HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-eus_Latn	9/18/2025	Basque (eu)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,219,485	55,906,127	37,094,503 (66.35 %)	1.5B	9,493,741,284	8.88 GB

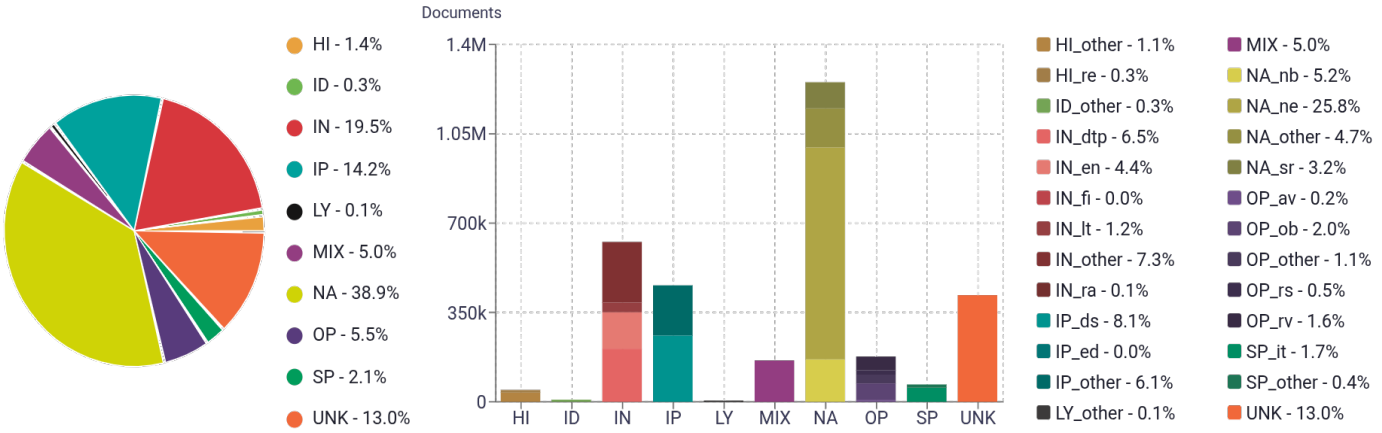
Top 10 domains

Domain	Docs	% of total
wikipedia.org	122K	3.79%
berria.eus	91K	2.82%
argia.eus	80K	2.47%
hitza.eus	73K	2.28%
goiena.eus	73K	2.26%
euskadi.eus	72K	2.23%
consumer.es	66K	2.04%
blogspot.com	52K	1.61%
eitb.eus	51K	1.57%
naiz.eus	30K	0.94%

Top 10 TLDs

Domain	Docs	% of total
eus	1.7M	53.61%
com	668K	20.74%
org	312K	9.70%
es	195K	6.06%
net	101K	3.13%
info	46K	1.44%
eu	24K	0.75%
biz	21K	0.66%
com.es	15K	0.47%
fr	11K	0.35%

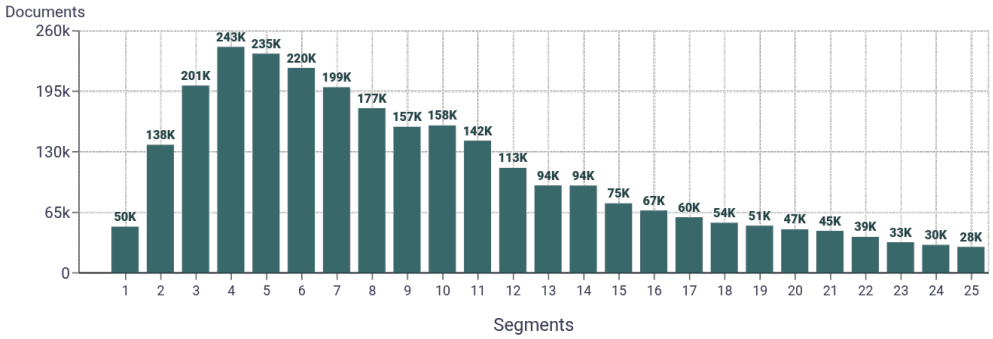
Register labels



MT:7.7% | 249K Documents

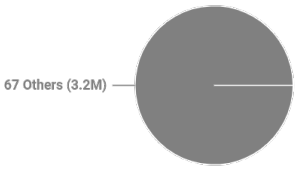
Documents size (in segments)

≤ 25 segments **85.35%** (2.7M documents)
> 25 segments **14.65%** (472K documents)



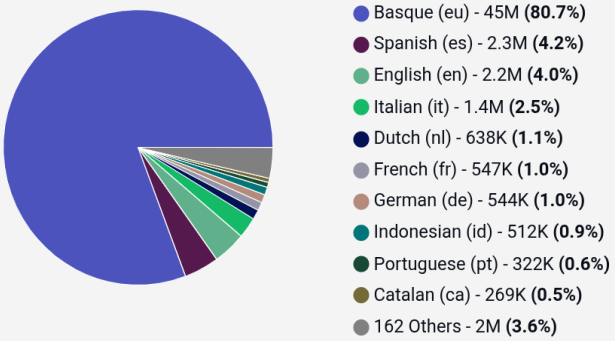
Document collections

CC = 93.69%
IA = 6.31%

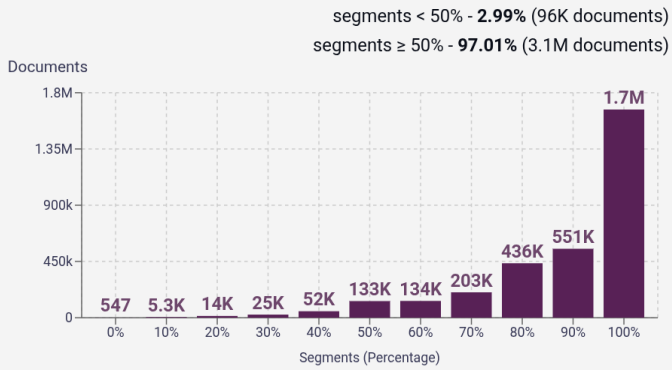


Language Distribution

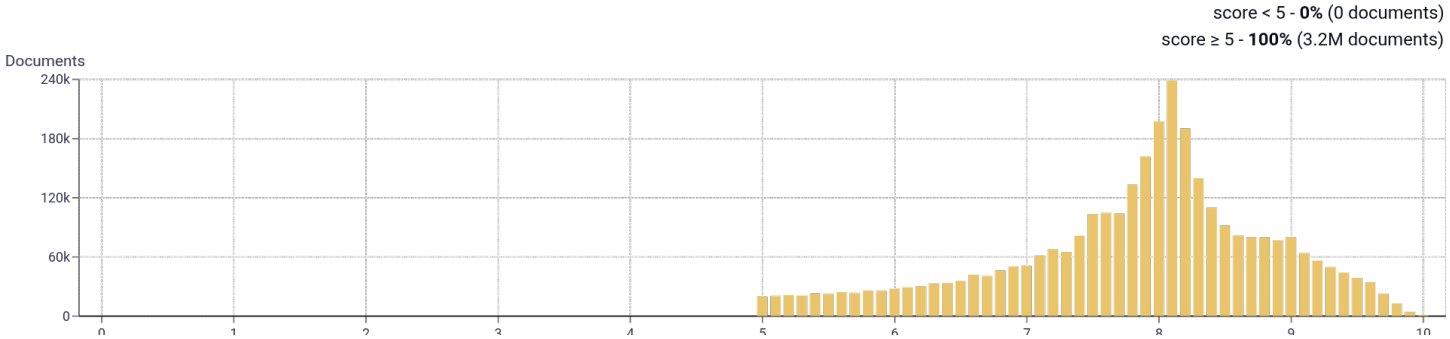
Number of segments in the Basque (eu) corpus



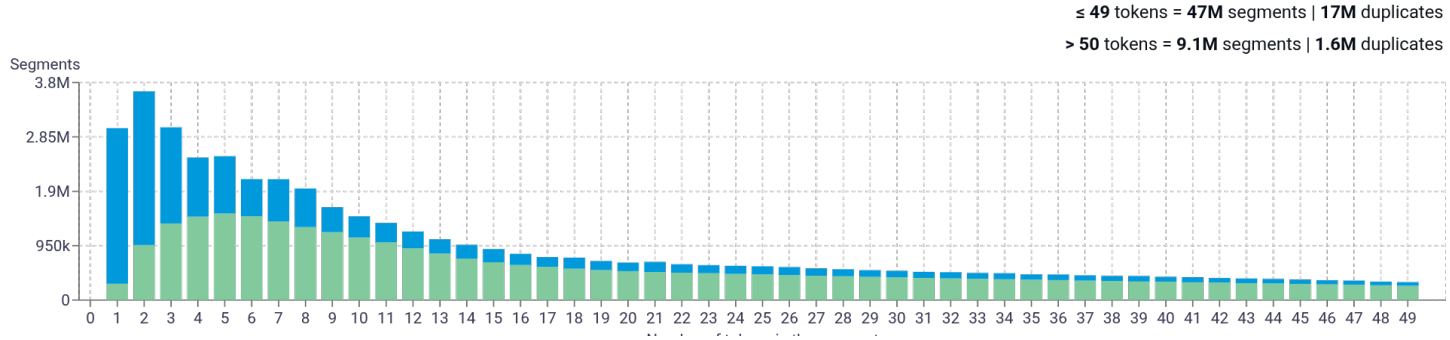
Percentage of segments in Basque (eu) inside documents



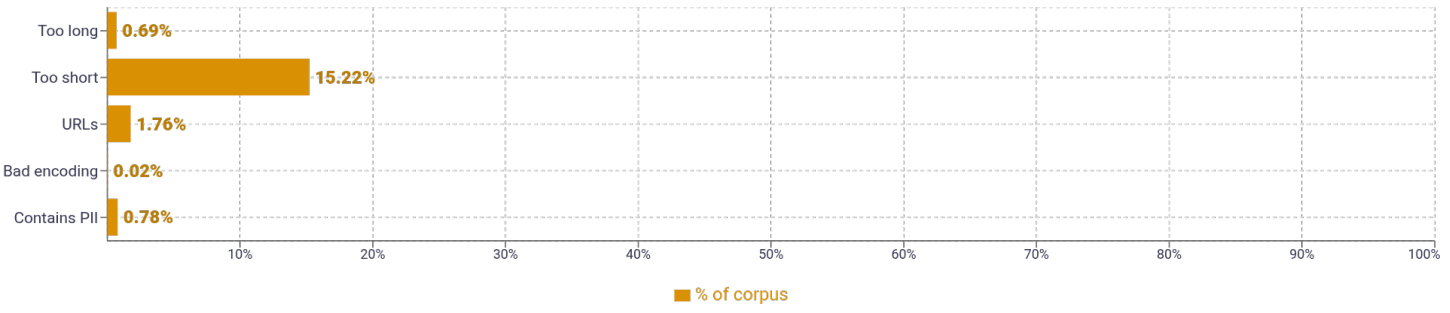
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	behar | 3,959,950 · de | 3,717,173 · izango | 3,501,494 · egiten | 3,045,197 · nahi | 2,694,449
2	de la | 504,972 · ahal izango | 496,538 · euskal herriko | 404,605 · parte hartu | 351,513 · iturburu kodea | 329,032
3	aldatu iturburu kodea | 328,509 · euskal autonomia erkidegoko | 77,923 · parte hartzen duten | 47,084 · bertan behera utzi | 36,874 · parte hartu nahi | 34,670
4	asteko gai hautatuekin osatutako | 24,426 · kronikak zure posta elektronikoan | 24,425 · gai hautatuekin osatutako albiste | 24,425 · hautatuekin osatutako albiste buletina | 24,424 · whatsapp edo telegram bidez | 22,859
5	iritziak eta kronikak zure posta | 24,425 · asteko gai hautatuekin osatutako albiste | 24,425 · gai hautatuekin osatutako albiste buletina | 24,424 · whatsapp edo telegram bidez jaso | 22,774 · azken ordukoak whatsapp edo telegram | 22,210

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				