Stylometry of Short Stories through Voyant Corpus Summary Tool: A Text Mining Study

Zafar Ullah[1]
Arshad Mahmood[2]

Abstract

Text mining tools quantify big data precisely to express stylistic features of any text within a few seconds. This study aims to extract quantified stylistic qualities computationally in five American short stories (1. Button, Button, 2. Clearing in the Sky, 3. Dark They Were and Golden Eyed, 4. Thank You M'am, 5. The Piece of String). Students of literature encounter complications in grasping precise stylistic qualities of a writer in a short span of time or even before reading the text since exact comprehension of stylistic qualities facilitates in understanding the target text, its form, literary function and readers' critiquing ability. Current study mines the selected five short stories to derive certain stylistic characteristics quantitatively and qualitatively. Summary panel, a component of Voyant text mining tools, is applied to unveil quantified and qualitative stylistic features of literary texts. One advantage of implementation of Voyant Summary tool for data generation is that its results are statistical, qualitative, accurate and efficient. Mixed method is applied to produce authentic and valid results of this research. On philosophical grounds, Knowledge Discovery Theory by Rakesh Aggrawal is applied to extract potentially worthwhile new knowledge patterns. Major findings of this study are that all short stories employ dialogic technique and the common theme "said" is found in all short stories. Three short stories among five demonstrate their vocabulary density from 0.307 to 0.362 which exhibit ratio of unique words and total words in a short story. Average length of sentences ranges from 9.9 to 12.9 words which express small and simple construction of sentences by the writer.

Keywords: *Stylometry, Voyant, Corpus Summary, Text Mining*

1. Introduction

As a finger print is unique, similarly each literary writer stands out for a peculiar literary style in his/her work. Technically, stylistics is a branch of applied

[1] Lecturer, Department of English Studies, NUML, Islamabad
[2] Head of English Department, NUML, Islamabad

linguistics and it studies linguistics and tonal style of any writer through text mining of his/her work. Stylometry or computational stylistics refers to the extraction of automated values and characteristics from any writer's works. Text mining minutely examines the text with digital tools and finds total words, unique words, vocabulary density, average length of sentences and recurrent themes with their occurrences. As text mining is done with digital tools, so current study applies Summary tool on five American short stories to extract knowledge patterns regarding qualitative and quantitative aspects. In this study, stylistic features are specified as total words, unique words, vocabulary density, average length of sentences and eight most occurring words along with their statistical weight. These most occurring words also refer to discussion of key concepts of their respective short story. Thus, stylometry and concept mining merge to express stylistic qualities of the selected five short stories.

Statistics and stylistics are tied in a strong bond. This study falls in the domain of text mining which mines digital text with Summary tool to explore statistical data, style related knowledge patterns and structural insight. Primarily, text mining or text analytics explores concepts, themes and key words with digital tools. Here this study precisely concentrates on stylistic aspects of five short stories. The focus is on quantified and qualitative data which open new vistas of literary style and quantified criticism.

This study intends to analyse stylistic qualities of the selected five short stories to present quantified stylistic qualities. The automated, quantified and finger print like stylistic qualities are generated with Summary tool which is designed by Sinclair and Rockwell in Canada in 2003. Recently, these tools have been upgraded to cope up with new digital hermeneutic and stylistic challenges (Sinclair & Rockwell, 2015). Validity and reliability of Voyant and Summary tool can be determined with reference to Graham, Milligan and Weingart (2013)'s quote that "Voyant is ideal for smaller corpuses of information or classroom purposes."

This study raises investigations about specific stylistic features of short stories: How does Summary tool features represent stylometry of five selected American short stories? How do the stylistic qualities of five American short stories match or differ with one another?

Significance of stylistics reflects through its advantages, needs and applications. To know stylistic features of any writer and his/her work is necessary to

understand and fully appreciate any piece of writing. Stylistics unveils linguistic features of the text in dominatingly quantified form. The study of stylistics has functional value because it facilitates the process of elucidation. This stylometric study is significant because text can't be adequately comprehended and evaluated without accomplishing stylistic investigation. Novelty of this study rests on the utilisation of Summary tool to explore literary stylistic values of short stories. Furthermore, it is also in high demand because use of computational tools for stylistic analysis is the least explored field and stylometric analysis of the selected short stories have not been done earlier. Current study extracts precise stylistic features qualitatively and quantitatively even without reading the text thoroughly. Its derived stylistic characteristics for example total words, unique words, vocabulary density, average length of sentences and statistical weight are accurate and authentic.

2. Literature Review

Different researchers employed various methods to determine stylistic qualities of authors with the help of text mining tools or their early forms. Initially, content analysis was considered as a benchmark to explore style of any writer (Krippendorff, 2003). Its major flaw was that it varied from person to person as content analysis changed by different analysts. In addition to it, quantification of style was missing and this niche was filled by present study. The use of computer was popularized in academia, hence, computational stylistic techniques were applied to perform swift analytical tasks (Stamatatos et al., 1999). Simultaneously, computers, algorithms and different models were designed and trained to explore literary style (Argamon et al., 2003; Zhang et al., 2002). Later, SVM (Support Vector Machine) approach for supervised learning was applied (Stamatatos, 2009) and N-gram approach was also used in stylometry. These studies were in harmony with the current research.

In the later stage, histograms for word-length were utilized to identify an author (Malyutov, 2006). Only one yardstick was used to measure a few stylistic qualities and other core parameters for example average length of sentence, total words, unique words, vocabulary density and repeated words were completely ignored. To fill this much needed gap, current study was initiated.

Moreover, authorship of text is a significant matter as it was the controversy of Shakespeare-Marlowe. This conflict of authorship can be resolved through Summary tool in stylometry. Inquiries regarding author identification employed computational stylistics or stylometry to great extent (Stamatatos et al., 2000; Van Halteren et al., 2005; Stamatatos, 2009).

Computational stylistics took interest in stylochronometry (Stamou, 2008) like carbon 14 test to judge age of any text. Voyant corpus summary just informs date of input in corpus summary and data visualization. Another way of searching stylochronometry is to find key words and to find their etymology which reveals the age of vocabulary for example the use of the word "hath" by Francis Bacon refers to Renaissance age and orthographic patterns of "Aprill" (April), "soote" (sweet) leads to 14th century English used by Geoffrey Chaucer.

Authorship of world famous novel *'Go Set a Watchman'* was claimed by fake writers, so its writing style was analysed computationally with frequencies and cluster analysis. Its cluster tree matched with corpus of 28 novels of Harper Lee. So, it was decided that Harper Lee was its bona fide author who had already written another novel *'To Kill A Mockingbird'* (Gamerman, 2015) because stylistic features of both works were identical. Ongoing study also discussed vocabulary density, corpus qualities and some stylistic elements which were peculiar features of any writer's literary style.

Tweedie, Singh and Holmes (1996) used Artificial Neural Networks (ANN) for computational study of literary style. Another similar ANN based stylometric study was conducted on three datasets of three literary figures Jonson (164 examples), Middleton (90 examples) and Shakespeare (168 examples) with Cascade-Correlation network architecture. It produced results that Shakespeare was the true author of Thomas Kyd's *'The Spanish Tragedy'* and Madison was the genuine author of the twelve disputed federalist papers (Waugh, Adams, & Tweedie, n.d.). To conclude, Artificial Neural Networks (ANN) were also used for author identification in computational stylistics domain. Contrary to it, the current study found such results as stylistic corpus summary, total words, unique words, vocabulary density, average length of sentences and key themes.

Support Vector Machine (SVM) was also applied to study stylistic qualities. Rabindranath Tagore (Bengali Nobel Laureate), Sarat Chandra and other writers' 150 short stories each were analysed in a stylometric study. The study employed SVM, decision tree and ANN to recognize segregated texts of aforementioned three writers (Chakraborty, 2012). Among them, SVM was proved to be the most accurate result producing system.

Web based tourism English, another subdivision of ESP (English for Specific Purpose), was compiled with the name of Tourism English Corpus (TEC), afterwards it was compared with Freiburg-LOB Corpus of British English

(FLOB). Both corpora were same in their lexical density. Their comparison shows that content words of TEC were more than FLOB. Average length of sentences in TEC was shorter than those in FLOB. TEC had more nouns and adjectives than FLOB. TEC utilized more verbs than FLOB (Kang & Yu, 2011, p. 129). As its deficiencies were concerned, TEC was a smaller corpus as compared to FLOB. It was having similarity with current study on the basis of average length of sentences and lexical density. Stylometric study is equally applicable on any genre or document.

A free Dutch corpus named CLiPS comprised 305,000 tokens and 1126 documents written by Dutch University students. It revealed personality, sentiments, author, age, gender and genre. SVM algorithm was applied in the research and its text was classified automatically (Verhoeven & Daelemans, 2014). This Dutch study overlooked vocabulary density, length of sentences, the most occurring words and these stylistic features were incorporated in Summary tool.

Statistics have been vastly used in the latest researches about language. Genre identification and authorial attributes were studied with amalgamation of traditional and topological characteristics. Applying Brown Corpus, stop words were removed from the text to attain lemmas (Amancio, 2015). This study exhibited the excessive use of complicated diagrams, advance statistical formulae and integration of Brown corpus into this data and these features were beyond comprehension of common social science students, however, current study tool is user friendly in terms of its comprehension and evaluation. Contrary to aforementioned complicated stylistic analysis, Summary tool is very easy and quick especially for tech retards. Furthermore, Summary tool also has a feature of stop words to refine textual results.

Text mining tools were devised for stylistic analysis. Stylometry with R extracted quantifiable features of French and English literature without any programming or integration of any extra tool since R package is useful for digital humanities learners. Moreover, it produced corpus, n grams and fascinating visuals for analyzing style of any text (Eder, Rybicki, & Kestemont, 2016). Most of the features of R package resemble Summary tool except most frequent words of the corpus. Summary tool fulfilled this deficiency and these most frequent words expressed hedges and repeated ideas of any writer.

Recently, in language technology, Stylo and Cluto tools have been used for automatic extraction of stylistic qualities from Polish language. They found out average word length, average sentence length, punctuation marks, word frequencies and style markers to discriminate stylistic qualities without any programming skill (Eder, Piasecki, & Walkowiak, 2017). Comparing Stylo and Cluto tools with Voyant tools, Stylo and Cluto have extra features of finding average length of words, punctuation marks, parts of speech tagging of selected words and style markers.

The Chinese researchers made corpus of Mo Yan and Zhang Wei writers to differentiate their quantified stylistic features. Statistical study revealed that their works differed in sentence patterns, length of words and delineation of social set up. This study found that Zhang Wei's linguistic style didn't change frequently while Mo Yan's stylistic features didn't vary (Li, Ji, & Xu, 2017, p. 1158).

In Sweden, Hemingway's work was revisited to evaluate common assumption and to explore stylometric features in treatment of women. Noun, adjectives and token numbers per thousand were counted to represent stylistic features (Sundberg, & Nilsson, 2018).
Recently, stylometric similarities have been explored in James Joyce's and O'Brien's works which concluded that the latter was the follower of the former. This study was conducted with stylistic clusters (most frequent words), delta analysis and average sentence length with visuals (O'Sullivan, Bazarnik, Eder, & Rybicki, 2018). This study showed visual and numeric data to prove similar stylometric features. Stylometric qualities led to quantified criticism and every aspect of literature was discussed precisely and correctly. It is not far from assuming that in future, generalised statements of literary critics going to be discarded and exact knowledge based perceptions and comments would be weighted in arena of academia. Current study fills this niche and manifests stylistic aspects of short stories and those features are useful in literary criticism and unveiling of linguistic topographies.

3. Methods

Philosophy inspired the entire research design and this research was based on positivism which affirmed machine oriented experimental research (The Writepass Journal, 2012, June 5). Evidently, theoretical grounding built foundations of every study. Current research was conducted in the light of Knowledge Discovery Theory which was propounded by world renowned Indian computer scientist, Rakesh Aggrawal. Knowledeg Discovery Theory emphasised

on "the extraction of implicit, previously unknown and potentially useful information from data" (Cabena, Hadjinian, Stadler, Verhees, Zanasi, 1998, p. 9). This theory promoted discovery of new dimensions of interesting knowledge patterns. Knowledge Discovery Theory incorporates data selection, data preparation, data cleansing and extraction of innovative useful knowledge patterns in the form of data visualization, statistics and word form.

This study applied mixed method approach since this research relied on statistical and qualitative data which were extracted on the basis of statistics. Generated results exhibited quantitative data in terms of total words, unique words, vocabulary density, average words per sentence and statistical weight of most occurring words. Qualitative data was derived in the form of most frequent words and their extraction was based on their occurrence in the text.

In the research strategy, text of each short story was uploaded on Voyant tools and panel of Summary was chosen to display and interpret stylistic data. Distant reading technique was employed to mine five selected American short stories from the syllabus of intermediate level English compulsory in Punjab, Pakistan. The generated data for each short story was exported as PNG image and it was inserted into discussion section of the paper. Each figure consisted of total words, unique words, vocabulary density, average length of sentences and most occurring vocabulary with their statistical value. Afterwards, numeric data were analysed to unveil stylistic patterns and features.

Systematic sample of first five short stories was selected from 1st year Intermediate English textbook because more than 1 million intermediate students study these short stories in Pakistan. Moreover, they are modern short stories, so they are liked by most of the modern youngsters. Selected texts are openly accessible and Voyant is an open access web source tool. So no ethical issue regarding human participation or permission from tool designer was involved in this study.

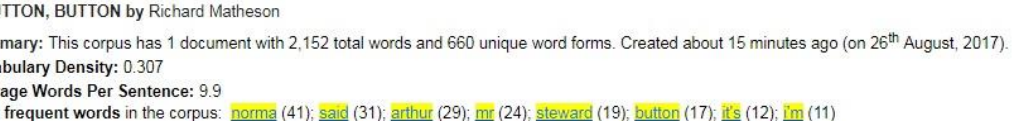
Data were generated both qualitatively and quantitatively to produce data for mixed method data analysis approach. Data were collected by composing first five short stories from Book I of intermediate English syllabus of Punjab Boards and Federal Board, Pakistan. As collected data were consisting of mixture of qualitative and quantitative data, so data were analysed through mixed method i.e qualitative and quantitative. Mixed Method Approach applied "combination of qualitative and quantitative approaches" (Durant, 2004, p.10). Therefore, it

incorporated domains of statistics, mathematics, corpus and literature. Quantitative aspects of stylistic qualities (Sebastiani, 2002) were necessary for getting exact quantitative answer. This need was fulfilled by total words, unique words, vocabulary density and average length of all sentences. On the other hand, qualitative data were collected in the form of most occurring words along with their statistical occurrence data. Its rationale was that Summary tool generated quantitative and qualitative data, so mixed method was the most appropriate method for mixed type of data since data and its analysis correlated with each other. Mixed method was justified that both methods crosschecked and validated each other. In addition to it, this mixed method eliminated researcher's personal inclinations about style of any writer.

Manually, it was a gigantic task to explore quantified stylistic features of each literary piece of writing. Furthermore, it was a very time consuming activity to count each word, then separating unique words from total words was also another gigantic task. Afterwards, counting average number of words in each sentence demands a lot of time. In addition, assigning statistical weight to each word required a long span of time, despite all efforts, human errors were still possible. To assist humans in the quantified and systematic discovery of knowledge about stylistic characteristics of any writer's work, Summary tool mined big data text of any writer to extract computational stylistic features or stylometry within a few seconds. Students of intermediate read short stories and faced problems of analysing stylistic qualities of short stories and the extraction of key themes in the form of most occurring words. Moreover, learners wanted to know about total words, unique words, vocabulary density, average length of sentences and the most frequent words in the corpus.

4. Results and Discussion

In this section, result of Summary tool is pasted in the form of a figure and its stylistic characteristics especially novelty of knowledge patterns are revealed. In reality, statistical analysis of each short story and the extraction of the most frequent words are new knowledge patterns. In the following lines, stylistic characteristics of each short story are interpreted.


1.BUTTON, BUTTON by Richard Matheson

Summary: This corpus has 1 document with 2,152 total words and 660 unique word forms. Created about 15 minutes ago (on 26th August, 2017).
Vocabulary Density: 0.307
Average Words Per Sentence: 9.9
Most **frequent words** in the corpus: norma (41); said (31); arthur (29); mr (24); steward (19); button (17); it's (12); i'm (11)

Figure 1 Corpus Summary of Button, Button

Stylometry precisely quantifies stylistic features of any literary work. Basically, Richard Matheson, the short story writer, writes 660 unique words and he repeats these unique words more than three times until they become 2152 total words. This repetition creates ease for beginners and intermediate level learners that is why this short story is included in intermediate English textbook. Its vocabulary density is 0.307 which is quite appropriate for intermediate level. Vocabulary density is generated by division of unique words by total words and its result is presented in round figure. Average words per sentence also unveil the use of long sentences (compound and complex sentences) or short sentences and individualized stylistic qualities. Summary tool shows the inclusion of average 9.9 words in a sentence and such small sentences enhance readability of the short story. It also unveils stylistic feature of the use of simple and small sentence in the short story to make it comprehensive for all age group readers especially for young readers. Moreover, this short story is a science fiction which is meant for children and adults alike. Therefore, repetitive vocabulary and small sentences make Richard Matheson's style reader friendly for all age group readers.

Most frequent words reveal knowledge patterns of major characters of the short story for example "Norma (41)", "Arthur (29)" and "Steward (19)" and all of them are the most occurring characters. Apart from human characters, most important non-living character is "button (17)" which refers to the title of the short story as well as it plays a decisive leading role in the short story. Steward entices Norma to push the button and Norma eagerly yearns to push the button unit to win an unknown murder reward of 50,000 dollars. On the other hand, her husband, Arthur, obstructs her way and rejects the idea of getting the amount at the cost of blood of an unidentified living soul in any corner of the world. Norma is willing to push the button and after long contemplation, eventually, she pushes the button for collective monetary gains and after some time, she receives the tragic news of her husband's accidental death and she remembers his insurance sum of 50,000 dollars.

Another word "said (31)" shows the use of dialogic style in this short story. Dialogues of three main characters, Norma, Arthur, Steward, develop the short story to its logical ultimate conclusion. Apart from it, phrase "I'm (11)" suggests the use of first person narrative technique in this short story. First person technique is a more reliable source of narration than third person narrative technique. To conclude, dialogues and first person narrative techniques are key stylistic qualities of Richard Matheson.

Figure 2 Corpus Summary Clearing in the Sky

'Clearing in the Sky' comprises 599 unique words and they are reused almost four times in this corpus, so total words are 2228. So its vocabulary density is 0.269 and it denotes that new words appear less in this corpus. As compared to the first short story, its vocabulary density is less but the average words in a sentence are more in number. Jesse Stuart's facile, repetitive and naturally easy vocabulary matches with the topic of nature.

The most frequent words highlight several stylistic qualities and these words are selected on the basis of their statistical weight. Firstly, two main unnamed characters have been mentioned namely "I (88)", "he (53)". They denote that first person and third person narrative techniques have been employed in this short story. Comparing both techniques, first person narrative technique is more reliable as compared to third person narrative technique which is the least reliable narration technique. One knowledge discovery is that characters are simple and unnamed because of father and son relationship. Furthermore, in American rural areas, calling the name of father is considered as an indecent act which probably motivates the short story writer to avoid the use of proper names. Another reason is that author Jesse Stuart serves as a teacher and he teaches morality to his readers. Secondly, the word "father (13)" has been used to show filial love and obedience level on the part of a son. Thirdly, the theme of "said (17)" reveals the frequent use of dialogues to develop plot of the short story and to interlink characters. Overall, Jesse Stuart uses dialogues and the mixed narrative technique of both first person and third person narrative techniques to delineate this nature based short story.

Key themes have been mentioned for instance "land (15)", "mountain (14)" and "path (12)". This short story incorporates multifaceted descriptive natural beauty and aesthetic sense of fineness of frequent references to mountain, steep paths on the mountain and fertile land on the mountain top where the old man grows alfalfa, potato, tomato and yam with unique taste.

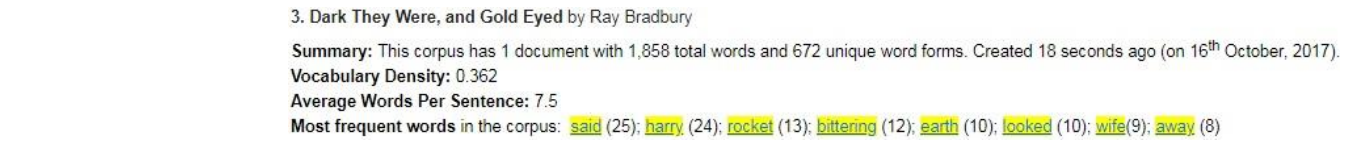


Figure 3 Corpus Summary Dark They were and Golden Eyed

Computational stylistics unveils stylistic features of the literary work. Analysing text mining features of this short story, total words of this short story are 1858 while unique words are 672 and it discovers that unique words have been repeated almost three times in this corpus. It shows that this is how the text becomes reader friendly and suitable for basic and intermediate level learners. Its vocabulary density is 0.362 and average words per sentence are 7.5 which indicate the use of small sentences. Therefore, even a basic level reader can easily comprehend Ray Bradbury's short stories.

Statistics are at work in humanities to pinpoint key themes from large dataset of texts. Like other stories, the most frequent word of this short story is "said (25)" and it guides about the excessive use of dialogues for the progression of ideas and events in this science fiction story. Other themes of "rocket (13)", "earth (10)", "away (8)" elaborate the main idea that rockets are propelling to Mars from the Earth because there is war on the earth, hence, travelling to Mars is the only refuge for humans. Mythologically, Mars is a god of war and humans should get refuge at Mars during nuclear world wars.

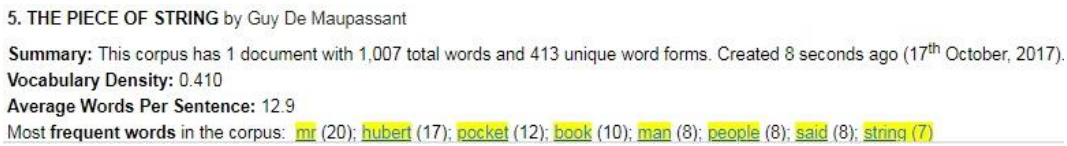
Most frequent words also expose major characters: "Harry (24)" "Bittering (12)" and his "wife (9)". Protagonist of the story is Harry who wishes to run away from earth to Mars to save lives of his family and later, he wants to come back to the earth. To accomplish this aim, he starts building rocket, though he remains unsuccessful in this highly technical endeavour. Again, after long stay at Mars, he completely adapts Martian environment; he behaves like true Martians; and he detests earth people considering them ridiculous people.

Figure 4 Corpus Summary Thank You M'am

Summary tool quantifies stylistic qualities of any piece of writing. In Langston Hughes' short story, 426 unique words have been used almost three times, that

total words become 1361. By dividing unique words with total words, its vocabulary density appears 0.313 which is understandable for both basic and intermediate level readers. On average, each sentence comprises 12.5 words and it is almost equal to average words per sentence in the 2nd and the 5th short story.

The selected five short stories are full of dialogues, so conversational style is dominating in the selected short stories, hence, the most occurring theme is word "said (28)". Statistical knowledge discovery and the most occurring words express the presence of main characters for example "boy (27)" and the heavy "woman (23)" named Mrs. Luella Bates Washington Jones. The boy tries to snatch the purse of the woman at night but she kicks him well, holds him by his shirt front, rattles him and drags him to her house. Then she asks him to wash his "face (9)" because spick and span appearance betokens good attitude and character. Later, she gives him homely food, ten dollars to buy blue suede shoes because its buying desire compelled him to snatch the purse of Mrs. Luella. She also advises him to refrain from evil acts to materialize worldly desires because "shoes got by devilish ways will burn your feet".


5. THE PIECE OF STRING by Guy De Maupassant

Summary: This corpus has 1 document with 1,007 total words and 413 unique word forms. Created 8 seconds ago (17th October, 2017).
Vocabulary Density: 0.410
Average Words Per Sentence: 12.9
Most **frequent words** in the corpus: mr (20); hubert (17); pocket (12); book (10); man (8); people (8); said (8); string (7)

Figure 5 Corpus Summary The Piece of String

Guy De Maupassant was a French short story writer and due to his fame, his stories were translated into several languages. The analysed short story is an English translation of his short story. He uses 413 unique words in this short story and these words are repeated a bit more than once until its total words become 1007. Owing to less repetition, its vocabulary density is 0.410 which betokens more difficult lexis than other short stories for intermediate level readers. This short story and two others have almost similar 12.9 words average length of sentences.

Statistics govern text mining process and extracts the most occurring words which are central themes of the short story. Like other 4 short stories, dialogues are also present in this short story as the word "said (8)" denotes. Statistics reveal major characters that "Hubert (17)" is blamed to steal the "pocket (12)" "book (10)", he swears and explains the situation that he has picked a "string (7)" from the mud but "people (8)" don't accept his point of view. Later on, though the pocket book is found on the road, even then people keep on calling him an old "liar" and

ridicule him frequently. Hubert expresses his own philosophy thus, "There is nothing so shameful as to be called a liar" and this feeling becomes the cause of his lethal depression and imminent death.

Quantitative analysis of these five short stories have been presented in a tabular form to exhibit a holistic overview.

Table 1 Quantitative Corpus Summary of Five Short Stories

	Short Stories. Book I	Total Words	Unique Words	Vocab Density	Length of Sent
Short Story. 1	**Button, Button**	2152	660	0.307	9.9
Short Story. 2	**Clearing in The Sky**	2228	599	0.269	12.8
Short Story. 3	**Dark They Were, And Golden Eyed**	1858	672	0.362	7.5
Short Story. 4	**Thank You, M'am**	1361	426	0.313	12.5
Short Story. 5	**The Piece of a String**	1007	413	0.410	12.9

The table 1 reveals that vocabulary density ranges from 0.269 to 0.410 which is suitable for intermediate level readers. Except 5th short story, other four short stories have one common feature of three times repetition of unique words and this characteristic unveils the use of easy and repetitive vocabulary items. This reappearance of words enhances the readability of basic level learners through interconnectivity of ideas, motifs and characters. Average length of sentences ranges from 9.9 words to 12.9 words per sentence. This length of sentences is quite appropriate for basic level learners.

Qualitative data of this study has been exhibited in the table 2:

Table 2 Qualitative Data of Five Short Stories

	Short Stories. Book I	Qualitative Data
Short Story. 1	**Button, Button**	Norma (41), said (31), Arthur (29), Mr (24), Steward (19), button (17), it's (12), I'm (11)
Short Story. 2	**Clearing in The Sky**	I (88), he (53), said (17), land (15), mountain (14), father (13), path (12), years (12)
Short Story. 3	**Dark They Were, And Golden Eyed**	Said (25), Harry (24), rocket (13), Bittering (12), earth (10), looked (10), wife (9), away (8)
Short Story. 4	**Thank You, M'am**	Said (28), boy (27), woman (23), got (10), face (9), door (8), run (8), going (7)
Short Story. 5	**The Piece of a String**	Mr (20), Hubert (17), rocket (12), book (10), man (8), people (8), said (8), string (7)

Extraction of the most occurring words of each short story shows eight key themes and they have been chosen on the basis of their statistical weight. The

most occurring words in the text have been specified through Summary tool. These themes comprise characters and central ideas of each short story.

5. Conclusion, Implications and Futuristic Vision

This study has quantified stylistic qualities of the five short stories. First major common stylistic characteristic is the stylistic preference of "said" in the five short stories and this theme occurs 31, 17, 25, 28, 8 times respectively. It has revealed knowledge that these five stories evolve with dialogic words of key characters. The word "said" focuses on the voice of character, intimacy, dialogic style and the personalized opinions. Second prominent finding is that most occurring key words unveil key characters of each short story for example "Norma, Arthur, Steward" in 1st 'Button, Button', "I, he" characters in 'Clearing in the Sky', "Mr. Bittering, wife" in 'Dark They Were and Golden Eyed', "boy, woman" characters in 'Thank You M'am' and "Hubert, people" in 'The Piece of String'. Third finding is that first four short stories have repeated unique words three times and their total words are three times more than unique words. This stylistic quality accelerates the speed of reading and causes ease for basic level readers. Fourth finding is that on average the selected short stories comprise 9.9 words per sentence to 12.9 words per sentence. This average sentence length enhances readability for all age groups. This finding matches with findings of Stylo and Cluto tools (Eder, Piasecki, & Walkowiak, 2017). Fifth finding is the higher vocabulary density which betokens the use of easier vocabulary while lower vocabulary density suggests the inclusion of difficult vocabulary in the short story. To conclude, computational stylistics (Stamatatos, 2009) mined stylistic features with text mining tools instantly.

The futuristic approach is to analyse some sample representative short stories of a writer and explore repetitive characteristics to prescribe stylistic qualities of a writer. Other future researches can be conducted on finding stylistic qualities of Nobel laureates in literature. To emulate their stylistic features, new generation writers can follow their stylistic qualities to follow their footprints to earn the most prestigious reward of Nobel Prize. Besides, obnoxious text message blame game of Ayesha Gulalai to mudsling PTI Chairman Imran Khan in Pakistan can be finalised with Summary tool.

This study can be implicated on the selection of reading texts for different level of learners for instance primary, intermediate and advanced level readers. Furthermore, literary, non-literary and subjective type essays and assignments can be evaluated precisely in the same way. This study has implications in several

domains. In academia, Summary tool can be applied to analyse assignments of students or subjective answers or essays of students. In the domain of forensic linguistics, an accused's statement can be analysed with summary tool that what and how many times he/she repeats any word. This summary tool can also be applied for author identification purpose.

References

Amancio, D. R. (2015). A Complex Network Approach to Stylometry, 1–21. https:// doi.org/10.1371/journal.pone.0136076

Argamon, S., Šarić, M., and Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* pages 475–480. ACM.

Chakraborty, T. (2012). Authorship Identification in Bengali Literature: a Comparative Analysis. Proceedings of COLING 2012: Demonstration Papers. pp. 41-50

Durant, G. B. (2004). A typology of research methods within the social sciences. *NCRM Working Paper*, 1-22. Retrieved from http://eprints.ncrm.ac.uk/115/

Eder, M., Piasecki, M., & Walkowiak, T. (2017). An open stylometric system based on multilevel text analysis. *Cognitive Studies | Études Cognitives,* (17). https://doi.org/10.11649/cs.1430

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*.

Gamerman, E. (2015). Data miners dig into 'Watchman'. *Wall Street Journal.* Page. 107,

Houvardas, J., & Stamatatos., E. (2006). N-gram feature selection for authorship identification. In *Proceedings of artificial intelligence: Methodologies, systems, and applications*(p. 77–86). Retrieved from https://pdfs.semanticscholar.org/ 104e/ 507b4db1174137944e44efb248735a312682.pdf

Juola, P. (2006). Foundations and Trends in Information Retrieval. In *Authorship attribution.* 1(3):233–334.

Kang, N., & Yu, Q. (2011). Corpus-based Stylistic Analysis of Tourism English. *Journal of Language Teaching and Research, 2*(1). doi:10.4304/jltr.2.1.129-136

Krippendorff, K. (2003). *Content analysis: An introduction to its methodology.* SAGE Publications.

Li, Y., Ji, W., & Xu, D. (2017). Quantitative style analysis of Mo Yan and Zhang Wei's novels. *Proceedings of the International Conference on Web Intelligence - WI '17*. doi:10.1145/3106426.3109045

Malyutov, M. B. (2006). Authorship attribution of texts: A review In General theory of information transfer and combinatorics. pages 362–380. Springer-Verlag, Berlin, Heidelberg.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The Federalist.* Stanford: CSLI Publications.

O'Sullivan, J., Bazarnik, K., Eder, M., & Rybicki, J. (2018). Measuring Joycean Influences on Flann O'Brien. *Digital Studies/Le champ numérique*, *8*(1).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys.* 34(1):1–47.

Sinclair, S., & Rockwell, G. (2015, June 29). Voyant Tools Documentation. Retrieved March 1, 2018, from http://docs.voyant-tools.org/?p=0&scripto_action=transcribe&scripto_doc_id=944&scripto_doc_page_id=945

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (1999). Automatic authorship attribution. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. pages 158–164. Association for Computational Linguistics.

Stamatatos, E. Fakotakis, N. & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics.* 26(4):471–495.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. J. Am. Soc. Inf. *Sci. Technol*., 60(3):538–556.

Stamou, C. (2008). Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing.* 23(2):181–199.

Sundberg, D., & Nilsson, J. (2018). Papa Revisited: A Corpus-Stylistic Perspective on the Style and Gender Representation of Ernest Hemingway's Fiction.

The Writepass Journal. (2012, June 5). How to write a dissertation: Methodology - the writepass journal: The writpPass journal. Retrieved August 13, 2017, from https://writepass.com/journal/2012/06/how-to-write-a-dissertation-methodology/

Tweedie, F. J., Singh, S. and Holmes, D. I. (1996). An Introduction to Neural Networks in Stylometry. *Research in Humanities Computing,* 5: 249–263.

Van Halteren, H. Baayen, H. Tweedie, F. Haverkort, M. & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics.* 12(1):65–77,

Van Hulle, D., & Kestemont, M. (2016). Periodizing Samuel Beckett's Works: A Stylochronometric Approach. *Style. 50*(2), 172-202. doi:10.5325/style. 50.2.0172

Verhoeven, B., & Daelemans, W. (2014). CLiPS Stylometry Investigation (CSI) corpus : A Dutch Corpus for the Detection of Age , Gender , Personality , Sentiment and Deception in text. In *The 9th International Conference on Language Resources and Evaluation (LREC)* (pp. 3081–3085).

Waugh, S., Adams, A., & Tweedie, F. (n.d.). Computational Stylistics using Artificial Neural Networks.

Zhang, T., Damerau, F., & Johnson, D. (2002). Text chunking based on a generalization of winnow. *Journal of Machine Learning Research.* 2:615–637.