HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
dan_Latn.jsonl.tsv	6/16/2025	Danish (da)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
33,841,355	872,886,142	340,944,296 (39.06 %)	24B	132,541,320,753	126.49 GB

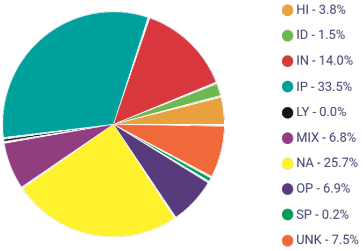
Top 10 domains

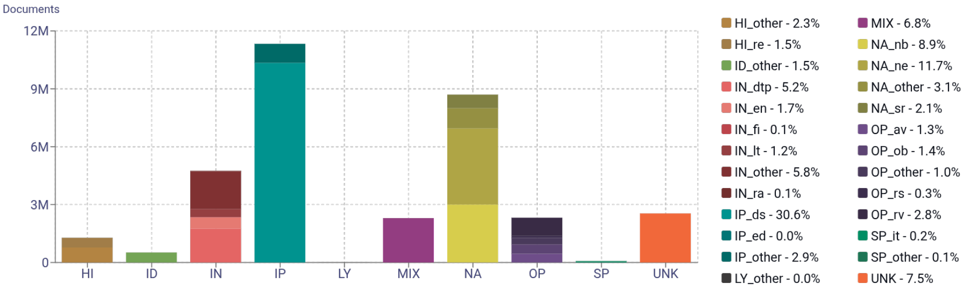
Domain	Docs	% of total
docplayer.dk	639K	1.89%
blogspot.com	541K	1.60%
wikipedia.org	458K	1.35%
billedeverden.com	444K	1.31%
blogspot.dk	271K	0.80%
tripadvisor.dk	239K	0.71%
dagens.dk	234K	0.69%
avisen.dk	155K	0.46%
wordpress.com	147K	0.43%
ekstrabladet.dk	144K	0.43%

Top 10 TLDs

Domain	Docs	% of total
dk	26M	76.44%
com	4.8M	14.11%
org	771K	2.28%
eu	644K	1.90%
net	350K	1.03%
nu	210K	0.62%
info	179K	0.53%
no	105K	0.31%
se	92K	0.27%
de	71K	0.21%

Register labels

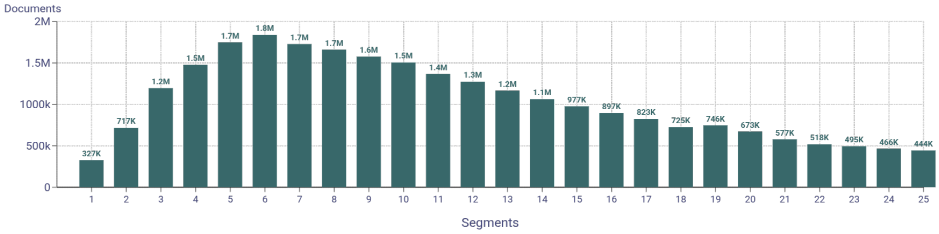




🤖 **MT**:6.2% | 2.1M Documents

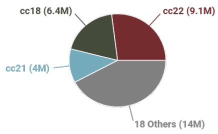
Documents size (in segments)

<= 25 segments **76.79%** (26M documents)
> 25 segments **23.21%** (7.9M documents)



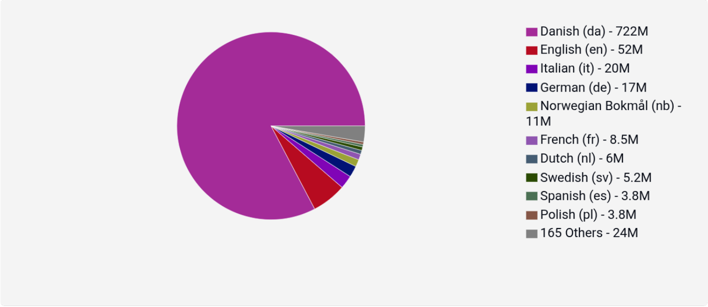
Documents by collection

CC = 69.09%
IA = 30.91%

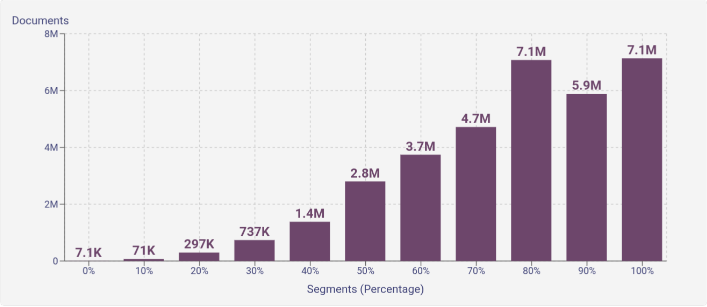


Language Distribution

Number of segments in the Danish (da) corpus

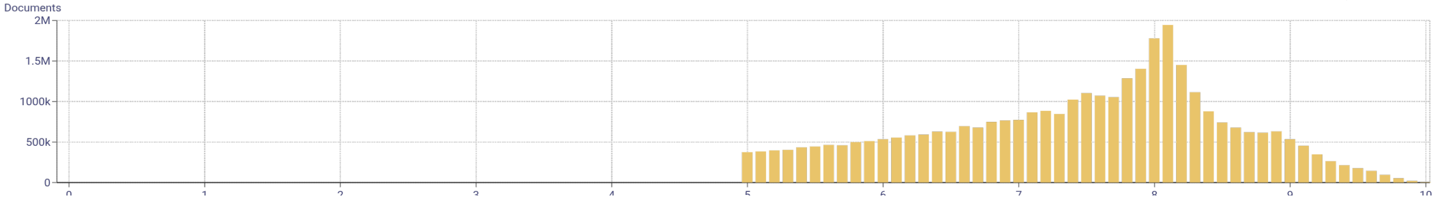


Percentage of segments in Danish (da) inside documents



Distribution of documents by document score

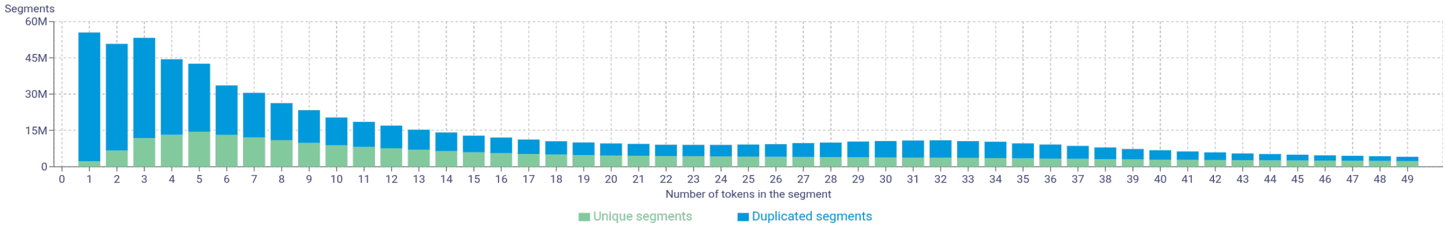
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (34M documents)



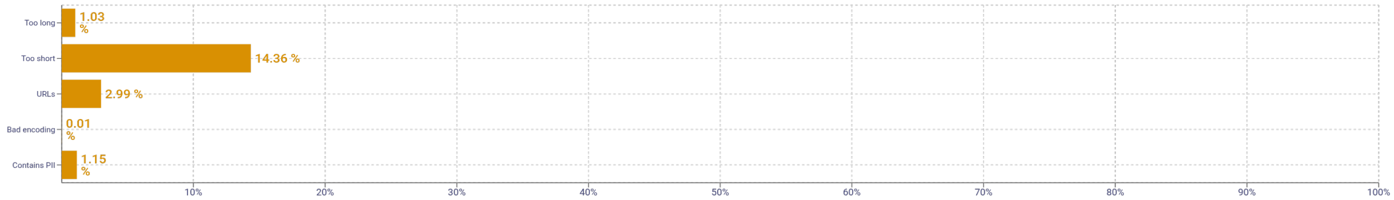
Segment length distribution by token

≤ 49 tokens = **262M** segments | **482M** duplicates

> **50** tokens = **129M** segments | **50M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kan | 151800263 mere | 100114187 så | 95416892 ved | 68939684 få | 33002482
2	læs mere | 11876884 blandt andet | 4161626 kan få | 3299614 så kan | 3147632 thai massage | 2550257
3	så du kan | 1415094 børn og unge | 1394061 først og fremmest | 999758 kan du se | 843334 gør det muligt | 738324
4	velkommen til at kontakte | 597981 stamoplysninger til brug ved | 543348 ved prøver til gymnasiale | 543094 prøver til gymnasiale uddannelser | 542511 undervisningsbeskrivelse stamoplysninger til brug | 492554
5	stamoplysninger til brug ved prøver | 543329 brug ved prøver til gymnasiale | 543093 ved prøver til gymnasiale uddannelser | 542511 prøver til gymnasiale uddannelser termin | 507566 undervisningsbeskrivelse stamoplysninger til brug ved | 492553

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