HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lim_Latn	9/18/2025	Limburgan

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
339,706	6,560,800	5,167,578 (78.76 %)	220M	1,101,581,541	1.04 GB

Top 10 domains

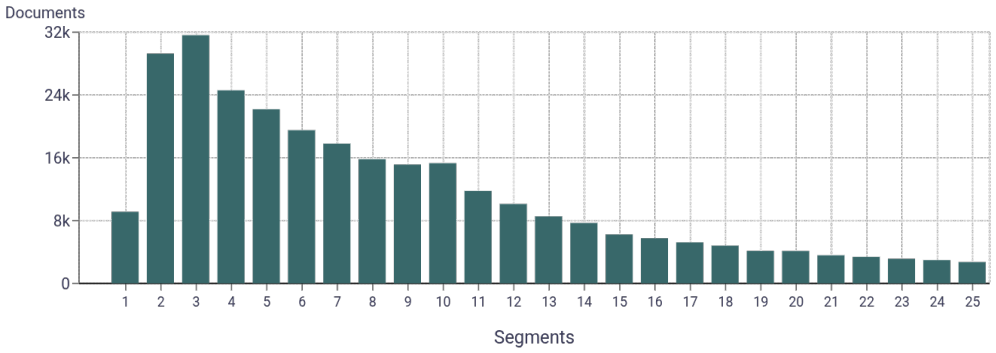
Domain	Docs	% of total
omropfryslan.nl	60K	17.57%
wikipedia.org	53K	15.72%
itnijs.frl	12K	3.61%
limburgslied.nl	4.7K	1.38%
vv-sds.nl	4.4K	1.28%
dbnl.org	3.4K	1.00%
demoanne.nl	3.4K	1.00%
martech.zone	3.1K	0.92%
androidsis.com	2.6K	0.76%
eturbonews.com	2.6K	0.75%

Top 10 TLDs

Domain	Docs	% of total
nl	163K	47.96%
org	65K	19.06%
com	55K	16.23%
frl	22K	6.39%
be	6.3K	1.86%
eu	4.9K	1.44%
net	4.3K	1.26%
zone	3.1K	0.92%
de	3K	0.89%
nu	1.2K	0.35%

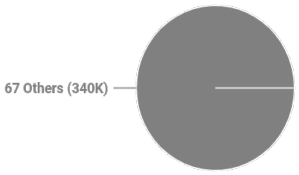
Documents size (in segments)

≤ 25 segments **83.81%** (285K documents)
> 25 segments **16.19%** (55K documents)



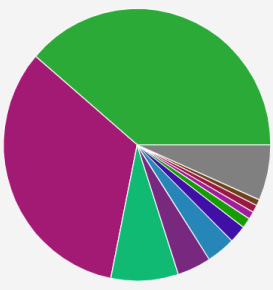
Document collections

CC = 94.88%
IA = 5.12%



Language Distribution

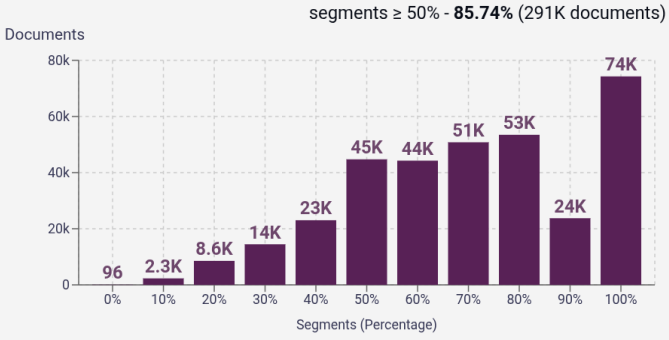
Number of segments in the Limburgan corpus



- Western Frisian - 2.5M **(38.7%)**
- Dutch - 2.2M **(33.1%)**
- English - 535K **(8.2%)**
- German - 269K **(4.1%)**
- French - 227K **(3.5%)**
- Limburgish - 138K **(2.1%)**
- Afrikaans - 80K **(1.2%)**
- Low German - 59K **(0.9%)**
- Italian - 58K **(0.9%)**
- Norwegian Bokmål - 48K **(0.7%)**
- 162 Others - 434K **(6.6%)**

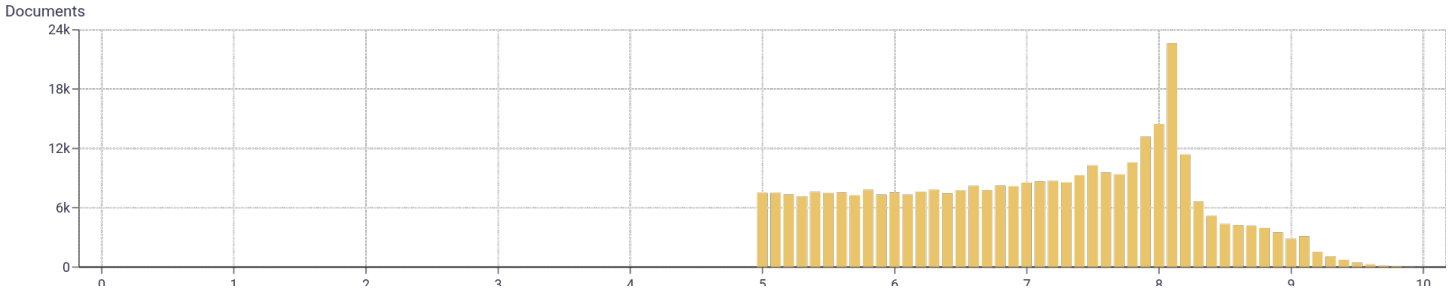
Percentage of segments in Limburgan inside documents

segments < 50% - **14.26%** (48K documents)
segments ≥ 50% - **85.74%** (291K documents)



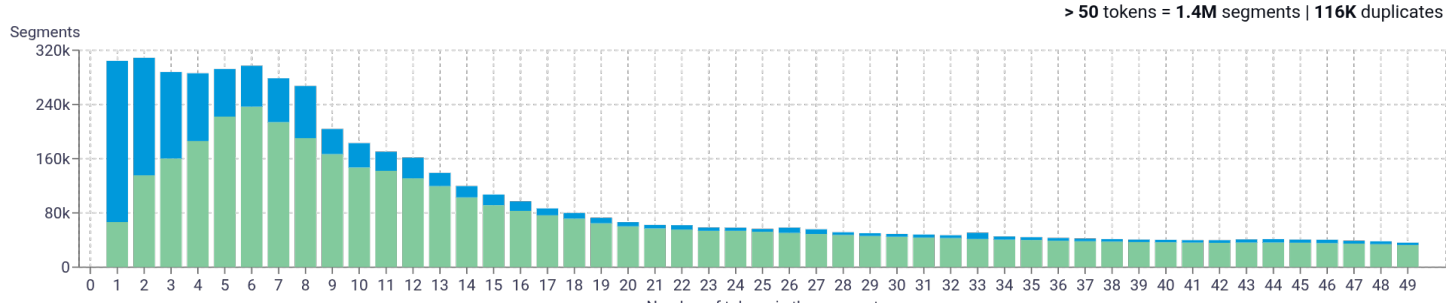
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (340K documents)

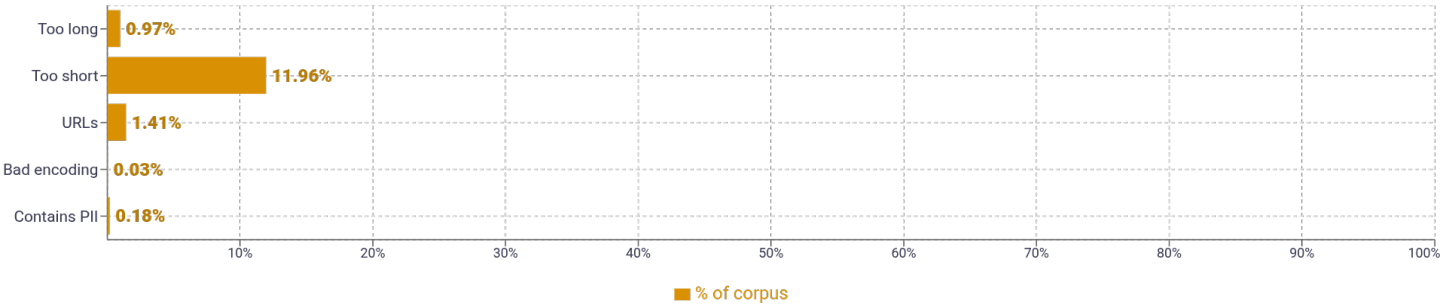


Segment length distribution by token

≤ 49 tokens = **5.2M** segments | **1.3M** duplicates
> 50 tokens = **1.4M** segments | **116K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	in | 4,692,072 · it | 3,601,487 · fan | 3,381,397 · yn | 2,373,430 · t | 2,167,952
2	fan it | 322,068 · yn it | 284,143 · fan in | 188,328 · is in | 186,934 · it is | 180,593
3	in de buurt | 47,477 · it is in | 29,660 · as jo in | 17,201 · is ien fan | 17,107 · body to body | 16,203
4	body to body massage | 11,593 · dit artikel is gesjreve | 10,865 · in amsterdamvereniging van eigenaars | 10,061 · wês de earste om | 8,238 · der binne noch gjin | 7,961
5	wês de earste om kommentaar | 8,226 · jo kinne de earste wêze | 7,865 · mar jo kinne de earste | 7,863 · der binne noch gjin opmerkingen | 7,863 · dit artikel is gesjreve in | 6,671

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				