HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
oci_Latn.jsonl.tsv	12/5/2024	Occitan (oc)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
189,906	4,194,906	1,267,789 (30.22 %)	133M	631,394,992	621.05 MB

Top 10 domains

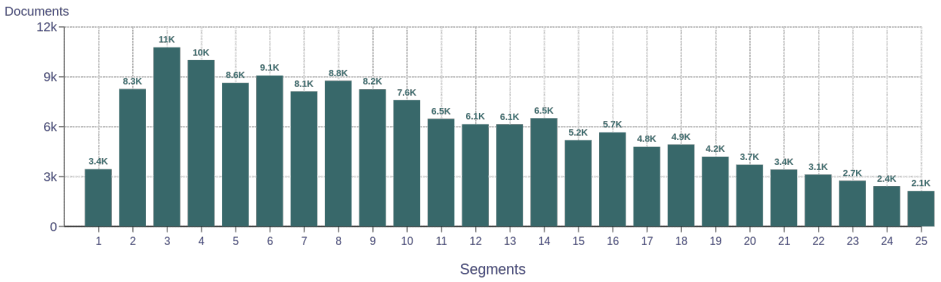
Domain	Docs	% of total
wikipedia.org	118K	62.25
jornalet.com	12K	6.16
blogspot.com	4.1K	2.14
vincent-lefrancois.com	3.1K	1.62
occitanparis.com	2.3K	1.21
aquodaqui.info	1.8K	0.94
blogspot.be	1.6K	0.82
sudouest.fr	1.5K	0.81
dordognelibre.fr	1.2K	0.63
espaci-occitan.com	1.1K	0.60

Top 10 TLDs

Domain	Docs	% of total
org	128K	67.17
com	36K	18.95
fr	11K	5.77
eu	4.6K	2.42
cat	2.8K	1.48
info	1.9K	1.02
be	1.6K	0.83
net	1.4K	0.75
com.mx	868	0.46
es	414	0.22

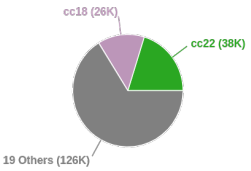
Documents size (in segments)

<= 25 segments **79.23%** (150K documents)
> 25 segments **20.77%** (39K documents)



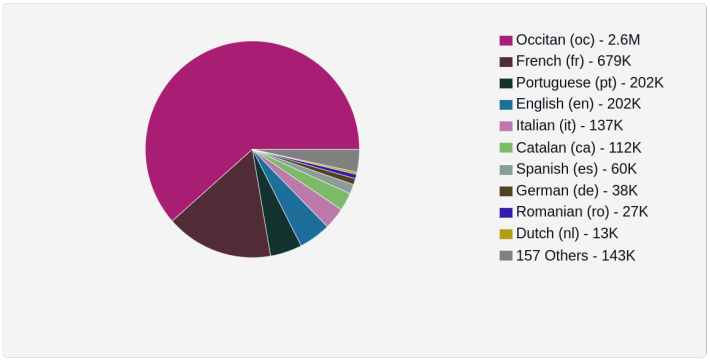
Documents by collection

CC = 65.70%
IA = 34.30%

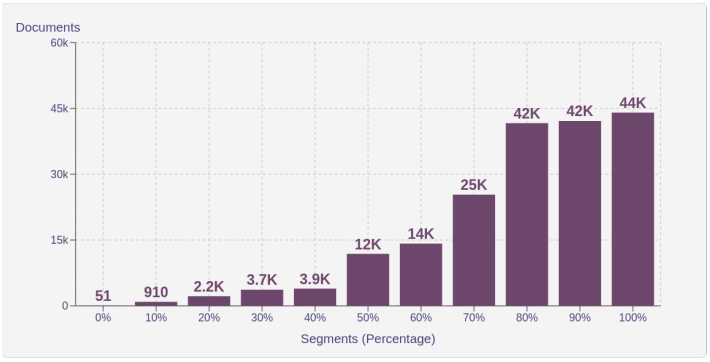


Language Distribution

Number of segments in the Occitan (oc) corpus

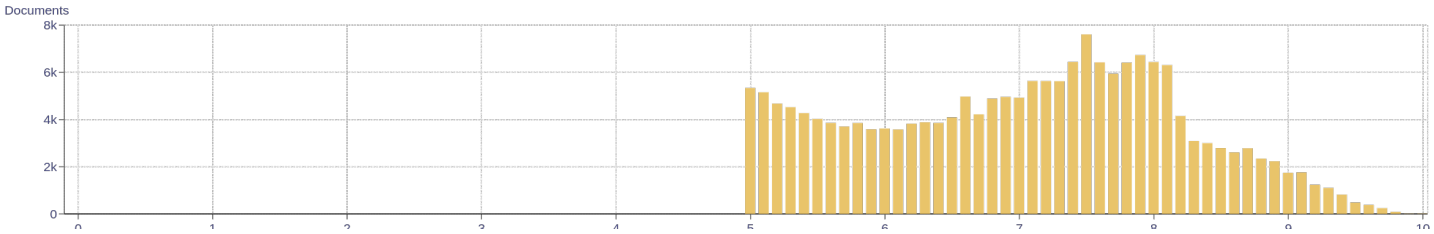


Percentage of segments in Occitan (oc) inside documents



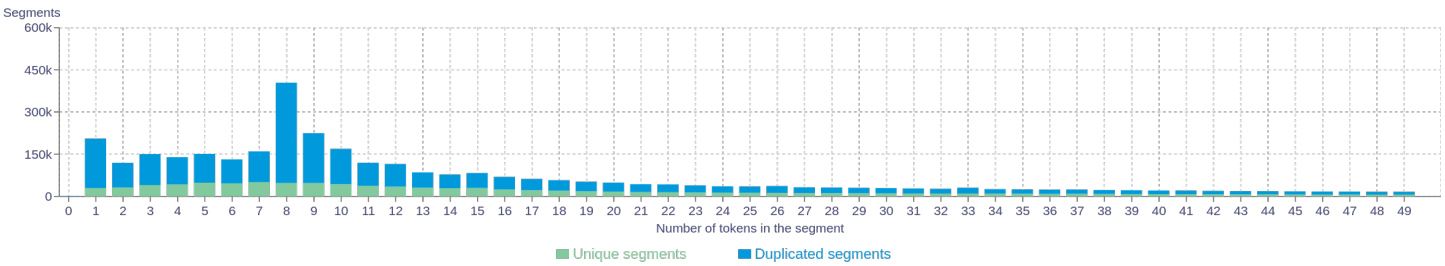
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (190K documents)

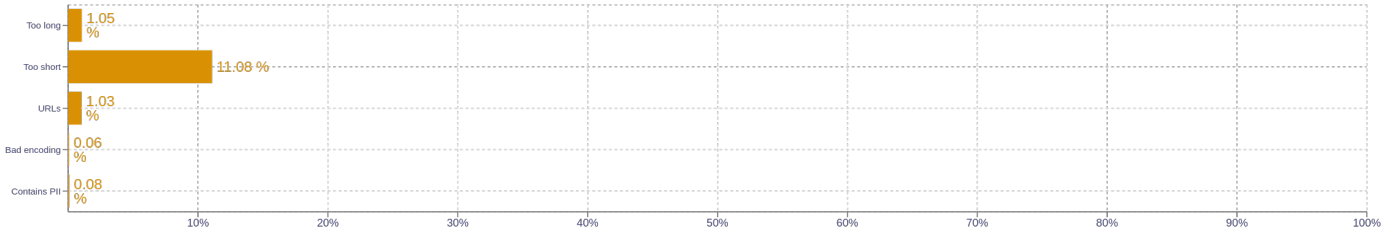


Segment length distribution by token

≤ 49 tokens = **972K** segments | **2.4M** duplicates
> 50 tokens = **803K** segments | **507K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	e | 2490409 · l | 2036095 · d | 1965628 · en | 1652824 · modificar | 1171034
2	e l | 65497 · e d | 61088 · en francés | 35377 · e en | 31816 · en occitan | 30570
3	modificar la font | 568442 · e la region | 26962 · luòcs e monuments | 15517 · e de l | 14941 · amb la comuna | 14525
4	situada dins lo departament | 22448 · ligadas amb la comuna | 14392 · mail et un site | 10245 · et un site internet | 10245 · aquesta pagina concernís l | 8190
5	personalitats ligadas amb la comuna | 14381 · mail et un site internet | 10245 · situada dins lo departament d | 5805 · des noms de lieux en | 5259 · noms de lieux en france | 5255

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