Semantic Analysis of Jurisprudential Zoroastrian Texts in Pahlavi: A Word Embedding Approach for an Extremely Under-Resourced, Extinct Language

Rashin Rahnamoun
Shahid Beheshti University, Tehran, Iran
rahnamounrashin@gmail.com

Ramin Rahnamoun
Central Tehran Branch, Islamic Azad University, Tehran, Iran
r.rahnamoun@iauctb.ac.ir

Abstract

Zoroastrianism, one of the earliest known religions, reached its height of influence during the Sassanian period, embedding itself within the governmental structure before the rise of Islam in the 7th century led to a significant shift. Subsequently, a substantial body of Zoroastrian literature in Middle Persian (Pahlavi) emerged, primarily addressing religious, ethical, and legal topics and reflecting Zoroastrian responses to evolving Islamic jurisprudence. The text *Šāyist nē šāyist* (Licit and Illicit), which is central to this study, provides guidance on purity and pollution, offering insights into Zoroastrian legal principles during the late Sassanian period. This study marks the first known application of machine processing to Book Pahlavi texts, focusing on a jurisprudential Zoroastrian text. A Pahlavi corpus was compiled, and word embedding techniques were applied to uncover semantic relationships within the selected text. Given the lack of digital resources and data standards for Pahlavi, a unique dataset of vocabulary pairs was created for evaluating embedding models, allowing for the selection of optimal methods and hyperparameter settings. By constructing a complex network using these embeddings, and leveraging the scarcity of texts in this field, we used complex network analysis to extract additional information about the features of the text. We applied this approach to the chapters of the *Šāyist nē šāyist* book, uncovering more insights from each chapter. This approach facilitated the initial semantic analysis of Pahlavi legal concepts, contributing to the computational exploration of Middle Persian religious literature.

1 Introduction

Zoroastrianism is one of the world's oldest religions, with origins dating back approximately three thousand years to the Avestan period. Historical texts suggest that King Goshtasp (Vishtaspa) Kiani was the first monarch to adopt Zoroastrianism; however, some scholars regard him as a mythical figure (Boyce, 2017). The Sassanian period, however, marked the height of Zoroastrian influence within governmental structures, establishing the religion as an integral component of the state (Tessmann, 2022). In the 7th century, the Sassanid Empire fell with the expansion of Muslim forces, marking a significant shift in Zoroastrianism's societal role. With the loss of governmental authority, Zoroastrian scholars placed greater emphasis on defending and preserving their beliefs. Consequently, a substantial portion of Zoroastrian religious literature was produced post-Sassanid Empire, during the Islamic Caliphate (Choksy, 1987).

The Persian language, a member of the Indo-Iranian subgroup of the Indo-European language family, has evolved through three main historical phases: Old, Middle, and Modern Persian (Windfuhr, 2013). Old Persian was the primary language during the Achaemenid Empire, originating in the Fars region of present-day Iran. Middle Persian, prevalent during the Sassanian period and early Islamic Caliphate, served as the primary language for religious and scholarly texts of that time.

Modern Persian emerged around the 10th and 11th centuries, spreading throughout the remnants of the Sassanian Empire's territory. Surviving Middle Persian texts primarily address Zoroastrian religious themes, covering topics such as philosophy, prophecy, ethics, exhortation, debate, jurisprudence, and law (Daryaee, 2018). As mentioned earlier, during the Sassanian era, religious debates between Zoroastrian jurists and their counterparts from other faiths became increasingly intense. Additionally, significant interpretative differences emerged within the Zoroastrian scholarly community itself. Following the fall of the Sassanian kingdom and the establishment of Islamic rule in former Zoroastrian territories, numerous questions arose concerning religious laws and practices. This trend began under the Sassanians, where various Zoroastrian legal scholars debated the correct

Proceedings of the New Horizons in Computational Linguistics for Religious Texts (Coling-Rel), pages 23–41
January 19, 2025. ©2025 Association for Computational Linguistics

interpretation of religious law (Janos, 2005).

One of the key works addressing these issues is *Šāyist nē šāyist* (Licit and Illicit), which provides religious rulings on topics such as purity and pollution, offering authoritative perspectives on common questions of the time. This book, along with other Middle Persian texts, reflects Zoroastrian scholars' inquiries into the relationship between Zoroastrian legal principles at the end of the Sassanian period and emerging Islamic jurisprudence (Janos, 2005).

In this study, to the authors' knowledge, machine processing of Book Pahlavi texts is explored for the first time. The selected text is a jurisprudential Zoroastrian text, with the goal of uncovering semantic relationships between words and phrases within the text. Word embedding methods were applied for this purpose. It is important to note that Pahlavi is an extinct and extremely under-resourced language[1], which poses unique challenges for machine processing.

To embed words in the text, a Book Pahlavi corpus was prepared, and word embeddings were generated from this corpus. To evaluate the embeddings, a dataset of vocabulary pairs was created, allowing for the comparison of different methods. Based on this comparative dataset, the optimal method and hyperparameter settings were selected. Using these models, semantic analysis of words in the Pahlavi legal text was performed, representing, to the authors' knowledge, the first time these processes have been conducted in this context.

We constructed a complex network of words from the *Šāyist nē šāyist* book by representing each word as a node, with edges created based on cosine similarity of their embeddings. Analyzing structural properties such as degree distribution, clustering coefficient, and degree centrality enabled us to explore word relationships and semantic patterns, offering deeper insights into this extremely under-resourced language.

2 Related Work

2.1 Middle Persian Language (Pahlavi)

Middle Persian, or Pahlavi, was the language used by Zoroastrian scholars for religious texts during the Sassanian period and the early Islamic Caliphate. At the same time, it was also employed for Christian and Manichaean writings, and Sassanian inscriptions (Boyce, 1990). However, since this article focuses specifically on Zoroastrian religious texts in the Pahlavi language, these other areas will not be discussed, despite their importance in the overall corpus of Pahlavi literature.

The Pahlavi language was written in various scripts, all derived from the Aramaic script, which served as the official court language during part of the Achaemenid period. One of these scripts is Manichaean, developed by the Prophet Mani, and several texts written in this script have survived (Goshtasb et al., 2021). Another is Pazand, written in the Avestan alphabet, introduced to make Zoroastrian texts in Pahlavi more accessible, as the original Pahlavi script posed challenges for readers of Avestan. The Pahlavi inscription script (MacKenzie, 2014), used for Sassanian inscriptions, consists of separate letters. Finally, the Pahlavi book script, used specifically for Zoroastrian religious writings, is the primary focus of this article (Cereti et al., 2005).

The Book Pahlavi language presents complexities at multiple levels. One challenge is that certain letters are connected through ligatures, making them difficult to read. Additionally, many letters in this script are ambiguous, allowing a single word to be interpreted in several different ways. This characteristic has led to multiple interpretations of Pahlavi Zoroastrian texts. Furthermore, adjacent letters often form shapes that can resemble various sequences of letters, adding further layers of interpretative difficulty. Another significant issue arises from the fact that Book Pahlavi is an extinct language that has not been spoken for centuries. Consequently, there are substantial disagreements over the meanings of certain words. Like many other languages, Book Pahlavi contains ambiguity in word meanings, with some words exhibiting homonymy or polysemy. At the sentence level, additional ambiguities make interpreting text meaning particularly challenging. Currently, three primary online sources provide access to Pahlavi texts. The first is the Thesaurus Indogermanischer Sprach- und Textmaterialien (TITUS) [2], which has made a substantial collection of Pahlavi texts publicly available for years (Gippert, 2002). The second source is the Parsig database, (Goshtasb et al.,

[1] The standard terminology refers to it as a low-resource language, though terms like "extinct" and "extremely under-resourced language" (Coto-Solano, 2022b) are also used, as these languages lack an active speaking community and are often ancient. Moreover, there is a shortage of datasets due to the scarcity of remaining resources available for these languages.

[2] https://titus.uni-frankfurt.de/

2021) [3]an open-access web-based resource that provides Pahlavi texts in their original script, along with transliterations, transcriptions, and translations into both Persian and English. Furthermore, since the Book Pahlavi character set has not yet been added to the Unicode standard [4], the texts in Pahlavi script use the site's own custom font.

The third source is the Zoroastrian Middle Persian Digital Corpus and Dictionary (MPCD) [5], which offers a substantial collection of Pahlavi texts that have been transliterated and transcribed (Neuefeind et al., 2022). It also includes English and German translations of these texts, as well as a Pahlavi-English dictionary. Notably, the MPCD provides a REST API on its website, allowing easy access to its data sources. Consequently, this research utilizes the MPCD database as a primary source of information.

2.2 Word Embedding Methods

Word embedding is a type of shallow neural network widely used in Natural Language Processing (NLP) tasks (Torregrossa et al., 2021). In this approach, words are represented as vectors and mapped to a lower-dimensional latent space that preserves semantic properties. Within this space, words with similar meanings are located close to one another, facilitating the capture of semantic relationships (Patil et al., 2023).

In this study, we use two popular static word embedding models: Word2Vec and FastText. Word2Vec (Mikolov et al., 2013) is a static word embedding algorithm that can be implemented with two different models: Continuous Bag of Words (CBOW) and Skip-Gram (SG). The core idea behind this algorithm is that words appearing in similar contexts are likely to have similar meanings. However, Word2Vec does not capture word order within a sentence, treating each word as an independent unit. In contrast, FastText (Bojanowski et al., 2017) incorporates subword information by representing each word as a set of character n-grams, making it sensitive to word morphology and capable of handling rare and misspelled words.

2.3 Word Embedding and Evaluations in Low-Resource Languages

Evaluations in word embeddings for low-resource languages is more challenging than for high-resource languages, as these models require large text corpora, which are often unavailable for low-resource languages (Arppe et al., 2023). However, studies have demonstrated promising results even with limited datasets for such languages (Coto-Solano, 2022a). While word embedding evaluation in high-resource languages is widely studied and supported by numerous evaluation benchmarks, it remains a challenging task in low-resource languages (Ngomane et al., 2023).

Two primary evaluation methods exist: intrinsic and extrinsic. Intrinsic evaluation includes techniques such as analogy and categorization tasks; in low-resource languages, OddOneOut and Top-k are particularly common (Stringham and Izbicki, 2020). Extrinsic methods involve using word embeddings within an NLP application (e.g., text classification) and evaluating performance based on the application's outcomes. However, since this study does not apply word embeddings in a specific extrinsic task, extrinsic evaluation is beyond our scope. Different evaluation methods have been successfully applied to a range of low-resource languages, yielding insightful results (Coto-Solano, 2022a; Ngomane et al., 2023; Lugli et al., 2022; Lakmal et al., 2020).

2.4 Texts and Complex Networks

Representing texts as graphs, such as knowledge graphs (Chen et al., 2020), is not a new concept. Previous works have also utilized complex network representations of text and applied statistical analyses, such as (Ferraz de Arruda et al., 2018), (Stanisz et al., 2024). One challenge with most complex network analysis algorithms is their high computational complexity, which makes them impractical for very large datasets. However, in our study, we focus on Pahlavi, an extinct and extremely under-resourced language, where available data is limited. This allows us to construct and perform complex network analyses, such as calculating clustering coefficients (Holland and Leinhardt, 1971), (Soffer and Vazquez, 2005) and degree centrality distributions (Zhang and Luo, 2017), to gain a deeper understanding of the language's features in written texts.

3 Methodology

In this paper, we first examine various word embedding methodologies applied to Pahlavi, an extinct and extremely under-resourced language in

[3] https://www.parsigdatabase.com/
[4] https://www.unicode.org/standard/unsupported.html
[5] https://www.mpcorpus.org/

Table 1: This table summarizes the key characteristics of the Pahlavi texts analyzed in this study, including the file names, total lines, total tokens, unique tokens, average sentence length, and the length of the longest sentence in each text. The data highlights the variability in text length and complexity, as well as differences in token diversity across the texts.

File Name	**Total Lines**	**Total Tokens**	**Unique Tokens**	**Avg Sentence Length**	**Longest Sentence**
Great Bundahišn	2,393	36,894	4,343	15.42	339
Dādestān ī dēnīg	1,144	18,278	3,325	15.98	138
Dādestān ī mēnōy ī xrad	737	10,887	1,672	14.77	98
Dēnkard 3	2,472	80,507	7,747	32.57	288
Dēnkard 6	2,382	27,612	2,329	11.59	45
Dēnkard 8	1,031	23,577	3,088	22.87	102
Dēnkard 9	1,019	35,627	3,811	34.96	318
Hazār dādestān	1,743	38,750	1,751	22.23	131
Nāmagīhā ī manuščihr	343	11,001	2,028	32.07	276
Nērangestān	2,525	30,114	3,739	11.93	100
Pahlavi Wīdēwdād	3,404	53,448	4,039	15.70	131
Pahlavi Yasna	2,832	58,649	3,663	20.71	228
Rivāyat of Ādurfarnbay	1,086	12,341	1,286	11.36	90
Šāyist nē šāyist	1,320	16,575	2,061	12.56	73
Wizīdagīhā ī zādspram	913	18,240	3,053	19.98	128
Zand ī pargard ī juddēwdād	3,015	35,600	2,641	11.81	52

which most surviving texts center around religious themes. Given that this is the first study of its kind on Pahlavi, there is no comparable prior research; therefore, we conducted our own evaluation tasks to determine the most effective embedding approach, as detailed in the following sections.

Additionally, we investigate the religious text *Šāyist nē šāyist*, a Zoroastrian Middle Persian compilation of diverse laws and customs concerning sin, ritual purity, and various ceremonial and religious practices (Müller, 1880). Using the most effective embeddings identified through our optimized hyperparameters, we construct a complex network of nodes and edges for this text, applying network analysis to illuminate its structural features. This process illustrates how computational methods can provide a more in-depth examination of this text.

3.1 Data Collection and Preparation

As previously noted, Book Pahlavi texts are written in a connected script that is challenging to read, and this alphabet is not yet supported by the Unicode standard. Consequently, the source[6] used in this study provides transcriptions of the texts rather than the original Book Pahlavi script. Thus, the input files of the analyzed texts contain transcriptions rather than the original Book Pahlavi characters. Table 1 presents key information and specifications of the texts used in this study.

The preliminary analysis of Table 1 indicates that the *Šāyist nē šāyist* text lacks sufficient volume to yield reliable results using various word embedding methods, a limitation confirmed during implementation. Consequently, other accessible Pahlavi texts with an adequate word count were utilized in this study. Tokenization of the texts was straightforward, as words were already separated by spaces. After preparing the text corpus, a method was needed to evaluate various word embedding techniques. This presented unique challenges in the context of the Pahlavi language, as no prior studies exist in this area. Using English benchmark datasets, such as *WordSim353*, was not effective. Pahlavi texts primarily focus on the Zoroastrian religion, and *Šāyist nē šāyist* specifically addresses religious rulings within Zoroastrianism. Although most words in these texts can be translated into English, their semantic similarity rarely aligns with the types of semantic relationships found in datasets like *WordSim353*. Following these considerations, this research employs a set of paired vocabulary items divided into two categories: related and unrelated words. The first category includes related words, encompassing several types: relational nouns and adjectives, such as *pašm* (wool) and *pašmēn* (woolen); singular and plural forms, like *yašt* (a type of prayer) and *yaštan* (praying); compound nouns and their root nouns, such as *dād* (justice) and *dādestān* (judgment); and synonyms, like *mōy* (hair) and *nāxun* (nail). The second category consists of unrelated words, for which it is expected that their generated vectors will not be similar. A total of 54 word pairs were selected

[6] https://www.mpcorpus.org/

and used to compare the performance of various methods and hyperparameter settings.

3.2 Word Embeddings

We conducted experiments on our corpus, as shown in Table 1, using two word embedding methodologies: Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). We tested various hyperparameters, including vector size, window size, and the Skip-Gram model. The results are presented in Section 4.

One common approach for word embedding in low-resource languages is Cross-lingual Word Embedding, where a high-resource language is used as the primary source to support a language with limited resources (Ngomane et al., 2023). This method has been successfully applied to many low-resource languages, yielding promising results. However, a key consideration for Zoroastrian texts in Pahlavi is that the word similarities between Zoroastrian religious concepts are vastly different from those in high-resource languages. Therefore, this approach is not pursued in this study.

3.3 From Text to Networks

Inspired by (Ferraz de Arruda et al., 2018), we developed our own text representation in a network format to enable more detailed analysis of the chapters of the *Šāyist nē šāyist* book, aiming to investigate its features more thoroughly. We define a graph $G = (V, E)$, where each node $v \in V$ represents a unique word in the corpus. The edges $e_{ij} \in E$ between words v_i and v_j are based on the similarity of their embeddings, as calculated with cosine similarity.

For each word v, its embedding $\mathbf{w}$ can be generated using either the Word2Vec or FastText model.

3.3.1 Cosine Similarity and Edge Creation

To determine whether an edge e_{ij} should exist between two words v_i and v_j, we calculate the cosine similarity between their embeddings $\mathbf{w_i}$ and $\mathbf{w_j}$:

$$cosine\_similarity(v_i, v_j) = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\|\mathbf{w_i}\| \|\mathbf{w_j}\|}$$

An edge e_{ij} is created between v_i and v_j if their cosine similarity is greater than a threshold T_M. In our experiments, we set T_M to values of 0.5, 0.65 and 0.75, but any other threshold can also be used based on the specific analysis requirements.

3.3.2 Graph Preprocessing

Due to the fact that nodes v with no connecting edges, e_{ij}, which are isolated, were removed from the graph, as our focus was on analyzing words with similar meanings. Additionally, nodes v with a degree of $deg(v) \leq 2$ were also removed to focus the analysis on more significant nodes.

3.4 Analytic of Complex Networks

In the analysis of complex networks, several key metrics help us understand the structure and significance of nodes within the graph. In this section, we define and calculate the degree distribution, clustering coefficient, and degree centrality for the network we constructed from *Šāyist nē šāyist* book.

3.4.1 Degree Distribution

In the context of our text-based network for the book, the degree of a node v (representing a unique word) is defined as the number of edges incident to v, which corresponds to the number of similar-word connections it has within the network. The degree distribution, denoted by $P(k)$, represents the probability that a randomly chosen word node has exactly k connections (Barabási, 2013). Mathematically, it can be expressed as:

$$P(k) = \frac{V_k}{V}$$

where:

- V_k is the number of nodes in the graph with degree k,
- V is the total number of nodes in the graph.

3.4.2 Clustering Coefficient

The clustering coefficient is a measure of how closely nodes in a network tend to form clusters. It quantifies the probability that the neighbors of a given node are also connected to one another, which can help in identifying communities within the network (Holland and Leinhardt, 1971), (Soffer and Vazquez, 2005). The words v that are strongly connected within communities due to their similarity, as indicated by the edges, are expected to have meaningful connections with other words.

For a given node v (representing a word in the text), the local clustering coefficient $C(v)$ is defined as:

$$C(v) = \frac{2e_v}{k_v(k_v - 1)}$$

where:

- e_v is the number of edges between the neighbors of node v,
- k_v is the degree of node v, or the number of neighbors connected to v.

The global clustering coefficient C of the entire network is then calculated as the average of the local clustering coefficients for all nodes:

$$C = \frac{1}{N} \sum_{v \in V} C(v)$$

where N is the total number of nodes in the network and V is the set of all nodes.

3.4.3 Degree Centrality

Degree centrality (Zhang and Luo, 2017) reflects the importance and influence of a node in a network, which in our case is a word, based on its connections that indicate similarity in meaning to other words. It highlights how significant a word is within the text by analyzing its semantic relations to others. It can be defined as:

$$C_D(v) = \frac{k_v}{V - 1}$$

where:

- k_v is the degree of node v,
- V is the total number of nodes in the network.

4 Experiments

In this section, we first experiment with different word embeddings, then construct a complex network using these embeddings. We apply network-based metrics to extract features from the *Šāyist nē šāyist* book, and subsequently analyze and discuss the results.

4.1 Results

In Table 2, we experimented with various configurations, including window size, vector size, and the SG value, to compare the performance of different word embeddings for FastText and Word2Vec, evaluating their effectiveness for the first time in Pahlavi (Bojanowski et al., 2017)(Mikolov et al., 2013).

- **Window size**: By defining the range of words that are taken into account around a target word, this parameter enables the model to record contextual information.

Table 2: Accuracy Comparison between FastText and Word2Vec Embeddings in percentage. All Word2Vec accuracy results are bolded. The best accuracy values for FastText and Word2Vec underlined. SG represents the Skip-Gram model (where 1 indicates Skip-Gram and 0 indicates CBOW). The vector size denotes the dimensionality of word embeddings, and the window size determines the range of neighboring words used for context.

Vector	Window	SG	FastText	Word2Vec
25	2	0	57.96	**64.40**
25	2	1	59.13	**67.78**
25	5	0	65.33	**68.88**
25	5	1	63.70	**70.66**
25	10	0	71.27	**74.36**
25	10	1	66.12	**70.16**
50	2	0	57.52	**62.86**
50	2	1	59.19	**68.52**
50	5	0	63.77	**68.84**
50	5	1	63.10	**68.74**
50	10	0	70.57	**73.51**
50	10	1	67.02	**67.90**
100	2	0	57.14	**60.52**
100	2	1	59.01	**67.97**
100	5	0	63.58	**67.84**
100	5	1	63.64	**69.69**
100	10	0	68.50	**73.28**
100	10	1	66.92	**68.25**

- **Vector size**: This is a reference to the word embeddings' dimensionality, or number of features.
- **SG (Skip-Gram) value**: The model type is specified by this option. The Skip-Gram model, which forecasts the words that surround a target word, is indicated by a value of 1. The Continuous Bag of Words (CBOW) model, on the other hand, predicts a target word based on its surrounding words and has a value of 0.

In the *Šāyist nē šāyist* book, a religious text containing 23 chapters, we selected and analyzed three chapters. The examination of the remaining chapters is similar in nature. These three chapters were chosen because they are conceptually related while also distinct from the others. The selected chapters include: the first chapter, which discusses sins; the third chapter, which addresses the rulings on women; and the twenty chapter , which focuses on advice and admonitions. The analysis is presented in a segmented manner with corresponding graphs which explained in Section 3.3 , and the results for chapters 1 in Fig. 1 and 20 in Fig. 2 are shown using Word2Vec, which demonstrates the

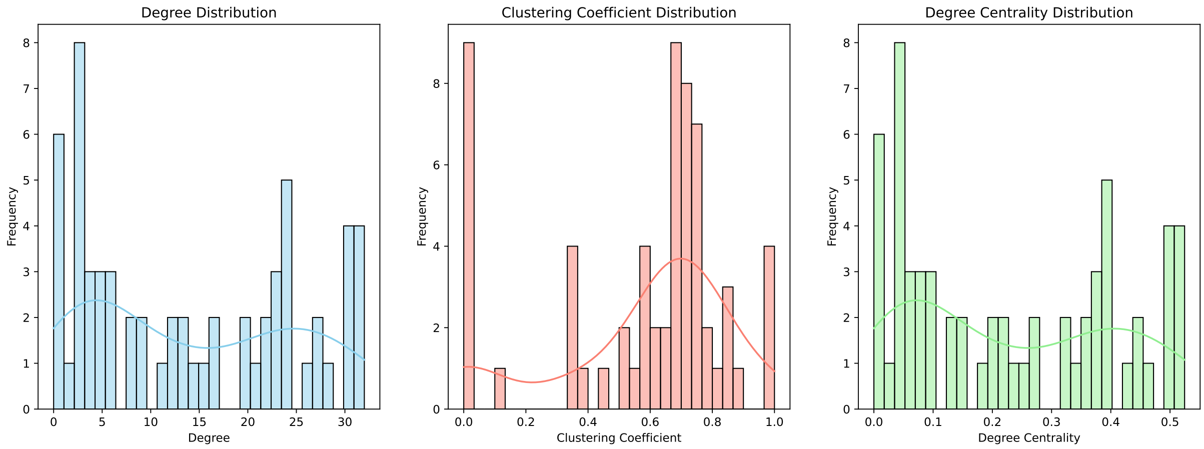

Figure 1: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.75$, Word2Vec.

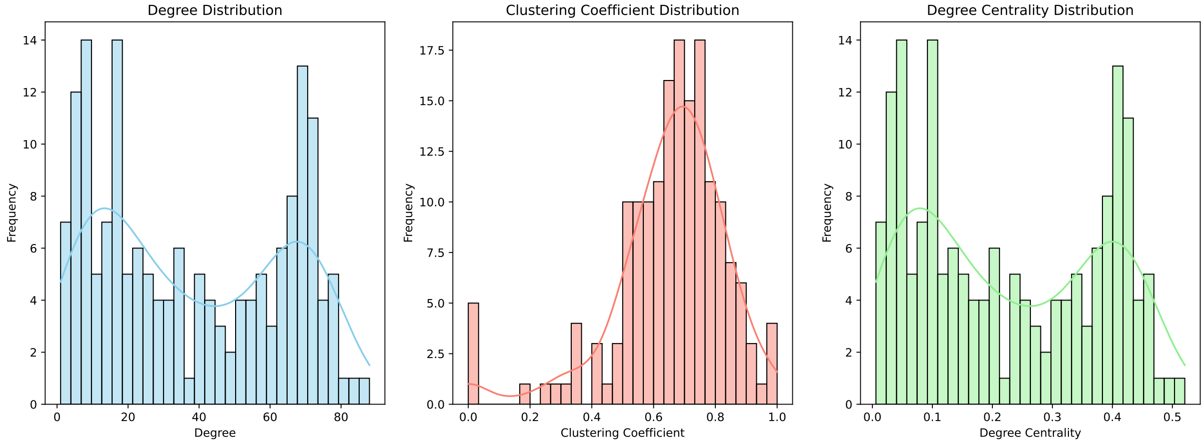

Figure 2: Clustering coefficient of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.75$, Word2Vec.

best performance with optimal hyperparameters in Table 2.

4.2 Discussion and Analysis

For the word embeddings presented in Table 2, we expected the best results when the vector size was minimized, the window size was maximized, and the Skip-Gram model was used. An interesting observation is that, for the Pahlavi language, the same hyperparameters in both Word2Vec and FastText produced the best results. Although Word2Vec slightly outperformed FastText in accuracy by approximately 3%, both models showed similar performance with these settings.

For a more detailed analysis, we examined the t-SNE visualization for both embeddings.The t-SNE visualization of FastText embeddings in two dimensions demonstrates this model's ability to capture the morphological structure of words in Pahlavi. In Fig. 3 in left side, a section of the t-SNE plot highlights words with the common suffix *ān*. For instance, *framān* (order) appears close to *mardōmān* (humans) in the embedding space, even though they are not semantically related. This outcome reflects FastText's sensitivity to the *ān* suffix, a common plural suffix in Pahlavi, which is frequently used in both singular and plural forms in our evaluation dataset. In another section of the t-SNE plot, shown in Fig. 3 in right side, words with the suffix *išn* are clustered near one another, forming a distinct group. While some words in this cluster have related meanings, others do not, reflecting the model's tendency to group words with similar morphological endings, regardless of semantic similarity.

The t-SNE visualization of Word2Vec embeddings in two dimensions demonstrates that this model can position semantically related words adjacently on the 2D map. In Fig. 4, *gētī* (the material world) and *mēnōy* (the spiritual world) appear close to each other, reflecting their related meanings. Another example is *yazadān* (god) and

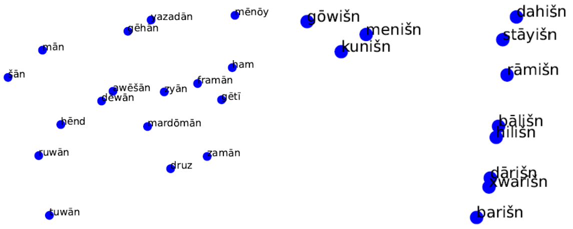

Figure 3: Parts of t-SNE visualization of FastText embeddings in two dimensions

zarduxšt ('Zarathustra'), which are also placed near one another due to their semantic association.

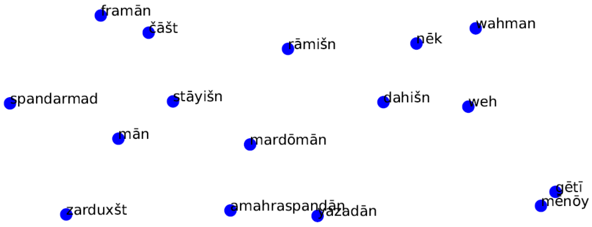

Figure 4: A part of t-SNE visualization of Word2Vec embeddings in two dimensions

The complete t-SNE visualization for both embeddings is provided in Appendix A.

By analyzing the book chapters, it is evident that Chapters 1 and 20 show significant differences in network structures and parameters. As shown in Figures 1 and 2, the clustering coefficient differs notably between these two chapters, indicating stronger, more meaningful connections and closer word associations in Chapter 20 compared to Chapter 1. Additionally, when we analyze degree centrality, we observe that words with stronger influence on contextual meaning appear more frequently in Chapter 20 than in Chapter 1. Moreover, for the degree distribution, high-degree words occur more frequently and in greater numbers in Chapter 20, suggesting that the words in Chapter 20 are more related and specific. In contrast, the words in Chapter 1 appear to be more general and less related to one another. All the graphs for these chapters can be found in the Appendix B.

5 Conclusions

In this paper, we present the first evaluation of word embeddings for Pahlavi, an extinct and extremely under-resourced Middle Persian language. After identifying suitable embedding parameters for this language, we constructed a complex network based on the chapters of the *Šāyist nē šāyist*, a religious book. The structure of this book, with its notably small chapters, facilitates easier analysis of its network. We extracted deeper features from each chapter, which could contribute to understanding this extinct language, particularly given the scarcity of remaining resources. While studies on high-resource languages often struggle with large graphs due to their high time complexity, our approach is different: the limited amount of available texts in Pahlavi allows us to leverage complex network analysis to gain a deeper understanding of the language, despite the typical challenges posed by big graphs in other contexts.

6 Limitations

In this study, an attempt was made, to the authors' knowledge, to embed words in the Pahlavi language for the first time, though significant limitations exist in this area. Zoroastrian text sources in Pahlavi face a lack of Unicode standardization, so Pahlavi texts are either unavailable in Pahlavi script across online sources or are presented in a custom font specific to the hosting site, complicating accessibility. This study employs transliteration, which places the challenges of word interpretation in the Pahlavi script onto the transliterator, potentially introducing inconsistencies. Like any language, Pahlavi has a unique morphological structure that requires stemming for each word. Although stemming has been extensively studied in modern Persian (a language closely related to Pahlavi), it remains complex, and thus this study does not address stemming.

Another major issue in the Pahlavi language is the lack of any dataset containing related or unrelated word groups with graded similarity or difference, as exists in English and modern Persian. Furthermore, other data types essential for evaluating word embedding methods, such as analogy datasets, are also absent in Pahlavi. This study presents a preliminary evaluation framework to compare different embedding models and adjust their hyperparameters; however, this dataset is basic and serves only for initial comparisons among methods. Consequently, it is evident that a more comprehensive dataset will be necessary for further progress in this field.

References

Antti Arppe, Andrew Neitsch, Daniel Dacanay, Jolene Poulin, Daniel Hieber, and Atticus Harrigan. 2023.

Finding words that aren't there: Using word embeddings to improve dictionary search for low-resource languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 144–155.

Albert-László Barabási. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Mary Boyce. 1990. *Textual sources for the study of Zoroastrianism*. University of Chicago Press.

Mary Boyce. 2017. Zoroastrianism. *A new handbook of living religions*, pages 236–260.

Carlo Giovanni Cereti et al. 2005. A middle persian dictionary: Project proposal. In *Orientalia Romana VIII: Middle Iranian Lexicography*, pages 181–190. IsIAO.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert systems with applications*, 141:112948.

Jamsheed K Choksy. 1987. Zoroastrians in muslim iran: selected problems of coexistence and interaction during the early medieval period. *Iranian Studies*, 20(1):17–30.

Rolando Coto-Solano. 2022a. Evaluating word embeddings in extremely under-resourced languages: A case study in bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

Rolando Coto-Solano. 2022b. Evaluating word embeddings in extremely under-resourced languages: A case study in bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

Touraj Daryaee. 2018. Middle persian (pahlavi). *A Companion to Late Antique Literature*, pages 103–121.

Henrique Ferraz de Arruda, Filipi Nascimento Silva, Vanessa Queiroz Marinho, Diego Raphael Amancio, and Luciano da Fontoura Costa. 2018. Representation of texts as complex networks: a mesoscopic approach. *Journal of Complex Networks*, 6(1):125–144.

Jost Gippert. 2002. Der titus-server: Grundlagen eines multilingualen online-retrieval-systems. *Historical Social Research/Historische Sozialforschung*, 27(1 (99):207–214.

Farzaneh Goshtasb, Masood Ghayoomi, and Nadia Hajipour Artarani. 2021. Corpus-based analysis of middle persian texts based on the pārsīg database. *Language Studies*, 12(1):255–280.

Paul W Holland and Samuel Leinhardt. 1971. Transitivity in structural models of small groups. *Comparative group studies*, 2(2):107–124.

Jany Janos. 2005. The four sources of law in zoroastrian and islamic jurisprudence. *Islamic Law and Society*, 12(3):291–332.

Dimuthu Lakmal, Surangika Ranathunga, Saman Peramuna, and Indu Herath. 2020. Word embedding evaluation for sinhala. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1874–1881.

Ligeia Lugli, Matej Martinc, Andraž Pelicon, and Senja Pollak. 2022. Embeddings models for buddhist sanskrit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3861–3871.

David Neil MacKenzie. 2014. *A concise Pahlavi dictionary*. Routledge.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Friedrich Max Müller. 1880. *The Sacred Books of the East: Pahlavi texts (pt. 1), translated by EW West*, volume 5. Clarendon Press, Oxford, UK.

Claes Neuefeind, Francisco Mondaca, Øyvind Eide, Iris Colditz, Thomas Jügel, Kianoosh Rezania, Arash Zeini, Alberto Cantera, Chagai Emanuel, and Shaul Shaked. 2022. Das zoroastrische mittelpersische digitales corpus und wörterbuch (mpcd). In *DHd*.

Derwin Ngomane, Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2023. Unsupervised cross-lingual word embedding representation for english-isizulu. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 11–17.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Sara Nadiv Soffer and Alexei Vazquez. 2005. Network clustering coefficient without degree-correlation biases. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(5):057101.

Tomasz Stanisz, Stanisław Drożdż, and Jarosław Kwapień. 2024. Complex systems approach to natural language. *Physics Reports*, 1053:1–84.

Nathan Stringham and Mike Izbicki. 2020. Evaluating word embeddings on low-resource languages. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 176–186.

Anna Tessmann. 2022. *The Wiley Blackwell Companion to Zoroastrianism*. John Wiley & Sons.

François Torregrossa, Robin Allesiardo, Vincent Claveau, Nihel Kooli, and Guillaume Gravier. 2021. A survey on training and evaluation of word embeddings. *International journal of data science and analytics*, 11(2):85–103.

Gernot Windfuhr. 2013. *The Iranian Languages*. Routledge.

Junlong Zhang and Yu Luo. 2017. Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*, pages 300–303. Atlantis press.

A t-SNE visualization

This section presents a complete t-SNE visualization of Word2Vec and FastText.

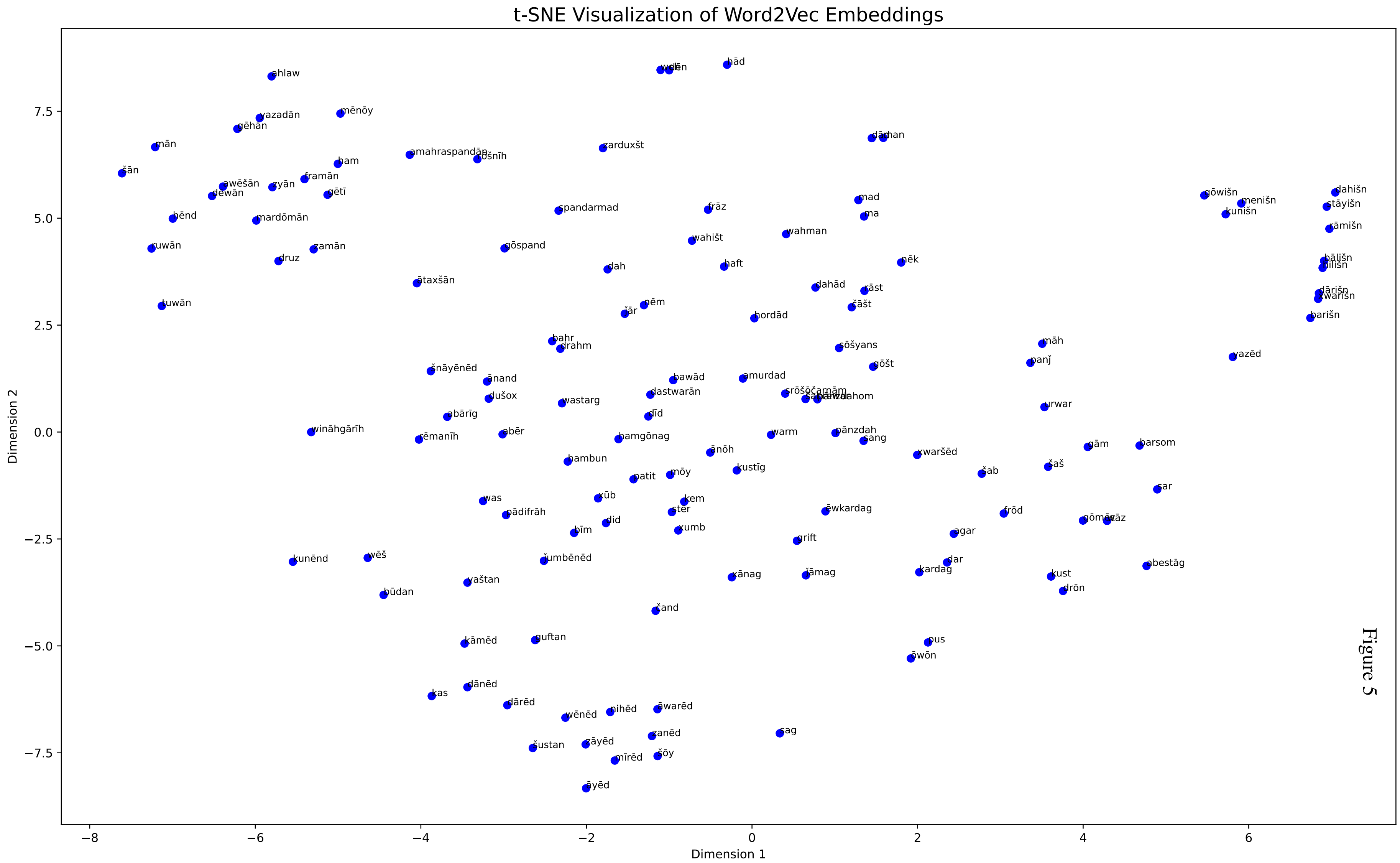
t-SNE Visualization of Word2Vec Embeddings
Dimension 1
Dimension 2
−8
−6
−4
−2
0
2
4
6
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
ahlaw
yazadān
mēnōy
gēhān
mān
šān
ham
framān
awēšān
zyān
dēwān
gētī
hēnd
mardōmān
ruwān
zamān
druz
tuwān
amahraspandān
rōšnīh
ātaxšān
zarduxšt
spandarmad
gōspand
dah
hād
frāz
wahišt
wahman
haft
nēm
jār
hordād
mad
ma
nēk
dahād
rāst
čāšt
sōšyans
gōšt
māh
panǰ
urwar
gōwišn
menišn
kunišn
dahišn
stāyišn
rāmišn
bālišn
hilišn
dārišn
xwarišn
barišn
yazēd
bahr
drahm
šnāyēnēd
ānand
dušox
abārīg
wastarg
bawād
dastwarān
amurdad
srōšōčarnām
dīd
wināhgārīh
rēmanīh
abēr
hamgōnag
warm
pānzdah
sang
ānōh
hambun
patit
mōy
kustīg
xwaršēd
šab
šaš
gām
barsom
sar
was
pādifrāh
xūb
kem
stēr
did
bīm
xumb
ēwkardag
grift
agar
frōd
gōmēz
abestāg
kunēnd
wēš
ǰumbēnēd
yaštan
būdan
xānag
ǰāmag
kardag
dar
kust
drōn
čand
kāmēd
guftan
pus
ōwōn
dānēd
kas
dārēd
wēnēd
nihēd
āwarēd
zanēd
sag
šustan
zāyēd
mīrēd
šōy
āyēd

Figure 5

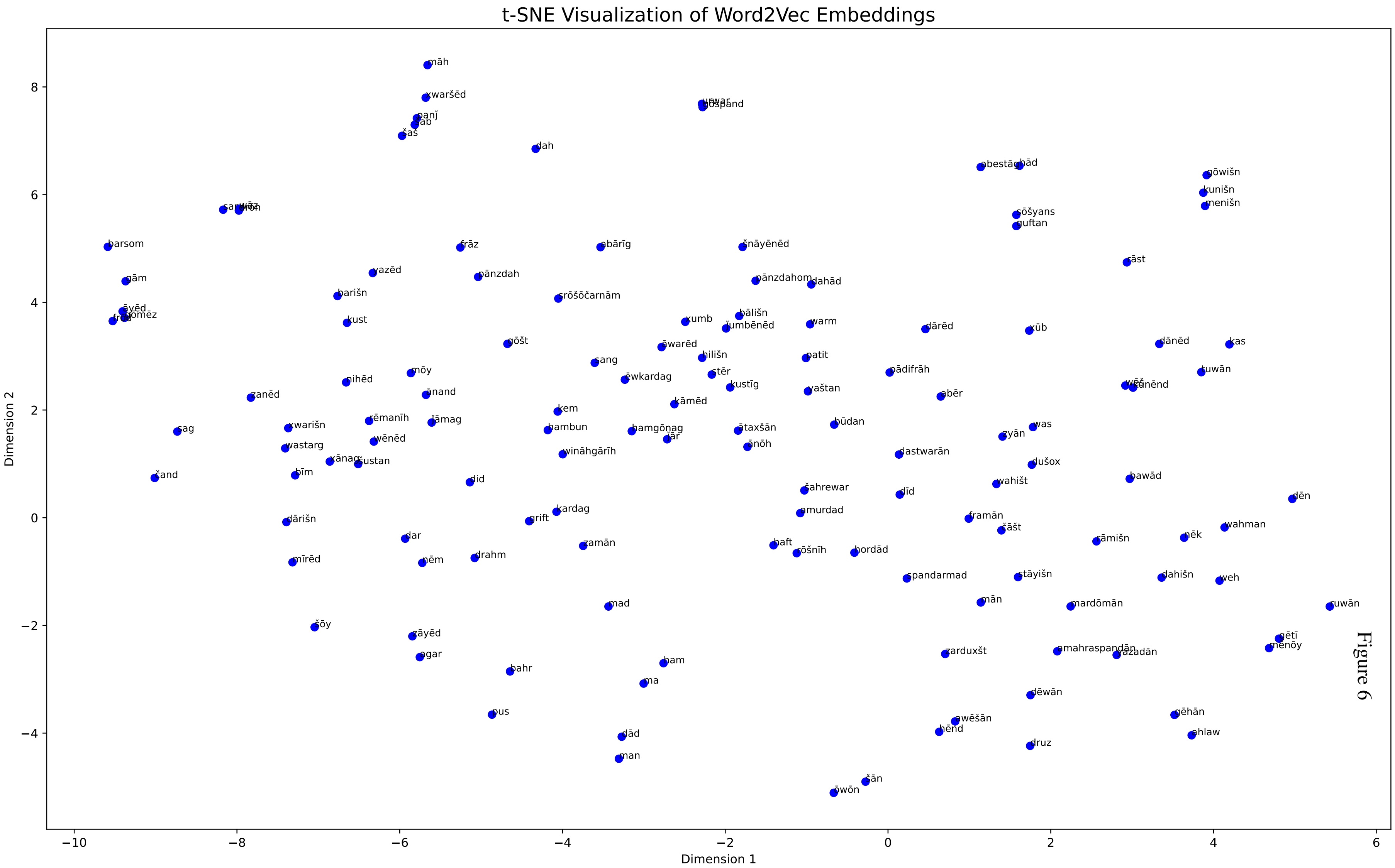
t-SNE Visualization of Word2Vec Embeddings
Dimension 1
Dimension 2
−10
−8
−6
−4
−2
0
2
4
6
8
6
4
2
0
−2
−4
māh
xwaršēd
urwar
gōspand
panǰ
šab
šaš
dah
abestāg
hād
gōwišn
kunišn
menišn
sar
wāz
drōn
sōšyans
guftan
barsom
frāz
abārīg
šnāyēnēd
rāst
yazēd
pānzdah
gām
barišn
srōšōčarnām
pānzdahom
dahād
āyēd
gōmēz
frōd
kust
xumb
bāliš̌n
ǰumbēnēd
warm
dārēd
xūb
dānēd
kas
gōšt
āwarēd
hilišn
patit
sang
mōy
stēr
pādifrāh
tuwān
nihēd
ēwkardag
kustīg
yaštan
wēš
kunēnd
zanēd
ānand
kāmēd
abēr
kem
rēmanīh
ǰāmag
būdan
was
xwarišn
hambun
hamgōnag
ātaxšān
zyān
wēnēd
ǰār
wastarg
ānōh
dastwarān
xānag
šustan
winābgārīh
dušox
bīm
did
šahrewar
wahišt
bawād
čand
dēn
dīd
kardag
amurdad
framān
dārišn
grift
wahman
čāšt
dar
rāmišn
nēk
zamān
haft
rōšnīh
hordād
mīrēd
nēm
drahm
spandarmad
stāyišn
dahišn
weh
mad
mān
mardōmān
ruwān
šōy
gētī
zāyēd
mēnōy
agar
zarduxšt
amahraspandān
yazadān
bahr
ham
ma
dēwān
pus
gēhān
awēšān
dād
hēnd
ahlaw
druz
man
šān
ōwōn

Figure 6

B Network Structures for Chapters 1, 3, and 20 of *Šāyist nē šāyist*

In this appendix, Chapters 1, 3, and 20 of the *Šāyist nē šāyist* book are presented, with similarity values of T_M set to 0.5, 0.65, and 0.75 for edge creation in the complex network. Both Word2Vec and FastText network structures are included.

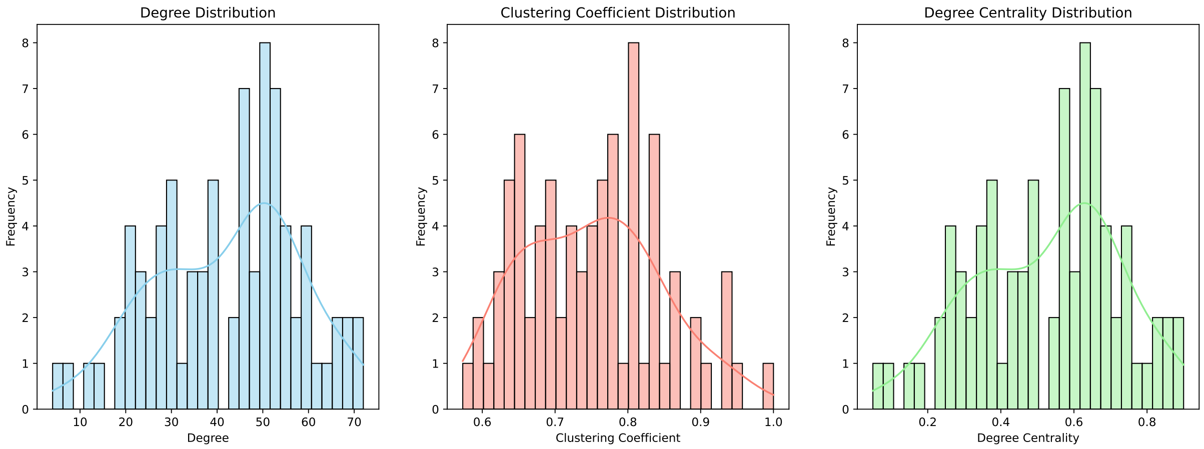

Figure 7: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.5$, Word2Vec.

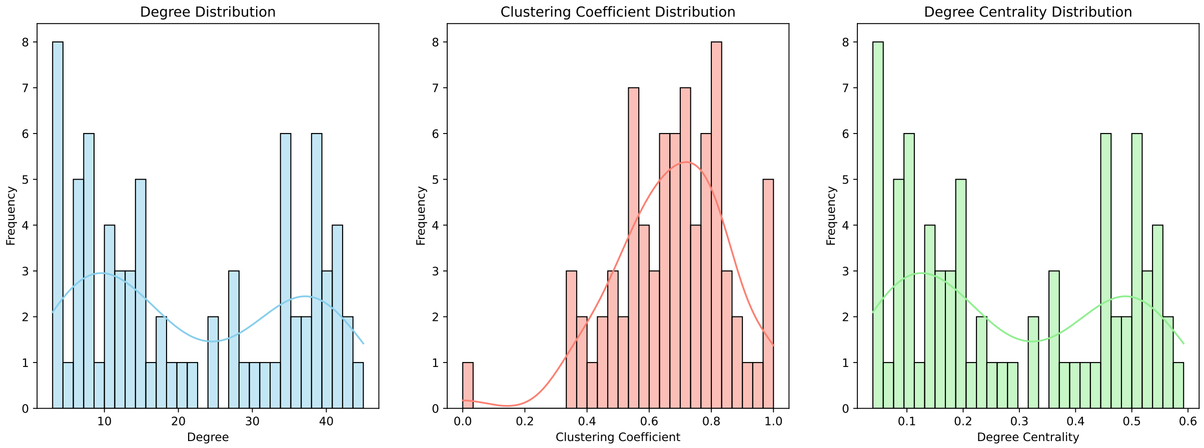

Figure 8: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.65$, Word2Vec.

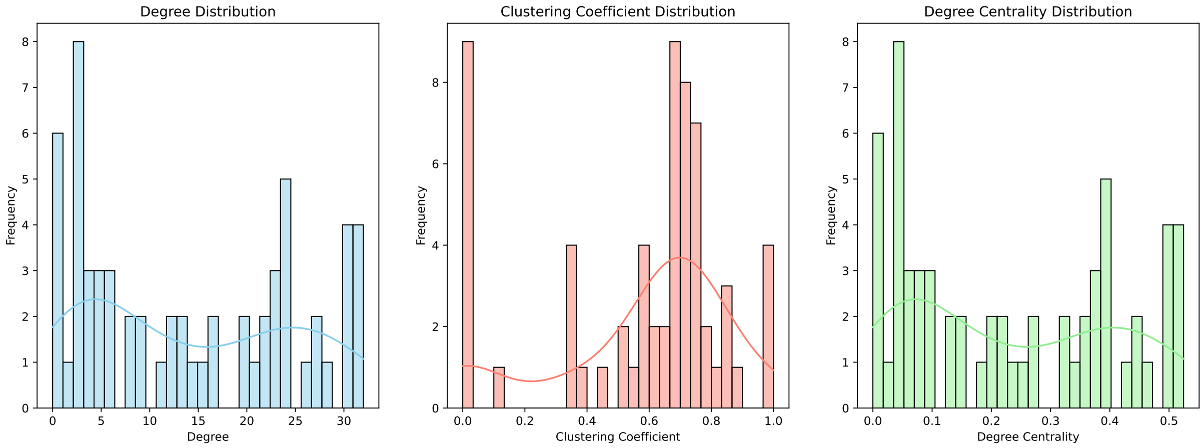

Figure 9: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.75$, Word2Vec.

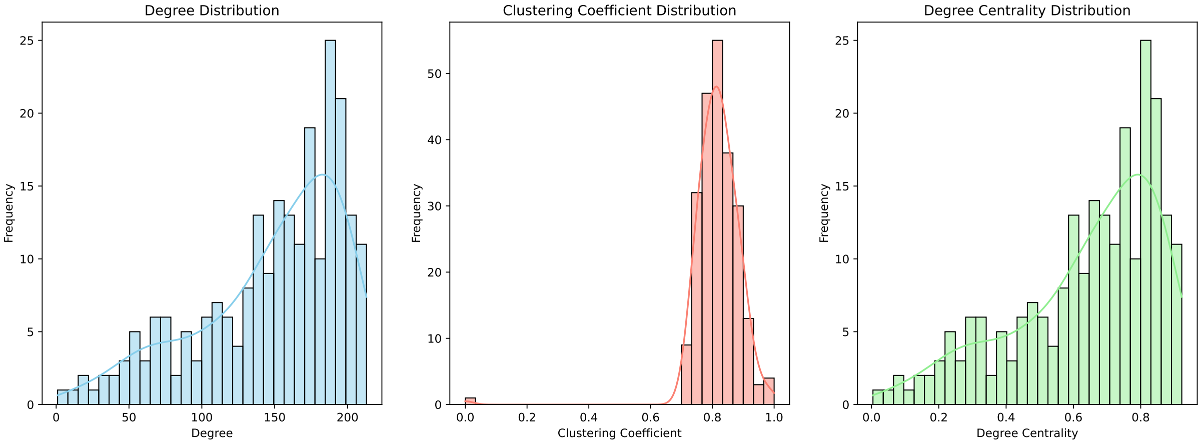

Figure 10: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.5$, Word2Vec.

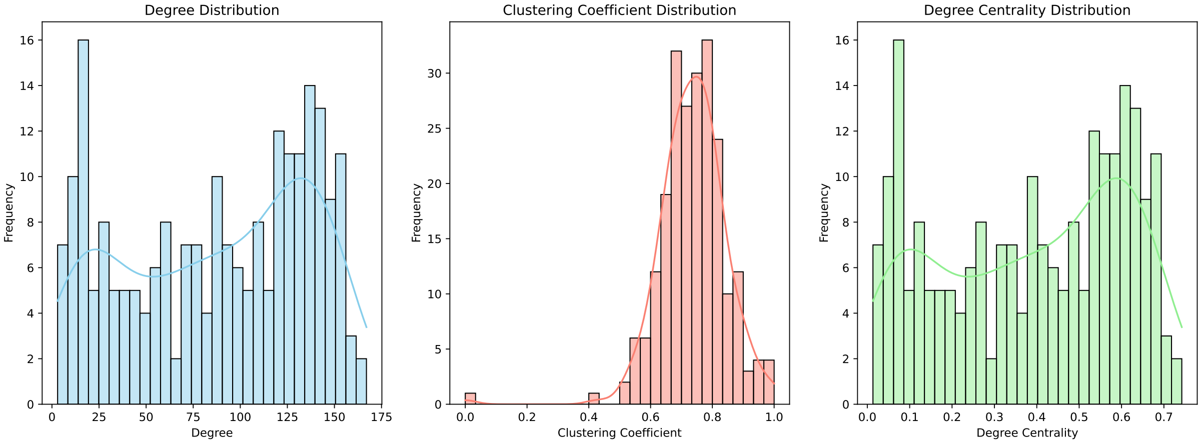

Figure 11: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.65$, Word2Vec.

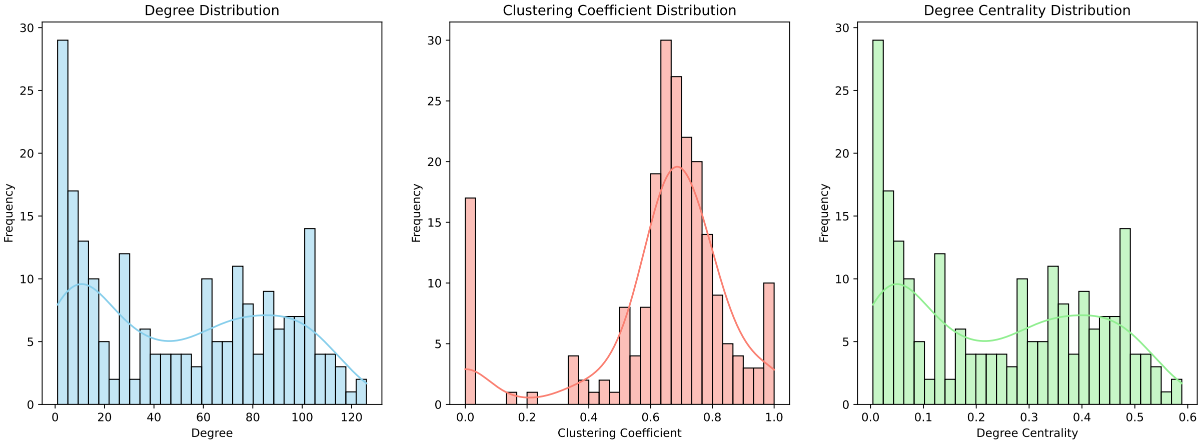

Figure 12: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.75$, Word2Vec.

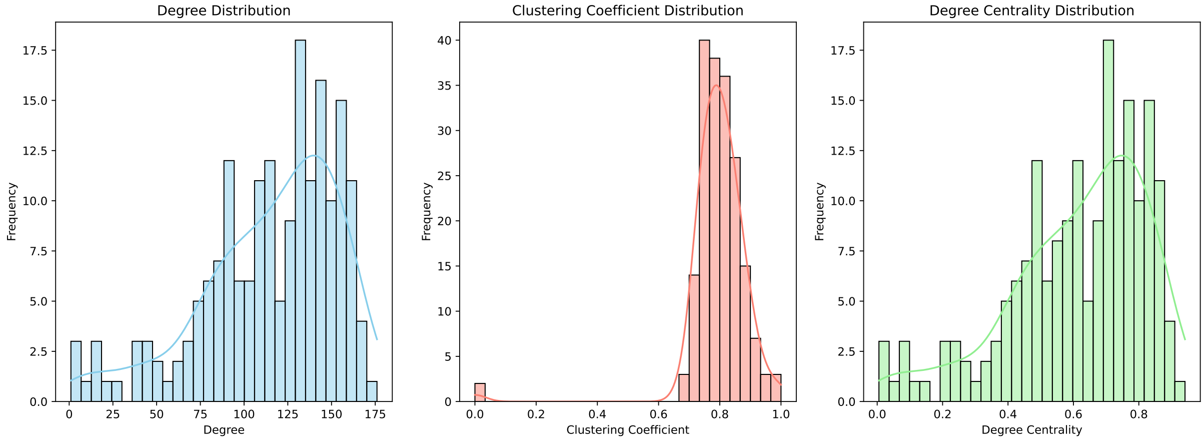

Figure 13: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.5$, Word2Vec.

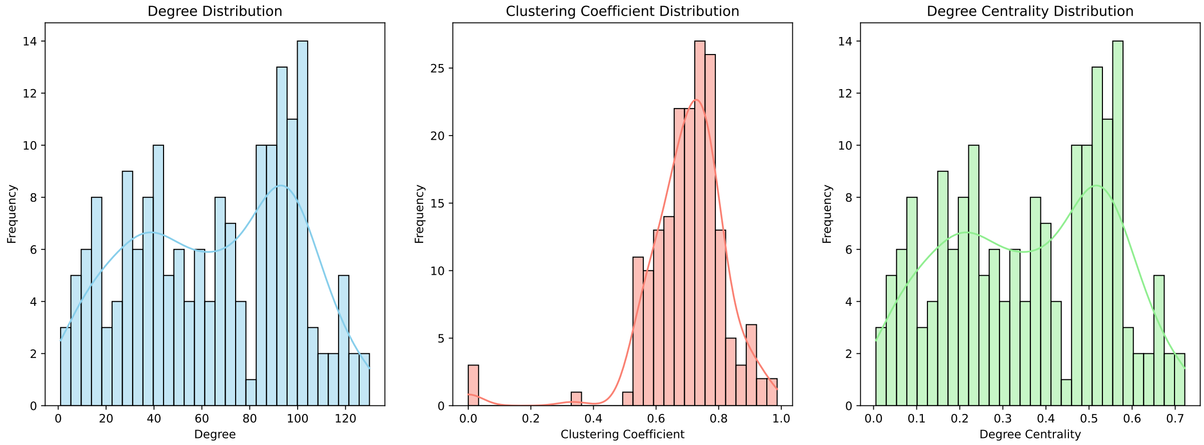

Figure 14: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.65$, Word2Vec.

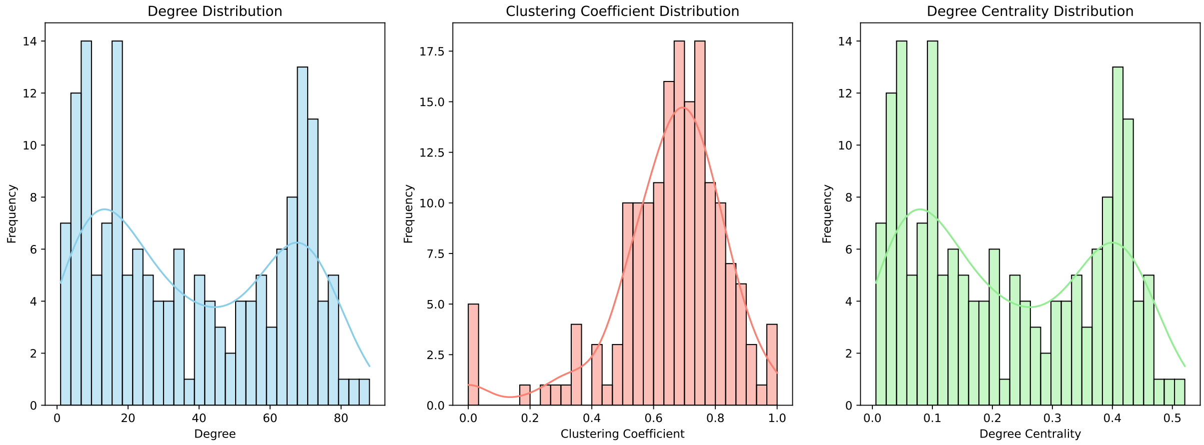

Figure 15: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.75$, Word2Vec.

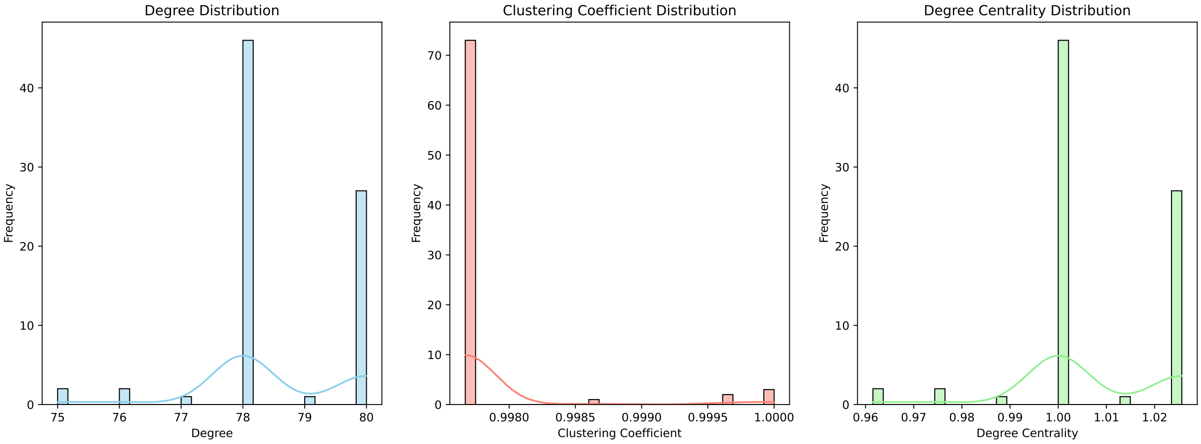

Figure 16: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.5$, FastText.

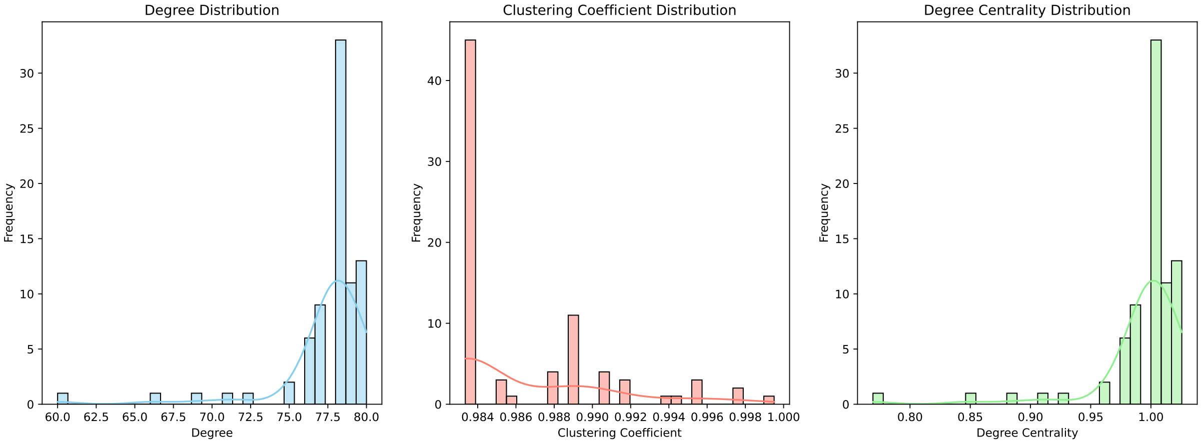

Figure 17: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.65$, FastText.

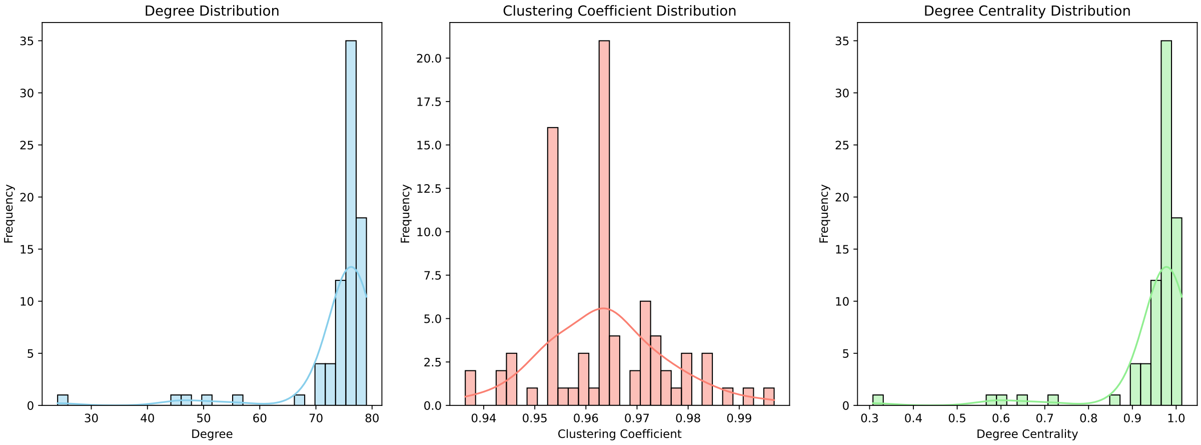

Figure 18: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 1), with $T_M = 0.75$, FastText.

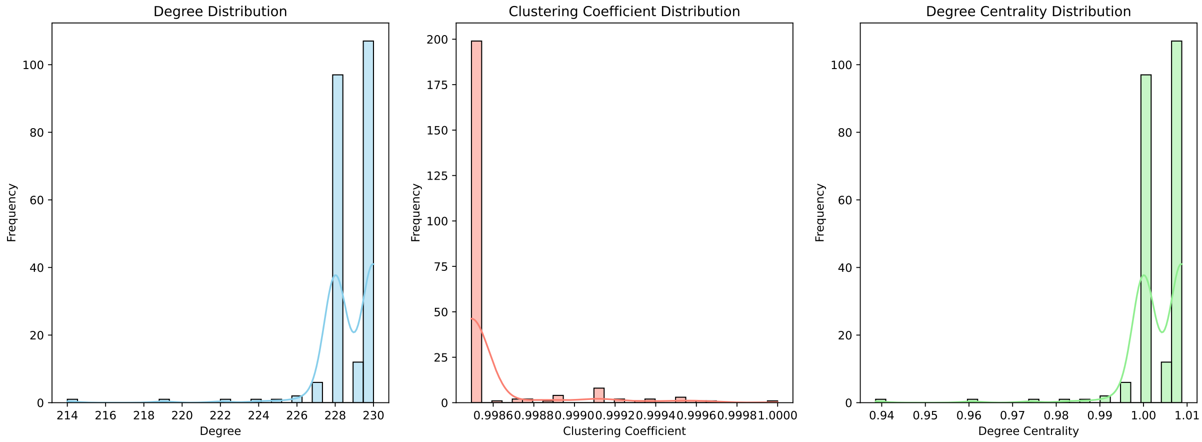

Figure 19: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.5$, FastText.

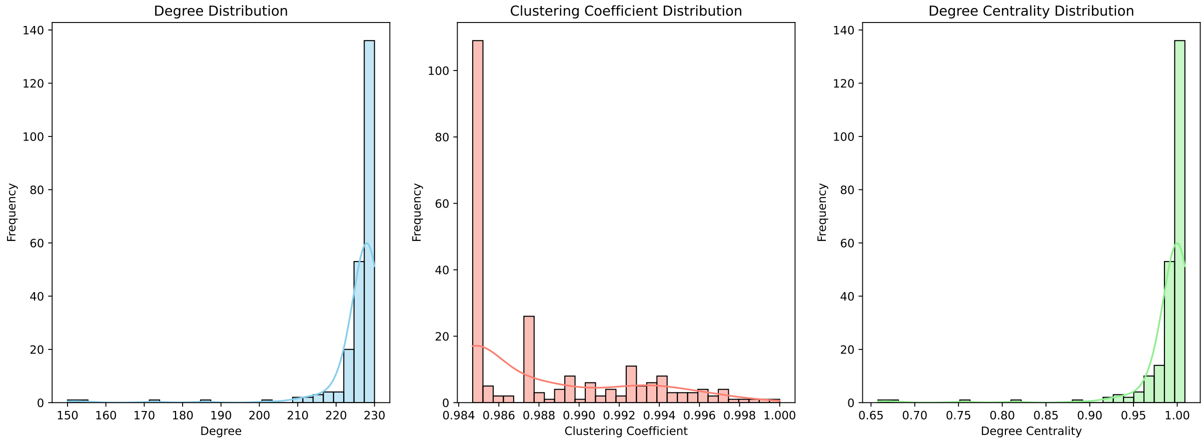

Figure 20: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.65$, FastText.

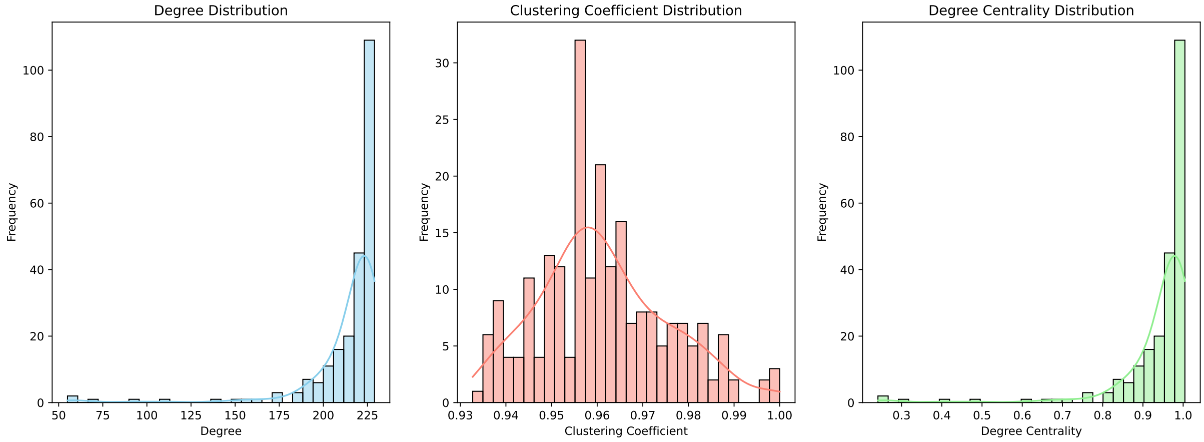

Figure 21: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 3), with $T_M = 0.75$, FastText.

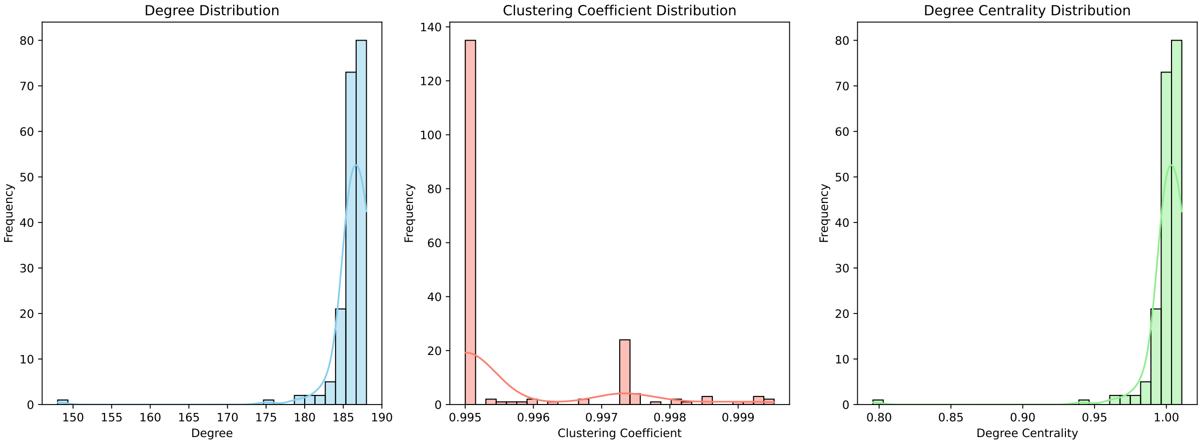

Figure 22: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.5$, FastText.

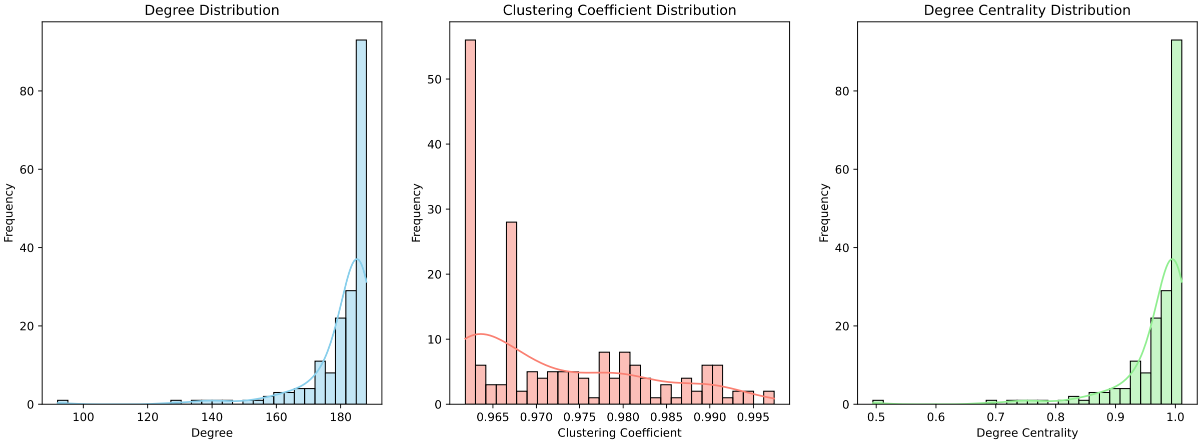

Figure 23: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.65$, FastText.

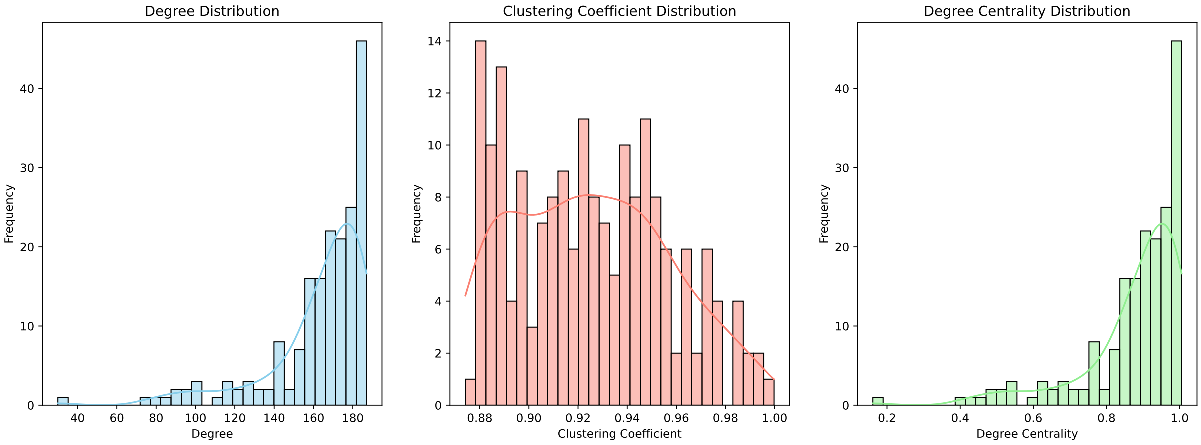

Figure 24: Degree distribution of the *Šāyist nē šāyist* book chapters (chapter 20), with $T_M = 0.75$, FastText.