

General overview

Corpus	Date	Language
hplt-v3-ltz_Latn	9/18/2025	Luxembourgish (lb)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
407,481	7,965,654	6,135,723 (77.03 %)	241M	1,331,272,155	1.28 GB

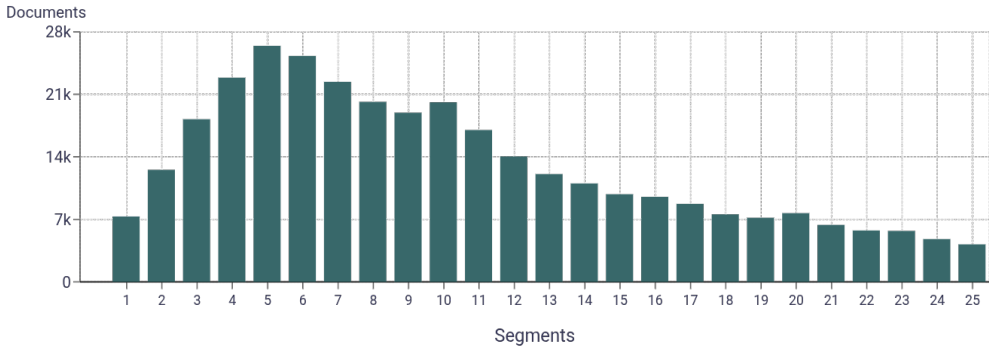
Top 10 domains

Domain	Docs	% of total
rtl.lu	37K	9.03%
wikipedia.org	28K	6.92%
100komma7.lu	14K	3.45%
moien.lu	12K	3.00%
eldo.lu	9.9K	2.42%
martech.zone	5.1K	1.25%
adr.lu	4.3K	1.05%
estadisticasgra...	2.8K	0.69%
meteoboulaide.com	2.8K	0.68%
piraten.lu	2.6K	0.63%

Top 10 TLDs

Domain	Docs	% of total
com	166K	40.84%
lu	154K	37.67%
org	38K	9.26%
net	6.4K	1.58%
eu	6.2K	1.51%
zone	5.1K	1.25%
de	4K	0.97%
fr	1.6K	0.39%
news	1.5K	0.36%
online	1.4K	0.35%

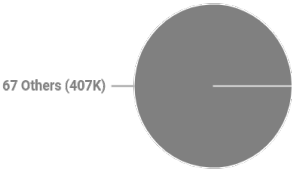
Documents size (in segments) ⓘ



≤ 25 segments **79.97%** (326K documents)
> 25 segments **20.03%** (82K documents)

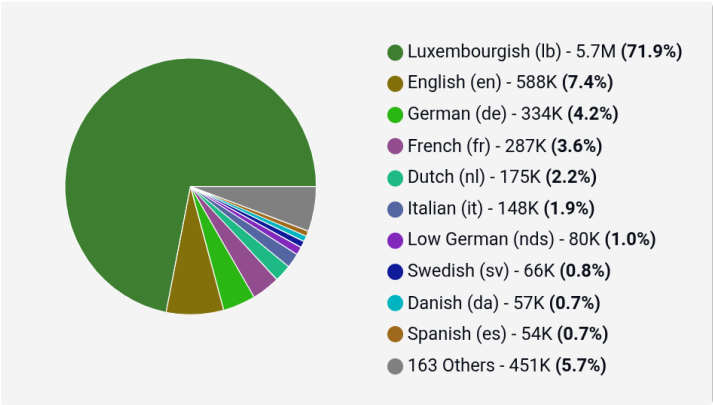
Document collections

CC = 95.19%
IA = 4.81%

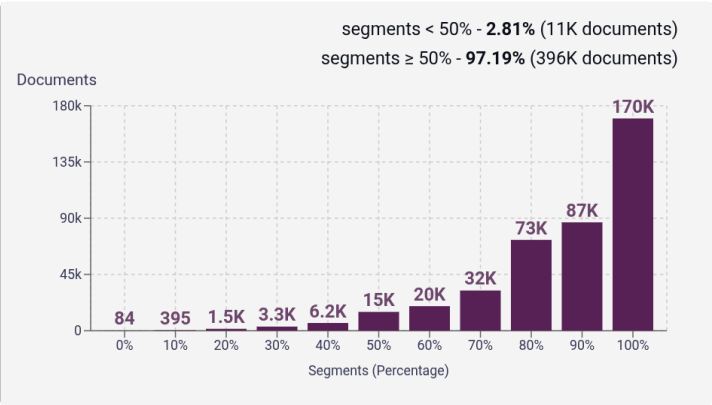


Language Distribution

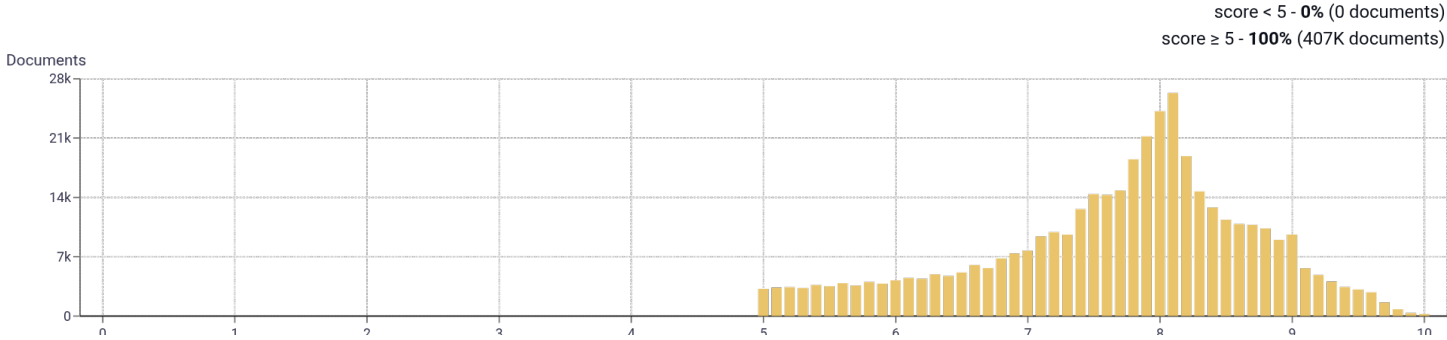
Number of segments in the Luxembourgish (lb) corpus



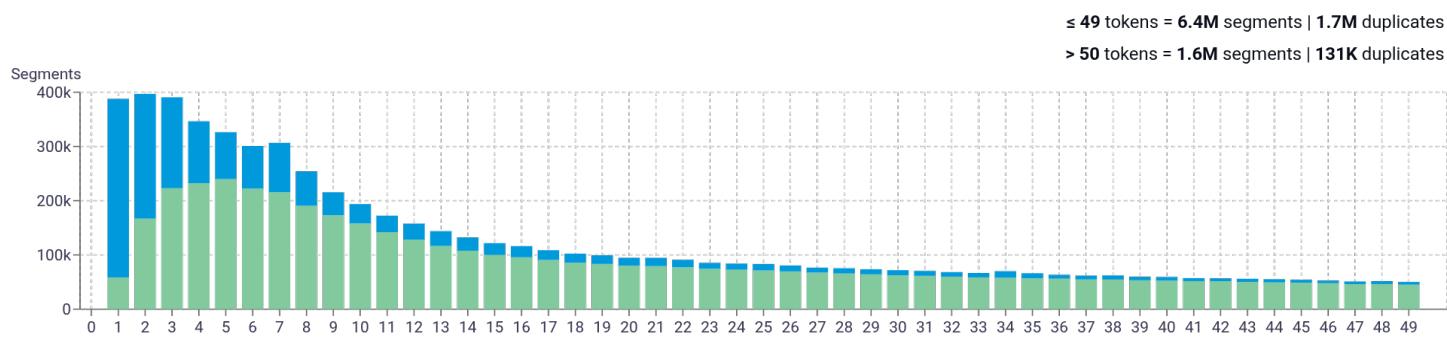
Percentage of segments in Luxembourgish (lb) inside documents



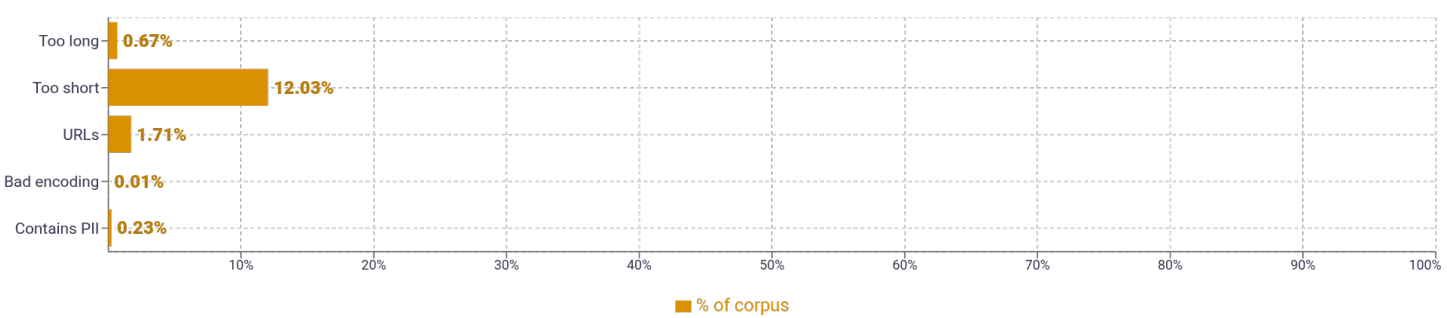
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	d 4,771,521wéi 1,281,942dës 412,048vill 407,480joer 362,740	
2	wéi d 73,923quelltext änneren 63,871well d 27,688iwwert d 26,716egal ob 26,569	
3	wéi och ëmmer 39,855einfach ze benotzen 12,710zur selwechter zäit 9,788lues a lues 8,838aart a weis 8,575	
4	éischt fir ze kommentéieren 9,885iwwerdeems déi lescht noriicht 5,055invité vun der redaktioun 3,281ëffentlech-rechtliche radio zu lëtzebuerg 3,269proposéiert programmer op lëtzebuergesch 3,252	
5	iwwerdeems déi lescht noriicht uginnt 5,055eenzegen ëffentlech-rechtliche radio zu lëtzebuerg 3,253déifsttemperaturen an der nächster nuecht 2,338nächstes nuecht leie bei ronn 2,337childhood story plus untold biografie 1,658	

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				