HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-oci_Latn	9/18/2025	Occitan (oc)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
106,458	2,070,795	1,441,385 (69.61 %)	75M	354,431,457	350.52 MB

Top 10 domains

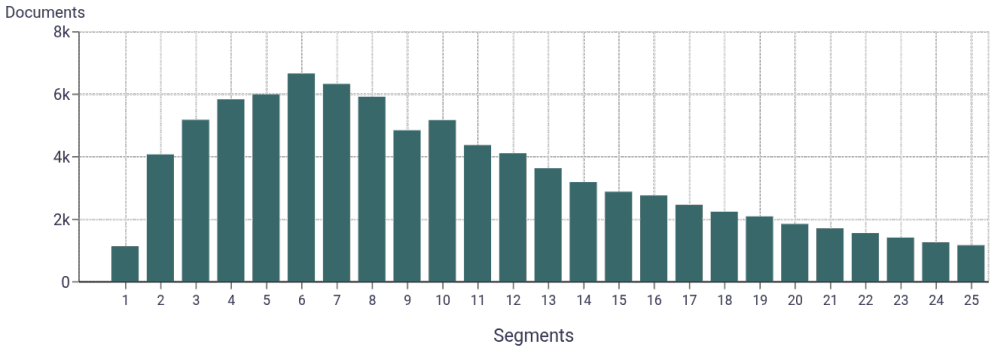
Domain	Docs	% of total
wikipedia.org	27K	25.49%
jornalet.com	17K	15.83%
blogspot.com	5.2K	4.92%
lodiari.com	3.9K	3.67%
occitanparis.com	2.3K	2.18%
sapiencia.eu	2.2K	2.07%
occitanica.eu	2K	1.84%
conselharan.org	1.6K	1.52%
ieo06.com	1.4K	1.35%
gencat.cat	1.3K	1.21%

Top 10 TLDs

Domain	Docs	% of total
com	42K	39.36%
org	40K	37.87%
fr	9.1K	8.50%
eu	6.4K	6.02%
cat	5K	4.69%
net	1.2K	1.09%
info	1K	0.96%
es	340	0.32%
it	331	0.31%
click	199	0.19%

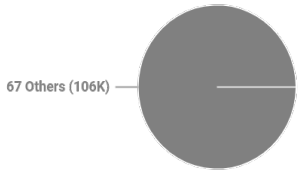
Documents size (in segments)

≤ 25 segments **82.6%** (88K documents)
> 25 segments **17.4%** (19K documents)



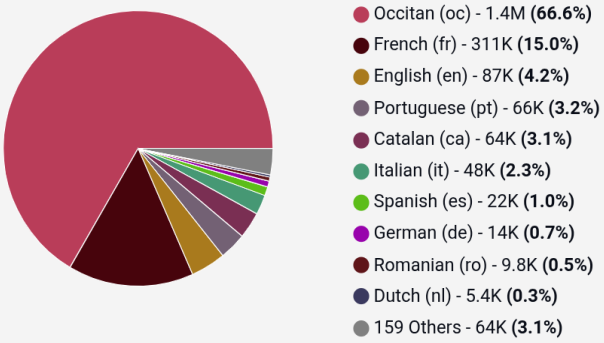
Document collections

CC = 92.03%
IA = 7.97%



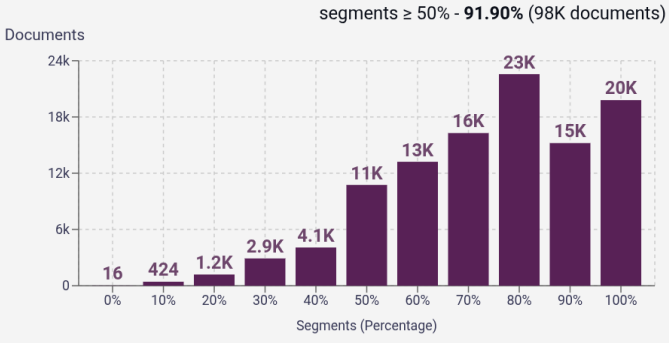
Language Distribution

Number of segments in the Occitan (oc) corpus



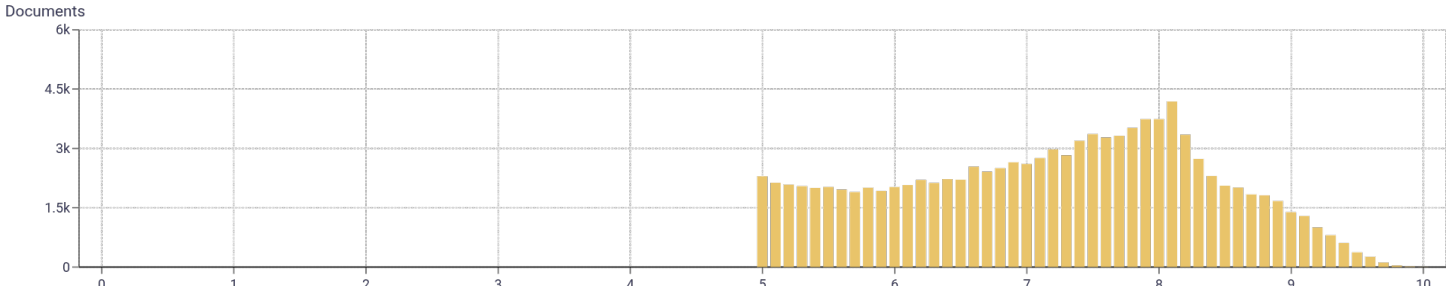
Percentage of segments in Occitan (oc) inside documents

segments < 50% - **8.10%** (8.6K documents)
segments ≥ 50% - **91.90%** (98K documents)



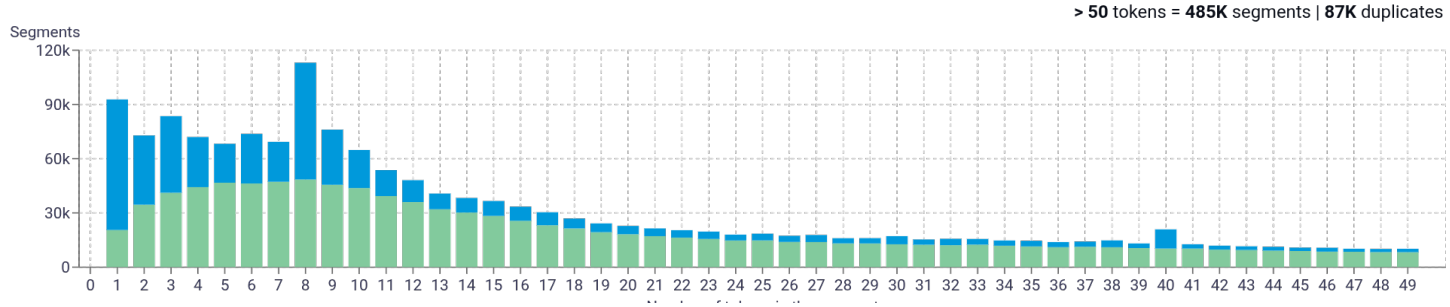
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (106K documents)

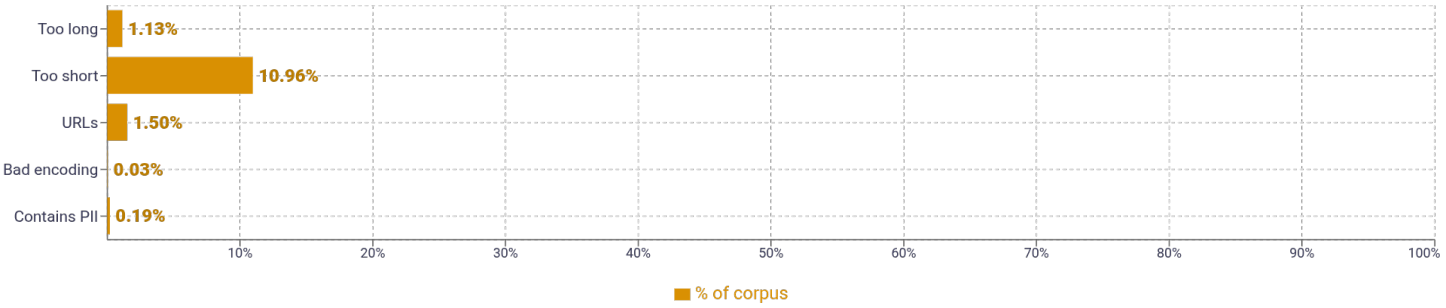


Segment length distribution by token

≤ 49 tokens = **1.6M** segments | **545K** duplicates
> 50 tokens = **485K** segments | **87K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	e | 1,552,081 · d | 1,186,140 · l | 1,166,485 · en | 902,002 · qu | 429,126
2	e d | 35,727 · e l | 35,082 · en occitan | 29,179 · mai d | 19,948 · e en | 18,283
3	modificar la font | 107,806 · o de l | 8,937 · mai que mai | 8,322 · recebre per e | 8,084 · clicar sul ligam | 8,073
4	encara clicar sul ligam | 8,049 · comentari es a mand | 8,049 · clicar sul ligam qu | 8,049 · cal encara clicar sul | 8,049 · anatz recebre per e | 8,049
5	vòstre comentari es a mand | 8,049 · terminar lo procès de validacion | 8,049 · encara clicar sul ligam qu | 8,049 · comentari es a mand d | 8,049 · cal encara clicar sul ligam | 8,049

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				