Interdisciplinary Journal of Linguistics
Volume [14] 2021, pp. 118-128

ASSAMESE CORPUS SCREENING FOR CLOSURE

Rajesha N.*

Rejitha K.S.*

Abstract

The language corpus is quite an essential component for Natural Language Processing. The agreeable size of corpus which can ensure the maximum coverage of the language text is a matter of concern. Whether or not a corpus is lexically saturated determines the future prospects of NLP applications built on it and the corpus design can be redrawn. The Lexical-Closure of the corpus is highly dependent on representation, script grammar, and characteristic features of a language. For this study power regression of statistics is used on LDC-IL Assamese text corpus to find out closure predictions.

Keywords: Assamese Corpus, Type-Token, Closure, NLP Applications

Introduction

A corpus is the representation of real world language. Although corpus is the reflection of natural language it should follow a set of methods and procedures for the language exploration. If a corpus covers significant part of contemporary language vocabulary then it can be considered as the representative of that language. Corpus building is a cost effective and time consuming process. The main qualities that a good corpus should have are quantity, quality, representation, equality, simplicity, retrievability and verifiability.

Balance of domains is considered to be a prerequisite while designing the corpus. Any claim of corpus balance is a matter of faith rather than reality, because there is no reliable way to measure the corpus balance scientifically. On the contrary, the notion relies heavily on intuition and best estimates. Quality, representation, equality, simplicity are the part of the corpus design. Retrievability and verifiability is a part of data structure.

* CIIL Mysore

The quantity, however, depends on language in question and purpose of corpus building. The agreeable size of corpus which can ensures the maximum coverage of the language text is a matter of concern. Whether or not a corpus is lexically saturated determines the future prospects of NLP applications built on it and the corpus design can be redrawn if needed. The Lexical-Closure of the corpus is highly dependent on representation, script grammar, and characteristic features of a language. Some languages may need very less corpus to cover the vocabulary and styles and some languages need more of it. As corpus builds, the lexical closure point can be calculated to predict how much corpus is needed to reach a lexically saturated quantity of corpus. The lexical closure of corpus analysis is done by type-token analysis.

This study suggests the complex nature of the corpus representation, saturation and volume of the Assamese corpus in theoretical and applied perspective. The LDC-IL Assamese Text Corpus [1] is one of the largest published Indian language text corpora for the study. It is one of the four published Indian language text corpus that are huge enough to be a member of one crore club, i.e. the corpus that has more than one crore words. Tamil, Hindi, Punjabi being the other three.

As it is reported in [2], the Assamese text data sampling strictly follows the generic guidelines of LDC-IL text corpus collection. The sampling method is well described in [3]. The LDC-IL followed a sampling method to collect the pages from books. For example, if the book has 100-200 pages every 10th page is selected as candidate page for sampling text, and if the book has 200-300 pages every 20th page of the book becomes the candidate for sampling. If any of the candidate page contains pictures, tables etc., then its next or previous page is selected for sampling text that possesses the text content. While selecting the book for sampling, the LDC-IL's motive is to select from wide variety of domains, thus the corpus can cover large part of vocabulary and should not miss out certain domains.

Assamese is an Indo-Aryan language. Unlike most Indo-Aryan languages that lack a native script, the Assamese language has its own script named after itself. It is also known as Oxomiya Akhor or Oxomiya Lipi, a variant of the Eastern Nagari script evolved from Kamarupia script. It is also used for Bengali and Bishnupriya Manipuri.

The LDC-IL Assamese text corpus is encoded in Unicode. It boasts 1,01,27,030 Tokens (words) in size worked up by

6,39,50,126 UTF characters, drawn from 1,084 different titles, thus the avg. token-length will be 6.31 UTF characters/tokens. As it is observed in [4], The Assamese avg. token-length falls between 6.01 of Bengali and 6.49 of Odia, typical of East-Indic languages. This comparatively higher than the North-Indic languages that typically falls in the range 5.01-5.60 UTF Characters/Token.

Since the sampling method is well defined and the categorization of source text material is made so that the balance of the corpus can be kept in check, in practical sense it is evident that text of some domain over-represented and some domain are under-represented in the published datasets. The [4] justifies, that is how the language texts are populated. 'It would be short-sighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' because the corpus used cannot be proved to be 'balanced'.' [5]. Reference [2] reports, The Aesthetics dominates the corpus, and the mass media mainly drawn from newspapers has 1/3rd of the share.

Table I. Domain Representation of Assamese Corpus

Domain	Word Count	Percentage
Aesthetics	52,33,452	51.68%
Mass Media	33,54,996	33.13%
Social Sciences	10,97,570	10.84%
Science and Technology	3,72,790	3.68%
Commerce	66,924	0.66%
Official Document	1,298	0.01%
Total	1,01,27,030	100.00%

It can be observed from [2] that there are many sub-domains which are poorly represented like Banking, Industry, Official Document, Criminology, Veterinary, Police Documents, Administration etc. Some sub-domains are not even got a chance to be a part of the corpus. In Indian scenario there is scarcity of text material in many fields. While collecting text from such domains lenience can be exercised to have a representation of types. The corpus too will have diverse types and in a much balanced state.

The domain-wise representation can be depicted as follows.

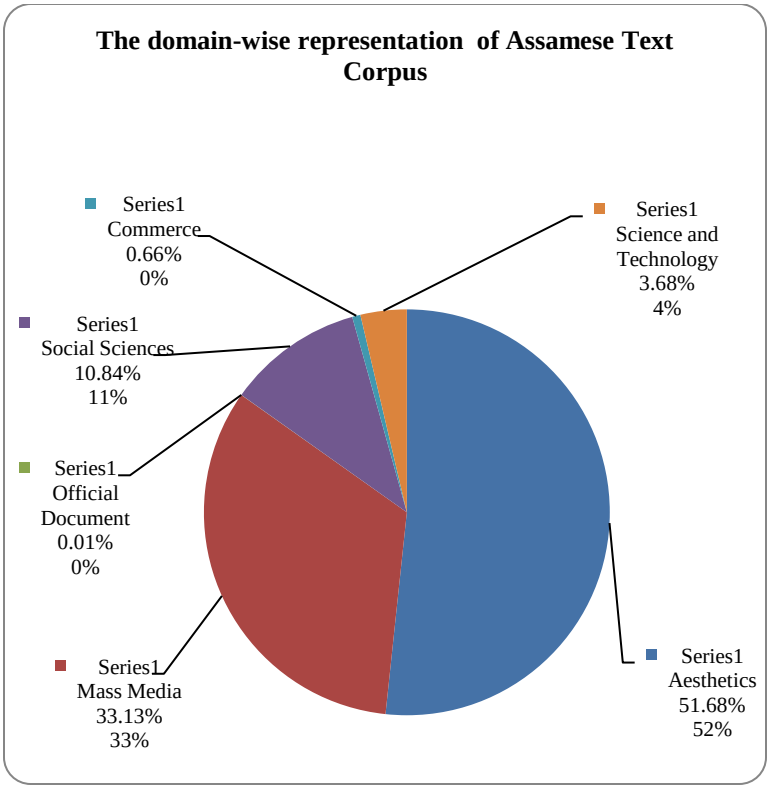


Fig. 1. Category Distribution of Assamese Corpus

As [7] observes 'a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language'. Since the Assamese corpus being a generic corpus has considerable size of one crore tokens drawn from various domains and the range of text types are more or less included closure analysis can be performed on this corpus.

Corpus/Closure

'Closure/saturation for a particular linguistic feature (e.g. size of lexicon) of a variety of language means that the feature appears to be finite or is subject to very limited variation beyond a certain point. To measure the saturation of a corpus, the corpus is first divided into segments of equal size based on its tokens. The corpus is said to be saturated at the lexical level if each addition of a new segment yields approximately the same number of new lexical items as the previous segment, i.e. when 'the curve of lexical growth has become asymptotic' or is flattening out. The

notion of saturation is claimed to be superior to such concepts as balance for its measurability.' [6].

As it is descriptively documented in [3] The Corpus saturation is affected predominantly by three factors 1) Representativeness of the corpus 2) Script grammar of the language 3) Characteristic features of language.

Approach to Saturation Analysis

As it is observed by [3] and [6] the type-token ratio (TTR), is the ratio obtained by dividing the types (the total number of different words) occurring in a text or by its tokens (the total number of words). A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite.

This test is a simple measure of lexical diversity of language which has been used in literary studies. The type/token ratio (TTR) varies widely in accordance with the size of the text -- or corpus of texts -- which is being studied. Since Assamese is a large corpus of Indo-Aryan family of languages and observed in [4] and [9] tends to lexical closure quite early as compared to Dravidian languages.

The calculated type-token ratio of Assamese corpus is 5, 60,510 (Types) / 1, 01, 27,030 (Tokens) = 0.055.

For the corpus closure analysis the conventional TTR is informative, of course, if one is dealing with a corpus comprising lots of equal-sized text segments.

The text fragments (XML Files) are randomly picked to carry out incremental type-token analysis. The random selection is done to keep the type-token analysis unbiased of any particular domain influence. Since LDC-IL corpus does not contain any white-space characters other than space, tokenization is performed splitting the corpus text across spaces. All punctuation marks that were part of token were truncated while deriving tokens, so as to extract pure Assamese words as tokens. Each distinct word is a type and each occurrence of a type counts as a token. It is important to mention that by types it is meant here as fully inflected word forms, not root forms. One lakh tokens is kept as the unit of benchmark at which acquired distinct types (words) are counted, and then these distinct words are subtracted from the distinct types acquired at previous benchmark to get the number of new types in this unit amount of tokens. This number is evaluated for the percentage growth in types at the given benchmark. The figures are tabulated and depicted to get a type-

token growth rate curve which shows how many new types will be found as the corpus size increases.

The following table shows the incremental type-token analysis of Assamese corpus.

Token Input	Distinct Types	Added Types/Unit	Percentage Distinct Type/Unit
100000	23,795	23,795	23.8
200000	38,609	14,814	14.8
300000	52,647	14,038	14.0
400000	64,417	11,770	11.8
500000	74,436	10,019	10.0
600000	83,125	8,689	8.7
700000	92,083	8,958	9.0
800000	1,01,389	9,306	9.3
900000	1,10,197	8,808	8.8
1000000	1,17,324	7,127	7.1
1100000	1,24,674	7,350	7.4
1200000	1,31,804	7,130	7.1
1300000	1,38,596	6,792	6.8
1400000	1,45,704	7,108	7.1
1500000	1,52,851	7,147	7.1
1600000	1,59,804	6,953	7.0
1700000	1,68,000	8,196	8.2
1800000	1,74,830	6,830	6.8
1900000	1,80,919	6,089	6.1
2000000	1,87,529	6,610	6.6
2100000	1,93,529	6,000	6.0
2200000	1,99,732	6,203	6.2
2300000	2,06,996	7,264	7.3
2400000	2,13,312	6,316	6.3
2500000	2,19,518	6,206	6.2
2600000	2,25,119	5,601	5.6
2700000	2,30,282	5,163	5.2
2800000	2,36,642	6,360	6.4
2900000	2,41,981	5,339	5.3
3000000	2,47,831	5,850	5.9
3100000	2,53,570	5,739	5.7
3200000	2,58,669	5,099	5.1
3300000	2,63,714	5,045	5.0
3400000	2,68,884	5,170	5.2
3500000	2,74,269	5,385	5.4
3600000	2,80,669	6,400	6.4
3700000	2,86,244	5,575	5.6
3800000	2,92,107	5,863	5.9
3900000	2,96,359	4,252	4.3
4000000	3,00,827	4,468	4.5
4100000	3,05,415	4,588	4.6
4200000	3,09,940	4,525	4.5
4300000	3,14,735	4,795	4.8
4400000	3,19,169	4,434	4.4
4500000	3,24,329	5,160	5.2
4600000	3,28,687	4,358	4.4
4700000	3,33,517	4,830	4.8
4800000	3,38,277	4,760	4.8
4900000	3,43,065	4,788	4.8
5000000	3,48,312	5,247	5.2
5100000	3,52,573	4,261	4.3
5200000	3,57,297	4,724	4.7
5300000	3,61,616	4,319	4.3
5400000	3,66,442	4,826	4.8
5500000	3,71,100	4,658	4.7
5600000	3,75,966	4,866	4.9
5700000	3,80,650	4,684	4.7
5800000	3,84,842	4,192	4.2
5900000	3,89,046	4,204	4.2
6000000	3,93,664	4,618	4.6
6100000	3,98,583	4,919	4.9
6200000	4,03,966	5,383	5.4
6300000	4,08,520	4,554	4.6
6400000	4,12,583	4,063	4.1
6500000	4,16,643	4,060	4.1
6600000	4,21,342	4,699	4.7
6700000	4,25,511	4,169	4.2
6800000	4,29,006	3,495	3.5
6900000	4,32,940	3,934	3.9
7000000	4,37,234	4,294	4.3
7100000	4,41,146	3,912	3.9
7200000	4,44,706	3,560	3.6
7300000	4,48,978	4,272	4.3
7400000	4,52,938	3,960	4.0
7500000	4,57,248	4,310	4.3
7600000	4,61,596	4,348	4.3
7700000	4,66,206	4,610	4.6
7800000	4,70,715	4,509	4.5
7900000	4,74,534	3,819	3.8
8000000	4,78,059	3,525	3.5
8100000	4,82,414	4,355	4.4
8200000	4,86,150	3,736	3.7
8300000	4,89,394	3,244	3.2
8400000	4,93,512	4,118	4.1
8500000	4,97,948	4,436	4.4
8600000	5,01,948	4,000	4.0
8700000	5,05,563	3,615	3.6
8800000	5,09,632	4,069	4.1
8900000	5,13,559	3,927	3.9
9000000	5,17,490	3,931	3.9
9100000	5,21,699	4,209	4.2
9200000	5,26,098	4,399	4.4
9300000	5,29,325	3,227	3.2
9400000	5,33,520	4,195	4.2
9500000	5,37,330	3,810	3.8
9600000	5,41,285	3,955	4.0
9700000	5,44,711	3,426	3.4
9800000	5,48,401	3,690	3.7
9900000	5,52,322	3,921	3.9
10000000	5,56,343	4,021	4.0
10100000	5,59,602	3,259	3.3
10127030*	5,60,510*	908*	3.4*

*Figures not used for type-token analysis

Table II. Type-Token Analysis of Assamese Corpus.

While token counts are pushed higher by repetition, type counts are pushed higher by lack of repetition. Some main factors that can influence the type count are:

1) degree of vocabulary restraint for simplification,

2) complexity of topic, and

3) frequency of topic change

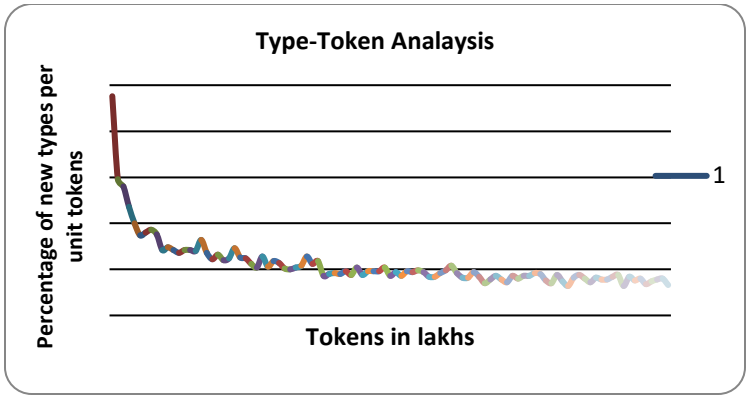
1) Degree of vocabulary restraint for simplification: Newspaper write-ups for common readers tend to simplify the topic. The news data does not have literary flourish, but it attracts people from all walks of life. Many unfamiliar domains, religious ideas, scientific principles etc. have to be conveyed to ordinary people. So the writer will have captured these domains in a simple and meaningful way. It needs proper usage of vocabulary, correct language structure and effective phraseology. The writer may use colloquial or non-standard terms or jargons to attract the readers. The words used need to be expressive and represents the feeling and attitude towards the events. This text is contemporary in nature. It is connected at discourse level and usually on a topic. The text may contain political news, editorials, or sports news. Since it is a newspaper extract, it contains words which are used in day-to-day life.

2) Complexity of topic: The complexity of the topic in the text under consideration in a unit can also affect the type count in the corpus of that particular unit. Generally, more complex topics require a more complex and diverse lexis. Since the corpus is generic and derived from various sources across various topics, the corpus naturally draws diverse lexicon. Some of the topics that were covered include: physics, chemistry, linguistics, technology, law, etc.

3) Frequency of topic change: In addition to the complexity of the topic, the actual frequency of the topic change can also have an effect on the type count of the unit. When many small topic comes in a unit of one lack words shows topics of diverse range then the type count in that unit shows higher range compare to a unit which covers similar or same topic.

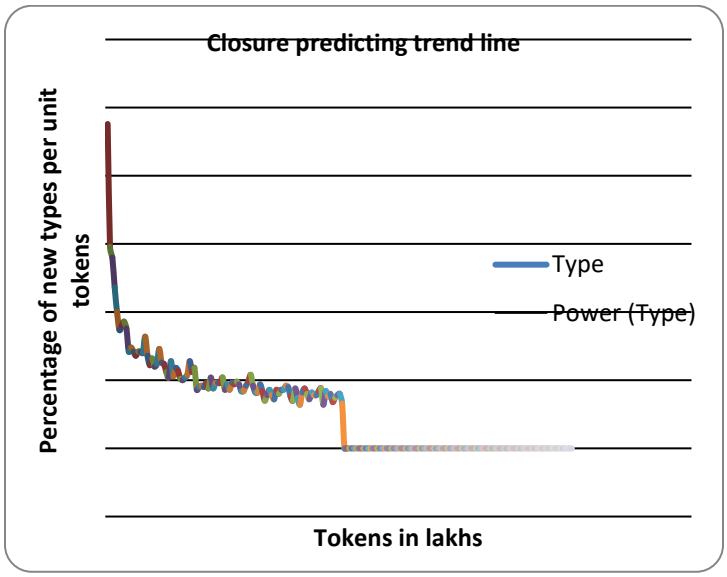
These could help explain the small bumps in Type Token Curve when the tabulated values are depicted.

On graphical scale the tabulated values can be described as below:



It can be observed that the curve is swinging between 3,000 to 4,000 types for each one lakh token addition of corpus, which means 3%-4% words are new words for every unit amount of input. Analysis of existing data plays a great role in corpus closure prediction. This can be a significant advantage as it enables a more structured approach towards the corpus collection decision making.

Power regression is a non-linear regression model, based on the equation: $y = \alpha x \beta$. To predict how far Assamese corpus needs to be collected to have a good coverage of words.



When the trend lines stretch to 2 crore token size, the power regression predicts around 2,500 distinct words/one lakh corpus input. It means 2.5% words are new words for every unit amount of input even when the corpus size is 2 crore words. Coefficient of Determination r^2 is 0.927 which is pretty good fit for the data provided.

Conclusion

Power regression method for estimating lexical closure point of the Assamese corpus shows pretty good fit for the text data. The Type-token curve is flattening below 5%, the Assamese is inching towards lexical closure. It needs to balance in every domain and acquire more text from under- represented domains. If it can acquire a balanced corpus of 2 crore words, it can be reviewed for lexical closure estimations.

References

Atkins, S., et al. “Corpus Design Criteria”. *Literary and Linguistic Computing*, 1992, pp. 1–16.

Biber, Douglas. “Representativeness in Corpus Design.*”* *Literary and Linguistic Computing*, Vol. 8, no. 4, 1993, pp. 243.

Bharadwaja, G. “Statistical Analyses of Telugu Text Corpora.” *Int. Dravidian Linguist*, vol. 36, no. 2, 2007, pp. 71–99.

Hussain, T., et al. "A Gold Standard Assamese Raw Text Corpus." *Compendium of Linguistic Resources in Indian Languages*, Central Institute of Indian Languages, Mysore, 2021, pp. 1-8.

K.S. Rejitha, Rajesha , N. “Type-Token Analysis on Tamil Corpus.” *Working Papers on Linguistics and Literature*, Department of Linguistics, Bharathiyar University, Coimbatore, 2021.

K.S. Rejitha, Rajesha N. “Saturation of Indian Language Corpora - Malayalam vs. Hindi.” *Working Papers on Linguistics and Literature*, Department Of Linguistics, Bharathiyar University, Coimbatore, Vol. XIII No.2, 2019, pp. 441-450.

Kimmo, Kettunen. “Can Type-Token Ratio be Used to Show Morphological Complexity of Languages?” *Journal of Quantitative Linguistics*, vol. 21, no. 3, 2014, pp.223-245.

Ramamoorthy, L., et al. “A Gold Standard Assamese Raw Text Corpus.*” Central Institute of Indian Languages*, Mysore. 2021, ISBN 978-81-948885-4-3

Xiao, Richard. Corpus Creation. *Handbook of Natural Language Processing*. CRC Press, London, 2010.