HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-lmo_Latn	9/18/2025	Lombard

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
116,732	1,607,907	1,249,957 (77.74 %)	62M	288,382,712	285.12 MB

Top 10 domains

Domain	Docs	% of total
rtr.ch	59K	50.28%
wikipedia.org	17K	14.69%
lausc.it	3.7K	3.14%
wikisource.org	3.3K	2.79%
provinzia.bz.it	1.5K	1.31%
noeles.info	1.3K	1.14%
engadinerpost.ch	1.2K	1.06%
provinz.bz.it	783	0.67%
playsuisse.ch	709	0.61%
admin.ch	610	0.52%

Top 10 TLDs

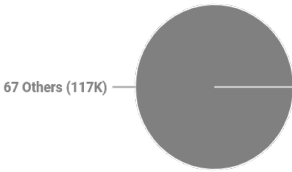| Domain	Docs	% of total
ch	76K	64.83%
org	22K	18.75%
it	8.9K	7.58%
com	2.9K	2.45%
bz.it	2.8K	2.44%
info	1.9K	1.62%
net	684	0.59%
tn.it	290	0.25%
eu	257	0.22%
is	169	0.14%

Documents size (in segments) ⓘ
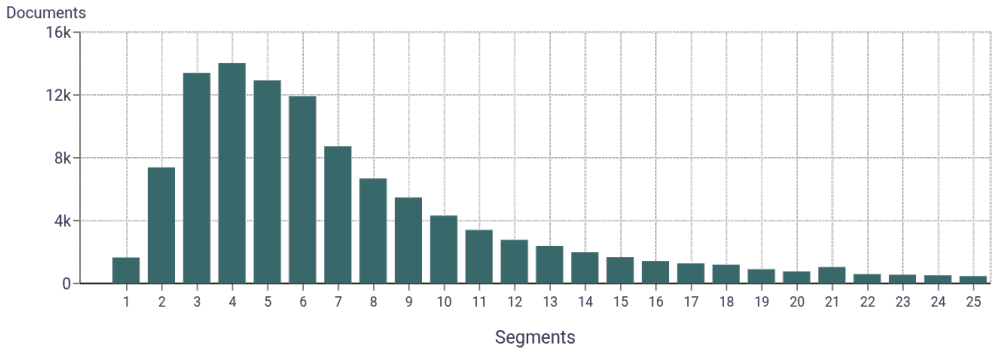
≤ 25 segments **92.03%** (107K documents)
\> 25 segments **7.97%** (9.3K documents)
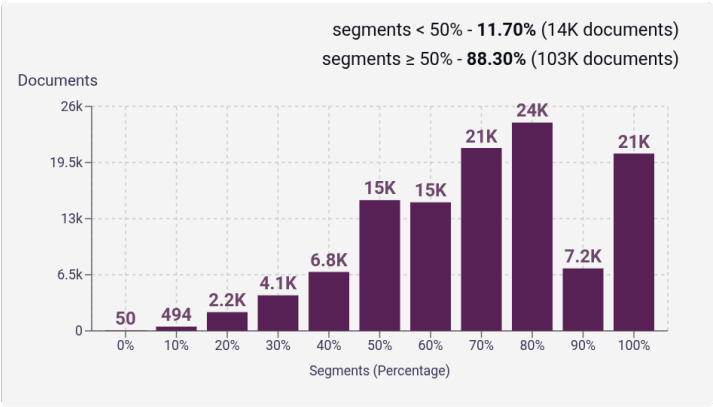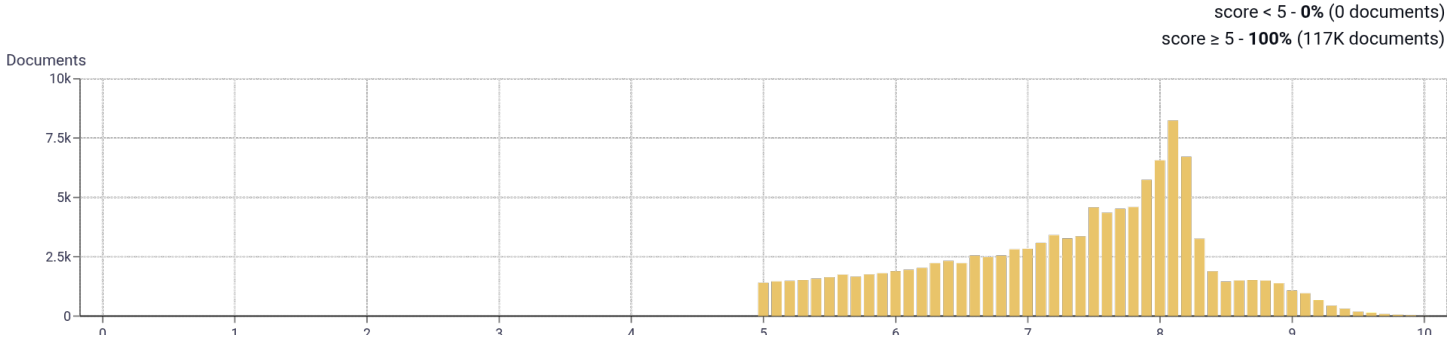

Document collections

CC = **97.17%**
IA = **2.83%**

67 Others (117K)

Language Distribution

Number of segments in the Lombard corpus

- Italian - 475K **(29.6%)**
- Romansh - 215K **(13.4%)**
- Catalan - 138K **(8.6%)**
- Spanish - 87K **(5.4%)**
- English - 83K **(5.2%)**
- German - 80K **(5.0%)**
- French - 66K **(4.1%)**
- Lombard - 66K **(4.1%)**
- null - 45K **(2.8%)**
- Piedmontese - 39K **(2.4%)**
- 163 Others - 315K **(19.6%)**
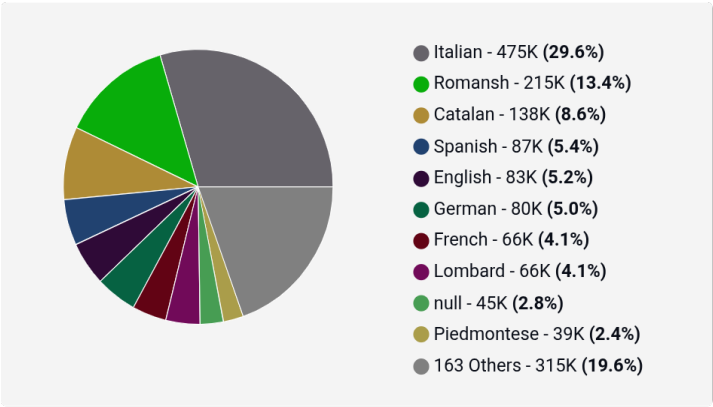
Percentage of segments in Lombard inside documents

segments < 50% - **11.70%** (14K documents)
segments ≥ 50% - **88.30%** (103K documents)

Distribution of documents by document score

Documents

Segment length distribution by token

Segments

Segment noise distribution

Too long	**1.40%**
Too short	**11.58%**
URLs	**1.41%**
Bad encoding	**0.02%**
Contains PII	**0.25%**

■ % of corpus
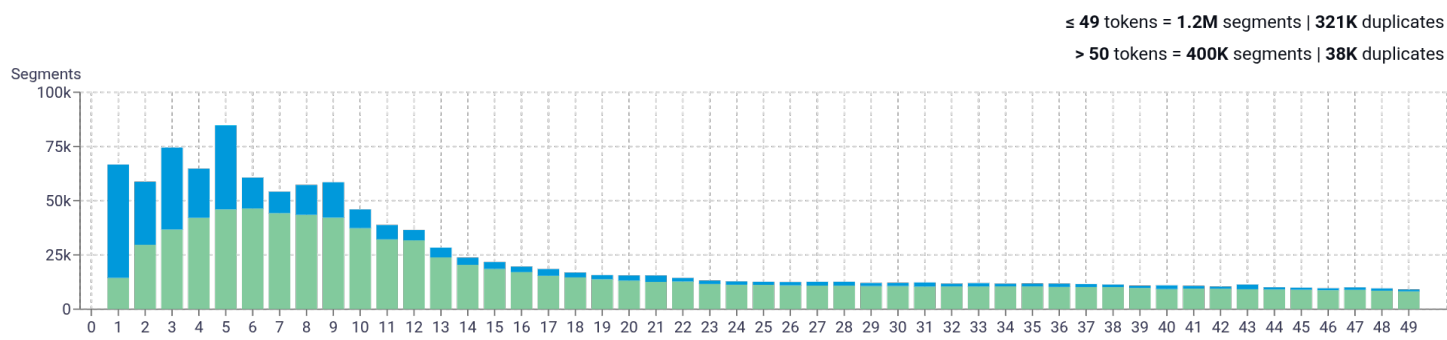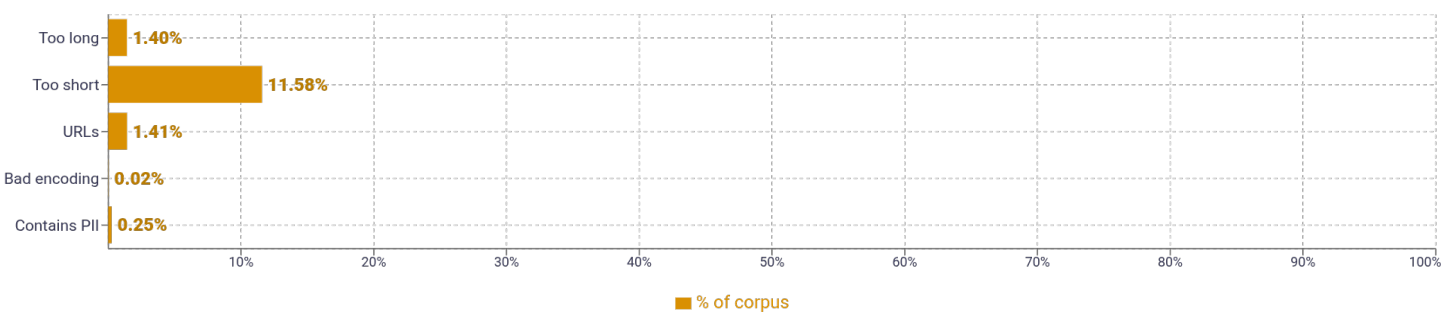
Frequent n-grams

SIZE	N-GRAMS					
1	l | 1,214,669	il | 705,062	en | 515,691	i | 440,497	d | 397,245	
2	en il | 80,270	en l | 37,332	ed il | 21,678	cun il | 20,858	cun l | 20,119	
3	qe qe qe | 18,653	modifitgar il code | 11,891	il cussegl federal | 7,663	en vigur dapi | 5,801	ultra da quai | 5,795	
4	qe qe qe qe | 18,643	en vigur dapi il | 5,358	œ œ œ œ | 4,189	versiun tenor la cifra | 3,164	i da la lf | 2,538	
5	qe qe qe qe qe | 18,633	œ œ œ œ œ | 3,247	cifra i da la lf | 2,538	i da la lf dals | 2,526	versiun tenor la cifra i | 2,318	

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				