HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ayr_Latn	9/18/2025	Central Aymara

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7,449	120,121	69,397 (57.77 %)	3.3M	19,688,536	19.94 MB

Top 10 domains

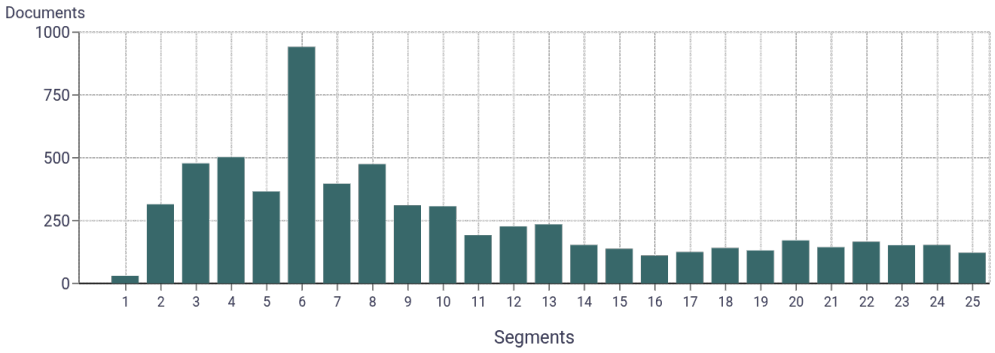
Domain	Docs	% of total
globalvoices.org	2.6K	34.43%
globalvoicesonl...	1.7K	23.13%
jw.org	449	6.03%
wikipedia.org	360	4.83%
phajsiwiphala.cl	335	4.50%
lyfta.app	314	4.22%
radiosangabriel...	296	3.97%
boliviatv.bo	211	2.83%
wol-children.net	148	1.99%
bibles.org	146	1.96%

Top 10 TLDs

Domain	Docs	% of total
org	5.4K	71.97%
org.bo	392	5.26%
cl	357	4.79%
com	355	4.77%
app	315	4.23%
bo	218	2.93%
net	169	2.27%
ru	61	0.82%
pe	46	0.62%
tv.bo	18	0.24%

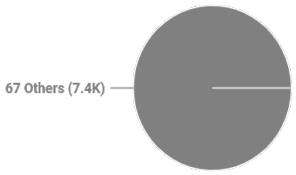
Documents size (in segments)

≤ 25 segments **87.06%** (6.5K documents)
> 25 segments **12.94%** (964 documents)



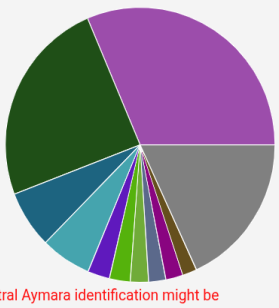
Document collections

CC = 77.77%
IA = 22.23%



Language Distribution

Number of segments in the Central Aymara corpus

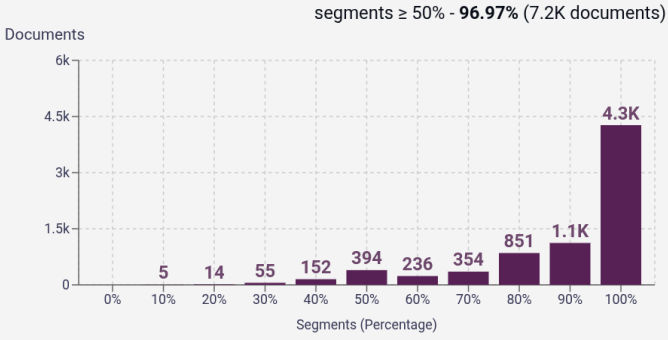


*Central Aymara identification might be inaccurate because it is not supported by FastSpell.

- Spanish - 38K **(31.4%)**
- English - 29K **(24.5%)**
- Swahili - 8.1K **(6.8%)**
- Indonesian - 7.3K **(6.1%)**
- Quechua - 3.2K **(2.6%)**
- Breton - 3K **(2.5%)**
- Filipino - 2.6K **(2.2%)**
- Italian - 2.5K **(2.1%)**
- German - 2.5K **(2.1%)**
- Basque - 2.1K **(1.7%)**
- 112 Others - 22K **(18.1%)**

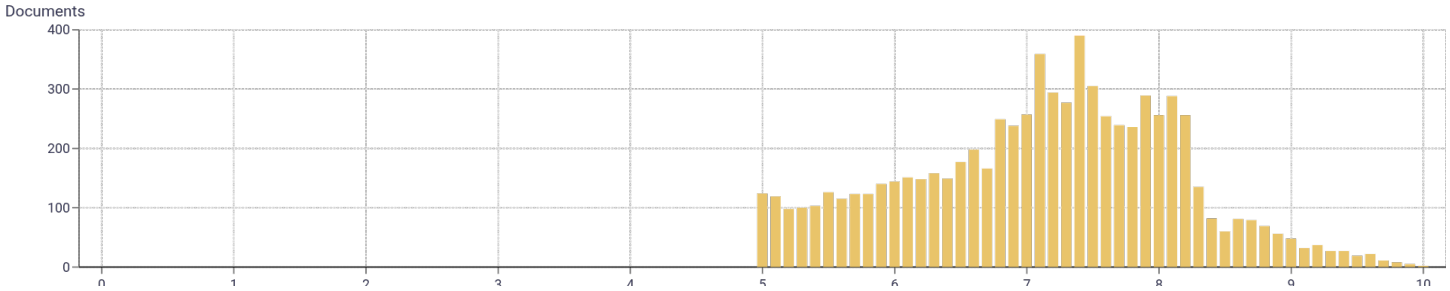
Percentage of segments in Central Aymara inside documents

segments < 50% - **3.03%** (226 documents)
segments ≥ 50% - **96.97%** (7.2K documents)



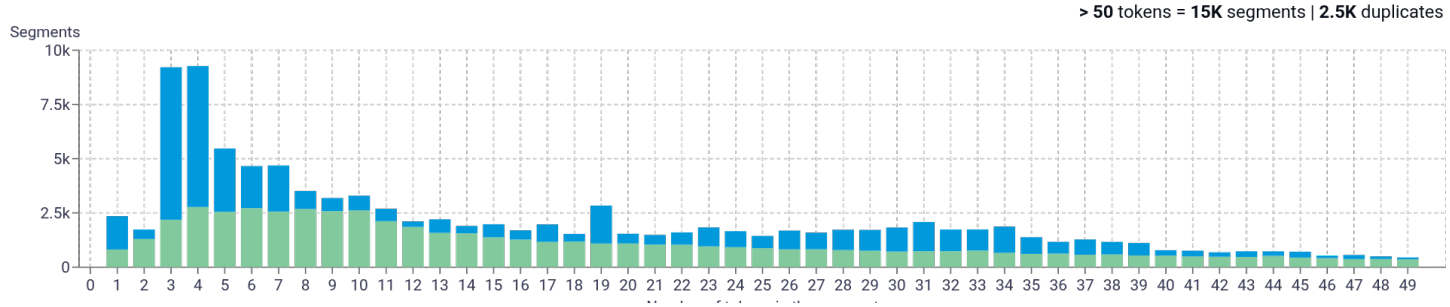
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (7.4K documents)

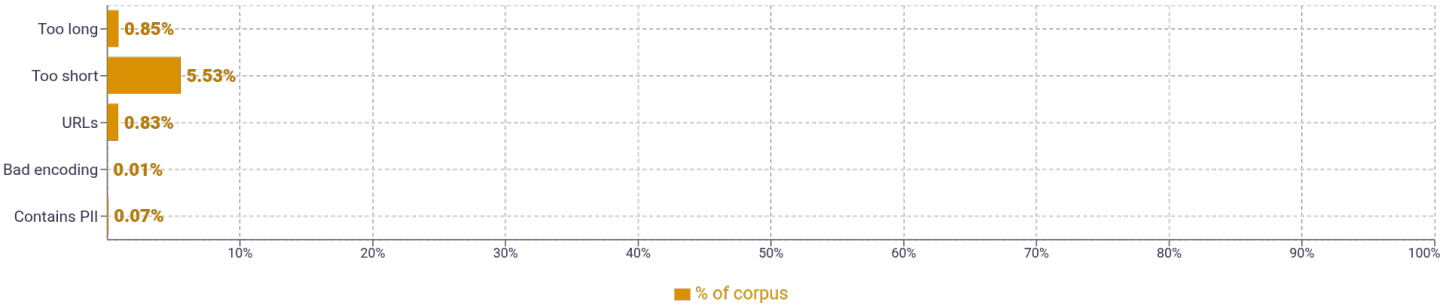


Segment length distribution by token

≤ 49 tokens = **105K** segments | **48K** duplicates
> 50 tokens = **15K** segments | **2.5K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	jan | 26,480 · jach | 13,266 · niźaśa | 7,119 · en | 6,767 · t | 6,537
2	central aymara | 3,995 · chimp askichaña | 2,260 · global voices | 1,876 · willka kuti | 1,665 · sata qallta | 1,357
3	may may markanakatpach | 961 · uraqpachan ukham may | 960 · trabajar sin importar | 960 · muy interesante trabajar | 960 · may markanakatpach irnqt | 960
4	uraqpachan ukham may may | 960 · sin importar las fronteras | 960 · muy interesante trabajar sin | 960 · may may markanakatpach irnqt | 960 · interesante trabajar sin importar | 960
5	uraqpachan ukham may may markanakatpach | 960 · trabajar sin importar las fronteras | 960 · muy interesante trabajar sin importar | 960 · askipuniw uraqpachan ukham may may | 960 · estar preparándose para el encuentro | 330

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				