HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-gaz_Latn	9/23/2025	Oromo (gaz)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
63,063	1,111,194	892,763 (80.34 %)	41M	250,067,199	243.39 MB

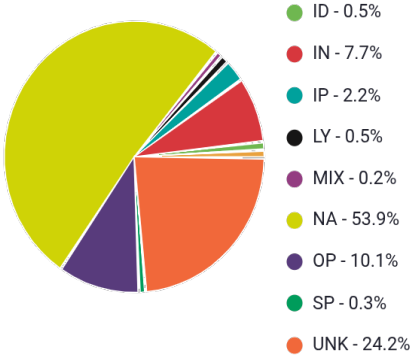
Top 10 domains

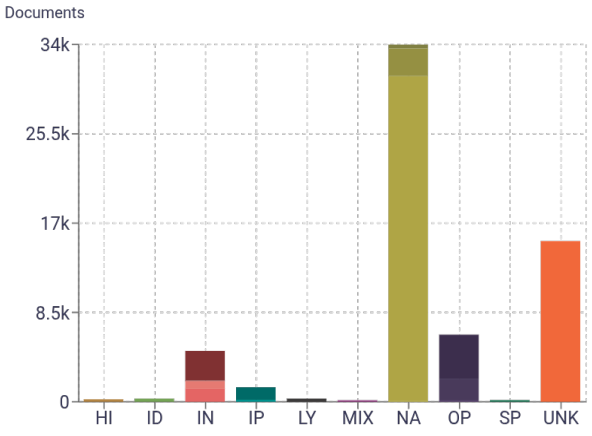
Domain	Docs	% of total
voaafaanoromoo.com	13K	20.02%
nuuralhudaa.com	4.4K	6.90%
kichuu.com	3.2K	5.04%
qeerroo.org	2.7K	4.29%
fanabc.com	2.7K	4.21%
ena.et	2.5K	4.02%
bilisummaa.com	2.5K	3.93%
bbc.com	2K	3.25%
ayyaantuu.org	1.4K	2.25%
addisstandard.com	1.3K	2.11%

Top 10 TLDs

Domain	Docs	% of total
com	42K	67.27%
org	12K	19.15%
et	3.4K	5.43%
is	1.1K	1.79%
gov.et	772	1.22%
net	597	0.95%
edu.et	353	0.56%
it	329	0.52%
no	293	0.46%
gov	228	0.36%

Register labels



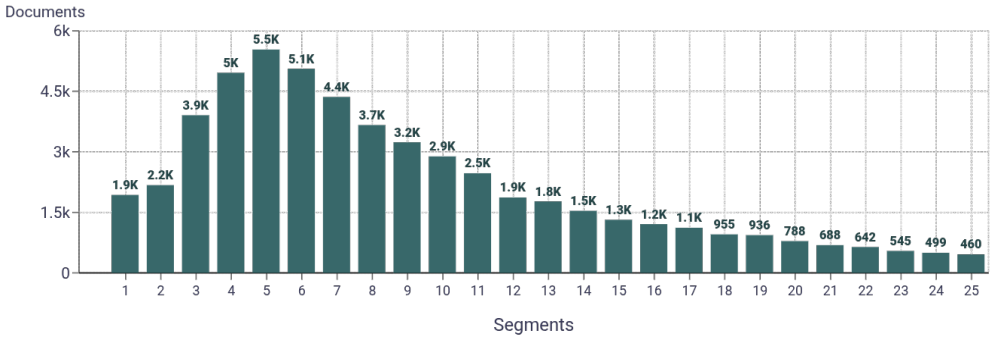


HI_other - 0.4% | NA_nb - 0.1%
ID_other - 0.5% | NA_ne - 49.0%
IN_dtp - 2.0% | NA_other - 4.2%
IN_en - 1.1% | NA_sr - 0.6%
IN_lt - 0.1% | OP_av - 0.0%
IN_other - 4.5% | OP_ob - 0.2%
IN_ra - 0.0% | OP_other - 3.3%
IP_ds - 0.3% | OP_rs - 6.6%
IP_other - 1.9% | OP_rv - 0.0%
LY_other - 0.5% | SP_it - 0.0%
MIX - 0.2% | SP_other - 0.2%
UNK - 24.2%

MT:5.5% | 3.5K Documents

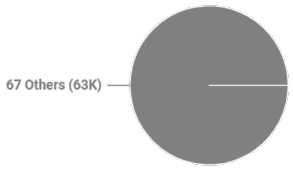
Documents size (in segments)

≤ 25 segments **86.46%** (55K documents)
> 25 segments **13.54%** (8.5K documents)



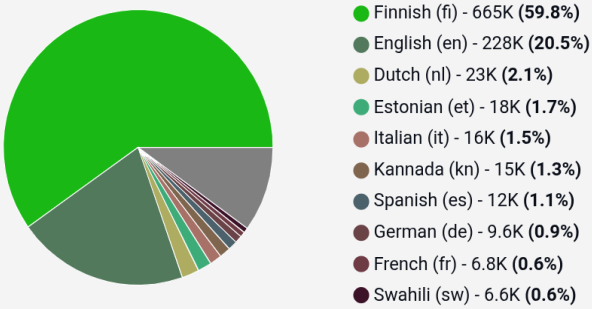
Document collections

CC = 92.73%
IA = 7.27%



Language Distribution

Number of segments in the Oromo (gaz) corpus

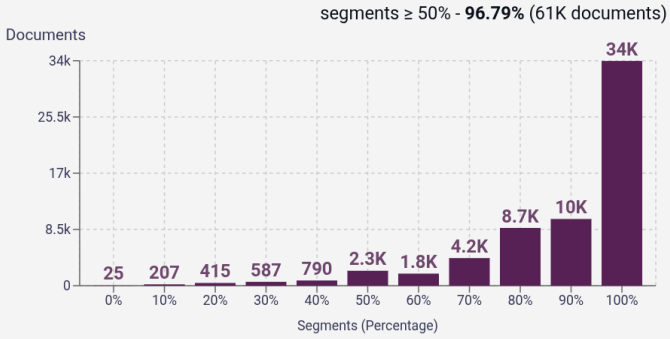


- Finnish (fi) - 665K **(59.8%)**
- English (en) - 228K **(20.5%)**
- Dutch (nl) - 23K **(2.1%)**
- Estonian (et) - 18K **(1.7%)**
- Italian (it) - 16K **(1.5%)**
- Kannada (kn) - 15K **(1.3%)**
- Spanish (es) - 12K **(1.1%)**
- German (de) - 9.6K **(0.9%)**
- French (fr) - 6.8K **(0.6%)**
- Swahili (sw) - 6.6K **(0.6%)**
- 160 Others - 111K **(10.0%)**

*Oromo (gaz) identification might be inaccurate because it is not supported by FastSpell

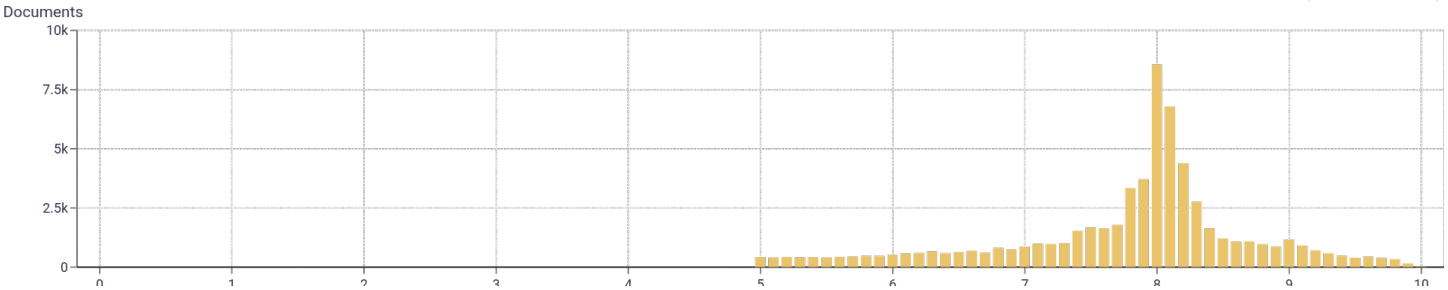
Percentage of segments in Oromo (gaz) inside documents

segments < 50% - **3.21%** (2K documents)
segments ≥ 50% - **96.79%** (61K documents)



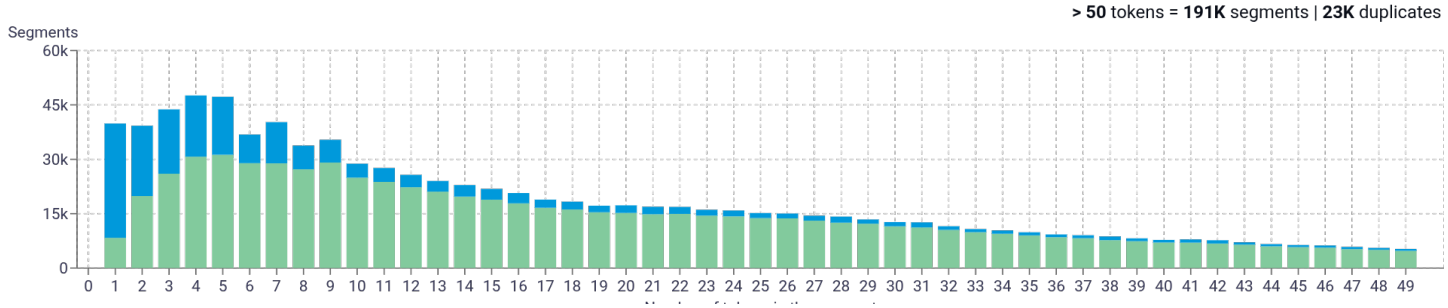
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (63K documents)

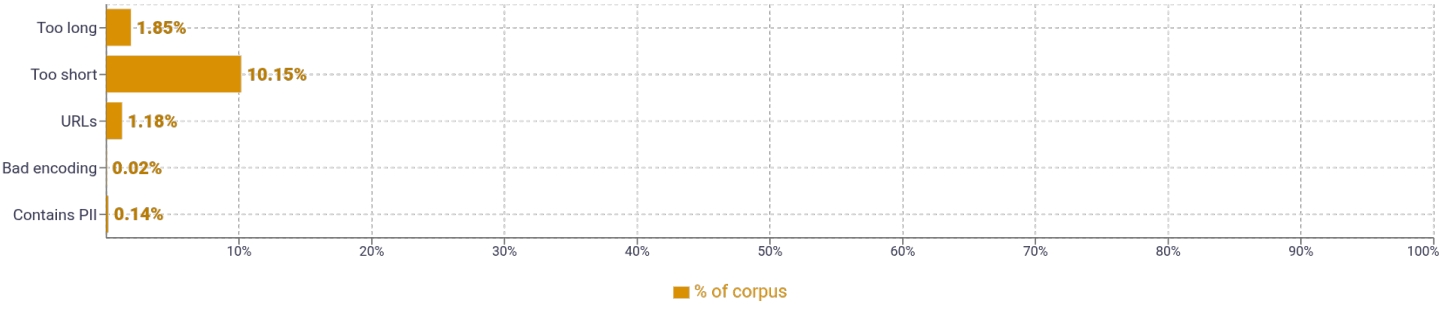


Segment length distribution by token

≤ **49** tokens = **920K** segments | **196K** duplicates
> **50** tokens = **191K** segments | **23K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ta | 310,829 · oromoo | 210,182 · irratti | 159,090 · keessatti | 155,574 · a | 146,157
2	afaan oromoo | 33,642 · adda addaa | 17,174 · bilisummaa oromoo | 16,827 · of the | 16,736 · ummata oromoo | 11,387
3	osoo hin taane | 7,258 · mootummaa naannoo oromiyaa | 4,533 · qeerroo bilisummaa oromoo | 2,880 · adda bilisummaa oromoo | 2,807 · irraa kan ka | 2,427
4	qofa osoo hin taane | 1,964 · rabbiin subhaanahu wa ta | 1,617 · not a registered user | 1,440 · already have an account | 1,439 · this comment as inappropriate | 1,232
5	report this comment as inappropriate | 1,232 · join facebook to connect with | 695 · and others you may know | 693 · facebook gives people the power | 629 · gives people the power to | 600

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				