

General overview

Corpus	Date	Language
hplt-v3-lim_Latn	9/18/2025	Limburgish (li)

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

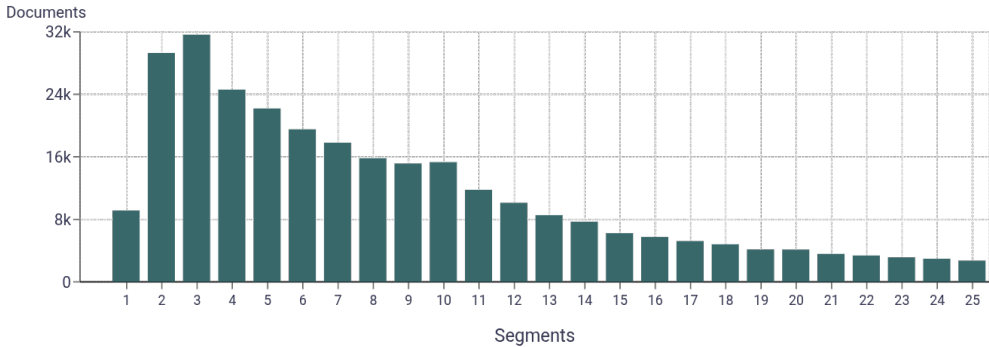
Top 10 domains

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%

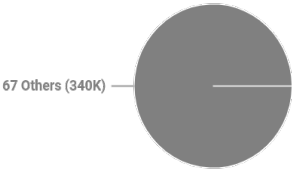
Documents size (in segments) ⓘ



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

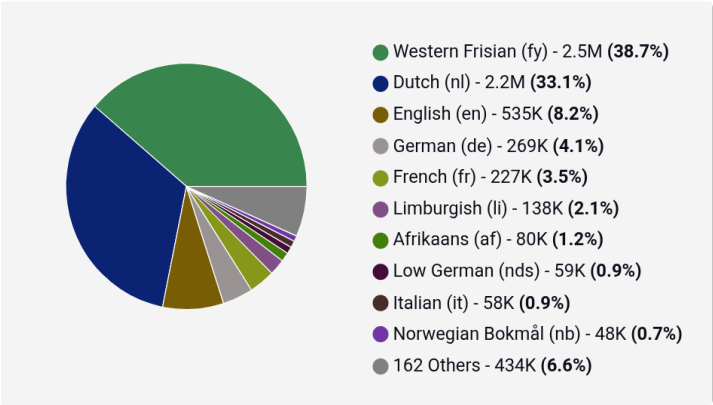
Document collections

CC = **94.88%**
IA = **5.12%**

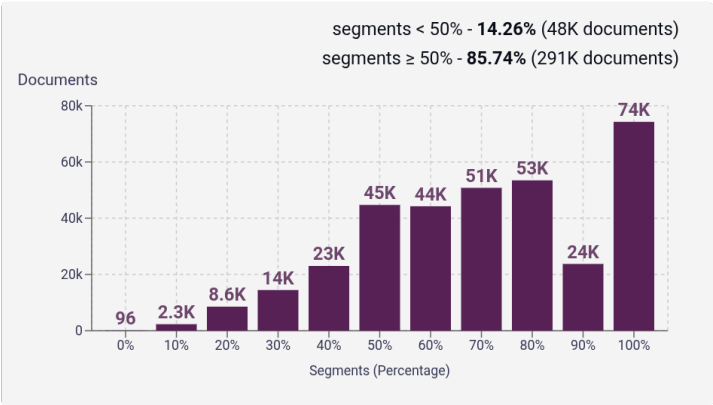


Language Distribution

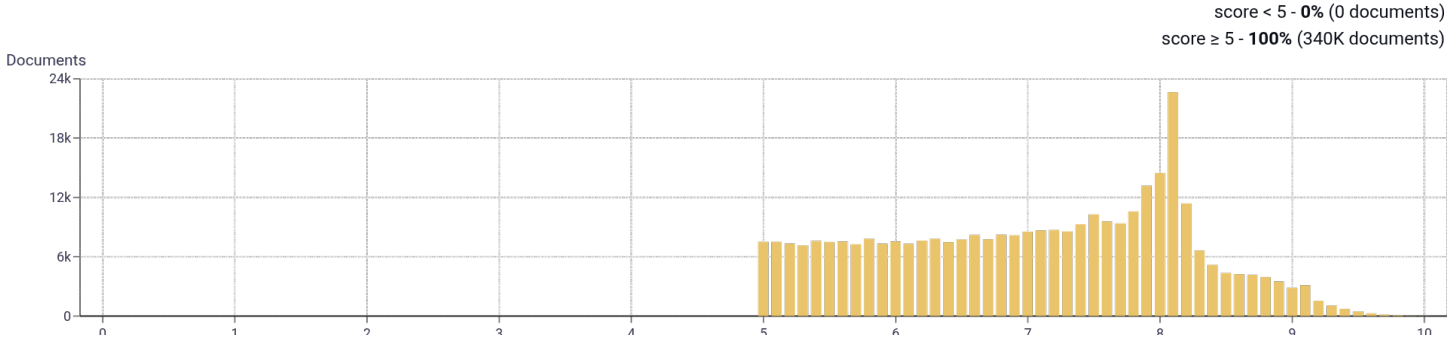
Number of segments in the Limburgish (li) corpus



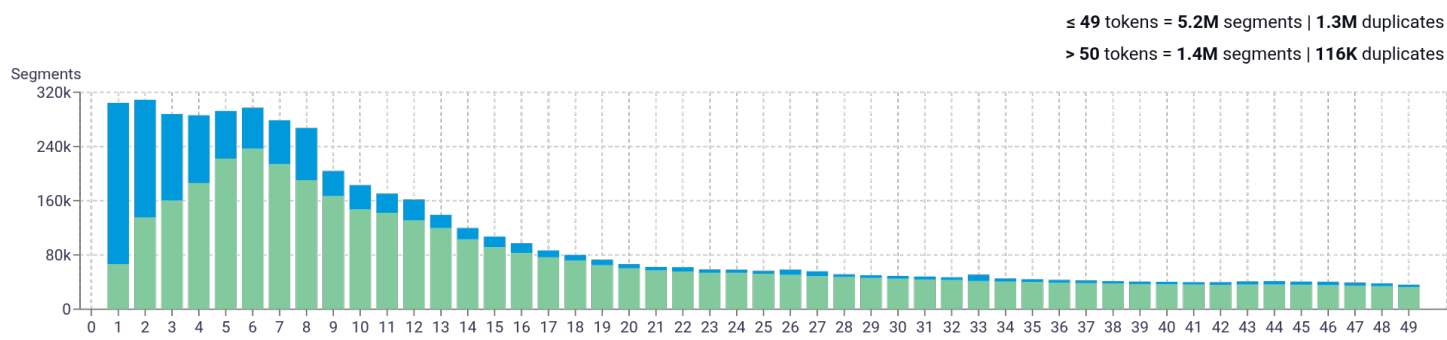
Percentage of segments in Limburgish (li) inside documents



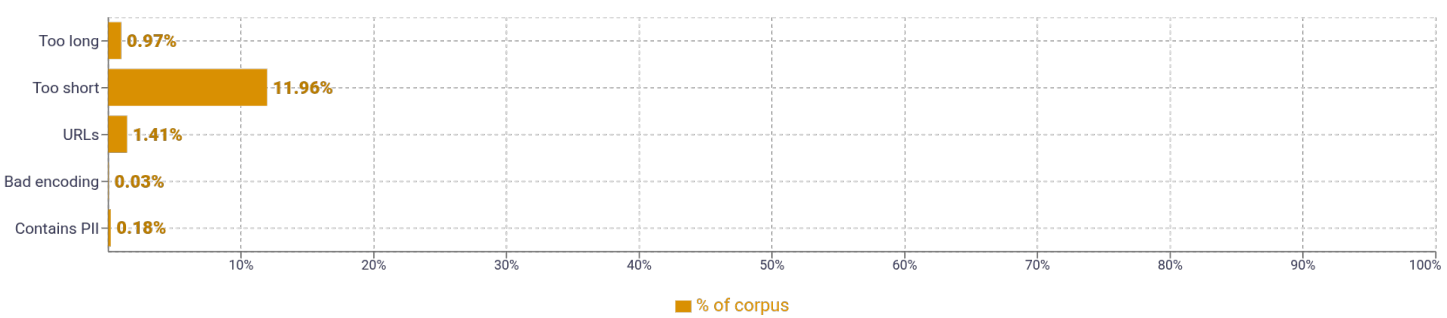
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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				