Statistical Analysis of Sentence Structures through ASCII, Lexical Alignment and PCA

Abhijeet Sahdev

New Jersey Institute of Technology

as4673@njit.edu

Abstract—While utilizing syntactic tools such as parts-of-speech (POS) tagging has helped us understand sentence structures and their distribution across diverse corpora, it is quite complex and poses a challenge in natural language processing (NLP). This study focuses on understanding sentence structure balance - usages of nouns, verbs, determiners, etc - harmoniously without relying on such tools. It proposes a novel statistical method that uses American Standard Code for Information Interchange (ASCII) codes to represent text of 11 text corpora from various sources and their lexical category alignment after using their compressed versions through PCA, and analyzes the results through histograms and normality tests such as Shapiro-Wilk and Anderson-Darling Tests. By focusing on ASCII codes, this approach simplifies text processing, although not replacing any syntactic tools but complementing them by offering it as a resource-efficient tool for assessing text balance. The story generated by Grok shows near normality indicating balanced sentence structures in LLM outputs, whereas 4 out of the remaining 10 pass the normality tests. Further research could explore potential applications in text quality evaluation and style analysis with syntactic integration for more broader tasks.

Index Terms—ASCII, PCA, NLP, linguistic analysis.

I. INTRODUCTION

As we progress into the intelligence age [1], celebrating the byproducts of plausible experiments that helped reinforce our faith in deep learning as the guiding methodology for future advancements, addressing large language models such as ChatGPT [2], Gemini [3] , etc., as compressed versions of the Internet that have transformed text generation, yet, one may always encounter grammatical fallacies in their response. There are various approaches to resolve this problem as comprehensively discussed in [4]. However, these methods and other traditional approaches rely on syntactic tools such as parts-of-speech (POS) tagging which again increase the complexity by utilizing word embeddings such as BERT [5] as we progress further in an NLP pipeline. The process depicted in [6] is also computationally intensive in obtaining the sentence structure.

Our study introduces a novel statistical approach that leverages ASCII codes and Principal Component Analysis which may have been overlooked in the recent years to assess sentence structure balance. Earlier approaches of PCA have been limited to classification tasks [7]. Here, the authors focus on representing natural language such as English in machine codes such as ASCII. It may not be intuitive to think that texts can be directly represented as numbers through a standardized mapping such as ASCII. For lexical align-

ment, seventeen lexical categories that include abbreviation, adjective, adposition, adverb, affix, conjunction, contraction, determiner, interjection, noun, other, phrase, predeterminer, preposition, pronoun, symbol and verb were also converted to ASCII. For example: affix is mapped to [65,102,102,105,120].

These lexical categories' ASCII representations are compressed using PCA. While doing so, there was no loss in explained variance, i.e, information was retained irrespective of the number of lexical categories chosen. Usually, ASCII values are used for basic tasks such as storing text, however, motivated by the this observation, the authors represented lexical categories as a single vector of 17 dimensions, Each sentence from our text corpora was sent through the same pipeline and finally aligned with this vector and their distributions were analyzed for normality, hypothesizing that balanced test - characterized by harmonious usage of lexical resources exhibited normal distribution. This hypothesis is tested on ten different diverse text sources along with one that was a combination of the previous ten. The sources varied from blogs, news, articles, movie reviews and a story generated by Grok, an LLM from xAI to evaluate its applicability between human and AI-generated texts. The results demonstrate that ASCII can be used for high-level text analysis, expanding its capabilities to statistical linguistic analysis.

II. BACKGROUND

A. *Principle Component Analysis*

It aims to project higher dimensional data to lower dimensions while ensuring that the variance in the original data is reflected as much as possible. This is captured using explained variance per component. The following are the steps to compute PCA [8] for a given data matrix Z.

- First the data is centered by subtracting the mean of each feature:

$$Z_{\text{centered}} = Z - \mu$$

where Z is the data matrix and μ is the mean of each feature (column).

- The covariance matrix is calculated as:

$$\Sigma = \frac{1}{n-1} Z_{\text{centered}}^T Z_{\text{centered}}$$

where n is the number of data points.

- The covariance matrix Σ is decomposed to find the eigenvectors (principal components) and eigenvalues:

$$\Sigma a = \lambda a$$

where a is an eigenvector, and λ is the corresponding eigenvalue.
- The eigenvectors are sorted in descending order by their corresponding eigenvalues. The top k eigenvectors are chosen as the principal components:

$$A_k = [a_1, a_2, \ldots, a_k]$$

where A_k is the matrix of the top k eigenvectors. We set $\bar{K}$ as one-dimensional or seventeen-dimensional, depending on the use case in our study.
- Finally, the data is projected onto the principal components:

$$X = Z_{\text{centered}} A_k \tag{1}$$

Where X is the matrix of the data in the reduced-dimensional space.

B. *Dot Product*

It helps us measure lexical alignment in our study of arrays with 17 dimensions, it is the sum of the product of individual components, represented as:

$$\bar{\mathbf{J}} \cdot \bar{\mathbf{K}} = \sum_{i=1}^{17} j_i k_i \tag{2}$$

III. METHODOLOGY

To compute the above measurements, the experimental setup is designed as shown in 1. The setup begins by fetching the seventeen lexical categories mentioned earlier and in the same order from an application programming interface. First, lexical categories are converted to ASCII codes and compressed to one dimension using Equation 1. There was no loss in explained variance in this case as shown in Figure 2. Compressed values are recorded here and they act as vector K in Equation 2. Next, seven different sources, including two blogs ([1], [9]), three articles ([10], [11], [12], a story generated using Grok ([13]) and movie reviews from Hugging Face ([14] with test, train and validations were selected for this study. Text corpora from these sources where sent through a pre-processing pipeline to extract sentences in ASCII codes. These ASCII coded sentences were then reduced using Equation 1 to seventeen dimensions. There was no loss in explained variance once again for every text corpus. Explained variance for the entire text corpus (when combined as one, it is a 2-D array of shape (29326,446)) is shown in Figure 3. The ASCII corpora, taken one source at a time, tokenized by sentences, converted to ASCII and compressed to seventeen dimensions make up $\bar{J}$ in Equation 2. The dot product was computed and analyzed using measures mentioned in the following sections. The entire jupyter notebook can be viewed at [15].
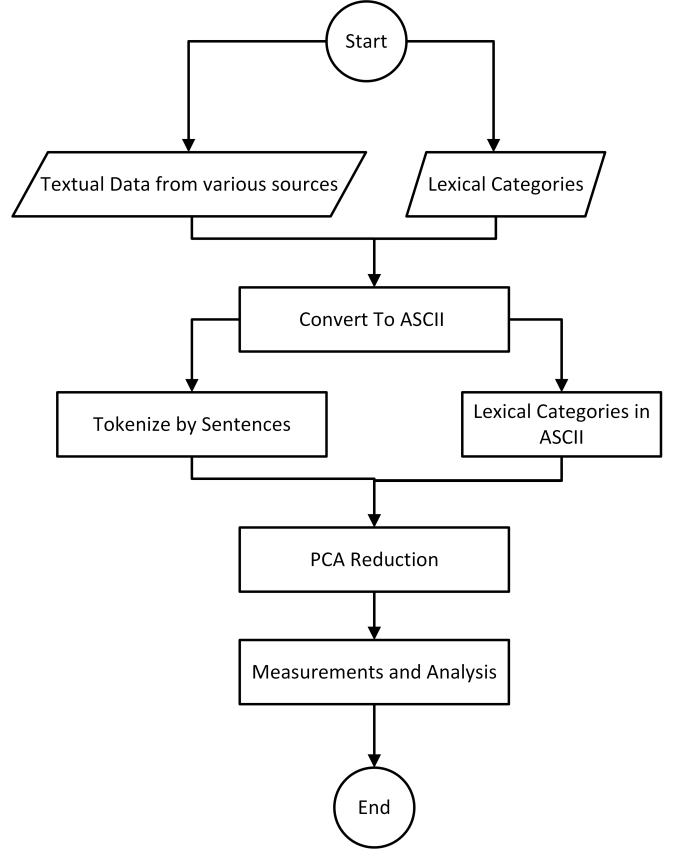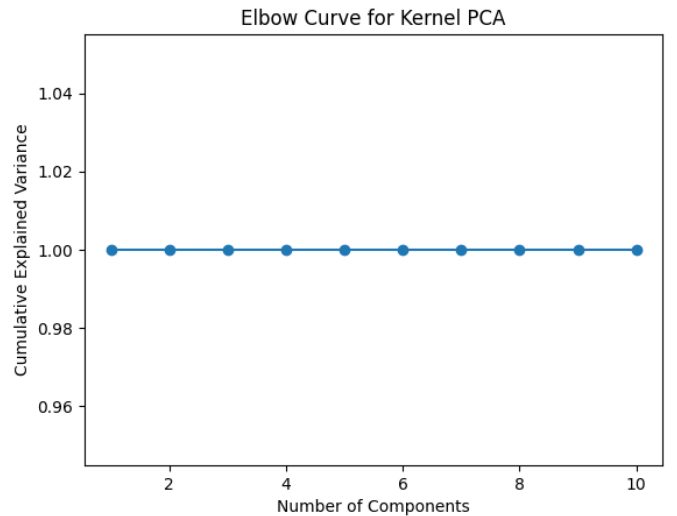

Fig. 1: Experimental Setup
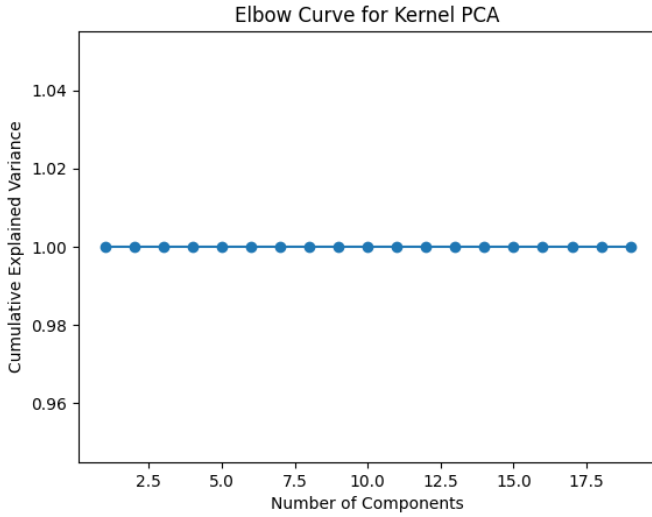

Fig. 2: Explained Variance

Fig. 3: Explained Variance for the Entire Text Corpus

IV. STATISTICAL MEASURES

A. Shapiro-Wilk Test

The authors used [16] to assess normality of the dot product distribution, which the Shapiro-Wilk test does by evaluating if a sample is from a normally distributed population by comparing the observed order statistics to expected values under normality. The test statistic S is as follows:

$$S = \frac{\left(\sum_{i=1}^{a} n_i g_{(i)}\right)^2}{\sum_{i=1}^{a} (g_i - \bar{g})^2} \tag{3}$$

where $g_{(i)}$ are the ordered sample values (i.e., the i-th order statistic), g_i are the original sample values, $\bar{g}$ is the sample mean, and n_i are coefficients derived from the expected values of order statistics of a standard normal distribution. A value of S close to 1 suggests the data is normally distributed, while a small S indicates deviation (less than 0.05) from normality. This test is used on datasets that are small or moderates.

B. Anderson-Darling Test

The Anderson-Darling Test [17] is another normality test that places more emphasis on the tails of the distribution compared to other tests like Shapiro-Wilk. It measures the distance between the empirical cumulative distribution function (*ECDF*) of the sample and the cumulative distribution function (*CDF*) of a normal distribution. The test statistic A^2 is defined as:

$$A^2 = -a - \frac{1}{a}\sum_{i=1}^{a}(2i-1)\left[\ln(F(t_{(i)})) + \ln(1 - F(t_{(a+1-i)}))\right] \tag{4}$$

where a is the sample size, $t_{(i)}$ are the ordered sample values, and $F(t_{(i)})$ is the CDF of the normal distribution evaluated at $t_{(i)}$. A larger A^2 value indicates a greater departure from normality. This test is especially sensitive to deviations in the tails. In practice, as implemented in Python's *scipy.stats.anderson* function, the test does not directly provide

(a) Blog_ia

(b) Guardian's article

(c) Article

(d) Blog_ref

(e) Generated using Grok

(f) Test Data

(g) Validation Data

(h) Train Data

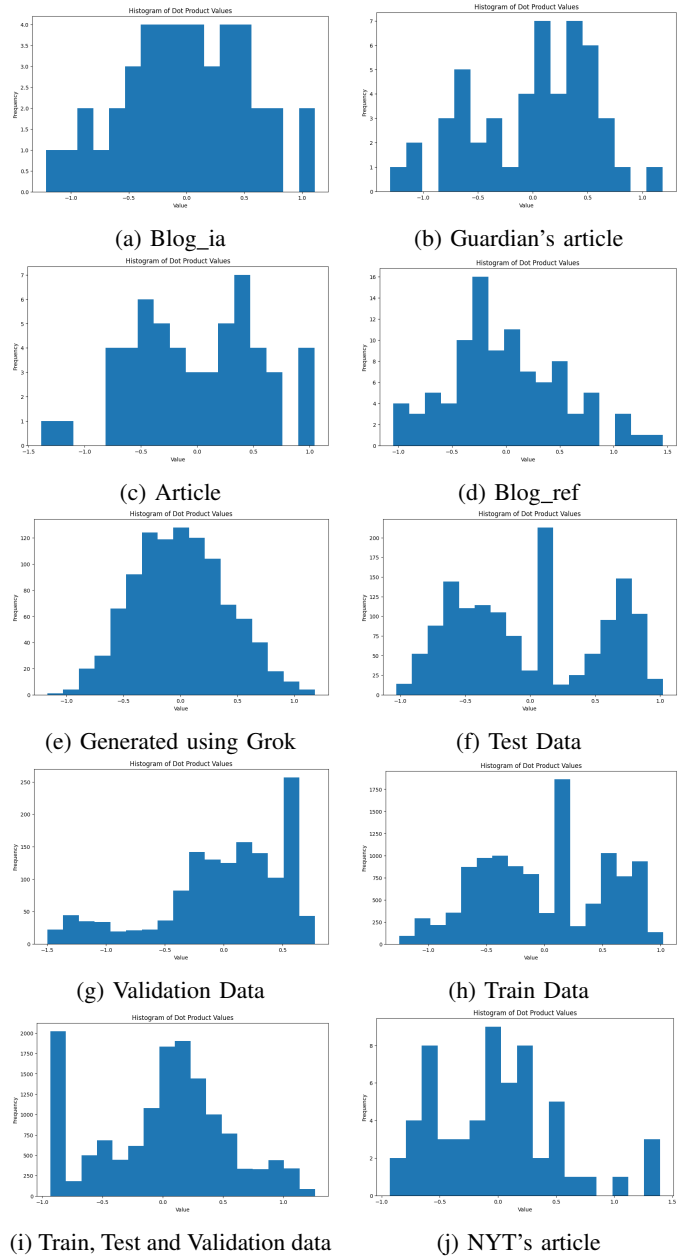(i) Train, Test and Validation data

(j) NYT's article

Fig. 4: Histogram of dot products per experiment

a p-value. Instead, the computed statistic A^2 is compared to critical values at specified significance levels (e.g., 5%, corresponding to $\alpha = 0.05$). If A^2 exceeds the critical value, the null hypothesis, i.e, the data follows a normal distribution is rejected, indicating non-normality; otherwise, it is accepted.

V. RESULTS

Table I depicts the results of the various experiments conducted along with their tests for normality. In either tests, four experiments clear the normality tests. The LLM generated story fails to clear the normality test after lexical alignment inspite of showing a bell curve in Figure 4e. Movie reviews dataset, either train 4h, test 4f, validation set 4g, don't clear the

TABLE I: Shapiro-Wilk 3 and Anderson-Darling Test 4 results for Normality Across Experiments

Experiment	Sample Size	Shapiro Statistic	Shapiro P-Value	Shapiro Normal	A-D Statistic	AD Critical Value (5%)	A-D Normal
Blog_ia	43	0.991 321	0.983 950	Yes	0.092 526	0.729	Yes
Guardian	50	0.964 547	0.137 800	Yes	0.757 544	0.736	No
Mus	54	0.977 142	0.387 746	Yes	0.410 529	0.739	Yes
NYT	60	0.953 110	0.021 862	No	0.680 246	0.743	Yes
Blog_ref	96	0.979 726	0.143 459	Yes	0.604 613	0.757	Yes
Grok	1007	0.995 795	0.007 514	No	1.076 552	0.784	No
Reviews_test	1402	0.934 546	2.47×10^{-24}	No	30.344 235	0.785	No
Reviews_val	1411	0.908 294	2.53×10^{-28}	No	35.084 494	0.785	No
Reviews_train	11 197	0.962 414	9.56×10^{-47}	No	130.846 467	0.787	No
Reviews_all	14 010	0.954 351	5.00×10^{-54}	No	191.833 786	0.787	No
Combined	29 326	0.931 265	1.36×10^{-76}	No	820.083 099	0.787	No

(a) Blog_ia (b) Guardian's article (c) Article (d) Blog_ref (e) Generated using Grok

(f) Test Data (g) Validation Data (h) Train Data (i) Train, Test and Validation data (j) NYT's article
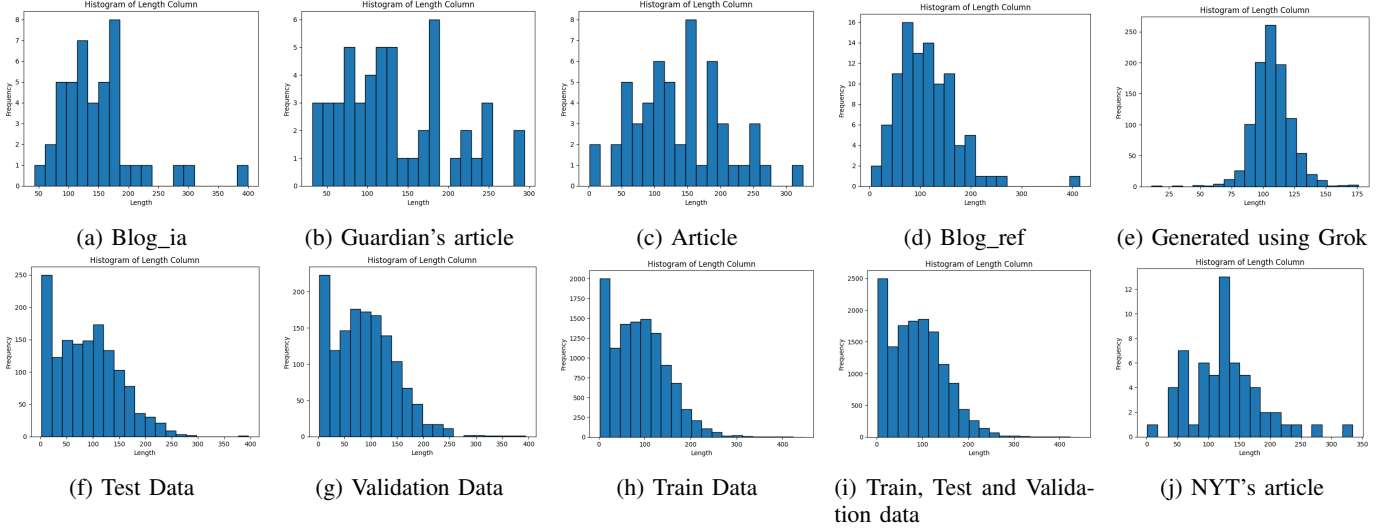
Fig. 5: Histogram of length of sentences

normality tests. Blogs and articles picked up from the Internet clear these normality tests indicating that their structural content in terms of the usage of lexical resources is well balanced.

VI. ANALYSIS AND DISCUSSION

Analyzing the text of [1], the authors noticed that the variation in sentence structures, i.e, the usage of simple, compound and complex sentences was balanced out throughout the article, which was the motivation behind lexical alignment and can be observed with clearing the normality test from I (Blog_ia) and Figure 4a. Similar trend was observed in [12] which also clears the statistical tests. Notice that the histogram of length of sentences for the text corpus that clear the test have multiple values for it in Fig 5. Taking Equation 2 and Figure 4 into consideration that highlights that a plethora of samples have a dot product value of zero, indicating orthogonality with the vector $\bar{K}$, however, the authors do understand the limitations of such stringent representations. The ones that don't clear the test seem to skewed in their distribution of length indicating a limitation in style of content as observed in Figures 5f, 5g, 4h, 4i which could be observed through variations in their dot product graphs as having submodular distributions (see Figures

4f, 4g, 4h, and 4i. As for the LLM generated story, viewing Figure 5e explains that it does not generate a sentence greater than 200 in length indicating no usage of a certain lexical resources too, which is found in Figures 5a, 5b, 5c and 5d. While the generated content did not have any errors either in spellings or meanings, there are instances where lexical categories occur in unnatural positions such as the usage of preposition 'of' to end a sentence, further research must be done. This approach proposed here complements the existing preprocessing NLP pipeline to analyze text corpora from a different context.

REFERENCES

[1] S. Altman, "Ia," https://ia.samaltman.com, 2025, accessed: 2025-03-11.
[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, and et al., "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774
[3] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, and et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530
[4] K. Omelianchuk, A. Liubonko, O. Skurzhanskyi, A. Chernodub, O. Korniienko, and I. Samokhin, "Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2404.14914

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[6] G. Churakov, "Pos-tagging to highlight the skeletal structure of sentences," 2024. [Online]. Available: https://arxiv.org/abs/2411.14393

[7] S. Jaffali and S. Jamoussi, "Text document dimension reduction using principal component analysis," 02 2012.

[8] J. Shlens, "A tutorial on principal component analysis," 2014. [Online]. Available: https://arxiv.org/abs/1404.1100

[9] S. Altman, "Reflections," https://blog.samaltman.com/reflections, 2025, accessed: 2025-03-11.

[10] T. Guardian, "The lesbian mothers scandal: Custody cases and families," https://www.theguardian.com/commentisfree/2024/dec/18/lesbian-mothers-scandal-custody-cases-families, 2024, accessed: 2025-03-11.

[11] T. N. Y. Times, "How to hide a wi-fi router to maintain a strong signal," https://www.nytimes.com/wirecutter/reviews/hide-wifi-router-maintain-signal/?campaign_id=290&emc=edit_wcd_20250127&instance_id=145906&nl=the-recommendation&regi_id=228774333&segment_id=189426&user_id=190509cfe7a0a183c26b1d1dd1d21fc6, 2025, accessed: 2025-03-11.

[12] NPR, "Museums turn to immersive tech to preserve the stories of aging holocaust survivors," https://www.npr.org/2022/04/10/1089652445/museums-turn-to-immersive-tech-to-preserve-the-stories-of-aging-holocaust-surviv, 2022, accessed: 2025-03-11.

[13] Jeet1912, "Cs634 presentation story," https://github.com/jeet1912/ms/blob/main/cs634/presentation/story.docx, 2025, accessed: 2025-03-11.

[14] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the ACL*, 2005.

[15] Jeet1912, "Final notebook for ds680 final term project," https://github.com/jeet1912/ms/blob/main/ds680/finalTermProject/finalNotebook.ipynb, 2025, accessed: 2025-03-12.

[16] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965. [Online]. Available: http://www.jstor.org/stable/2333709

[17] A. trujillo ortiz, R. Hernandez-Walls, K. Barba-Rojo, and A. Castro-Perez, "Andartest anderson-darling test for assessing normality of a sample data." 04 2007.