 34,032,338</div> <div>tantra massage 9,437,979</div> <div>sex massage 6,994,120</div> <div>blandt andet 5,790,323</div>	
3	<div>så du kan 4,452,211</div> <div>giver dig mulighed 2,020,261</div> <div>først og fremmest 1,976,171</div> <div>body to body 1,716,188</div> <div>gør det muligt 1,665,840</div>	
4	<div>ved køb over kr. 1,097,309</div> <div>body to body massage 924,493</div> <div>velkommen til at kontakte 804,264</div> <div>leder efter den billigste 720,293</div> <div>endnu ikke nogle anmeldelser 715,421</div>	
5	<div>oftest ud med at finde 700,144</div> <div>yderligere beskrivelse og flere billeder 672,096</div> <div>kan det være en god 469,861</div> <div>ved at købe dine varer 461,420</div> <div>gennemgået de mest populære webshops 405,411</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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				