

General overview

Corpus	Analytics date	Language
ayr_Latn.jsonl.tsv	10/3/2024	Aymara (ayr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
9,223	188,527	48,021 (25.47 %)	4.2M	24.66 MB	24,900,277

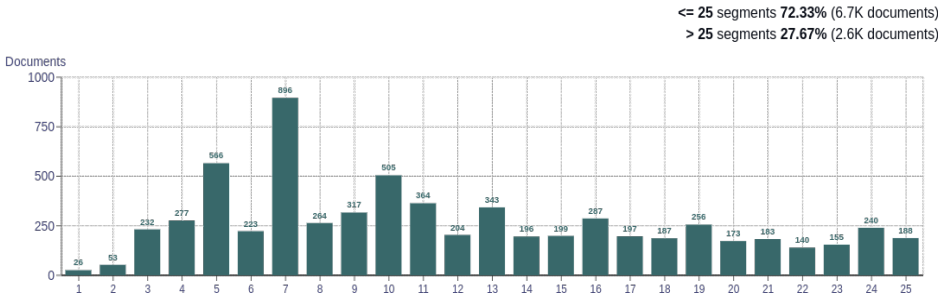
Top 10 domains

Domain	Docs	% of total
globalvoicesonline.org	3.5K	37.67
globalvoices.org	3.1K	33.79
wikipedia.org	983	10.66
jw.org	545	5.91
wol-children.net	191	2.07
bibles.org	146	1.58
radiosangabriel.org.bo	137	1.49
boliviatv.bo	103	1.12
acuerdonacional.pe	99	1.07
agenciapulsar.org	44	0.48

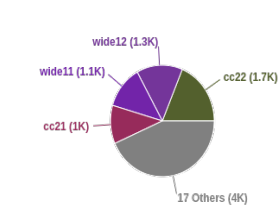
Top 10 TLDs

Domain	Docs	% of total
org	8.4K	90.76
net	201	2.18
org.bo	140	1.52
com	138	1.50
pe	113	1.23
bo	107	1.16
is	31	0.34
cl	24	0.26
es	14	0.15
gob.bo	12	0.13

Documents size (in segments)

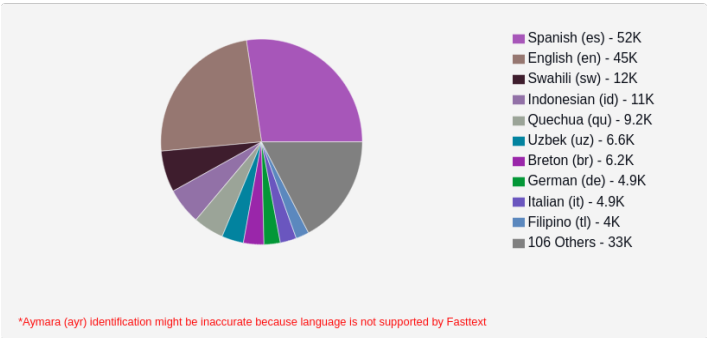


Documents by collection

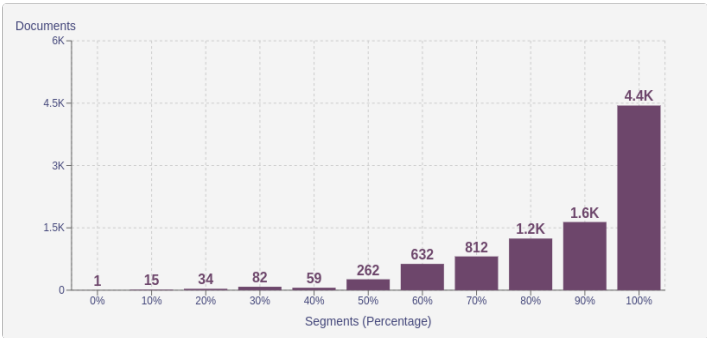


Language Distribution

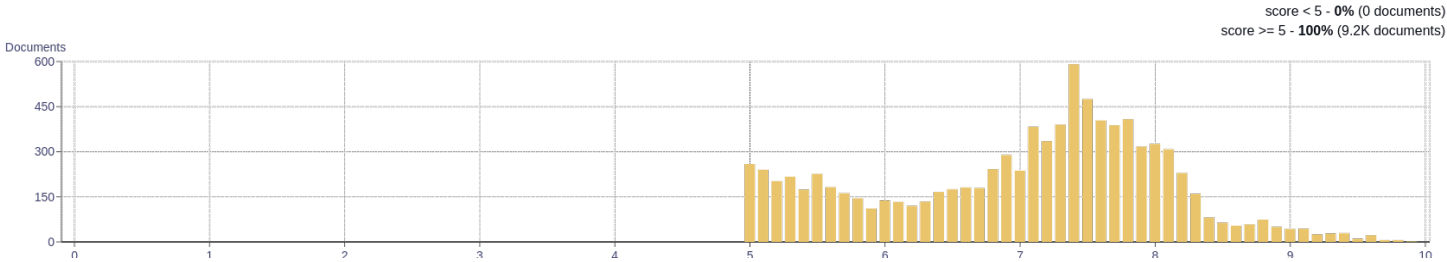
Number of segments



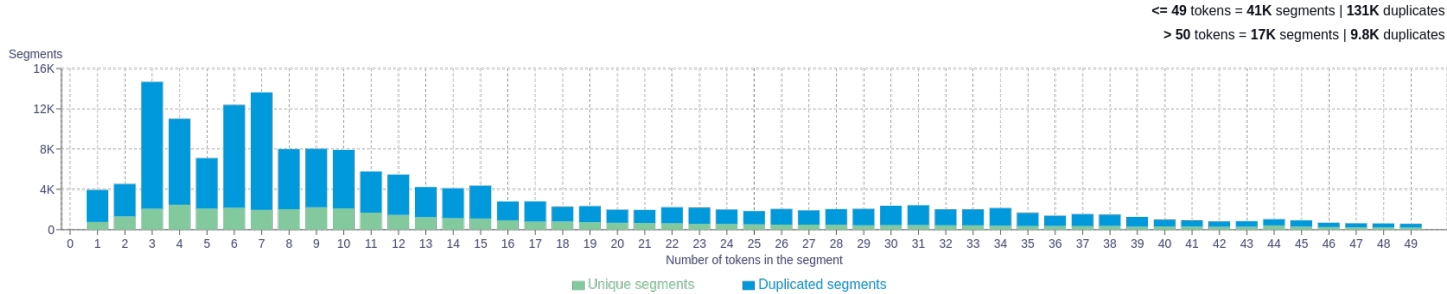
Percentage of segments in Aymara (ayr) inside documents



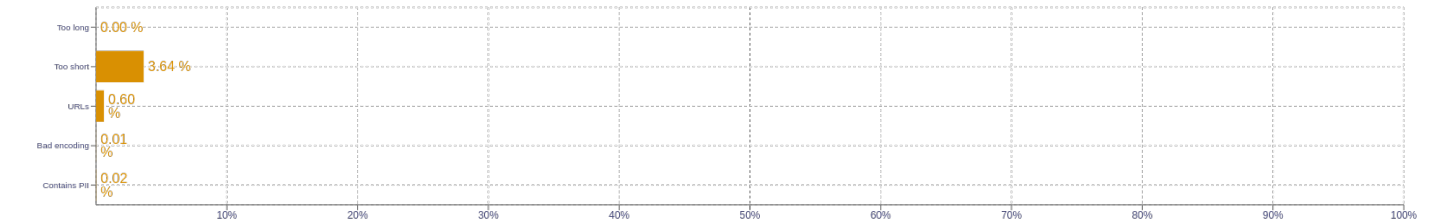
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>jan 33203</div> <div>en 12275</div> <div>jach 12228</div> <div>markan 11886</div> <div>t 11164</div>
2	<div>chimp askichaña 8176</div> <div>global voices 4665</div> <div>sata qallta 2834</div> <div>willka kuti 2818</div> <div>mara t 2142</div>
3	<div>markap amachana santun 497</div> <div>jay jay jay 490</div> <div>global voices aymarata 382</div> <div>global voices podcast 355</div> <div>libertad de expresión 345</div>
4	<div>one day on earth 328</div> <div>jay jay jay jay 328</div> <div>winx winx winx winx 320</div> <div>sat sat sat sat 255</div> <div>isturyanaka aka tuqita latin 210</div>
5	<div>winx winx winx winx winx 310</div> <div>sat sat sat sat sat 254</div> <div>isturyanaka aka tuqita latin america 210</div> <div>universidad pública de el alto 175</div> <div>isturyanaka aka tuqita citizen media 168</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>