

General overview

Corpus	Date	Language
hplt-v3-swe_Latn	9/18/2025	Swedish

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
97,717,372	2,477,567,519	1,357,912,986 (54.81 %)	67B	372,734,478,126	361.46 GB

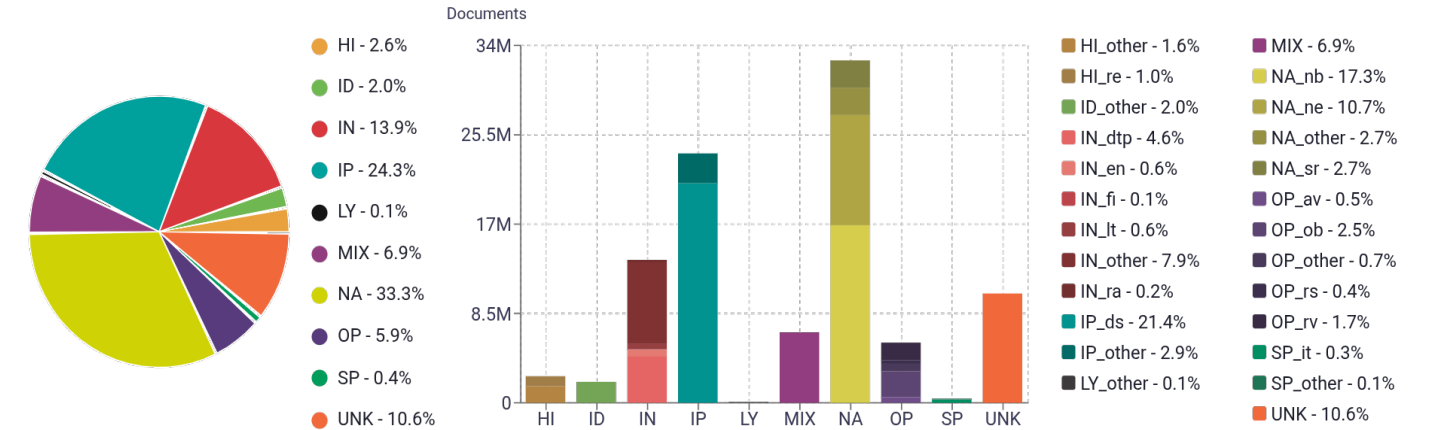
Top 10 domains

Domain	Docs	% of total
web.app	5.9M	6.07%
blogspot.com	3.7M	3.75%
blogg.se	3.5M	3.58%
netlify.app	1.8M	1.87%
wordpress.com	1.7M	1.73%
firebaseapp.com	1.7M	1.71%
blogspot.se	1M	1.06%
sverigesradio.se	741K	0.76%
docplayer.se	676K	0.69%
aftonbladet.se	483K	0.49%

Top 10 TLDs

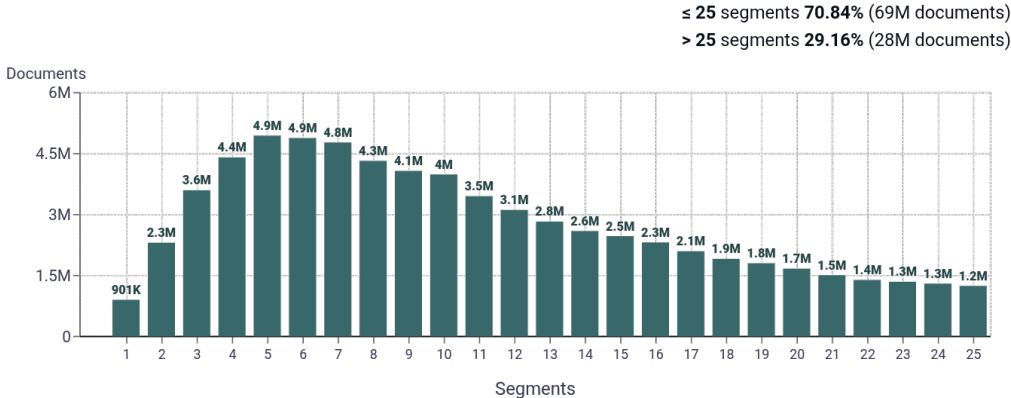
Domain	Docs	% of total
se	56M	57.56%
com	20M	20.68%
app	7.8M	7.95%
eu	2.4M	2.43%
nu	2.3M	2.37%
org	2M	2.04%
fi	1.7M	1.74%
net	917K	0.94%
info	673K	0.69%
ru	288K	0.29%

Register labels

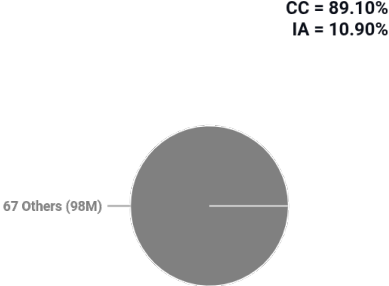


MT:8.4% | 8.2M Documents

Documents size (in segments) ⓘ

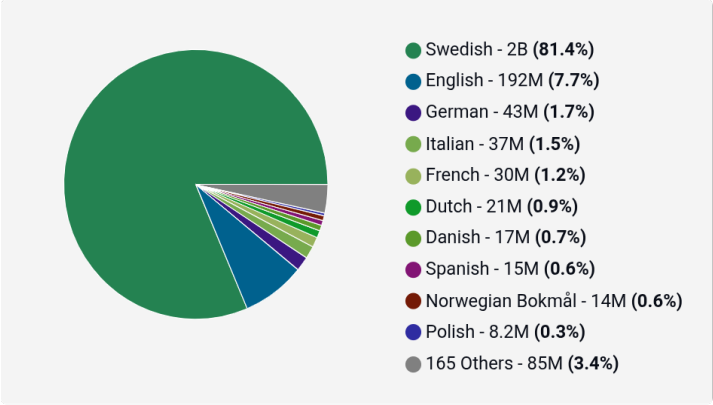


Document collections

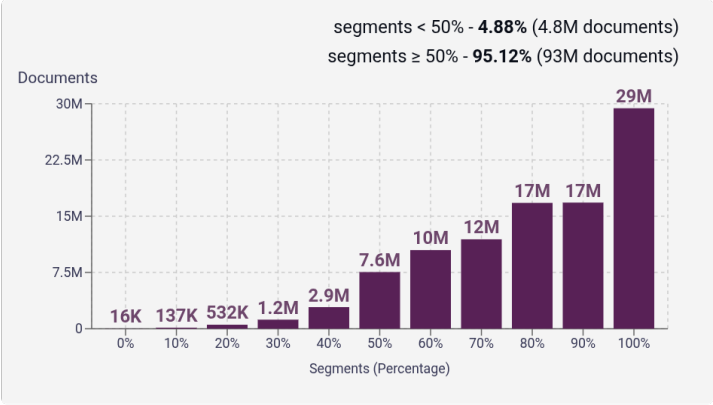


Language Distribution

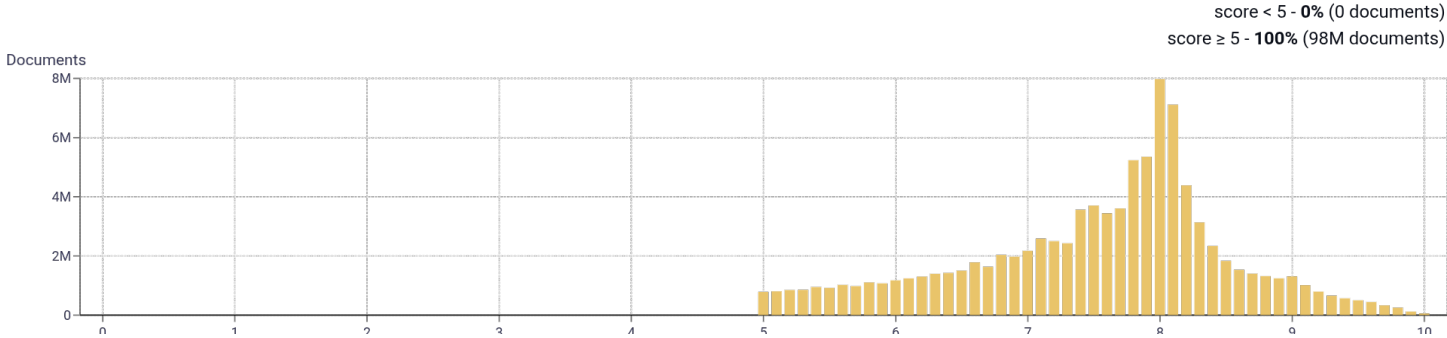
Number of segments in the Swedish corpus



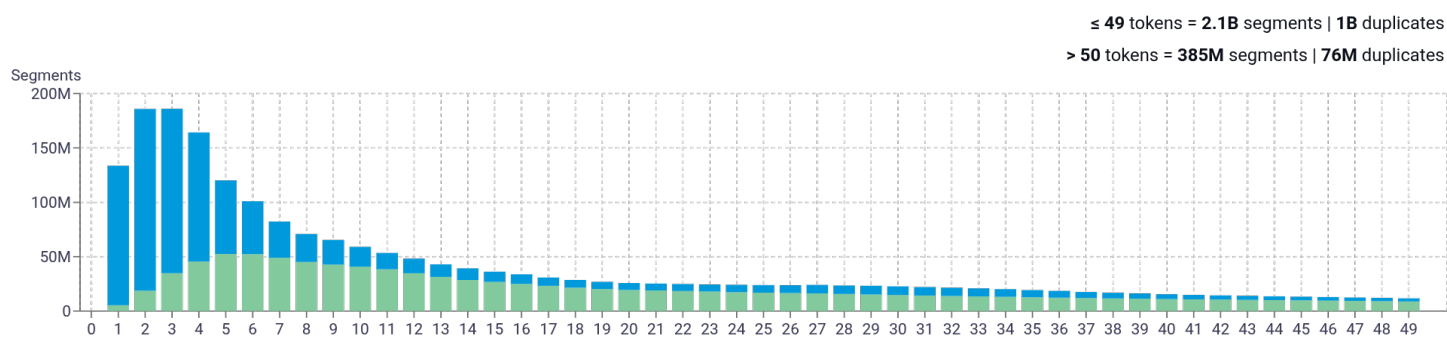
Percentage of segments in Swedish inside documents



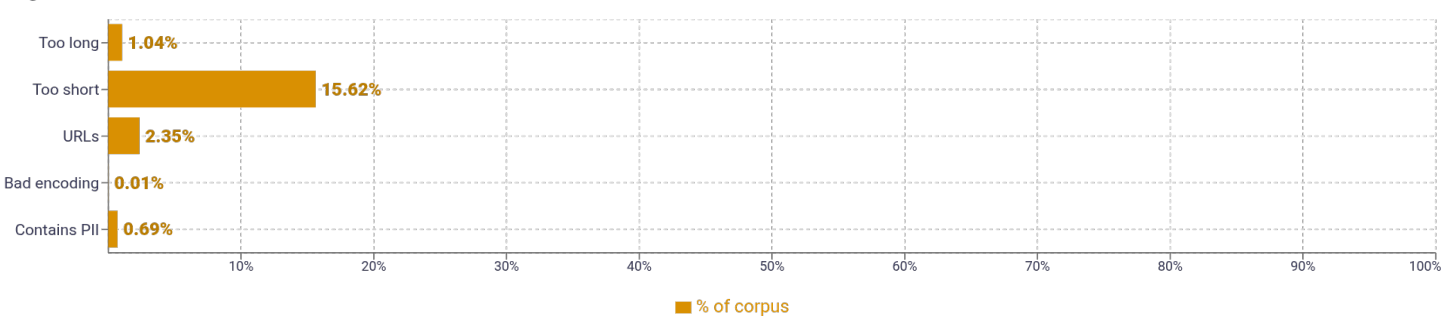
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ska   154,413,170   gratis   140,519,447   mer   136,259,061   massage   127,252,802   kommer   123,922,688	
2	läs mer   46,506,057   thai massage   23,195,695   erotisk massage   15,981,074   bland annat   15,567,912   escort tjejer   10,469,360	
3	skicka en kommentar   2,940,509   knulla med omedelbart   1,967,935   kvinnor som söker   1,707,823   å andra sidan   1,563,637 gör det möjligt   1,547,510	
4	body to body massage   1,007,289   runt om i världen   799,000   hotellet den senaste timmen   691,199   gå ner i vikt   677,539 kvinnor som söker män   607,880	
5	tittade på det här hotellet   707,788   below is the raw ocr   411,967   is the raw ocr text   411,966   from the above scanned image   411,804 do you see an error   410,838	

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	ntp
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				