

Voyanting Hermeneutic Patterns From James Hilton's Novel Goodbye, Mr. Chips: A Text Mining Study

Zafar Ullah¹, Arshad Mahmood²

¹Instructor, Virtual University Islamabad, Pakistan. ORCID: <https://orcid.org/0000-0003-3773-1467>

²Professor National University of Modern Languages (NUML) Islamabad, Pakistan.

Abstract

As the originality of this research is concerned, text mining is the least explored study in Pakistan. The chief problem is that novels and their readings are uninteresting and time-consuming for digital natives of the Digi modern age. So, queries have been raised to explore key themes, collocation patterns, interlinks among various themes and characters, stylometry, and disambiguation of word sense through bidirectional context. To respond to these queries, a mixed-methods approach has been utilized with triangulation of Rakesh Aggrawal's Knowledge Discovery Theory and Rockwell and Sinclair's Hermeneutica Theory to explore knowledge patterns through Voyant text mining tools. Major findings of this research reveal the extraction of digital hermeneutic knowledge patterns by the exposure of key themes through Cirrus tool, the search of the most frequent standard collocation patterns through Phrases tool, the revelation of knowledge graph linking key themes and issues with Links tool, the unveiling of stylometric features with a Summary tool and disambiguation of confusing words with Context tool. Moreover, humanly unsearchable precise, accurate knowledge patterns and digital hermeneutic patterns have been discovered by learners and researchers in the twinkling of an eye. This research has implications on teaching, learning, visual studies, and analysis of small and big data texts because knowledge patterns along with statistical weight are discovered instantly.

Keywords: Voyant, text mining, Knowledge graphs, Stylometry, Collocation

Introduction

This study aims to extract knowledge patterns automatically with Voyant text mining tools, and these patterns are supported with accurate visuals, corpus, statistics, and accurate information. Now voyaging from general to particular, the multidisciplinary field of Digital Humanities (henceforth, DH) encompasses data mining, educational data mining, and text analytics/ text mining. DH is defined thus: the use of digital sources to extract semantic shades, societal

aspects, cultural values (Terras, Nyhan, & Vanhoutte, 2016). Other sub-disciplines of DH are elucidated thus: Data mining (DM) automatically extracts information trends and patterns with algorithms (Oracle, 2017). Text mining (henceforth, TM), or text analytics means the extraction of meaningful, interesting knowledge patterns from unstructured texts. Further delimiting DM, Educational Data Mining (EDM) is “a learning science” from the academic domain and for educational purposes. It strives to develop the present predicament of education, learning and teaching (Calders, & Pechenizkiy, 2012) through text mining.

The major research problem is that Pakistani textbooks are in traditional, static, and boring paper form while modern students spend most of their time with technological gadgets since they are digital natives of this Digi modern age. Learners acquire knowledge patterns with great toil and even then, those pieces of information are imprecise and inaccurate. Moreover, learning and teaching are considered as “neglected stepchild” in digital humanities (Gold, 2012). Simultaneously, big data and extensive production and profusion of data also necessitate the conduct of text mining studies. The remedy of the problem lies in shifting from close reading to distant reading (Moretti, 2013) with Voyant text mining tools which produce exact and interactive corpus, visuals, and knowledge patterns to learn in a very interesting and interactive manner, hence, it enables learners to discover unique knowledge patterns.

The rationale behind conducting this study is the scarcity of text mining research in Pakistan, to fill this gap, the present study has been designed. This study also introduces the distant reading technique (Moretti, 2013) through Voyant tools to explore knowledge and information with accuracy and precision from big data instantly.

This study is significant because it automatically extracts knowledge patterns with Voyant text mining tools and these patterns are supported with accurate visuals, corpus, statistics, and accurate information (Vanchena, 2012). Voyant is used in advanced countries to study big data texts and mine hermeneutic patterns.

The major objective of this study is to derive structured, harmonious, and new knowledge patterns from large unstructured data in the shortest possible time as the concept of smart learning and teaching with Technological Pedagogical Content Knowledge (TPACK) has been propagated in this post-modern technological era. Consequently, their “digital wisdom”, cognitive abilities, learnability, interest, and research-based learning aptitude develop in full magnitude.

Research Questions

The current research addresses these research queries:

1. How does cirrus expose major motifs in the novel *Goodbye, Mr. Chips*?
2. What sort of collocation patterns has been produced to acquire standard collocation?
3. How does the knowledge graph interconnect various issues and characters of the selected novel?
4. How does text mining summary exhibit stylometric attributes of the novelist?
5. How does the context of certain keywords resolve the ambiguity of problematic words?

The aforementioned research questions lead to produce qualitative, quantitative, and visual data for instance in question 1, thematic word clouds are in qualitative and in visual form while its

statistics are in numeric form. In question 2, Collocation patterns were in qualitative form but their count was in quantitative form. Therefore, such generated data required a mixed method for data analysis. As the newness of this research is concerned, the use of Voyant tools for reading and deriving academic and digital hermeneutic patterns from the novel, *Goodbye, Mr. Chips* is the first study in Pakistan. Apart from these elements, extracted knowledge patterns and interactive visuals through Voyant text mining tools are innovative and replete with academic and pedagogical intellect.

Literature Review

Developing an underpinning and premise, this segment shed light on previous research works on text mining of novels with Voyant tools, some significant corpus analyses of ESL textbooks, and previous EDM studies.

Previous Text Mining Research on Novels

Voyant text mining tools had been applied in more than 28 research projects. Jannidis analyzed 350 German novels and revealed knowledge patterns regarding culture and history with the help of Voyant tools. In another study, Yeates (2013) worked on 1500 apocalyptic fictions in his Ph.D. dissertation. He explained fictional material with frequency trends and word clouds. Apart from it, two other novels, ‘*The Mill on the Floss*’ by George Eliot and ‘*Sherlock Holmes*’ by Arthur Conan Doyle were analyzed with Voyant tools. Complete work of Shakespeare, Jane Austen, and J.K. Rowling’s Harry Potter series (Sinclair, & Rockwell, 2015) were analyzed with Voyant tools to discover interesting knowledge patterns. These studies just employed one or two tools but the present study employs five tools for detailed visual, qualitative, and quantitative data analysis.

Previous Studies on EDM

Data mining is a holistic study of all types of data but how educational data was mined for the service of academia is a point of discussion in this part. Ueno (2004a) emphasized the use of text mining for learning and teaching purposes. Chen, Li, Wang & Jia (2004) suggested automated processes to write e-textbook, so content should be crawled from websites. Simultaneously, web mining for academic purposes was proposed. Tane, Schmitz & Stumme (2004) searched and compiled identical data by use of clustering technique in EDM. Hammouda & Kamel (2010) worked on text mining of documents.

Romero and Ventura (2007) wrote a review paper covering a decade ranging from 1995 to 2005 and found 81 works of EDM. Among them, seven text mining works were present in the list. Use of visuals, classification, statistics, clustering, and association rules were common features in them.

Pena-Ayala, Domingues, and Medel (2009) found text mining tools, models by reviewing 91 studies of EDM and Computer-Based Education System (CBES). In another study, Backer and Yacef (2009) studied 45 seminal works on EDM, knowledge patterns, student learning models, and impacts of learning. This academic investigation also concentrated on the extraction of knowledge patterns through Data Mining Knowledge Discovery (KDD).

Likewise, the current study also unveiled knowledge patterns with KDD from a novel taught in intermediate.

The most extensive review work was done by Romero & Ventura (2010) by studying 235 research works. They were divided into categories for instance graphic data in 35 works; instructional material for students in 52 works; performance of students in 76 works; modeling for learners in 28 studies; unsuitable behaviors of learners in 23 studies; mind maps in 10 studies etc.

Some important research works on EDM are being delineated here: Becker et al. (2000)'s work was done to classify traditional education; Tang et al. (2000)'s work on text mining AIWBE system; Luan (2002)'s research on clustering; Shen et al. (2002)'s work on visualization LCM system; Minaei-Bidgoli & Punch (2003)'s study on classification Web-based course; Shen et al. (2003)'s work on clustering Web-based course and sequence pattern; Romero et al. (2004)'s research on association AIWBE system; Ueno (2004a)'s study on text mining Web-based course; Ueno (2004b)'s seminal work on web-based course; Avouris et al. (2005) are designing of statistic Web-based course; Feng et al. (2005)'s research on prediction AIWBE system; Dringus and Ellis (2005)'s work on text mining LCM system; Hammouda & Kamel (2010)'s designing of text mining in Web-based syllabus (Romero & Ventura, 2007, p.141).

Previous Corpus-Based Studies of EFL Textbooks

Since studies on text mining of textbooks and novels are very limited, some corpus analyses of English textbooks have been done. The rationale for discussing previous literature on corpus use in textbooks studies is that text mining first developed corpus and then created other visuals. Consequently, text mining is more comprehensive as compared to corpus studies.

Johns (1991) initiated the idea of exploitation of corpus for learning and teaching as a researcher interacted with data and corpus. Learning through research guided that corpus is a beneficial study instrument in process of learning and teaching (Meunier & Gouverneur, 2007, p.153). This research technique motivated students who should explore real language patterns from corpora through learning by doing activities.

In some other studies, Biber et al (1999, p.992, 993) worked on three to five-word lexical bundles in ESL textbooks and he found the presence of 1:10:100 in one million word text. It resulted in one-word learning, then phrases learning, and later sentence learning stages in language teaching methods. So similar material was being introduced in harmony to easy to difficult and known to unknown teaching and learning approaches.

Some other corpus experts Burnard and McEnery (2000), Sinclair (2004), Connor and Upton (2004) studied uses of corpus for TEFL. Mukherjee and Rohrbach, (2006) and O'Keeffe, McCarthy, and Carter (2007) gathered 1st and 2nd language learners' corpus for learning and teaching L2. After learning some years, it has been noticed that there remained a wide gap between the linguistic competence of natives and non-natives. Its major reason was incompatible material from TEFL textbooks. The inclusion of natives' linguistic items in TEFL textbooks certainly lessened the linguistic demarcation between natives and non-natives.

Grammar and lexical bundles were studied through corpus (Gabrielatos, 1994; Biber et al., 2004; Romer, 2004a and 2004b; Koprowski, 2005; Meunier & Gouverneur, 2007). Genre

studies of ESL textbooks were conducted to segregate different elements of register and English for Academic Purposes (Swales 2002; Paltridge 2002; Biber et al. 2002). The wrong use of registers leads to serious pragmatic and semantic failure; hence, communication fails and the purpose of learning a language flops. Phraseology was also researched from textbooks through corpus (Biber et al. 2004; Koprowski 2005; Meunier & Gouverneur 2007; Gouverneur 2008). Similarly, the present research also discussed the 15 most common phrases from the novel.

Some previous textbook analyses were conducted on automated corpus methods (Biber et al., 2004; Chujo, 2004; Anping, 2005; Romer, 2004b, 2006; Meunier & Gouverneur, 2007 and Gouverneur, 2008). They benefitted already built corpora in the analysis of textbooks. Current research initiated the emerging trend of making new corpus through text mining of the novel, *Goodbye, Mr. Chips*.

First time in the 21st century, Biber et al. (2002) compiled the first textbook corpora whose title was TOEFL 2000 Spoken and Written Academic Language Corpus. It consisted of 27 million words of oral and written utterances recorded in American alma maters. They strived to show the interrelationship of textbook language and daily conversational language on parameters of lexical bundles. Furthermore, lexical bundles are significant because they transmit the ideology of language users too. Apart from it, learning lexical bundles augmented written and oral linguistic fluency.

Then Romer (2004a) prepared a 100,000 words second textbook corpus, German English as a Foreign Language Textbook Corpus (GEFL TC). She revealed differences in modal auxiliaries and continuous sentences in English textbooks. To understand linguistic and grammatical discrepancies, the corpus-based comparison is essential. In advanced countries, the corpus is used in classrooms for self-paced and authentic learning and teaching. In the same year, Chujo (2004) compared the lemmas of English textbooks with the wordlist of the British National Corpus (BNC). To attain a word list for a specific profession, it is subtracted from BNC general word list. The remaining particular words are a specialized word list of certain occupations. The very next year, Anping (2005) worked on a comparative study between modal auxiliaries and progressives by preparing a corpus of 100,000 words of Chinese and foreign English teaching textbooks.

Meunier & Gouverneur (2007) and Gouverneur (2008) prepared 700, 0000 words TeMa Corpus of textbooks to explore phrases with two adjectives. In 2007, Gouverneur studied the use of two adjectives and Gouverneur (2008) researched the usage of two verbs “make” and “take” qualitatively and quantitatively in the TeMa corpus derived from three ESL textbooks.

Gabrielatos (2005) posited that textbook corpora were “pedagogic” in their application. Biber (2004)’s work was on 84 lexical bundles and their ideology to understand the TEFL register. He found that referential bundles were less existent in textbooks while they were frequently present in teaching discourse. Thus, this research improved the quality of TEFL books.

Discussing further effects of corpus on ESL textbooks, new corpus-based textbooks are being published, hence, Cambridge added a sticker on the title of corpus-based books. Another change is that our textbook corpus is entirely different from everyday spoken English, so there is incongruity and to bridge this discrepancy, new linguistic items are being added in English ESL and TEFL textbooks to teach real and pragmatic English language. If language material is

deviant from real language, it means learning and teaching are going astray. Inclusion of delexicalised verbs in 'Headway' textbook, presence of reported speech in 'Cutting Edge' English textbooks, and incorporation of vague words in five volumes textbook 'Innovations' (Dellar & Hocking, 2000) are practical outcomes of corpus studies.

Methodology

This section shed light on Saunders, Lewis, & Thornhill's (2012) research on and research methodology has been written accordingly. Philosophy determines the research paradigm and this research opted for post-positivist philosophy which leads research design to deductive research approach, machine-based empirical research, and relativistic perspective (The Writepass Journal, 2012, June 5). Its data generation and data analysis followed a mixed-methods approach

As the theoretical premise is concerned, Knowledge Discovery Theory in Data Mining (KDD) was applied in this study since it was multidisciplinary in nature and is comprised of machine learning, NLP, artificial intelligence, and statistical information (Fayyad, Shapiro, Smyth, & Uthurusamy, 1996). Rakesh Aggrawal, a renowned Indian computer scientist, is a pioneer of KDD (IBM Research, n.d.). KDD is defined that it reveals hidden and unknown knowledge patterns with the help of data (Cabena, Hadjinian, Stadler, Verhees, Zanasi, 1998). Thus, knowledge patterns construct knowledge on canonized paradigms and its "positive externalities" (McDonald, 2012) are essentially beneficial for learners. Its prime objective is to convert crude or dispersed data into an interesting knowledge pattern (Rahayana, Siberschatz, 1998). Besides, KDD and TPACK model (technology, pedagogy, and content) (Koehler & Mishra, 2009) established an underpinning to integrate technology into the academic text to produce better comprehension, learning, and teaching.

Hermeneutic Theory highlights the following attributes:

- i. Its roots lie in the context of the text.
- ii. In the domain of computer programming, it did not examine the actual background program which was executed.
- iii. It manipulates data for a better understanding of data.
- iv. Different sorts of data interlink and strengthen each other for verification and comprehension.
- v. These tools and visuals lead to critical thinking.
- vi. If these tools fail, even then they expose new knowledge patterns.
- vii. These tools are interactive in nature, so they can be extended to meet the needs of research (Rockwell, & Sinclair, 2016, p. 166).

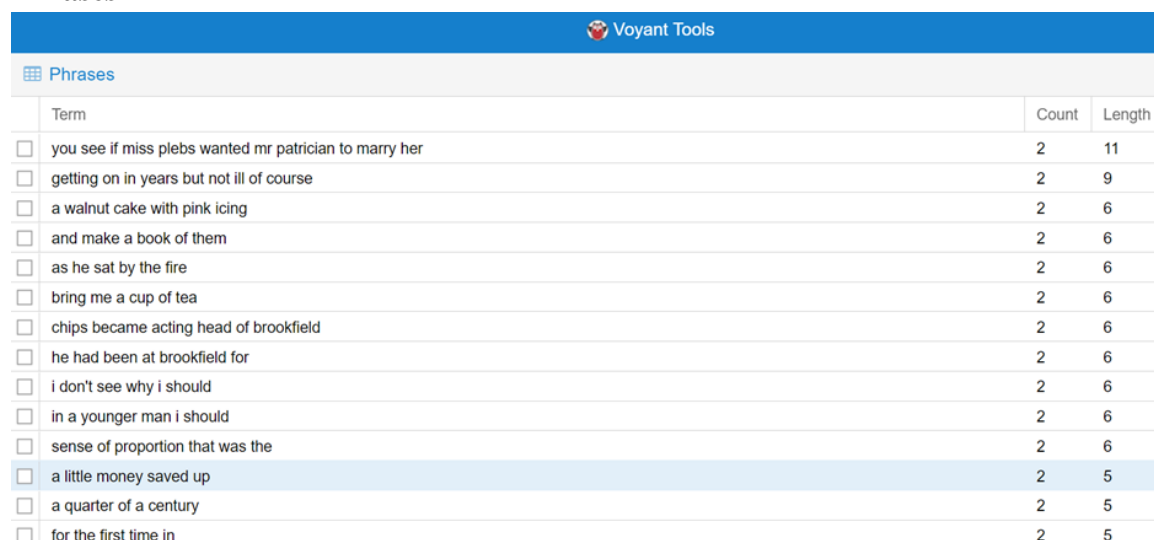
A brief primer of the Voyant tool is that it was designed by Stefan Sinclair (McGill University, Canada) and Geoffrey Rockwell (Alberta University, Canada) in 2003. Delimiting Voyant tools, Cirrus/word cloud of frequent vocabulary, Phrases, Links, Summary and Contexts tool was used to generate visual and tabular data for analysis. Cirrus tool generated word cloud with statistical data from 25 to 500 words. Phrases tool exhibited corpus and collocation patterns. Links tool showed the interconnectivity of words and themes in a visual pattern. The summary tool presented a complete overview of text mining, number of words, number of unique words, and vocabulary. Context's tool showed the context of most occurring

(39)” referred to students at Brookfield. Another masculine character, “Ralston (25)”, a modern principal of Brookfield, had a row with Chips and the latter was the victor in the conflict.

The name of “Katherine (9)”, a girlish wife of Mr. Chips, was used only 9 times and it suggested that her role and life were very short but the personal pronoun “she” was used 67 times. It meant that “she” was with him indirectly throughout the novel but her physical presence in the novel was transient. Another feminine character for instance Mrs. “Wickett (20)” in whose house Chips stayed after retirement, dominated after the retirement phase of the novel.

Another gender pattern was explored in that only two female characters, Katherine and Mrs. Wickett’s names were found while there are dozens of male names in the novel. Furthermore, the word “boys” was used 39 times. It revealed that males dominated novels and females stayed for a very short period. Again, a gender bias was exposed that there was no personal feminine name for Mrs. Wickett. Themes of “laughter (19)” and “joke (19)” were used in the text and the novel exhibits comic elements in the form of jokes. Another theme “remember (36)” referred to the nostalgic state of mind of central figure Mr. Chips. To conclude, the above-mentioned word cloud revealed major topics of discussion and characters in the novel.

Phrases



Term	Count	Length
<input type="checkbox"/> you see if miss plebs wanted mr patrician to marry her	2	11
<input type="checkbox"/> getting on in years but not ill of course	2	9
<input type="checkbox"/> a walnut cake with pink icing	2	6
<input type="checkbox"/> and make a book of them	2	6
<input type="checkbox"/> as he sat by the fire	2	6
<input type="checkbox"/> bring me a cup of tea	2	6
<input type="checkbox"/> chips became acting head of brookfield	2	6
<input type="checkbox"/> he had been at brookfield for	2	6
<input type="checkbox"/> i don't see why i should	2	6
<input type="checkbox"/> in a younger man i should	2	6
<input type="checkbox"/> sense of proportion that was the	2	6
<input type="checkbox"/> a little money saved up	2	5
<input type="checkbox"/> a quarter of a century	2	5
<input type="checkbox"/> for the first time in	2	5

Figure 2. Phrases

Phraseology enhanced fluency in speaking and writing while mismatch in bigrams distorted communication and linguistic norms. Excluding nonsense collocations, standard bigrams from the first 15 phrases have been selected for instance “Brookfield bells” (Adj+N), “Brookfield boys” (Adj+N), “Brookfield history” (Adj+N). To learn a foreign language, usually, the learning process starts from one word to phrases and then typical sentences. These phrases accelerated learners’ fluency in all language skills. Apart from language learning, they are very useful in searching hedges. Furthermore, characters disclosed their ideologies during conversation since dialogues of characters manifested their inner selves.

Links

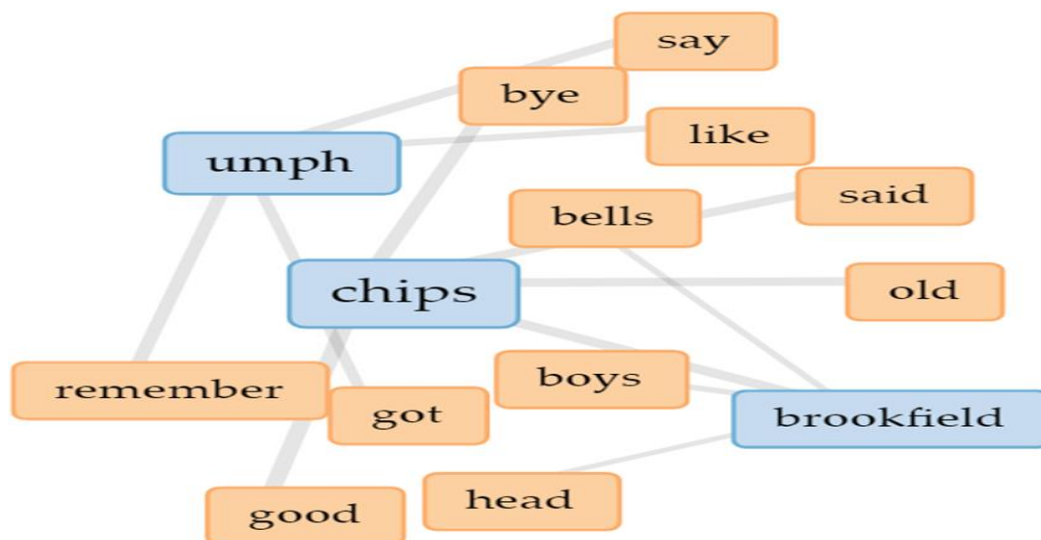


Figure 2. Links

Artificial Neural Networks (ANN) function like the human brain, so the strong relationship of various issues was evident that “Chips” and “Brookfield” had an integral relationship and it ran in his veins like blood. Another ANN showed that Brookfield “School” was deeply associated with “boys”, school “bells” since boys studied there and they followed bells to lead an organized life.

The utterance of nonverbal expression “umph” was frequently uttered by Mr. Chips, that’s why words “say” and “said” were associated with the words “umph” and “Chips”. The title of the novel was Goodbye, Mr. Chips, and the word “Chips” was also linked with “bye” which occurred six times in the corpus. The novel of the title was spoken by Katherine and Linford boy during their conversations with “Chips”. Another ANN about the word “head” and “Chips” was also related to each other because Chips worked as an acting head of school and except Ralston, all heads of Brookfield always respected Chips.

Chips and “like ” relationship were presented here. Word “like ” occurred 65 times in the corpus while the word dislike occurred just twice. It referred that Chips liked “boys” and “Brookfield' ’ excessively and reciprocally Brookfield boys liked him. The word “remember” referred to his nostalgic feelings about “Brookfield”, “boys”, “head” and Katherine because his life was replete with old memories at Mrs. Wickett’s residence.

In a nutshell, the Links tool showed the interconnectivity of characters, thoughts, and themes. The interrelationship is integral for understanding the internal weaving of a novel, links of characters, bonding of themes since the critique of the novel revolved around the aforementioned links. Nonverbal expression “umph” was linked with Chips, not with Ralston or Katherine. Mrs. Wickett and Katherine’s link was never shown because they never interacted with each other.

The debate also arises about the utility of just finding the number of repetitions of words or phrases. The power of rhetoric and emphasis lies in repetition as Martin Luther King junior

repeats the phrase “I have a dream” in his renowned speech. The number of words transforms verbal analysis into statistical and clear proof of some key ideas.

Summary

This corpus has 1 document with 16,758 total words and 3,160 unique word forms.

Created 28 seconds ago.

Vocabulary Density: 0.189

Average Words Per Sentence: 16.4

20 Most frequent words in the
corpus: chips (155); umph (120); brookfield (86); old (68); school (60)

Figure 3. Summary

Summary tool disclosed computational stylistic attributes of the novelist, James Hilton. He used 3160 unique words and after almost five times repetition, he composed 16,758 words to complete the novel. Vocabulary density showed after how many words a new word occurred. Mathematically, it was derived by division of total words by unique words and this process was termed as Inverse Absolute Vocabulary density (Simpson, 2000, May). If vocabulary density was higher, it meant the novelist had used new words frequently in the text. If vocabulary density was low as this corpus summary showed only 0.189, it suggested that James Hilton used simple, easy and repetitive vocabulary throughout the novel.

To extend this discussion, whenever any teacher desires to select a text for his/her students, vocabulary density should be checked. For beginners, low vocabulary density should be recommended while high-density vocabulary books, lessons, or stories should be selected for advanced level. So, the text mining summary facilitated the selection of suitable books for the right level of learners. Moreover, the same technique could be applied to check the quality of research and literary works, exam papers, and homework.

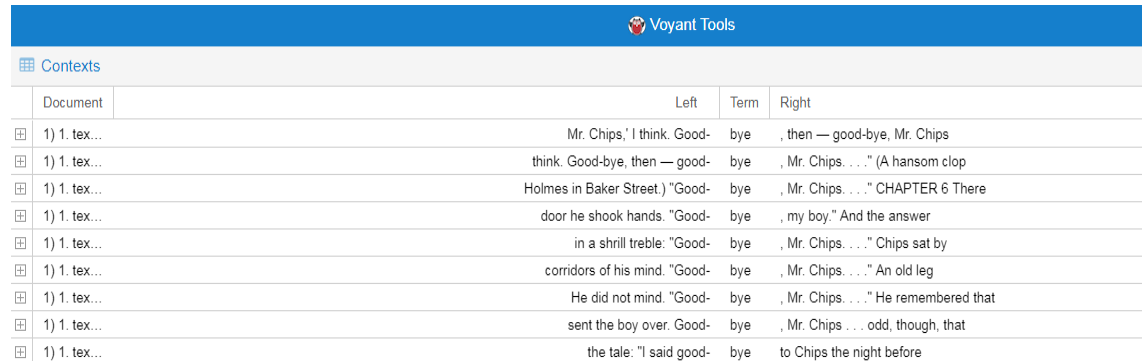
The novelist used an average of 16.4 words per sentence in this novel. Before text mining, just a generalized statement was given that this novelist wrote smaller or longer sentences but their length was not quantified. Consequently, the Summary tool quantified the style of the novelist computationally and this new quantitative dimension was attained by the text mining process. Fewer words per sentence showed frequent use of simple sentences while longer sentences indicated the use of compound and complex sentences in the text. Another application was that the same Summary tool analysis could be used for the assessment of sentences during exams.

Lastly, 5 most common words with their occurrence number were given to show a thematic glimpse of the novel.

Text mining summary not only elaborated total words but also showed unique words. By compiling and analysing total works, the total vocabulary of a writer was quantified. One step ahead, if it was compared with the present dictionary, coinage of new words could also be assessed.

Contexts

Word “wire” was used twice. Once it was used in the literal sense with “wire guard” and the second time it was used for Ralston for his dynamics and fiery nature with “live wire and fine power transmitter”. As a result, the Context tool clarified and disambiguated the semantic layer by showing its context.



Voyant Tools			
Contexts			
Document	Left	Term	Right
1) 1. tex...	Mr. Chips; I think. Good-	bye	, then — good-bye, Mr. Chips
1) 1. tex...	think. Good-bye, then — good-	bye	, Mr. Chips. . . . (A hansom clop
1) 1. tex...	Holmes in Baker Street.) "Good-	bye	, Mr. Chips. . . . CHAPTER 6 There
1) 1. tex...	door he shook hands. "Good-	bye	, my boy." And the answer
1) 1. tex...	in a shrill treble: "Good-	bye	, Mr. Chips. . . . Chips sat by
1) 1. tex...	corridors of his mind. "Good-	bye	, Mr. Chips. . . . An old leg
1) 1. tex...	He did not mind. "Good-	bye	, Mr. Chips. . . . He remembered that
1) 1. tex...	sent the boy over. Good-	bye	, Mr. Chips. . . . odd, though, that
1) 1. tex...	the tale: "I said good-	bye	to Chips the night before

Figure 5. Context

The context of the first word of the title of the novel, “Goodbye” was searched and it was found 9 times in this corpus. The First 3 times, it was uttered by Katherine just a night before their marriage. Then 2 times, Linford, a small child, uttered this word. The last 3 times, Mr. Chips thought of both Katherine and Linford. Voyant tool facilitated elucidation of title, its thematic, contextual, and sequential occurrence. Instead of reading the entire novel to find out its title, just in a few seconds, precise and useful knowledge patterns were discovered. Thus, the context of keywords took the readers directly to various uses of a word.

Conclusion

This pioneering study with Voyant tools established a beneficial premise to use it as a reading and learning tool in Pakistan as students from 22 foreign universities for instance University of North Texas, USA, Stanford University, USA, Michigan State University, etc are studying textbooks with the help of Voyant tools.

Major findings of this research were mining of hermeneutic knowledge patterns by exposure of key themes through word cloud; extraction of most occurring standard collocation patterns; the revelation of linking issues; disclosure of vocabulary density, number of unique words, text mining summary; and clearance of semantic ambiguity of confusing words. Another discussion was that we should analyse the text with Voyant tools without reading or after reading the text. If the analysis was done after reading the text, how did this tool benefit us? Voyant tool worked in both ways. If there is big data for example all novels in English literature from all ages, obviously nobody can read them all together with the counting of words and themes. Voyant scrutinized all sources without reading them minutely. If we look into five skins, the second skin is the reader. The implied message is to read the original text to dig out the true meaning. To conclude, digital tools are to facilitate hermeneutic processes on a scientific and statistical basis with human interpretive support. Furthermore, digital tools build interactive visuals and knowledge patterns to comprehend large texts in the shortest possible time.

The contribution of this study is to establish innovative, attention-grabbing, and valuable ways of distant reading of novels and other pieces of literature as Moretti (2013) suggested to read large volumes most shortly and accurately. Moreover, reading with Voyant is interesting and Cirrus, Phrases, Links, Summary, Contexts tools made visuals and information striking and useful for deep learning and teaching. This study advocated changing traditional reading and learning style to Voyant based and digital hermeneutic embedded learning style to delve deeper into the realm of scientific study and knowledge discovery. Future research projects can be digital mining of digital books, image mining, audio-video mining, web mining, topic mining, multimedia mining, and cross-comparison of mined data.

References

- Anping, H. (2005). A corpus-based evaluation of ELT textbooks. Paper presented at the joint conference of the American Association of Applied Corpus Linguistics and the International Computer Archive of Modern and Medieval English, 12–15 May 2005, University of Michigan.
- Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future vision. *Journal of Educational Data Mining*, 1(1), 1–15.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36(1): 9–48.
- Biber, D., Conrad, S. & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.
- Burnard, L. & McEnery, T. (eds). 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Papers from the Third International Conference on Teaching and Language Corpora. Frankfurt: Lang.
- Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining. From Concept to Implementation*. Upper Saddle River, NJ: Prentice Hall.
- Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter*, 13(2), 3. doi:10.1145/2207243.2207245
- Chen, J., Li, Q., Wang, L., & Jia, W. (2004). Automatically generating an e-textbook on the web. In *International conference on advances in webbased learning* (pp. 35–42).
- Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests. Using a BNC lemmatized high frequency word list. In *English Corpora under Japanese Eyes*, J. Nakamura, N. Inoue, N. & T. Tabata (eds), 231–249. Amsterdam: Rodopi.
- Connor, U. & Upton, T. (eds). (2004). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi.
- Dellar, H. and D. Hocking. (2000). *Innovations*. Hove, Language Teaching Publications.
- 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/>
- Fayyad, U. M., Shapiro, G. P., Smyth, P., & Uthurusamy, R. (Eds.). (1996). *Advances in knowledge discovery and data mining*. Menlo Park, Canada: AAAI Press.

- Gabrielatos, C. (1994). Collocations, pedagogical implications and their treatment in pedagogical materials. Ms, Research Centre for English and Applied Linguistics, University of Cambridge. Available at <http://www.gabrielatos.com/Collocation.htm>
- Gabrielatos, C. (2005). Corpora and language teaching: Just a fling, or wedding bells? *TESL-EJ* 8(4): A1, 1–37.
- Gold, M. K. (2012). *Debates in the digital humanities* (1st ed.). University of Minnesota Press.
- Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In *Phraseology in Foreign Language Learning and Teaching*, F. Meunier & S. Granger (eds). Amsterdam: Benjamins. Gouverneur, C. Forthcoming. *Phraseology in Foreign Language Learning and Teaching: A Corpus-based Study of EFL Textbooks*. PhD dissertation, Universite Catholique de Louvain.
- Graham, S., Milligan, I., & Weingart, S. (2013). *Voyant Tools | The Historian's Macroscope: Big Digital History*. Retrieved August 11, 2017, from http://www.themacroscope.org/?page_id=639
- Hammouda, K., & Kamel, M. (2010). Data mining in e-learning. In *E-learning networked environments and architectures: A knowledge processing perspective* (pp. 374-404).
- IBM Research. (n.d.). *Knowledge Discovery and Data Mining - IBM*. Retrieved August 15, 2017, from http://researcher.ibm.com/researcher/view_group.php?id=144
- Johns, T. (1991). Should you be persuaded – Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing [ELR Journal, 4]* (pp. 1–16).
- Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education*. 9(1), 60-70.
- Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal* 59(4): 322–332.
- McDonald, D. D. (2012, March 14). Value and benefits of text mining. Retrieved March 16, 2017, from <https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>
- Meunier, F. & Gouverneur, C. (2007). The treatment of phraseology in ELT Textbooks. In *Corpora in the Foreign Language Classroom*, E. Hidalgo, L. Querada & J. Santana (eds), 119–139. Selected papers from the Sixth International Conference on Teaching and Language Corpora. (TaLC), University of Granada, Spain, 4–7 July, 2004. Amsterdam: Rodopi.
- Moretti, F. (2013). *Distant reading*. London: Verso.
- Mukherjee, J. & Rohrbach, J. M. 2006. Rethinking applied corpus linguistics from a language pedagogical perspective: New departures in learner corpus research. In *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*, B. Kettemann & G. Marko (eds), 205–232. Frankfurt: Lang.
- O'Keeffe, A., McCarthy, M. & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: CUP.

- Oracle. (2017). What Is Data Mining? Retrieved July 25, 2017, from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCJ
- Paltridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes* 21: 125–143.
- Peña-Ayala, A., Domínguez, R., & Medel, J. (2009). Educational data mining: a sample of review and study case. *World Journal of Educational Technology*, 2, 118–139.
- Rahayana, S., & Siberschatz, A. (1998). On the discovery of interesting patterns in association rules. In *Proceedings of the 24th VLDB Conference* (pp. 368-379). New York, NY.
- Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: Computer-assisted interpretation in the humanities*. London, England: MIT Press.
- Romer, U. (2004a). Textbooks: A corpus-driven approach to modal auxiliaries and their didactics. In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.), 185–199. Amsterdam: John Benjamins.
- Romer, U. (2004b). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In *Corpora and Language Learners*, G. Aston, S. Bernardini & D. Stewart (eds), 151–168. Amsterdam: John Benjamins.
- Romer, U. (2006). Looking at looking: Functions and contexts of progressives in spoken English and ‘School’ English. In *The Changing Face of Corpus Linguistics. Papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24)*, A. Renouf & A. Kehoe (eds), 231–242. Amsterdam: Rodopi.
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33, 135-146.
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on systems, man, and cybernetics, part C: applications and reviews*, 40(6), 601–618.
- Saunders, M., Lewis, P., & Thornhill, A. (2012). *Research methods for business students* (6th ed.). Pearson Education Limited.
- Simpson, Z. B. (2000, May). Project Gutenberg Vocabulary Analysis. Retrieved August 18, 2017, from <http://www.mine-control.com/zack/gutenberg/>
- Sinclair, S., & Rockwell, G. (2015). *Principles of Voyant Tools | Voyant Tools Documentation*. Retrieved May 29, 2017, from <http://DOCS.VOYANT-TOOLS.ORG/CONTEXT/PRINCIPLES/>
- Sinclair, J. (ed.). (2004). *How to Use Corpora in Language Teaching [Studies in Corpus Linguistics 12]*. Amsterdam: John Benjamins.
- Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In *Academic Discourse*, J. Flowerdew (ed.), 150–164. Harlow: Longman.
- Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic resource management for the web: An e-learning application. In *Proceedings of the WWW conference*, New York, USA (pp. 1–10).

- Terras, M. M., Nyhan, J., & Vanhoutte, E. (2016). Selected definitions from the day of digital humanities 2009-2012. In *Defining digital humanities: A reader*. London, England: Routledge.
- The Write pass Journal. (2012, June 5). How to write a dissertation: Methodology - the write pass journal: The writp Pass journal. Retrieved August 13, 2017, from <https://writepass.com/journal/2012/06/how-to-write-a-dissertation-methodology/>
- Ueno, M. (2004a). Data mining and text mining technologies for collaborative learning in an ILMS “samurai”. In ICALT.
- Vanchena, L. A. (2012). Reading German Culture, 1789–1918 [Distant Readings/Descriptive Turns: Topologies of German Culture in the Long Nineteenth Century. 21st St. Louis Symposium on German Literature & Culture, Washington University in St. Louis, March 29–31, 2012.] | Vanchena | JLT online Conference Proceedings. Retrieved March 25, 2017, from <http://www.jltonline.de/index.php/conferences/article/view/502/1306>
- Yeates, R. (2013, May 2). Voyant Tools | Post-Apocalyptic Cities. Retrieved March 25, 2017, from <https://postapocalypticcities.wordpress.com/2013/05/02/voyanttools/>