HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
nld_Latn.jsonl.tsv	8/18/2025	Dutch (nl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
138,651,388	3,073,709,787	1,268,908,825 (41.28 %)	81B	448,143,342,044	419.09 GB

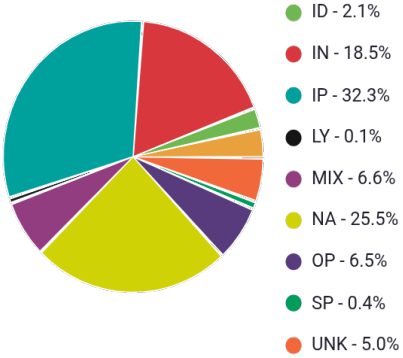
Top 10 domains

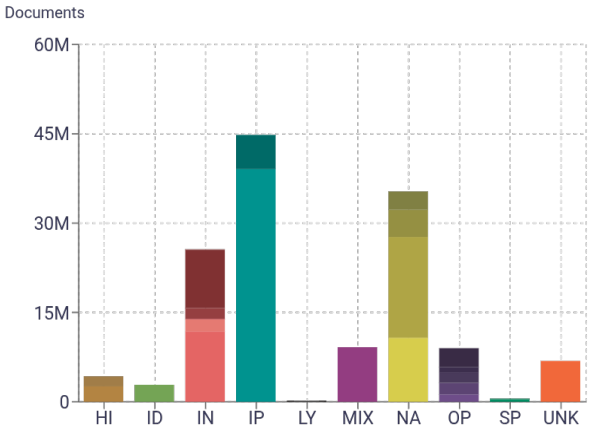
Domain	Docs	% of total
blogspot.com	1.9M	1.37%
wikipedia.org	1.7M	1.26%
blogspot.nl	1.1M	0.80%
knack.be	982K	0.71%
wordpress.com	895K	0.65%
docplayer.nl	697K	0.50%
nrc.nl	535K	0.39%
blogspot.be	488K	0.35%
tripadvisor.nl	369K	0.27%
viva.nl	344K	0.25%

Top 10 TLDs

Domain	Docs	% of total
nl	91M	65.64%
com	19M	13.37%
be	17M	12.12%
org	3.4M	2.48%
net	1.9M	1.38%
eu	1.9M	1.35%
nu	1.1M	0.80%
info	781K	0.56%
de	298K	0.21%
tv	172K	0.12%

Register labels



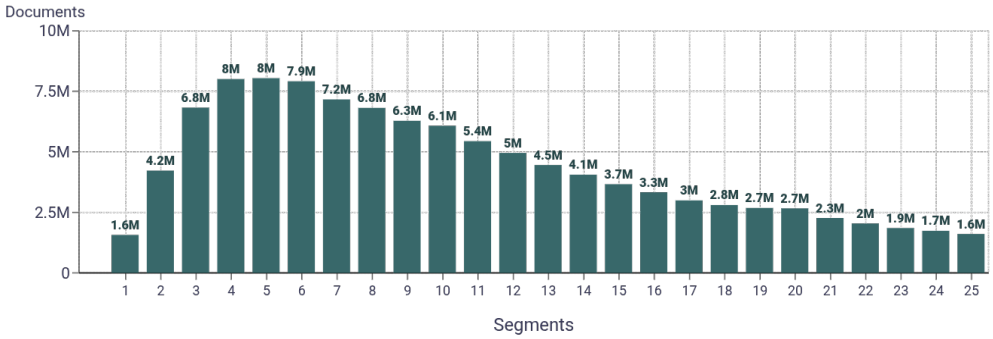


HI_other - 1.9% | MIX - 6.6%
HI_re - 1.2% | NA_nb - 7.8%
ID_other - 2.1% | NA_ne - 12.2%
IN_dtp - 8.5% | NA_other - 3.3%
IN_en - 1.5% | NA_sr - 2.2%
IN_fi - 0.1% | OP_av - 0.9%
IN_lt - 1.3% | OP_ob - 1.4%
IN_other - 7.0% | OP_other - 1.3%
IN_ra - 0.1% | OP_rs - 0.6%
IP_ds - 28.2% | OP_rv - 2.3%
IP_ed - 0.0% | SP_it - 0.3%
IP_other - 4.1% | SP_other - 0.1%
LY_other - 0.1% | UNK - 5.0%

MT:2.3% | 3.1M Documents

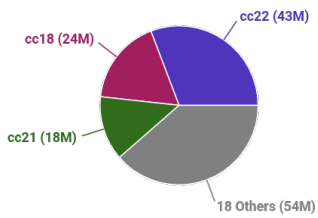
Documents size (in segments)

≤ 25 segments **78.98%** (110M documents)
> 25 segments **21.02%** (29M documents)



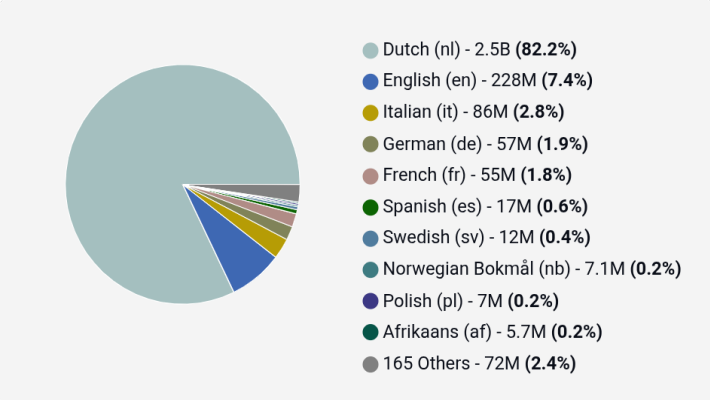
Document collections

CC = 70.96%
IA = 29.04%

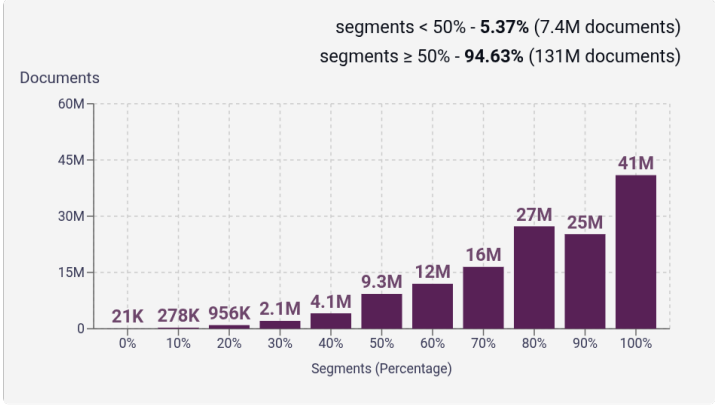


Language Distribution

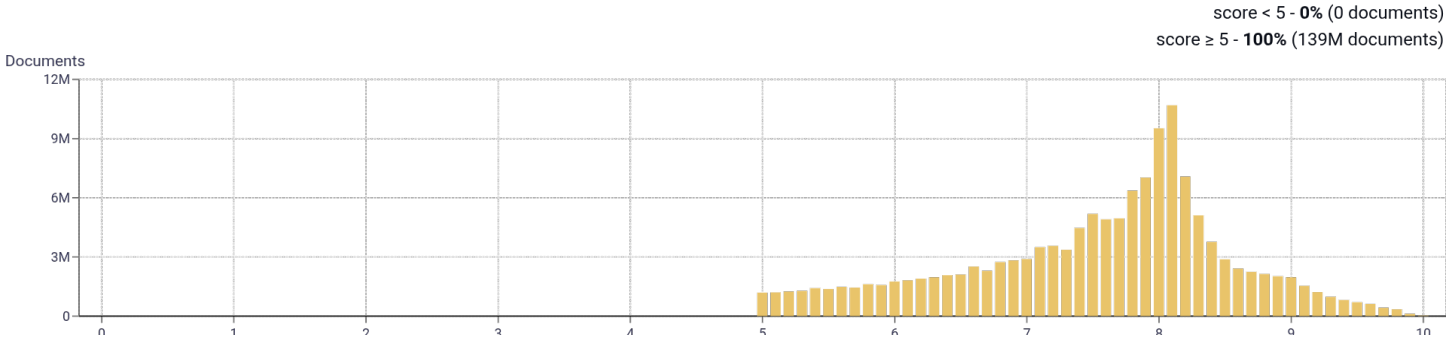
Number of segments in the Dutch (nl) corpus



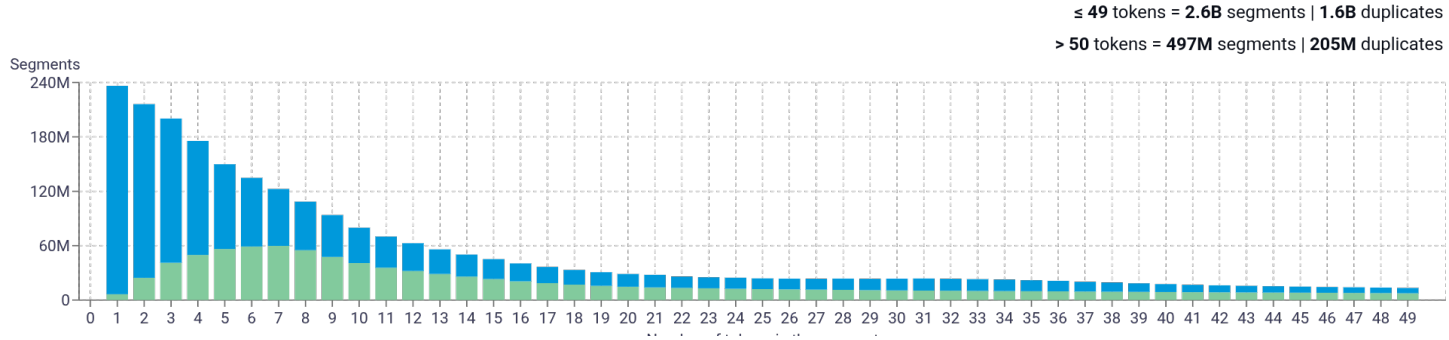
Percentage of segments in Dutch (nl) inside documents



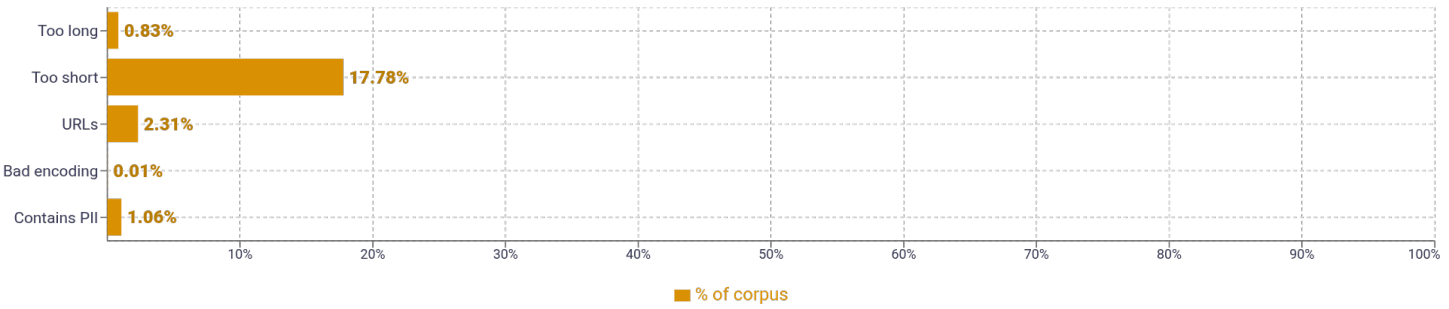
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	we | 242,264,018 · s | 136,165,714 · wel | 130,805,493 · informatie | 115,910,493 · onze | 115,370,728
2	lees verder | 8,952,929 · den haag | 7,602,187 · nadere informatie | 6,972,363 · jaar geleden | 6,508,500 · vorig jaar | 5,629,790
3	af en toe | 4,090,840 · neem dan contact | 2,207,727 · gebruik te maken | 1,894,157 · toe te voegen | 1,573,793 · bedoeld in artikel | 1,551,828
4	bezoek website meer informatie | 1,024,311 · website meer informatie bekijk | 1,021,857 · tweede kamer der staten-generaal | 874,042 · bent u op zoek | 806,125 · contact op te nemen | 746,944
5	contact op met dit bedrijf | 1,144,769 · bezoek website meer informatie bekijk | 1,021,856 · college van burgemeester en wethouders | 496,736 · e-mailadres wordt beveiligd tegen spambots | 496,588 · contact met ons op via | 464,979

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				