HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-npi_Deva	9/24/2025	Nepali

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,211,445	76,231,876	52,684,140 (69.11 %)	2.5B	15,011,305,825	37.19 GB

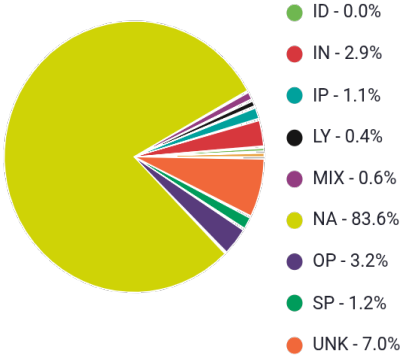
Top 10 domains

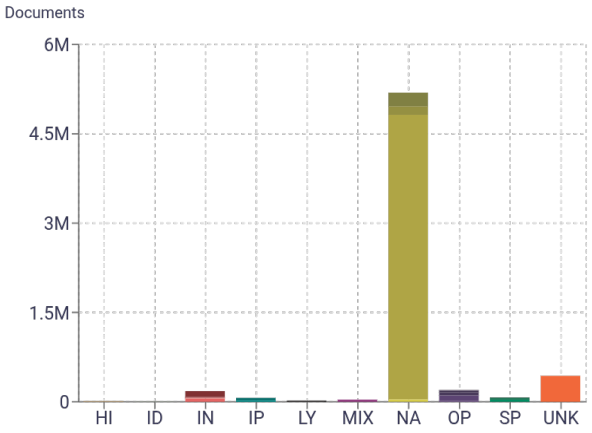
Domain	Docs	% of total
onlinekhabar.com	84K	1.35%
ratopati.com	83K	1.33%
ekantipur.com	65K	1.05%
eadarsha.com	49K	0.78%
nagariknetwork.com	47K	0.76%
enepalese.com	47K	0.76%
ujyaaloonline.com	45K	0.72%
setopati.com	43K	0.69%
abhiyan.com.np	39K	0.62%
khabardabali.com	34K	0.55%

Top 10 TLDs

Domain	Docs	% of total
com	5.6M	90.55%
com.np	166K	2.68%
org	112K	1.80%
net	67K	1.08%
tv	49K	0.80%
gov.np	49K	0.79%
org.np	27K	0.44%
news	20K	0.33%
com.au	15K	0.25%
in	7.6K	0.12%

Register labels



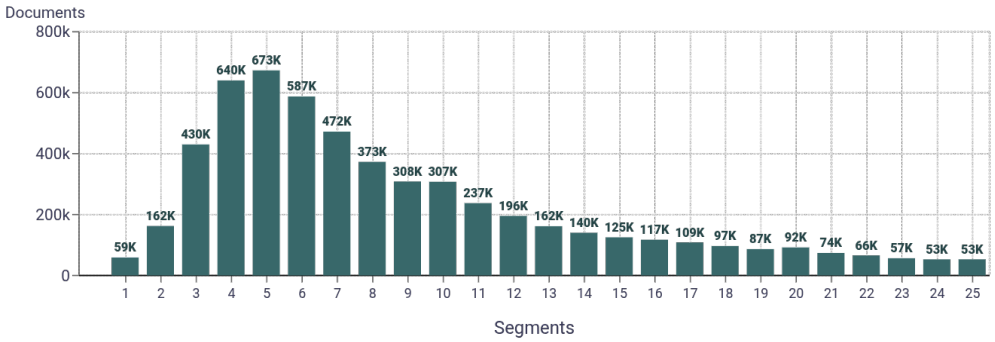


HI_other - 0.1% | MIX - 0.6%
HI_re - 0.0% | NA_nb - 0.7%
ID_other - 0.0% | NA_ne - 76.9%
IN_dtp - 0.8% | NA_other - 2.3%
IN_en - 0.2% | NA_sr - 3.7%
IN_fi - 0.0% | OP_av - 0.2%
IN_lt - 0.2% | OP_ob - 1.5%
IN_other - 1.6% | OP_other - 0.9%
IN_ra - 0.0% | OP_rs - 0.4%
IP_ds - 0.4% | OP_rv - 0.2%
IP_other - 0.7% | SP_it - 0.6%
LY_other - 0.4% | SP_other - 0.6%
UNK - 7.0%

MT:3.3% | 206K Documents

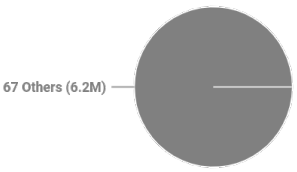
Documents size (in segments)

≤ 25 segments **91.38%** (5.7M documents)
> 25 segments **8.62%** (536K documents)



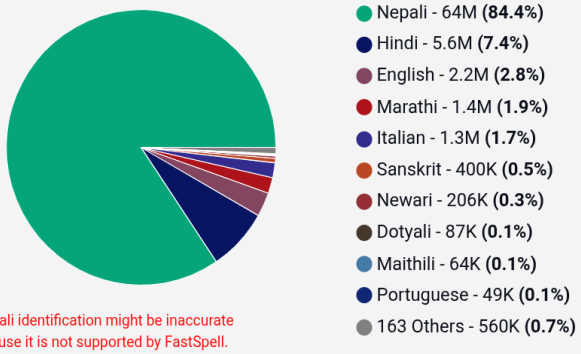
Document collections

CC = 93.52%
IA = 6.48%



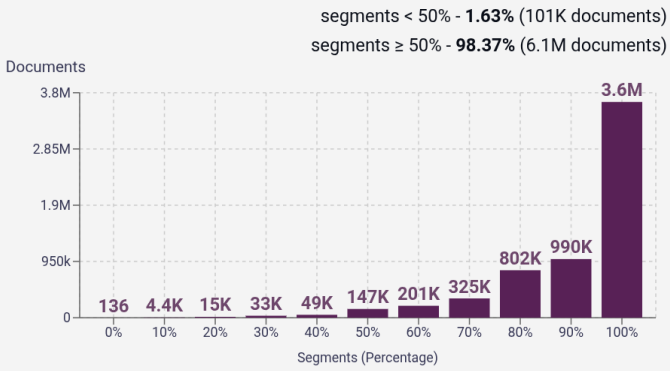
Language Distribution

Number of segments in the Nepali corpus

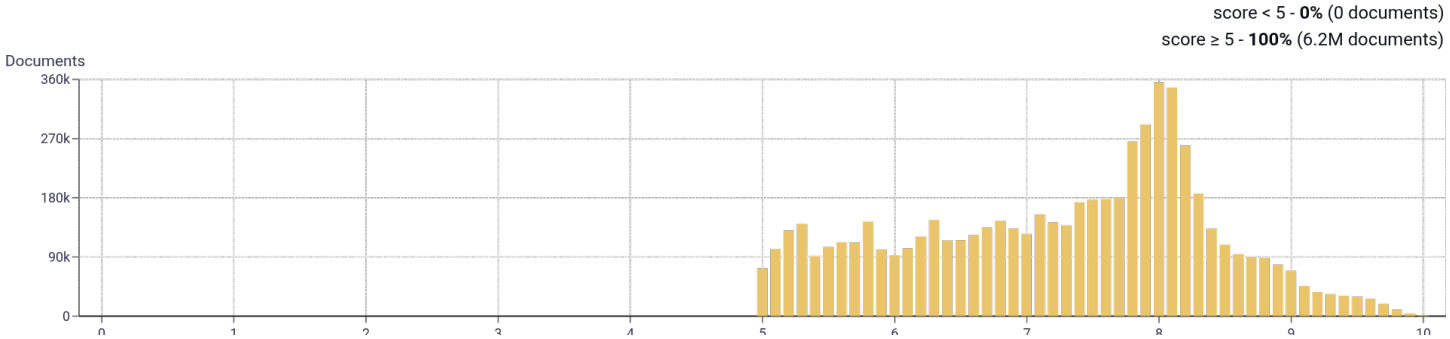


*Nepali identification might be inaccurate because it is not supported by FastSpell.

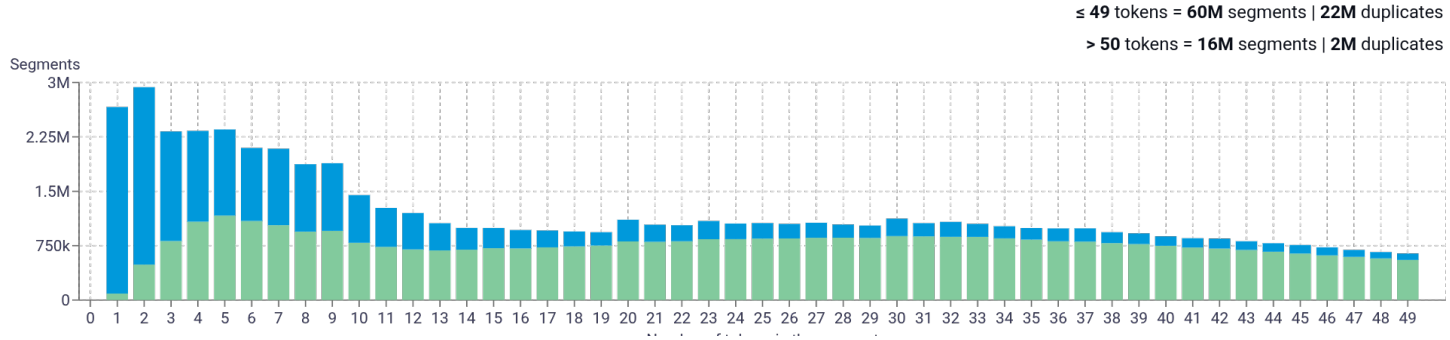
Percentage of segments in Nepali inside documents



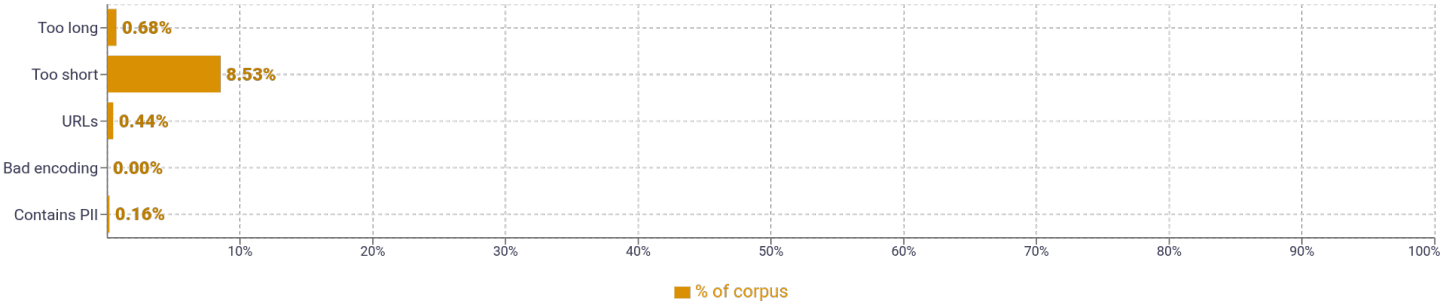
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	नेपाल | 5,289,629 · काम | 4,318,701 · नेपाली | 4,216,072 · हजार | 3,655,437 · भएका | 3,630,121
2	प्रहरी कार्यालय | 456,420 · वडा नं | 453,105 · केपी शर्मा | 414,725 · जिल्ला प्रहरी | 376,808 · जानकारी दिनुभयो | 346,298
3	जिल्ला प्रहरी कार्यालय | 260,253 · प्रधानमन्त्री केपी शर्मा | 239,951 · केपी शर्मा ओलीले | 186,250 · प्रमुख जिल्ला अधिकारी | 177,981 · नेपाल कम्युनिष्ट पार्टी | 116,147
4	प्रधानमन्त्री केपी शर्मा ओलीले | 121,508 · खबर पढेर तपाईलाई कस्तो | 74,091 · तपाईलाई कस्तो महसुस भयो | 73,951 · पढेर तपाईलाई कस्तो महसुस | 73,945 · अध्यक्ष केपी शर्मा ओलीले | 50,464
5	खबर पढेर तपाईलाई कस्तो महसुस | 73,942 · पढेर तपाईलाई कस्तो महसुस भयो | 73,940 · प्रकाशित कुनै समाचारमा तपाईको गुनासो | 43,738 · समाचारमा तपाईको गुनासो भए हामीलाई | 42,255 · माओवादी केन्द्रका अध्यक्ष पुष्पकमल दाहाल | 36,203

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				