

General overview

Corpus	Analytics date	Language
nso_Latn.jsonl.tsv	12/3/2024	Northern Sotho (nso)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,066	143,306	85,222 (59.47 %)	6.1M	26.86 MB	27,357,543

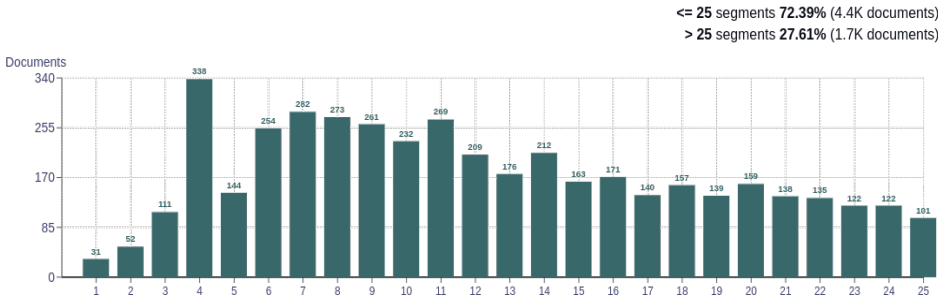
Top 10 domains

Domain	Docs	% of total
jw.org	4K	65.15
biblesa.co.za	494	8.14
wikipedia.org	315	5.19
southafrica.co.za	258	4.25
oxforddictionaries.com	116	1.91
nalibali.org	97	1.60
sars.gov.za	63	1.04
indabukoyakho.com	34	0.56
fundza.mobi	30	0.49
hockeygods.com	28	0.46

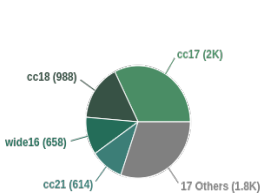
Top 10 TLDs

Domain	Docs	% of total
org	4.5K	73.36
co.za	946	15.60
com	369	6.08
gov.za	89	1.47
org.za	62	1.02
net	34	0.56
mobi	32	0.53
ac.za	22	0.36
io	12	0.20
eu	10	0.16

Documents size (in segments)

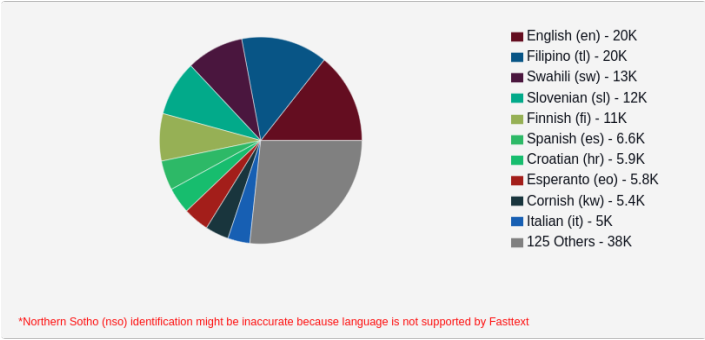


Documents by collection

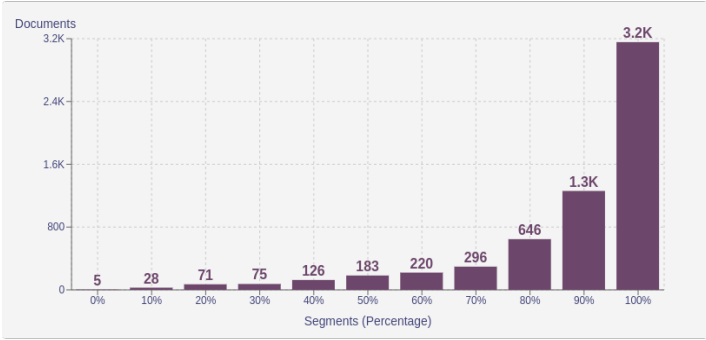


Language Distribution

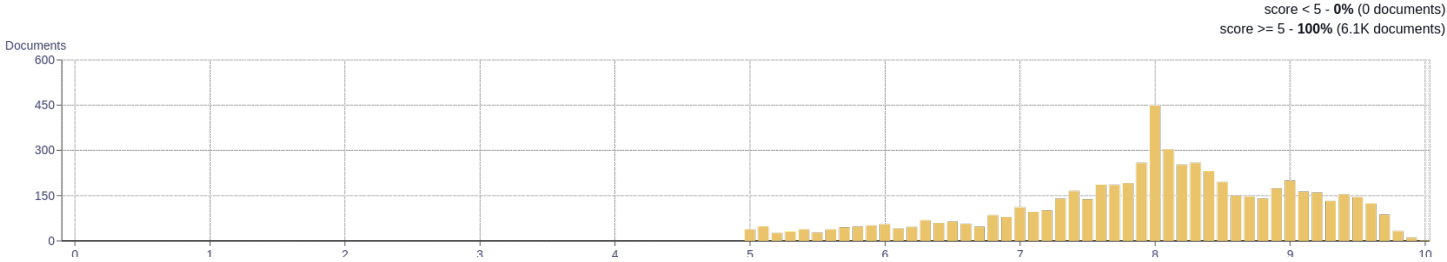
Number of segments



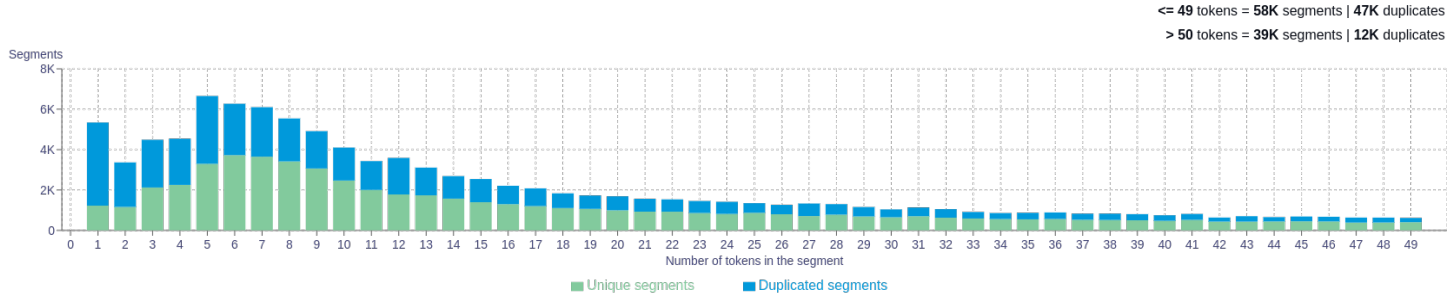
Percentage of segments in Northern Sotho (nso) inside documents



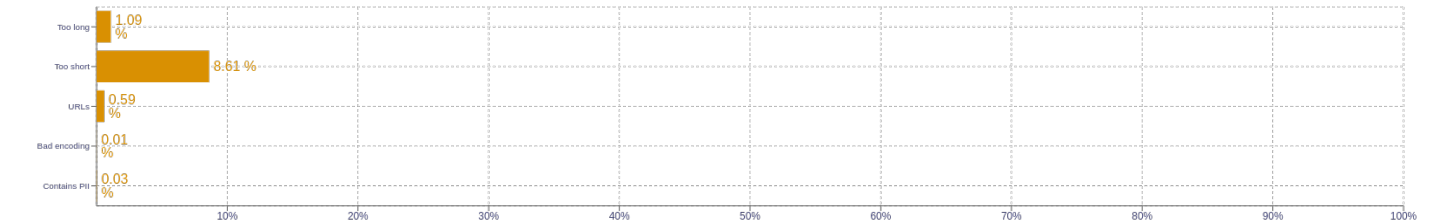
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ya   124023</div> <div>re   69684</div> <div>wa   64276</div> <div>tša   55880</div> <div>ge   55289</div>
2	<div>baka la   7434</div> <div>yo mongwe   6799</div> <div>wa gagwe   5292</div> <div>tše dingwe   5088</div> <div>ya gagwe   4606</div>
3	<div>dihlatse tša jehofa   3798</div> <div>beibele e re   1765</div> <div>mongwe le yo   1660</div> <div>yo mongwe wa   1618</div> <div>baob ba bego   1577</div>
4	<div>lega go le bjalo   3766</div> <div>ge e le gabotse   1772</div> <div>mongwe le yo mongwe   1659</div> <div>yo mongwe le yo   1650</div> <div>sengwe le se sengwe   772</div>
5	<div>yo mongwe le yo mongwe   1649</div> <div>die go die go die   880</div> <div>ithute ka mo go oketšegilego   825</div> <div>phetolelo ya lefase le lefsa   591</div> <div>bophelo bjo bo sa felego   554</div>

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>