HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
swe_Latn.jsonl.tsv	7/5/2025	Swedish (sv)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
66,812,510	1,754,434,421	692,201,258 (39.45 %)	45B	249,421,899,758	242.25 GB

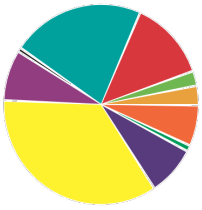
Top 10 domains

Domain	Docs	% of total
blogspot.com	2.9M	4.41%
blogg.se	2.5M	3.77%
blogspot.se	2M	3.03%
wordpress.com	1.4M	2.15%
web.app	1.1M	1.59%
wikipedia.org	857K	1.28%
docplayer.se	597K	0.89%
netlify.app	592K	0.89%
runeberg.org	589K	0.88%

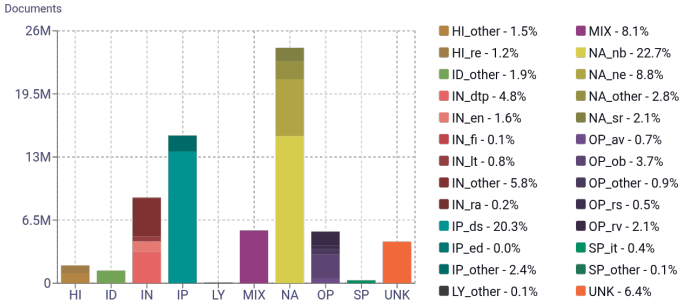
Top 10 TLDs

Domain	Docs	% of total
se	41M	62.10%
com	13M	20.11%
org	2.3M	3.49%
nu	1.9M	2.80%
app	1.7M	2.48%
eu	1.3M	1.97%
fi	1.3M	1.93%
net	645K	0.96%
info	509K	0.76%

Register labels



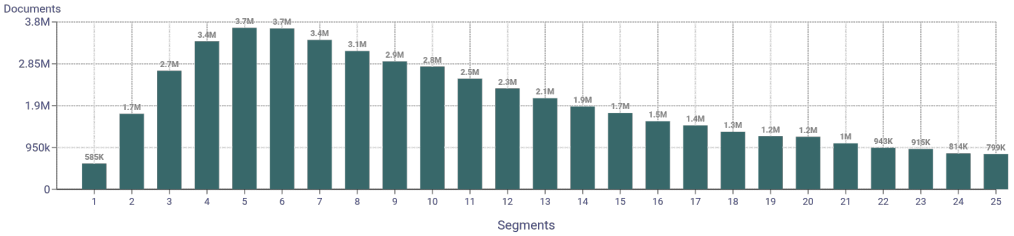
HI - 2.7%
ID - 1.9%
IN - 13.2%
IP - 22.8%
LY - 0.1%
MIX - 8.1%
NA - 36.3%
OP - 8.0%
SP - 0.5%
UNK - 6.4%



MT:5.2% | 3.5M Documents

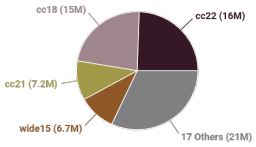
Documents size (in segments)

<= 25 segments **100%** (50M documents)
> 25 segments **0%** (0 documents)



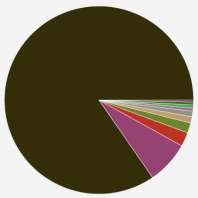
Document collections

CC = 68.51%
IA = 31.49%



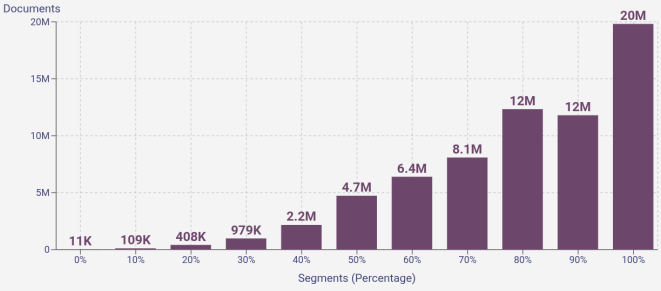
Language Distribution

Number of segments in the Swedish (sv) corpus



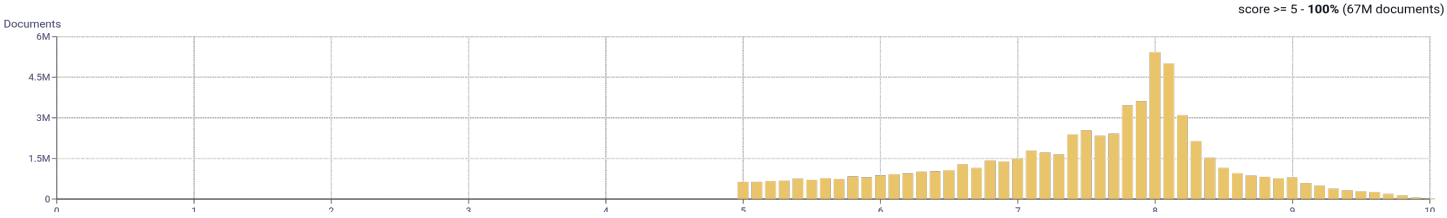
Swedish (sv) - 1.4B
English (en) - 124M
Italian (it) - 41M
German (de) - 30M
French (fr) - 18M
Dutch (nl) - 16M
Spanish (es) - 9M
Danish (da) - 8.9M
Norwegian Bokmål (nb) - 8.6M
Portuguese (pt) - 5.6M
10

Percentage of segments in Swedish (sv) inside documents



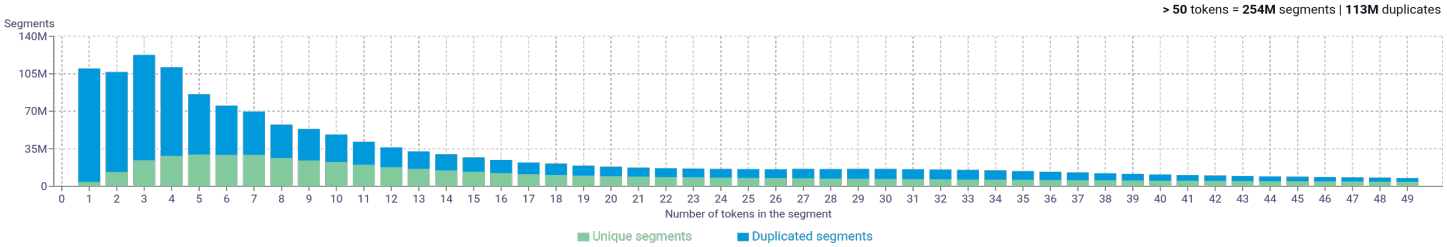
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (67M documents)

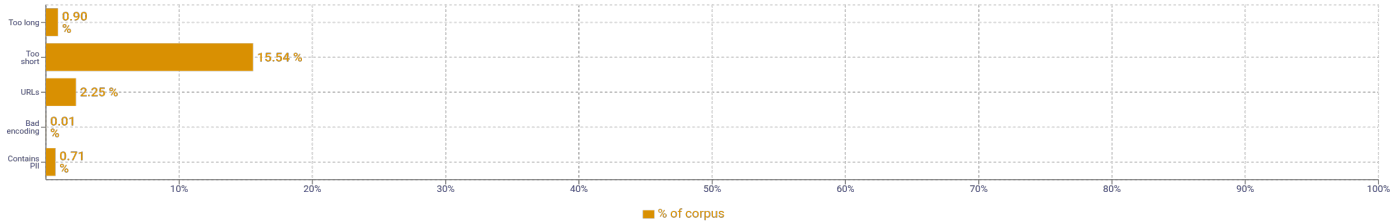


Segment length distribution by token

≤ 49 tokens = **1.5B** segments | **953M** duplicates
> 50 tokens = **254M** segments | **113M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	`mer | 131227510` `ska | 115609327` `kommer | 85612146` `finns | 77703608` `in | 75066238`
2	`läs mer | 15906155` `bland annat | 10628604` `helt enkelt | 5194682` `erotisk massage | 4489559` `hela tiden | 4347814`
3	`barn och ungdomar | 776446` `barn och unga | 672038` `diskriminering och kränkande | 615289` `först och främst | 605924` `genom att använda | 590854`
4	`diskriminering och kränkande behandling | 604494` `do you see an | 576174` `below is the raw | 576109` `därför är det viktigt | 268510` `body to body massage | 240043`
5	`do you see an error | 576170` `below is the raw ocr | 576108` `commons har media som rör | 172710` `been proofread at least once | 164558` `ankomstdatumet till hotellet är ogiltigt | 148771`

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