

General overview

Corpus	Date	Language
hplt-v3-tel_Telu	9/18/2025	Telugu

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,235,742	81,941,576	52,483,068 (64.05 %)	2.3B	14,990,621,788	36.74 GB

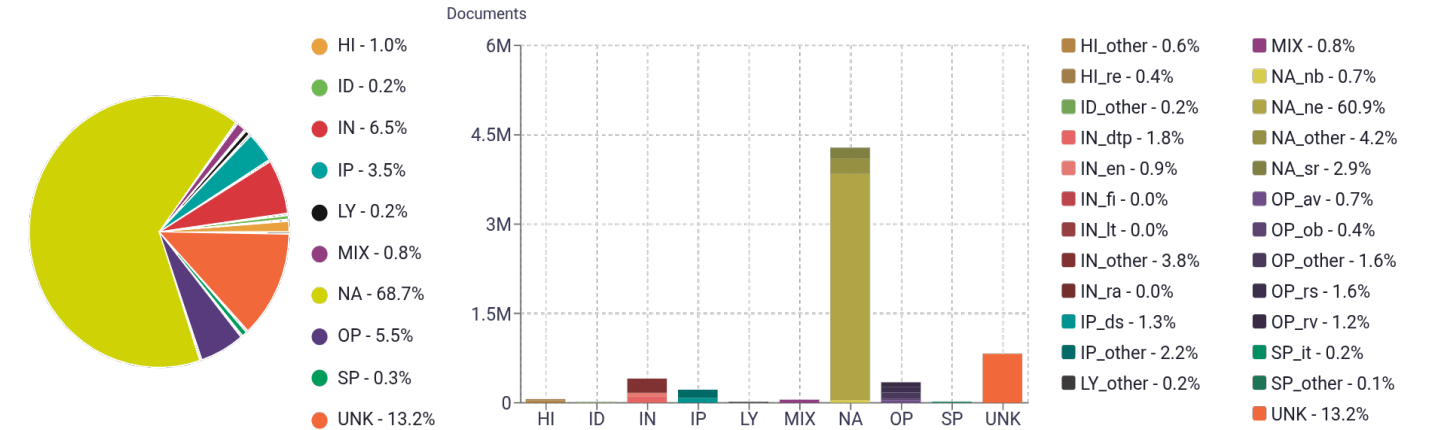
Top 10 domains

Domain	Docs	% of total
sakshi.com	357K	5.73%
eenadu.net	280K	4.48%
asianetnews.com	186K	2.98%
andhrajyothy.com	180K	2.89%
ntnews.com	155K	2.49%
news18.com	121K	1.94%
oneindia.com	103K	1.65%
samayam.com	91K	1.47%
tv9telugu.com	91K	1.47%
filmibeat.com	90K	1.44%

Top 10 TLDs

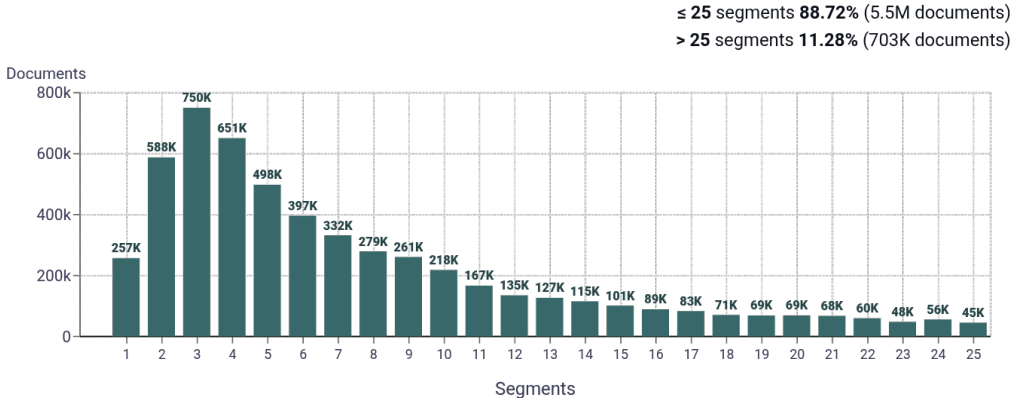
Domain	Docs	% of total
com	4.8M	76.73%
in	495K	7.93%
net	436K	7.00%
org	195K	3.13%
news	57K	0.91%
co.in	48K	0.77%
tv	28K	0.45%
page	22K	0.36%
gov.in	20K	0.32%
info	16K	0.26%

Register labels

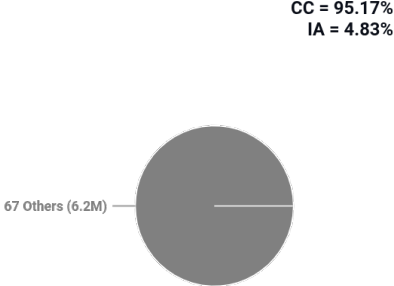


MT:6.0% | 375K Documents

Documents size (in segments)

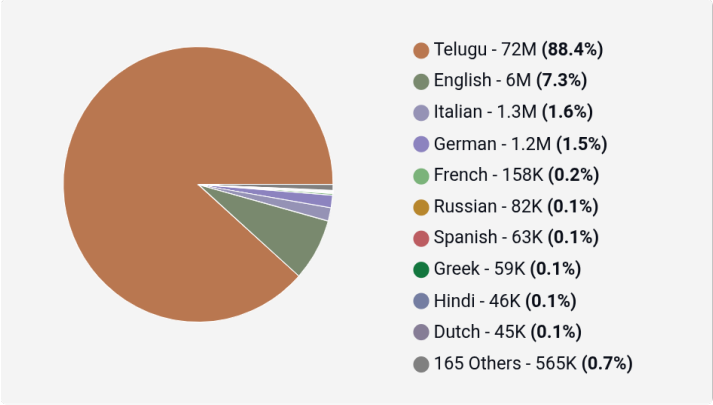


Document collections

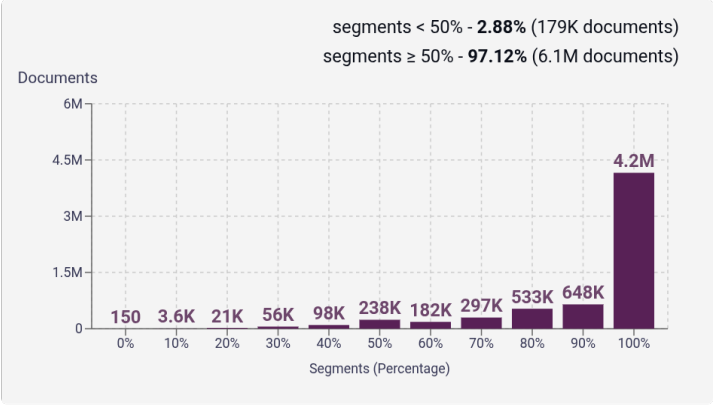


Language Distribution

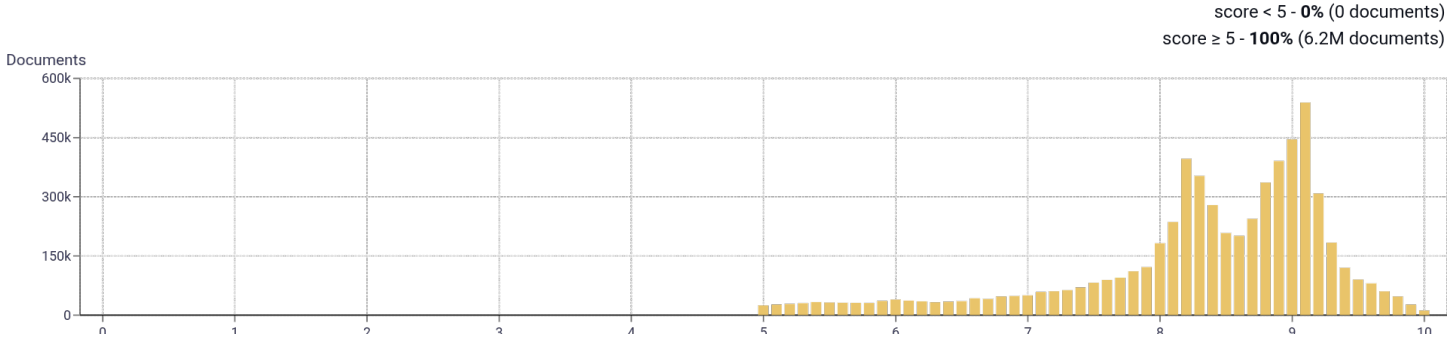
Number of segments in the Telugu corpus



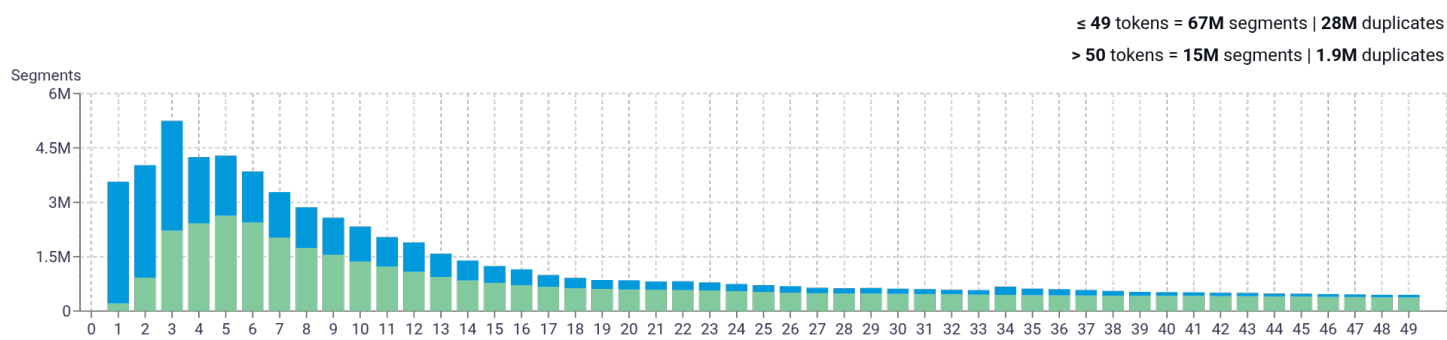
Percentage of segments in Telugu inside documents



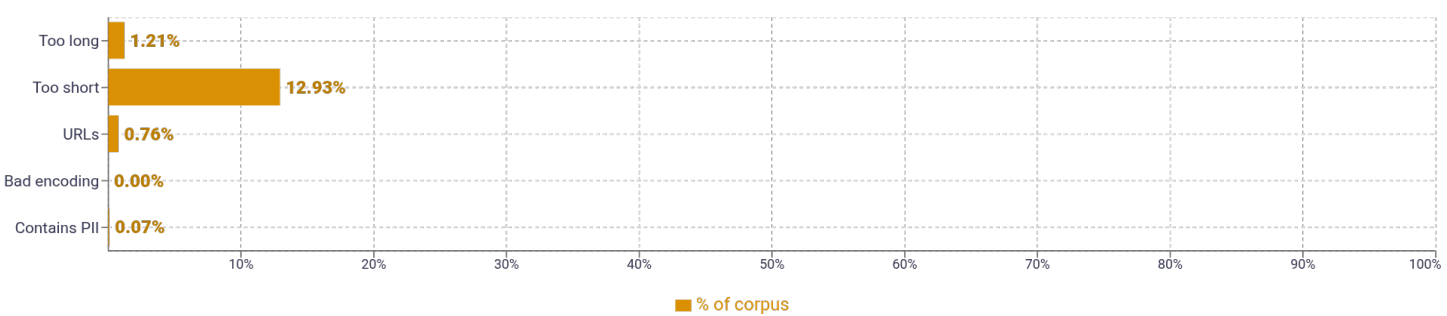
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS					
1	అని   5,635,869	సుంచి   5,520,106	కోసం   4,637,700	చాలా   3,958,456	మీ   3,919,781	
2	read more   892,707	సోషల్ మీడియాలో   497,755	సంగతి తెలిసిందే   441,338	విషయం తెలిసిందే   394,523	చాలా మంది   381,759	
3	post a comment   127,485	సోషల్ మీడియాలో వైరల్   86,978	ఆధ్యాత్మిక ప్రదేశాలను తొరచునుంది   85,983	నుండి ఈ ఆధ్యాత్మిక   85,982	travel కర్ణాటక జూన్   85,940	
4	నుండి ఈ ఆధ్యాత్మిక ప్రదేశాలను   85,976	వాస్తవమైన ఈమెయిల్ ఐడీలతో ఆభిప్రాయాలను   75,004	రీతిలో ఉన్న కామెంట్స్ ప్రచురించలేం   75,004	ఫీక్ ఐడీలతో పంపించే కామెంట్స్   75,004		
5	నుండి ఈ ఆధ్యాత్మిక ప్రదేశాలను తొరచునుంది   85,975	వాస్తవమైన ఈమెయిల్ ఐడీలతో ఆభిప్రాయాలను వ్యక్తీకరించాలని   75,004	కించపరిచే రీతిలో ఉన్న కామెంట్స్ ప్రచురించలేం   75,004	ఈమెయిల్ ఐడీలతో ఆభిప్రాయాలను వ్యక్తీకరించాలని మనవి   75,004	వ్యక్తం చేసే ఆభిప్రాయాలను ఎడిటోరియల్ టీమ్   75,003	

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 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>, <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				