HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-nld_Latn	9/18/2025	Dutch

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
200,688,712	4,245,883,008	2,305,008,403 (54.29 %)	116B	638,984,277,678	597.59 GB

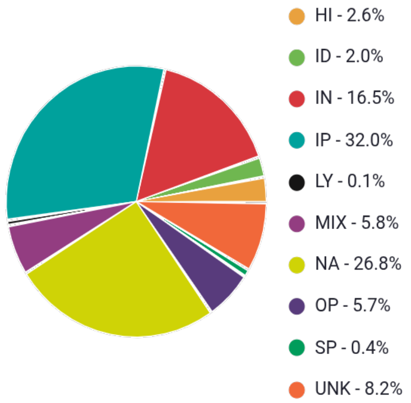
Top 10 domains

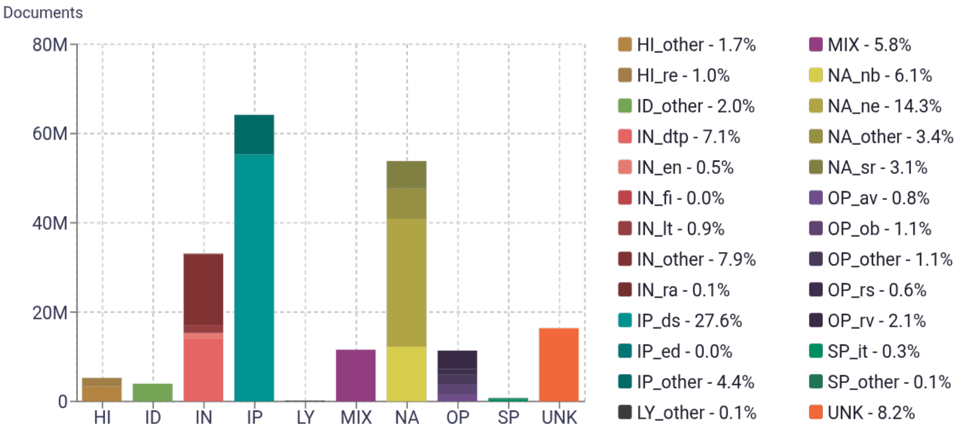
Domain	Docs	% of total
blogspot.com	2.2M	1.08%
wordpress.com	1.3M	0.63%
nrc.nl	858K	0.43%
docplayer.nl	785K	0.39%
knack.be	773K	0.38%
wikipedia.org	600K	0.30%
ad.nl	526K	0.26%
blogspot.nl	516K	0.26%
nu.nl	507K	0.25%
hln.be	502K	0.25%

Top 10 TLDs

Domain	Docs	% of total
nl	132M	65.59%
com	28M	14.13%
be	24M	11.86%
org	3.1M	1.54%
net	2.6M	1.28%
eu	2.6M	1.27%
nu	1.4M	0.72%
info	982K	0.49%
de	588K	0.29%
ru	421K	0.21%

Register labels

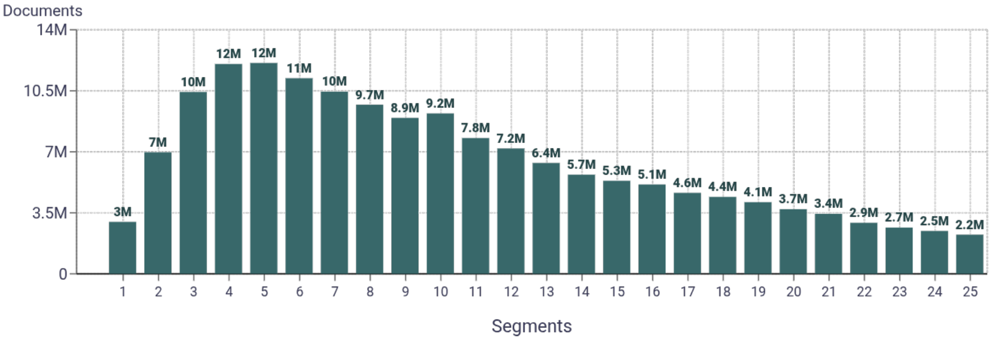




MT:5.3% | 11M Documents

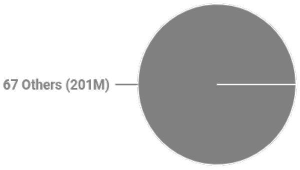
Documents size (in segments)

≤ 25 segments **80.79%** (162M documents)
> 25 segments **19.21%** (39M documents)



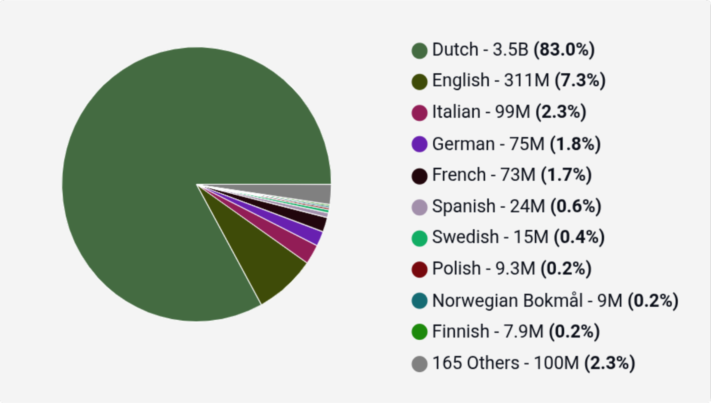
Document collections

CC = 90.72%
IA = 9.28%



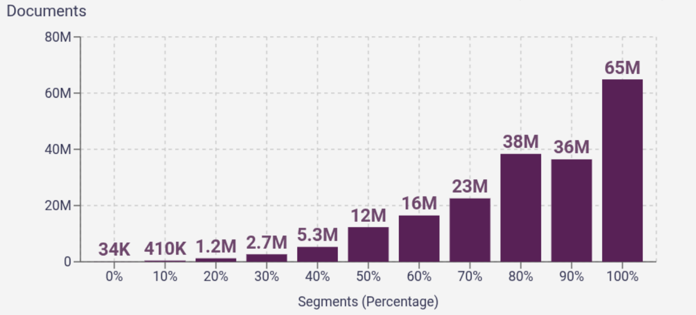
Language Distribution

Number of segments in the Dutch corpus



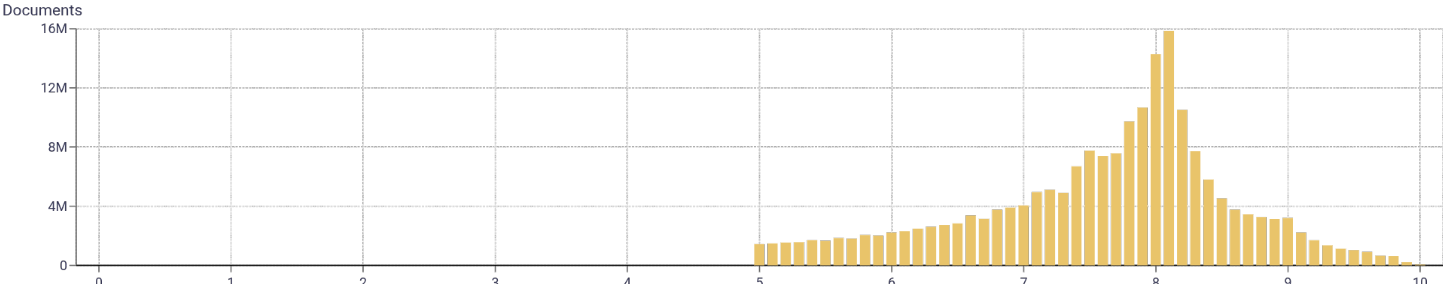
Percentage of segments in Dutch inside documents

segments < 50% - **4.81%** (9.6M documents)
segments ≥ 50% - **95.19%** (191M documents)



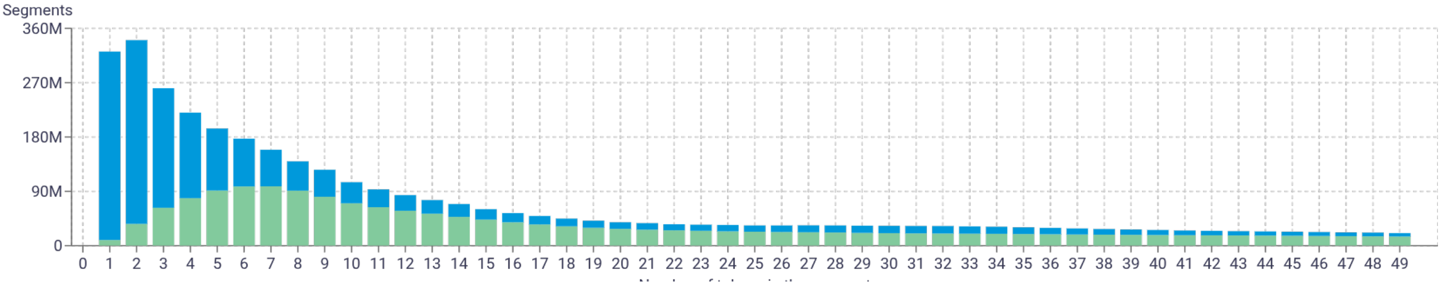
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (201M documents)

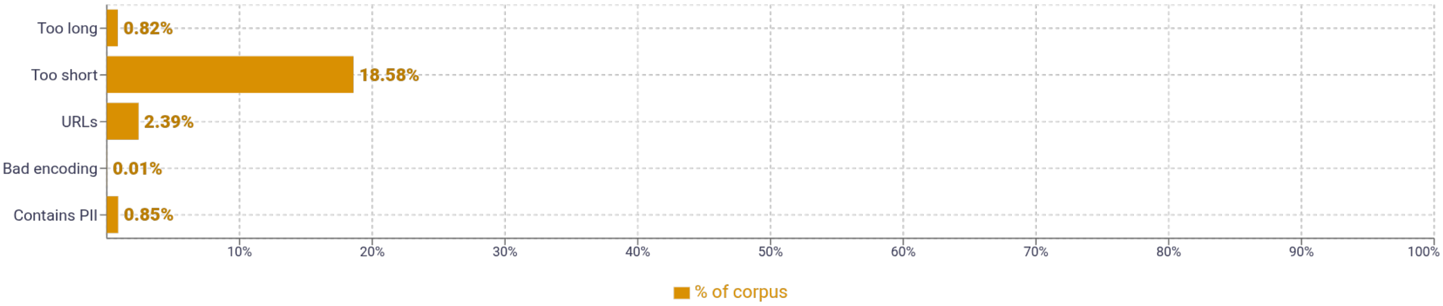


Segment length distribution by token

≤ **49** tokens = **3.5B** segments | **1.8B** duplicates
> 50 tokens = **730M** segments | **157M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	we | 347,116,541 s | 198,635,510 wel | 170,478,771 onze | 165,295,453 jaar | 153,150,188
2	nadere informatie | 42,939,969 online casino | 12,997,223 erotische massage | 12,562,376 lees verder | 11,060,282 den haag | 10,844,140
3	af en toe | 5,248,065 ervoor te zorgen | 4,055,592 no deposit bonus | 3,801,084 casino no deposit | 3,294,287 neem dan contact | 3,076,119
4	casino no deposit bonus | 3,243,685 online gokkast spelen gratis | 2,414,200 gratis en met geld | 2,409,508 log in of maak | 1,823,552 onbeperkt toegang tot showbytes | 1,773,156
5	spelen gratis en met geld | 2,409,310 gratis onbeperkt toegang tot showbytes | 1,772,911 account aan en mis niks | 1,771,805 niks meer van de sterren.- | 1,765,166 bezoek website meer informatie bekijk | 1,262,665

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.	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				