UNVEILING KNOWLEDGE PATTERNS IN INTERMEDIATE ENGLISH TEXTBOOKS THROUGH VOYANT TEXT MINING TOOLS: A DIGITAL HUMANITIES STUDY

By

Zafar Ullah

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

January, 2019

Unveiling Knowledge Patterns in Intermediate English Textbooks through Voyant Text Mining Tools: A Digital Humanities Study

By

Zafar Ullah

M.Phil., National University of Modern Languages, 2014

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
In **English**

To

FACULTY OF ARTS AND HUMANITIES

NATIONAL UNIVERSITY OF MODERN LANGUAGES, ISLAMABAD

NATIONAL UNIVERSITY OF MODERN LANGUAGES FACULTY OF ARST AND HUMANITIES

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with the overall exam performance, and recommend the thesis to the Faculty of Arts and Humanities for acceptance:

Thesis Title: Unveiling Knowledge Patterns in Intermediate English Textbooks through Voyant Text Mining Tools: A Digital Humanities Study

Submitted By: Zafar Ullah **Registration #:** 598/PhD/Eng/S-16(Ling)

Dr. Arshad Mahmood
Name of Research Supervisor

Signature of Research Supervisor

Dr.Inayat Ullah
Name of HoD

Signature of HoD

Dr. Muhammad Uzair
Name of Dean (FAH)

Signature of Dean (FAH)

Prof. Dr. Muhammad Safeer Awan
Name of Pro-Rector Academics

Signature of Pro-Rector Academics

Maj. Gen. Muhammad Jaffar HI (M) (Retd)
Name of Rector

Signature of Rector

Date

CANDIDATE'S DECLARATION FORM

I Zafar Ullah

Son of Saleem Ullah

Registration # 598/PhD/Eng/S-16 (Ling)

Discipline English Linguistics

Candidate of Doctor of Philosophy at the National University of Modern Languages do hereby declare that the thesis **Unveiling Knowledge Patterns in Intermediate English Textbooks through Voyant Text Mining Tools: A Digital Humanities Study** submitted by me in partial fulfillment of PhD degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution.

I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

Signature of Candidate

Zafar Ullah

Date

Name of Candidate

ABSTRACT

Thesis Title: Unveiling Knowledge Patterns in Intermediate English Textbooks through Voyant Text Mining Tools: A Digital Humanities Study

The contemporary digital era faces the challenge of extracting knowledge patterns from big diversified data which are difficult to read with the traditional “close reading” method. Likewise, traditional paper textbooks are considered inadaptable and less appealing, so their reading becomes uninteresting, time-consuming and less knowledge- investigative. This dissertation on text mining primarily aims to discover interactive knowledge patterns, innovative and idiosyncratic knowledge bearing dimensions through “distant reading”. To address the research problem, intermediate English textbooks have been analysed with five Voyant tools: Summary, Cirrus, Phrases, Links and Contexts. The main focus of the analysis is the transformation of static traditional Pakistani intermediate English textbooks into interactive data visuals of Summary, Cirrus, Phrases, Links and Contexts. Theoretical triangulation integrates Knowledge Discovery Theory and Hermeneutica Theory. Accordingly, the textbooks have been analysed with mixed methods to explore new interactive knowledge patterns. Results have been displayed in the form of data visualization, tabular, qualitative and quantitative data. The current research finds that Summary tool precisely quantifies stylometric features of total words, unique words, vocabulary density, average sentence length and the most frequent themes in each piece of writing. Cirrus discovers most of the key themes and characters. Phrases tool extracts 168 which are the most repeated standard collocation patterns. It was also found that Links tool interrelates almost all key ideas with one another through accurate Knowledge Graphs. Further, Context’s tool disambiguates word sense by discriminating their context, contextual meanings and parts of speech. The current study contributes by resolving the research problem, saving time with distant reading and adding aesthetic appeal for Voyant users. Finally, pedagogical implications of the current study introduce autonomous learning and teaching of textbooks, corpus building, visual generation, interesting knowledge pattern discovery and the data unification for libraries. Moreover, the current study also diverts students, teachers and publishers to digital text mined learning, teaching and publishing.

TABLE of CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Adv	Adverb
AI	Artificial Intelligence
ANN	Artificial Neural Network
AOLA	Austrian On-Line Archive
Apo	Apostrophe s
Art	Article
Aux	Auxiliary Verb
CALL	Computer Assisted Language Learning
CCAT	Center for Computer Analysis of Texts
Conj	Conjunction
DAE	Diploma of Associate Engineering
DCC	Digital Curation Centre
Det	Determiner
DM	Data Mining
DH	Digital Humanities
DSS	Decision Support System
DV	Data Visualization
EDA	Exploratory Data Analysis
ELT	English Language Teaching
ESL	English as a Second Language
ESP	English for Specific Purpose
Ex There	Existential There
FBTEE	The French Book Trade in Enlightenment Europe
HCY	Humanities Computing Yearbook
HTTP	Hyper Text Transfer Protocol
IBM	International Business Machines
Id	Idiom
Inf V	Infinitive verb

Int	Interjection
KG	Knowledge Graph
KDD	Knowledge Discovery Theory and Data Mining
KPK	Khyber Pakhtunkhwa
KWIC	Key Word In Context
LDA	Latent Dirichlet Allocation
LSP	Learn Smart Pakistan
MALL	Mobile Assisted Language Learning
ML	Machine Learning
Mod	Modal Verb
MWEs	Multi-Word Expressions
N	Noun
NDNP	National Digital Newspaper Program
NLP	Natural Language Processing
Nu	Number
NUML	National University of Modern Languages
OCP	Oxford Concordance Program
OLEs	Object Linking and Embeddings
PADI	Preserving Access to Digital Information
PNG	Portable Network Graphics
Phr V	Phrasal Verb
Phr	Phrase
PIECTZU	Pakistani Intermediate English Textbook Corpus Zafar Ullah
PP	Prepositional Phrase
Prep	Preposition
Prn	Pronoun
Prop Adj	Proper Adjective
POS	Parts of Speech Tagging
Ref Prn	Reflexive Pronoun
Rel Prn	Relative Pronoun
SEO	Search Engine Optimization

SGML	Standard Generalized Markup Language
SVM	Support Vector Machine
TEI	Text Encoding Initial
TM	Text Mining
USB	Universal Serial Bus
UWN	University World News
V	Verb
VRML	Virtual Reality Modeling Language
VUE	Visual Understanding Environment
XML	Extensible Markup Language
WIPO	World Intellectual Property Organization
WSD	Word Sense Disambiguation

ACKNOWLEDGEMENT

First of all, I am thankful to ALEEM (The All-Knowing) Allah Almighty Who helped me to accomplish this task of research.

My research supervisor, Prof. Dr Arshad Mahmood (Director Publications), always politely suggested and guided me to refine this research work. My co-supervisor, Dr Muhammad Noman Malik, Head of Computer Science Department, NUML also guided me to elevate the standard of this study according to modern paradigms. I am highly thankful to Canadian Prof. Geoffrey Rockwell (University of Alberta, Canada) and Prof. Stefan Sinclair (University of McGill, Canada), the two designers of Voyant tools, who promptly responded to my all emails in detail.

I extend my sincere gratitude to my foreign co-supervisor, Prof. Carolyn Penstein Rose, who replied to my questions and assisted me with her quick feedback, suggested books, videos, lectures and other academic content. She was always available for the help of an unknown person selflessly. Later, she provided me with the golden opportunity to stay at Carnegie Mellon University (CMU), the USA for five months and to interact with CMU’s worthy scholars who enlightened my vision and scholarship.

I am thankful to Dr Qayyum Khattak, Mr Mudassar from NUML, Shahzeb Shafi, Farial Mir, Wishma Ihsan, Aiman Qudoos, Alishba, Amna Ajmal for facilitating me in proofreading. Dr Hadia from NUML assisted me in the comprehension of statistics and mathematics. Mr Zia from NUML solved technical problems, and Mr Mughees from NUML assisted me in verifying collocation categories.

Some silent agents supported me with their prayers, pieces of advice and untold numerous sacrifices. My parents, my spouse, my sweet angels (children) and sisters sacrificed on many occasions and did not utter even a single word of complaint. They missed my company without disturbing and telling me. Once my daughter told me to come home, and I told her that I could not come because I was writing a big book. She told me innocently to come home, stay with us for ten thousand months and bring all your pages with you and she will write all pages quickly for me. May God reward all aforementioned persons with the best worldly and eternal reward! Amen

DEDICATION

This dissertation is being dedicated to The All-Knowing Allah Almighty Who originates and sustains old and new vistas of knowledge from His unfathomable and immeasurable treasures of revelation, rationality, knowledge, truth and reasoning for hermeneutics. I just pray

> “What is dark in me, illumine
> What is low, raise and support”
>
> (Milton)

CHAPTER 1

INTRODUCTION

1.1 An Overview

Text mining/ text analytics is a subdomain of digital humanities (DH), an intersection of humanities and digital tools to reveal digital hermeneutic patterns. Text analytics unveils the unknown interactive knowledge patterns in the form of data visualization. Text mining communicates profound insight and deep learning (a machine learning method for data representation) (Hearst, 1999). This section begins by outlining the definitions of key concepts in the title of the dissertation, statement of the problem, research objectives, research questions, research design, significance, scope of text mining and structure of the study, as shown in figure 1.

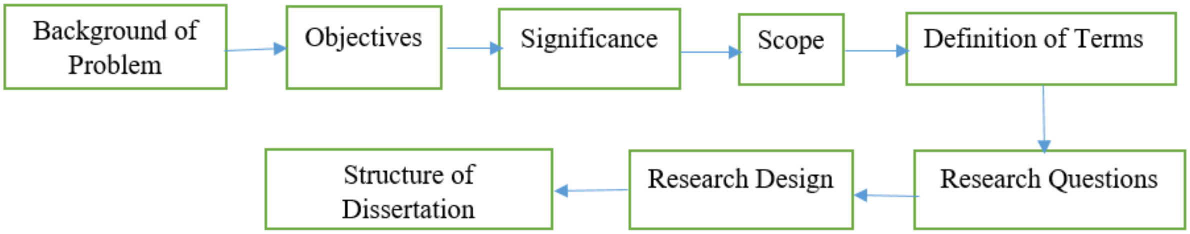


Figure 1 Flowchart of Introduction Chapter

Big data are heaping up with the passage of time and close reading of entire data becomes very difficult in a short time. Furthermore, students also lose interest in close reading and consider it a normative activity, while they spend most of their time using technological gadgets. So, it necessitates the introduction of distant reading and text mining of intermediate English textbooks to present an example of knowledge discovery through digital tools and how these knowledge patterns fulfil several academic and linguistic needs.

1.2 Background of Research Problem

With every passing day, teenagers buy new technological gadgets and the number of e-readers is increasing, while educational content in Pakistan is found only in hard copies of textbooks. It may cause a lack of motivation and interest in academic content and textbooks. Moreover, knowledge sources are so big and scattered that they have become unmanageable with close reading. Therefore, the solution to these problems is distant reading and text mining of big data unveils meaningful knowledge patterns. To divert students to digital academia, digital pedagogy and digital content have been promoted through the current study. Educational Data Mining (EDM) has emerged recently as an interdisciplinary study, and its roots lie in the domains of computational linguistics, Natural Language Processing (NLP), Artificial Intelligence (AI), statistics, Machine Learning (ML) and knowledge discovery.

1.2.1 Traditional and Digital Learners

Pakistani society has been divided into several economic classes which determine their learning styles and choice of academic institutions. Lankshear and Knobel (2011) classify two types of students: one group studies with traditional close reading and note-taking style, while the second group prefers to learn through digital content, gadgets and tools. The former belongs to the middle class and the latter belongs to the affluent class which can easily afford the technological gadgets and fees of such elite institutes. In addition, some of the Pakistani modern private schools and colleges have introduced CALL (Computer Assisted Language Learning) and MALL (Mobile Assisted Language Learning) with multimedia, tabs and computers in their classrooms. However, most of the Pakistani government schools and colleges are deprived of digital facilities, infrastructure and training; hence, their students require digital text and tools in and out of their classrooms for advanced level learning. Consequently, the current study is the practical manifestation of text mining and distant reading of intermediate English textbooks with five Voyant tools: Summary, Cirrus, Phrases, Links and Contexts.

Keeping in view another dimension, school studies on the traditional pattern and home study with technology can be combined with the effective use of DH which introduces “processes of learning that are deeper and richer than the forms of learning to which they are exposed in schools” (Gee, 2004, p. 107). As a result, computer-based education (CBE) imparts knowledge to students while promoting autonomous learning (Cristobal, & Sebastian, 2013). Its main advantage is that different IQ level students can extract their desired knowledge patterns according to their own learning pace.

1.2.2 Statement of the Problem/ Challenge

Textbook contents and their activities are full of "static" learning material (Romero, Ventura, Delgado, & De Bra, 2007), for this reason, learners consume more time than digital learning. Static material is inadaptable, unstructured or semi-structured, insufficient to fulfill all academic needs of digital learners. That is why some students memorize the given content and exercises with frequent practice, but their cognitive abilities, creativity, and comprehension levels have not been enhanced according to their academic level. Moreover, technology addicted students or "digital natives" (Prensky, 2001) of Digi Modern Age want to discover knowledge patterns and data visualizations interestingly through digital hermeneutics and autonomous methods of digital humanities during their homestay. This evidence strengthens the point thus: "45% of the world's internet users are below the age of 25" and "in most of the world's least developed countries, young people are nearly three times more likely than the general population to be using the internet" (United Nations, 2018). Furthermore, digital natives of the contemporary age are too busy to read voluminous books and scattered yottabyte (10^{24}) data to extract knowledge patterns. Following digital pedagogy of advanced universities, new digital humanities methods and tools must be incorporated into the learning process because printed books do not play their vital role in the construction and transfusion of knowledge in modern digital ways (Burdick, Drucker, Lunenefeld, Presner, & Schnapp, 2012). To transform this "static" learning material into dynamic interactive data visualization, a text mining process has been conducted to enhance "digital wisdom" (Prensky, 2012), which reflects in knowledge patterns of stylistic features, corpus summary word cloud, collocation patterns/ n-gram, KGs, and KWIC for word sense disambiguation.

Discussing challenges on technical grounds, students of humanities are not well versed in coding, computer language Python and the use of IDE as a code editor; therefore, they are unable to unveil knowledge patterns from big data or any text. That is why computer programmers can easily work on Python NLTK to extract collocations/ n-grams, context and summary; Python, Pandas, and Numly libraries to extract word clouds; Pyplot library to develop KGs. The use of Voyant tools has resolved this technical problem because Python libraries run these in-built codes in their tools to generate data visualization and knowledge patterns without any coding or programming skill. One postulate of the theory states that “It is not like black boxes” (Rockwell, & Sinclair, 2016, p. 166). It means that Voyant tools do not examine the actual background programme which is executed. Voyant tools concentrate on the discovery of knowledge patterns for hermeneutic patterns.

Pakistan is just at its initial stage of text analytics; hence, the current study aims to mine text mining of intermediate English textbooks with the triangulation of Knowledge Discovery Theory and Hermeneutica Theory. To conclude, the current study aims to discover uniformed, textual, quantified, structured, and visual knowledge patterns from an unstructured text.

1.3 Significance of Study

Human beings evolve from the modern world to the postmodern and digital modern world, so textual big data have scattered on countless websites, books and other sources, but it is a big challenge to extract useful knowledge patterns in the shortest possible time. Thus, the academic challenges of digital natives have also changed, and Voyant like tools have become a necessity to explore knowledge patterns from big data. Furthermore, many modern students lose interest in printed textbooks and resort to digital content as Higher Education Commission has started to promote digital books and digital libraries. To address these challenges, digital text mining tools have become a necessity to connect academia and digital advancements through text analytics/ text mining.

As the newness of the current research is concerned, it is the introductory study in DH in Pakistan. In the world, some corpus-based textbook analyses have been done, but a few research projects have applied digital tools for voyanting (the use of Voyant tools to mine text, study, teach, conduct research) of English literature and ESL textbooks in the world. To fill this gap, this study has been conducted.

Its significance is evident with its manifold academic contributions. Prime stakeholders of this research are students, teachers and the publishing industry. More than 1 million intermediate students can become direct beneficiaries of this research by preparing their subjective and objective papers. Moreover, automated, self-paced, digital and dynamic study concepts have been promoted with the current research. Generating corpus summary, finding any key word, collocation pattern/ n-grams, knowledge graph and KWIC is swift and easy. Learning syntactic patterns in KWIC can become fun; and learning visual vocabulary in the form of word clouds can enhance learners’ interest and motivation. Finding context, searching answers to objective papers, playing with textbooks and learning from digital contents are key advantages of text mining. Digital native teachers and students are better able to get numerous benefits from these knowledge patterns and interactive visuals. To this end, they utilize this digital text in their classrooms as well as in their homes. Students and teachers can employ self-made corpora in the classrooms as they are being used in top ranking universities of the world. Consequently, a new wave of digital learning can be generated by the current study especially during pandemic circumstances.

Pedagogical implications of this study are significant; for instance, Summary tool quantifies stylistic features and corpus features of any literary text. Cirrus tool presents key themes and characters which are used for previewing purposes of a book. Phrases tool extracts collocations that are used for linguistic fluency. Links tool generates knowledge graphs to find the interrelationship of various themes and characters to facilitate hermeneutics. Finding answers to objective papers and compiling material on one topic has become convenient with Contexts tool.

NLP experts and software engineers have used different Python libraries to extract knowledge patterns and social scientists and learners of humanities feel difficulty to use Python libraries, for instance, Pandas, Numly, WordCloud to generate word clouds/ Cirrus; Python NLTK is used to search collocations/ n-grams, contexts and summary of corpus; and Pyplot library is used to generate knowledge graphs. Voyant tools and this study are significant because a humanities background student or teacher can easily extract various aforementioned knowledge patterns with the Voyant suite without any coding or programming skills since all codes are inbuilt in Voyant tools.

Commercial significance in the publishing industry is that publishers would be able to generate countless standard helpful material, exercises for publication of exercise books, sample objective papers, vocabulary lists and collocation/ n-grams or lexical bundles. As interactive visuals enhance learning and memorization skills, the use of data visualization can be enhanced in printed and digital textbooks which are gaining immense popularity. Thereupon, these interactive visuals can increase the readership, popularity and sale of digital books. Above all, it will set new trends of studying textbooks with data visualization and interdisciplinary approaches; therefore, the current study materialises the intellectual vision of polymath who is dexterous in multidisciplinary Renaissance studies (Heydenreich, 2019).

1.4 Scope of the Study

This dissertation consists of text mining of intermediate English textbooks whose 82,487 words corpus is named PIE TCZU (Pakistani Intermediate English Textbook Corpus Zafar Ullah). These textbooks have been mined to facilitate more than 1 million intermediate students from Punjab, Pakistan and thousands of ESL (English as a Second Language) teachers and publishers.

As the expansion of boundaries and extensive implications are concerned, it commences from text mining of English textbooks, but it has extended its scope and implications to the publishing industry, ELT (English Language Teaching) material designing for teaching and testing, forensic linguistics, data visualization, library science, cybersecurity and business intelligence.

1.5 Definitions of Key Terms

Some key terms of this dissertation have been explained in the following lines:

Knowledge Patterns: This study presents existing and idiosyncratic knowledge patterns which open new hermeneutic panoramas because “The unlocked information can lead to new knowledge, improved understanding” and “positive externalities” (McDonald, 2012). “Knowledge patterns are one way to formalize and describe lessons learned and the best practices (i.e., proven experience) about structuring knowledge, the design of Knowledge Management systems, or the development of underlying ontologies” (Rech, Feldmann, & Ras, 2012, p. 578).

Voyant Tools: Voyant tools organize texts as an understandable refined and desired knowledge patterns. As far as research tools are concerned, five Voyant tools: Summary, Cirrus, Phrases, Links, and Contexts have been used in the current study. Text mining means to derive useful knowledge patterns from the selected text and to illustrate them in the form of Cirrus/word cloud, collocations/ n-grams, knowledge graphs (KG), corpus summary and context of key words.

Corpus: Corpus has been defined thus: “Corpus is a collection of texts stored in an electronic database” (Baker, Hardie, & Mcenery, 2006, p. 48). The pedagogic corpus is used for learning and teaching purposes and Hunston (2002) posits that it raises awareness about the educational use of texts.

Summary: Summary tool quantifies all text into total words, unique words, the average length of sentences, vocabulary density and the most repeated terms of the text.

Stylometry: Stylometry/ computational stylistics/ Stylometrics means "to use statistical analyses to investigate stylistic patterns in order to determine (most probable) authorship of (literary) texts: it is concerned, therefore, very much with style as idiolect… Linguistic features commonly examined in stylometry include word length; sentence length; connectives; collocations" (Wales, 2011, p. 402).

Cirrus: Cirrus/ word cloud means a type of light cloud in the sky (Oxford University Press, 2021). It is a multi-coloured rectangular interactive image to display key themes and the statistical weight of the uploaded text.

Phrases: Different terms, for instance, Phrases/ Collocations/ n-grams, bigrams, trigrams, quadgrams, multiword expressions, formulaic language and standard phraseology have been used to describe different aspects of collocation. "The occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening" (Sinclair, 1991, p. 170). Moreover, collocations/ n-grams have been used as an "instrumentation for meaning" (Louw, 2010, p. 79), and they are used to enhance fluency in all language skills.

Links: Links/ KGs are human brain-like nets of numerous neurons with input and output nodes, so that information can flow through them to establish a hermeneutic pattern. KG is like the study of connectivism of characters and ideas to understand learning ecologies in an online environment (Boitshwarelo, 2011). Google knowledge graphs have earned worldwide fame to reveal new insights and domains.

Contexts: Contexts/ Key Word In Context (KWIC) refers to the search of any key word to study its bidirectional (previous and after) context for word sense disambiguation (WSD). The current hermeneutic study mechanism functions under digital humanities (DH) which means an intersection of digital technologies and humanities to unveil knowledge patterns.

Digital Hermeneutics: "Digital hermeneutics (is) understood as the encounter between hermeneutics and digital technology, particularly the Internet" (Capurro, 2010, p. 1). It was a philosophical movement in 2019. Advent of the internet, added the prefix of digital with hermeneutics.

1.6 Background and Definitions of Digital Humanities

Before 2004, the term “humanities computing” was in vogue, but after 2004, first time Susan Schreibman, Ray Siemens and John Unsworth termed it as ‘digital humanities’ (Schreibman, 2013). In 2008, the Office of digital humanities was established in Washington DC (Rockwell, & Sinclair, 2016, p. 74). From its advent, digital humanities has encompassed all domains of knowledge from applied sciences to social sciences.

The technology works in the background of digital humanities to facilitate the process of extracting useful knowledge patterns from big data in the shortest possible time; for instance, Shakespeare’s 37 plays can be read through distant reading techniques or Voyant tools. Furthermore, different data visualizations attract readers’ attention and lead them to a more profound and precise understanding of hermeneutic patterns. The extraction of corpus summary, key themes, collocations/ n-grams, knowledge graphs, and context have been used for learning and teaching purposes.

Various definitions describe different aspects of digital humanities (DH). Viterbo defines DH precisely: “Humanities gone digital and vice versa” (Gibbs, 2016, p. 295). Bobley comprehensively defines DH, “Under the digital humanities rubric, I would include topics like open access to materials, intellectual property rights, tool development, digital libraries, data mining, born-digital preservation, multimedia publication, data visualization, digital reconstruction, the study of the impact of technology on numerous fields, technology for teaching and learning, sustainability models, and many others” (Gibbs, 2016, p. 293). Another definition of digital humanities is presented by John Unsworth who says, “Using computational tools to do the work of the humanities” (Gibbs, 2016, p. 293). Therefore, it involves digital ideation, digital problem assessment, digital tool theorization, tool designing, and implications for societal welfare.

Geoffrey Rockwell, one of the designers of Voyant tools, defines DH as “The use of digital tools and methods in humanities study and dissemination” (Gold, 2012, p. 69). Johanna Drucker, one of the renowned scholars of DH, opines that DH “is the study of ways of thinking differently about how we know what we know” (Drucker, 2009). Combining both definitions, the use of digital tools unveils new scholarly knowledge patterns by merging different disciplines. There is a strong nexus among text analytics, computational linguistics, natural language processing (NLP), data visualization (DV), machine learning (ML), data mining (DM), text mining (TM), statistics and probability (a measure of likelihood) in digital humanities (DH).

Isaac Chuang of the Los Alamos National Laboratory, Neil Gershenfeld of (MIT, and Mark Kubinec of the University of California produced the first quantum computer having 2-qubit in 1998 (Holton, Coffeen, 2021). In the contemporary era, computers have evolved into quantum computers with capabilities of massively faster speed than normal computers to solve complex problems. There is a shift from binary principle to quantum bits (quibits), hence, quantum computers are 10 million times faster than binary computers. Quantum computers are four times faster than the speed of light. Solutions are quite precise and their accuracy is much better than normal computers. All processes would be cheaper than supercomputers. These quantum computers will transform digital humanities into quantum humanities (Barzen, & Leymann, 2020). Quantum applications will enable us to analyse and extract precise highly valuable knowledge patterns from big databases instantly.

1.7 Background and Definition of Text Mining

Text mining or text analytics is the branch of digital humanities. Text analytics is a modern term and text mining is an old term. Text mining (TM) derives high-quality information and interesting knowledge patterns through statistics. Text mining functions are text categorization, text clustering, ontologies, entity extraction, named entity recognition, taxonomy, concept extraction, document summarization, and entity related modelling. Text mining is defined as “the use of large online text collections to discover new facts and trends about the world itself” (Hearst, 1999). To conclude, TM aims to explore innovative as well as hidden facts, trends, knowledge patterns and data visualization from big textual data.

Being a subfield of digital humanities, text analytics deals with major linguistic issues regarding word; in a particular word cloud, word counts, word collocations/ n-grams, word links, total and unique words, vocabulary density, the most occurring words and word contexts. To accomplish these tasks, five Voyant tools have been applied to extract interactive knowledge patterns associated with the text.

Like an evolution, human text analysis shifts to text mining and educational data mining. Text is defined as a coherent written segment that can be used for critical analysis. A textbook is a standard, essential and frequently read material for understanding any subject at a particular level. As new textbooks have been introduced, their analyses have also pointed out their deficiencies and solutions. Initially, human textbook analyses commenced in the 1980s (Williams, 1983), and evaluators have explored the appropriateness and utility of textbooks for the specific age group, cultural harmony, gender, grammar competence and teaching of language skills. Such textbook analysis studies have also been conducted in Pakistani universities. Later, technology has been integrated into all spheres of life; human text analysis has also shifted to computer-aided text analysis or text mining. Digital text analytics or data mining has been applied to analyse texts with digital tools. Eventually, in Educational Data Mining (EDM), learning material and textbooks have been analysed with digital tools. EDM mines text with agile hermeneutics (software and tools to explore interactive knowledge patterns) (Calders, & Pechenizkiy, 2012). Further delimiting, the current text mining study mines intermediate English textbooks with the use of five Voyant text mining tools. The data visualization has been interpreted with the triangulation of Knowledge Discovery Theory (KDD) and Hermeneutica Theory.

1.7.1 Previous Text Mining Projects

Text mining projects have been run by developed university faculty, students and publishers who have large databases to analyse texts. First text mining centre in the world is The National Centre for Text Mining (NaCTeM) managed by the University of Manchester, UK. It focuses on texts from social sciences and biomedical (The National Centre for Text Mining, 2016).

The University of Alberta, Canada catalogues applications of text analysis through The Text Analysis Portal for Research named TAPoR (TAPoR, McMaster University, University of Alberta, 2015). University of California, Berkeley is working on text analytics processes in Biology through BioText (The BioText Project, 2016). Tufts University in Medford and Somerville, Massachusetts, USA introduces Visual Understanding Environment (VUE) to present visual tools for research. This project develops Zotero, a tool for managing bibliography and SEASR (Software Environment for the Advancement of Scholarly Research), a digital analysis tool (Nissenbaum, 2010). Following the same projects, Language Engineering Department has been established at UET, Lahore, Pakistan to design various software for Pakistani languages.

In the beginning, different types of data have been mined in business, biology, medical and other fields, but Educational Data Mining (EDM) emerged in 1995, and it got a boost in 2004. Next year, WebCAT company introduces the term "academic analytics" for EDM, and later "action analytics" appears in academia (Norris, Baer, Leonard, Pugliese, & Lefrere, 2008).

1.7.2 Text Analytics and Big Data

Presently, different researchers come across big databases of yottabytes (10^{24}) because of information explosion; so, to derive meaningful patterns from huge data, digital data mining techniques and speedy algorithms (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996) have been applied. Data mining verifies existing models and builds predictive models to fulfil prospects. Moreover, data mining extracts corpus summary, Cirrus, collocations/ n-grams, knowledge graphs and KWIC (Key Word In Context) from the input text. Therefore, text mining refers to the transformation of raw data or unstructured data into a structured, quantified, meaningful and interactive data visualization (Castro, Vellido, Nebot, & Mugica, 2007; Romero, & Ventura, 2007) knowledge discovery, hermeneutics and knowledge patterns which can be the prime objectives of reading any text.

1.7.3 Numeric and Visual Data in Text Analytics

Comparing numeric and visual data, the former is static and less appealing, whereas the latter is interactive, comprehensible and fascinating for learners. One step ahead, interactive visuals are more precise, memorable and comprehensive than mere words (Two Crows Corporation, 1999). Interactive visual data can be modified according to the quantity and need of research, but numeric data cannot be modified in the run time. The interactive quality of data visualization transforms into static images due to print on pages; however, this interactivity can be maintained only in soft files.

1.7.4 Text Mining and Textbooks

English textbooks derive the least advantages from DH; and Huntston "admit(s) that current textbooks are not corpus-based" (Huntston, 2010, p. 181) and many educationists are unaware that "how exactly the corpus is used to flesh out the linguistic contents of their textbooks" (Huntston, 2010, p. 181). This is the predicament of advanced countries, whereas in developing countries like Pakistan, text mining of English textbooks is the least explored research domain; hence, the present study fills this much-needed gap in text analytics. Use of the only corpus to explore text is an old method, but text analytics along with corpus and multiple data visualization is the latest and more valuable method. Consequently, the current study discovers knowledge patterns with interactive visuals as a summary of the corpus, Cirrus, collocation patterns/ n-grams, KG and KWIC.

1.8 Objectives of the Present Research

The current study aims to accomplish the following research objectives in the text mining of intermediate English textbooks:

i. To produce a summary of text mining to extricate the quantified information about stylometry, vocabulary density, the average length of sentences and the most frequent words in the corpus.

ii. To generate Cirrus/ word clouds to unveil the prominent motifs and characters.

iii. To point out collocation patterns/ n-grams to extract the most frequent standard phraseology.

iv. To create knowledge graphs to explain the interconnectivity of various themes and characters for digital hermeneutics.

v. To explore the bidirectional context of ambiguous words to comprehend the contextual word sense.

1.9Research Questions

The current study addresses the following research questions for each unit of Intermediate English textbooks:

i. How does text mining summary discover stylometric features from intermediate English textbooks?

ii. How does an interactive word cloud/ Cirrus reveal major themes and characters from intermediate English textbooks?

iii. What types of collocation patterns/ n-grams have been unveiled to extract the standard phraseology with its parts of speech?

iv. How do knowledge graphs present the interrelationship of various key themes and characters for digital hermeneutics?

v. How does the context of certain problematized words disambiguate the word sense by showing interactive bidirectional context?

1.10 Major and Sub Arguments

Primarily, this dissertation argues that distant reading can discover interactive knowledge patterns in the form of a corpus summary, word cloud, collocations/ n-grams, knowledge graphs, and KWIC through Voyant text mining tools. It discusses several sub-arguments: Cirrus/ word cloud reveals major themes and characters; Phrases tool exposes collocation patterns/ n-grams based on "rhetorical power of repetition" (Rockwell, & Sinclair, 2016, p. 106) and what "the author wants to emphasize" (Rockwell, & Sinclair, 2016, p. 106); Links tool joins different themes and characters in Knowledge Graphs (KGs); Summary tool analyses all features of the corpus, stylistic features of writers and their works in a quantified manner; and Contexts tool clarifies grammatical and semantic ambiguities of any problematised word with interactive bidirectional context.

1.10.1 Rationale for Undertaking the Research

Usually, students of intermediate are teenagers and they are much addicted to technological gadgets; consequently, they spend much time with technology. Diverging their more interest to academia, it is quite rational to integrate academic requirements into digital tools. The current study implicates to enhance Voyant tool learners' distant reading skill, digital scholarship, autonomous learning and exam results by extracting various knowledge patterns. Distant reading through Voyant tools has become fast, easy and fascinating. The usage of Voyant tools sharpens digital scholarship because they can easily explore knowledge patterns from big data. Furthermore, Voyant tools give impetus to learn autonomously, since various students have different IQs, learning styles and reading tastes. Moreover, Contexts tool facilitates finding precise answers to objective papers from all textbooks and setting sample objective papers for themselves. They can also compile relevant information for subjective type questions. Furthermore, learners can extract collocations/ n-grams to improve their fluency and accuracy in language. Consequently, these skills facilitate the process of knowledge discovery and hermeneutic skills.

Least research on text mining in Pakistan motivates me to initiate this research. Presently, there is a dearth of text analytics works with digital humanities tools, though some previous corpus linguistics M. Phil level research works on FIR's, federal secretariat language and newspapers have been conducted at NUML, Islamabad, Pakistan.

1.11 Theoretical Framework, Research Data and Voyant Tools

The current study employs mixed methods, an amalgamation of qualitative and quantitative methods in data collection and data analysis, to unveil interactive knowledge patterns. It also briefly discusses the theoretical framework, research data and Voyant "knowledge bearing" digital tools (Rockwell, & Sinclair, 2016, p. 10). Before discussing Knowledge Discovery Theory, the concept of wisdom should be clarified. Wisdom does not compile information chunks; rather, it is "the thinking that led you to assertions" (Rockwell, & Sinclair, 2016, p. 135), and it derives principles of knowledge patterns.

1.11.1 Theoretical Framework as Triangulation of KDD and Hermeneutica Theory

Designing of all digital tools is based on some theories and text mining tools are manifestations of DH theories. Triangulation of Knowledge Discovery and Data Mining (KDD) and Hermeneutica Theory builds a fundamental relationship to reveal some common features; hence, both of them have been applied to current text mining study to explore "knowledge patterns" (Cristobal, & Sebastian, 2013) with visuals of Summary, Cirrus, Phrases, Links, and Contexts. Both theories have been explained precisely in the following paragraphs.

Rakesh Agrawal, a renowned computer scientist at Microsoft, theorized knowledge discovery theory and later, it has been expanded. "In active data mining paradigm,… rules are discovered, …the history of the statistical parameters associated with the rules is updated… we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct to retrieve rules whose histories exhibit the desired trends" (Agrawal, & Psaila, 1995, p. 1). Summarization involves methods for finding a compact description for a subset of data (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). Knowledge Discovery Theory is also defined as "the extraction of implicit, previously unknown and potentially useful information from data" (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). It transforms random or unstructured data into structured and valuable information interactively. Knowledge discovery process performs six steps shown in figure 2. i. Data selection; ii. Pre-processed data; iii. Digitization of data; iv. Data mining; v. Interpretation and evaluation; vi. Knowledge discovery through patterns.

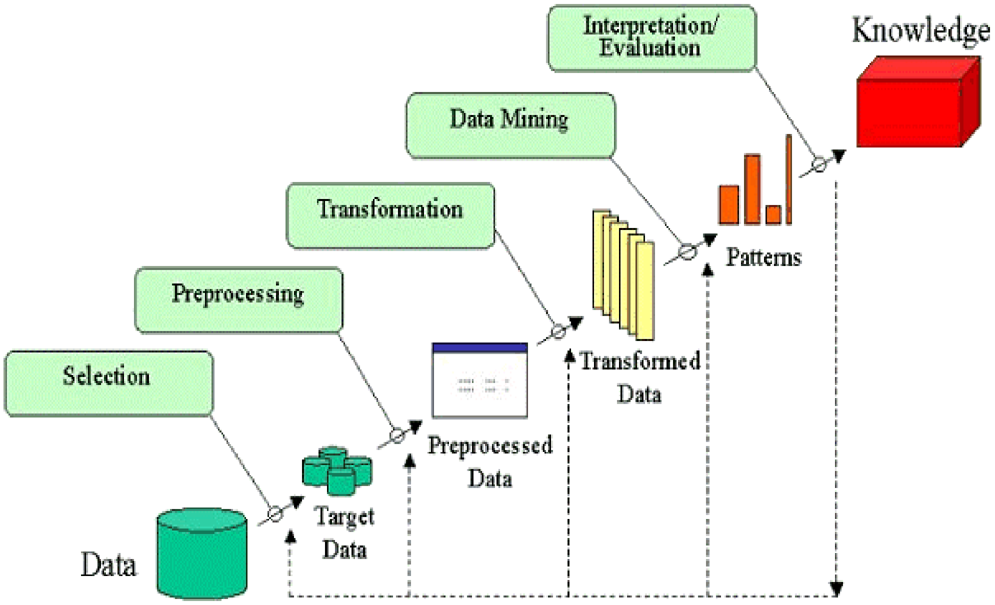


Figure 2 Knowledge discovery (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996).

Renowned digital humanities scholars Stefan Sinclair and Geoffrey Rockwell have designed and theorised Voyant tools. They have propounded Hermeneutica Theory (Rockwell, & Sinclair, 2016, p. 166) for hermeneutic interpretation of data visualization generated by Voyant tools. They explicated its following features:

i. “Hermeneutica Theory is embedded in a context.”

ii. “It is not like black boxes.” (In the domain of computer programming, it did not examine the actual background programme which was executed.)

iii. “Manipulation is in service of exploration and understanding.”

iv. “It is supplemented by other materials.”

v. “Knowledge bearing tools provoke reflection.”

vi. “Hermeneutic tools fail in interesting ways.”

vii. “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166).

1.11.2 Triangulated Theoretical Framework and Interdisciplinarity

Triangulated theoretical framework interrelates various fields of computer, computational linguistics, ML, statistics, NLP, AI, text mining and digital humanities. It is commonly called data mining which is grounded in databases, statistics and ML (Fayyad, Shapiro, Smyth, & Uthurusamy, 1996). This theory has focused on rules of association, characteristics, classification, clustering, linked data, named entity recognition and prediction. It employs techniques and some other theories: evidence theory, cloud model of mathematics, fuzzy sets, rough sets, neural network, genetic algorithm, SOLAM (data mining with online processes), visualization, exploratory learning and spatial inductive learning (Li, & Wang, 2005).

1.11.3 Research Data

Research data comprise of intermediate English textbooks which cover six genres of literature, for instance, short stories, poems, one-act plays, essays, biographical essays and one novel.

1.11.4 Introduction of Voyant Tools

New digital tools are capable of deriving knowledge patterns from both humanities and sciences. These tools deliver academic benefits of autonomous learning, collaborative teaching, research and publishing. They promote more dynamic and digital learning and teaching methods than traditional printed texts (Burdick, 2016). This study manifests different knowledge patterns with five Voyant text mining tools.

Voyant tools facilitate the process of text mining in order to reveal insightful interactive knowledge patterns which are prime objectives of the current study. They were designed by Sinclair and Rockwell (2015b) in Canada in 2003 and further upgraded in 2013. Principally, they introduce simplified text analytics tools for learning, teaching, research and publishing purposes. They are user-friendly even for technophobes, since no coding or computer programming is required to execute them. A user from humanities can just upload text in any format; and just with one click, it starts diversified text mining processes on five panels simultaneously.

Voyant tools consist of 25 tools, and among them, there are 17 corpus tools, nine document tools, 14 visual tools, eight grid tools, and three other tools. Present research concentrates on the functionality of only five tools: Summary, Cirrus, Phrases, Links and Contexts. Summary tools reveals total words, unique words, vocabulary density, the average length of sentences, the most frequent words and these features are required for stylomtric studies; Cirrus tool exhibits colourful and multi-sized word cloud with statistical data; Phrases tool presents tabular data of collocations/ n-grams in terms of length and occurrence, and they can be arranged in an ascending or descending order; Links tool develops KGs to study hermeneutics and interconnectivity of different themes and characters; and Contexts tool disambiguates word sense by showing interactive bidirectional context for example word “miss” whether it refers to a lady or to think about somebody passionately. Thus, Contexts tool clarifies ambiguous words and homographs. Another valuable quality of Voyant data is its interactivity and ubiquitous nature, enabling online access by holding the CTRL button and clicking on the most frequent words in Summary data. Moreover, they are easily exportable in the form of link and PNG image.

1.11.5 Delimitations of Research Data

In Pakistan, several different English textbooks are taught at different levels and educational boards. The current study delimits intermediate English textbooks taught and examined in nine Punjab Boards (Rawalpindi, Sargodha, Faisalabad, D.G. Khan, Lahore, Gujranwala, Sahiwal, Multan, Bahawalpur) and Punjab Board of Technical Education, Lahore. Five educational boards have mentioned the total number of their students: BISE Multan 74,491 students, BISE Bahawalpur (57, 339), BISE Faisalabad 97,528 BISE Sargodha 53,246 and Gujranwala 141, 726 students. Overall, the number of intermediate students in five boards is 4224, 330 (The Dawn, 2019). Suppose students of the other four BISE boards and technical education boards are also counted. In that case, there are approximately one million intermediate students who will be beneficiaries of this study. These students are not direct participants of this study, since it is not a user study.

Only five tools: Summary, Cirrus, Phrases, Links and Contexts have been delimited for voyanting (the use of Voyant tools to mine text, study, teach, conduct research). Summary tool is used for stylometry (Ullah, & Mahmood, 2019, pp. 1-17); Cirrus tool is used for word clouds; Phrases tool is functional for collocation patterns/ n-grams; Links tool is employed for the extraction of KGs; and Contexts tool is applied for searching KWIC to disambiguate word sense. Further delimitation of data of these tools is the specification of 25 themes in Cirrus, but in the novel, 95-word Cirrus is made due to its length. First, 15 most occurring collocations/ n-grams, ten most occurring words in the Summary panel were delimited. The five most occurring words in the poems have been delimited in the Summary panel because of their brevity. The first 15 KWIC have been displayed for further discussion

1.12 Structure of the Dissertation

The first chapter, Introduction, gives an overall view of the entire dissertation. Chapter two, Literature Review, reviews prominent previous research projects chronologically with a paradigm funnel approach. Moreover, previous data mining theories, their association with this study, and triangulation of Knowledge Discovery Theory and Hermeneutica Theory have been discussed. Chapter three, Research Methodology, sheds light on the complete procedure of data requirement, data generation and data analysis along with their rationales. Chapter four, Data Analysis, analyses the generated corpus of each lesson from intermediate English textbooks with five Voyant tools. Chapter five named Conclusion expresses major findings of the current study, major contributions, fulfilment of objectives and the future recommendations.

1.13 Conclusion

This chapter has introduced foundation ideas regarding definitions, background and key arguments of the current research and it chalks out the layout of the entire research. Moreover, research questions and objectives lead to the required documents, required data, and interpretation to unveil interactive knowledge patterns. Five Voyant text mining tools have been employed to mine intermediate English textbooks in the current dissertation. These tools generate data visualization and knowledge patterns from textual data for better and fast comprehension of the text.

The following chapter covers previous literature on and around the dissertation topic.

CHAPTER 2

LITERATURE REVIEW

2.1 Prologue

This chapter examined previous literature on and around key words of dissertation title, for instance, a broad spectrum of digital humanities (DH), data mining, educational data mining, text mining, corpus-based and corpus-driven analyses of EFL textbooks, data visualization, knowledge patterns, and Voyant tools for text mining along with previous research works on Summary (of the corpus, stylometry), Cirrus (word cloud), Phrases (collocation patterns/ n-grams), Links (knowledge graph), and KWIC (Key Words In Context). This section commenced with the chronological progression with Paradigm Funnel Approach (figure 3), showing the bigger picture of DH and then delimiting it to Voyant text mining research projects and tools for the study of intermediate English textbooks. Paradigm Funnel Approach had been explained thus: "the balls in the funnel are categories of works that are relevant to your investigation, but do not specifically address what you are doing. They will have more to do with your topic than with your thesis statement. They tend to contain a lot of works per category, which makes sense …" (Hofstee, 2006, p. 96). The figure 4 flow chart followed chronological development and Paradigm Funnel approach which was "used to structure an insightful literature review and to generate enlightened research thinking. It is especially useful for doctoral students and other researchers faced with a heterogeneous body of literature" (Berthon, Nairn, & Money, 2003).

Figure 2-1. The funnel method of structuring a literature review

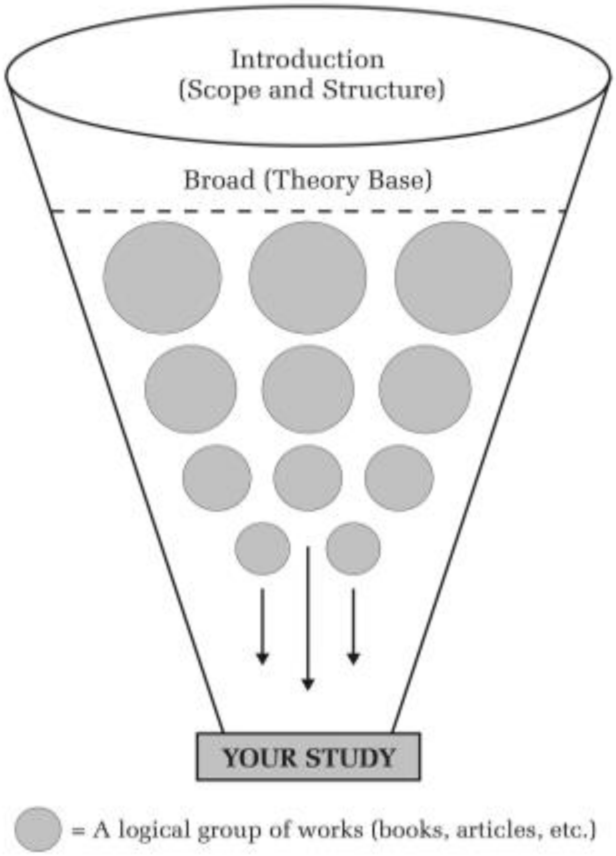


Figure 3 Paradigm Funnel Approach (Hofstee, 2006)

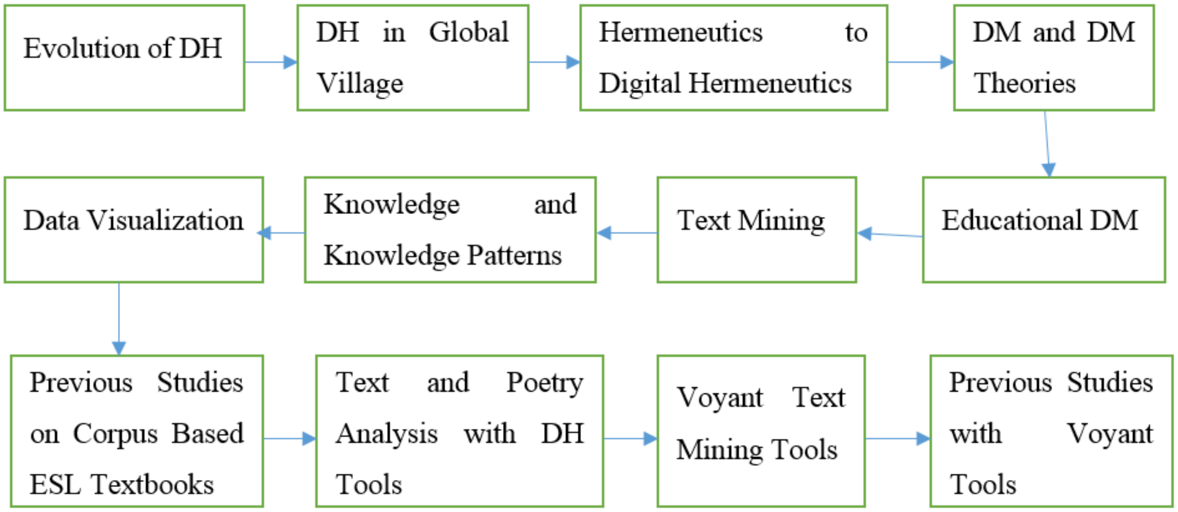

Evolution of DH
DH in Global Village
Hermeneutics to Digital Hermeneutics
DM and DM Theories
Educational DM
Text Mining
Knowledge and Knowledge Patterns
Data Visualization
Previous Studies on Corpus Based ESL Textbooks
Text and Poetry Analysis with DH Tools
Voyant Text Mining Tools
Previous Studies with Voyant Tools

Figure 4 Flowchart of Literature Review Chapter

2.2 Journey from Humanities to Digital Humanities

Humanities covered subjects of social sciences, for instance, literature, art, philosophy, logic. Humanities added the adjective of “digital” when computers were used in the study of humanities. Empiricism was the hallmark of applied sciences, but with the boost of technology, humanities also changed from theoretical to empirical studies with an aid of technology, logic and scientific tools. That is why humanities subjects were also called scientific studies such as linguistics, the scientific study of language.

What Immanuel Kant considered as philosophy in his book named *‘The Conflict of the Faculties’ (1798)*; in later ages, it was named as humanities (Readings, 1996), digital literacy and information literacy. The digital age, digital intellect and the use of technology modified the concept of knowledge, wisdom and intelligence. Subsequently, knowledge was changed into information in the 21st century. Following this influence, several social science subjects were digitized, termed computational social sciences or e-social sciences. Results of this change led to large scale data collection in social sciences and their analyses with precision and dexterity to prove these subjects scientific (Lazer, Pentland, Adamic, Aral, Barabasi, Brewer, & Jebara, 2009). To conclude, the use of technological tools in data collection and data analysis transformed humanities into sciences.

Modern human scientific knowledge had its traces in the Renaissance Age (1500 to 1650) in which classical knowledge was revived. Furthermore, various translations of Greek, Latin, Arabic and French played their vital roles to disseminate knowledge across boundaries. Simultaneously, humanism started to nurture; and human beings became the centre of attention for all subjects. Later on, the use of the printing press standardized the English language and other languages. Cheap printed books, low rates of postal services, colonialism expedited and disseminated knowledge of humanities all over the world.

English colonialism also played a significant role in circulating fruits of the renaissance to the entire world. Therefore, there was a shift from humanism to different other disciplines; hence, interdisciplinary, cross-disciplinary and transdisciplinary studies emerged in the firmament of academia. Different subjects were split on the name of specialization and super specialization to excel extensively. Post circumstances of WWII also promoted new linguistic departments and study centres in different universities of the world. DH reshaped traditional knowledge with new digital tools for a better understanding of knowledge. Under the spell of Marxism, growing industrialism changed academia into industry. Consequently, the education industry got a boom with computational linguistics, text mining and DH. Above all, the internet and Wi-Fi in computers and mobile phones revolutionized the concept of DH in the entire global world. As a result, digital humanities got deeper roots in several academic hubs, libraries and universities.

2.3 Evolutionary Phases in Digital Developments

With the flow of time, technology started to influence almost all spheres of life. In the beginning, its advancements were slow, but later it initiated an information explosion with new gadgets, tools and software. The chronological developments in the digital age had been delineated in the following time intervals.

2.3.1 Era from 1881 to 1930

In 1881, Hollerith introduced a punch card system and for the first time, they were used in the 1890 US census to compute statistical data. In 1923, Dr Arthur Scherbius invented the digital codifying machine named Enigma, which Germans widely used in WWII. In 1928, IBM introduced a rectangular punch card (Digital Preservation Management, 2013). To summarise, DH started with these punch cards to represent data with holes.

2.3.2 Era from 1931 to 1940

As the pioneer of DH is concerned, during the 1930s and 1940s, Josephine Louise Miles (June 11, 1911 – May 12, 1985) conducted two distant reading works on the use of adjectives by Romantic poets and the phrasal forms of poetry during eras of 1640s, 1740s, and 1840s (Buurma, & Heffernan, 2018).

In 1938, the first-time word “digital” was used for computerized data and next year, “Bomba” digital decoder machine was used by Allied forces to decode German forces’ WWII messages (Digital Preservation Management, 2013). Thus, text mining and forensic linguistics appeared without proper terms.

2.3.3 Era from 1941 to 1950

The relationship of man and technology had been established in this quote, “Man was child of God and technology was child of man, I think that God regards technology the way a grandfather regards his grandchild" (Busa, 1980, pp. 87-88). From the 1940s, an amalgamation of humanities and digitization commenced. In the beginning, social scientists were reluctant to use technology trustfully; hence, they used punched cards and handwritten cards which were very troublesome to carry when Italian Father Roberto Busa (1913-2011) loaded his work in trucks to Italy, and the shifted data were preserved on magnetic tapes. He prepared a handwritten and typed concordance of 'Thomas *Aquinas'* comprehensive digital work namely *'Index Thomisticus'*. In 25 years, he punched and processed 2200 words in the manual text mining process. Even in this manual process, three major drawbacks were persisting; hence, his project was shifted to IBM electronic computers (Busa, 1980). With the collaboration of IBM, Father Roberto Busa started his computer-aided text analysis of *'Index Thomisticus'* with punch cards, and finally compiled 800000 cards with concordance data. He also established a school of keypunch operators, and these experts were in high demand in the industry at that time. First-generation tool preparation was based on Busa's tools for counting and concordance in the 1960s and 1970s (Russell, 1965). To conclude, DH was promoted in Europe by Italian Father Roberto Busa's work with the collaboration of IBM in 1949, and it took the next thirty years to complete. He prepared the index of 10,631,980 words after the labour of 34 years. Its first volume was published in 1974.

In 1945, the first electronic computer named ENIAC was introduced and in the same year, Grace Hopper designed a Bug computer (Digital Preservation Management, 2013). In 1950, the US Federal Records Act incorporated machine-readable records as its official record. Thus, all digital materials got the status of a legal record; for this reason, this act motivated others to transform hard data into machine-readable and portable data (Digital Preservation Management, 2013) to give eternal life to the data.

In the 1950s, Josephine Louise Miles (June 11, 1911 – May 12, 1985), the pioneer of DH, became the project director for building a Concordance of John Dryden's poetic works. This project was based on index cards, but Miles took technical assistance from the Electrical Engineering department at the University of California at Berkeley to accomplish the concordance project, and for this end, she used punched cards and card-reading computers. After six years of hard work, her first text mining work was published. Rachel Sagner Buurma and Laura Heffernan declared it as "the first literary concordance to use machine methods" (Buurma, & Heffernan, 2018). Josephine Louise Miles's concordance was published seventeen years before the first volume of Roberto Busa's '*Index Thomisticus'*, which was wrongly considered the first digital humanities work (Buurma, & Heffernan, 2018).

2.3.4 Era from 1951 to 1960

In 1951, UNIVAC I, the first commercial computer, was launched for common people. In 1952, Grace Hopper introduced Compiler to program languages the first time, and IBM introduced the first scientific computer named IBM 701, and in the same year, the Norwegian Computing Center for the Humanities (now HIT) was founded. In 1955, IBM presented the first disk drive named RAMAC. Realizing utility and significance of technology, USA defence department adopted technology to train its staff to meet emerging cyber challenges of the world and to accomplish this aim, it introduced Advanced Research Projects Agency (ARPA) to defend virtual boundaries of USA (Digital Preservation Management, 2013).

2.3.5 Era from 1961 to 1970

Information technology evolved into data mining, and later data collection and development of databases started in the 1960s with the old filing system. Then Alvar Ellegard used machines for vocabulary calculations (Ellegard, 1962). In the same year, Parrish (1962) worked on the concordance of Matthew Arnold's poems in the USA, and next year, Wisbey (1963) prepared indexes for Early Middle High German texts in the UK. In those days, lexicons were being compiled with computers and corpus linguistics. In 1962, University of Michigan prepared first data archive (Digital Preservation Management, 2013), and in 1963, Andrew Morton, a Scottish clergyman, wrote an article in the newspaper, and claimed that St. Paul composed only four epistles, and his claim was based on computerized data analysis (Morton, 1965). Besides, there was a controversy over bona fide authorship of twelve papers and at that crucial time, DH solved the issue and revealed that Madison was the true author of those controversial papers (Mosteller, & Wallace, 1964). In conclusion, author identification can be done with text mining.

In 1963, the Centre for Literary and Linguistic Computing was established in Cambridge, and the next year, IBM patronized the Literary Data Processing Conference at Yorktown Heights to share problems and solutions (Schreibman, 2013). These initial conferences groomed the discipline of DH and created awareness in academic circles. Next year, IBM's Cambridge Research Lab started its function as the first virtual machine time-sharing system, and in 1965, DIGITAL's PDP-8, minicomputers were introduced. In 1968, USA libraries introduced digital catalogues (Digital Preservation Management, 2013) to locate any book in multiple ways swiftly and accurately. Thus, DH became popular in library science; hence, texts and archival data were digitized.

In 1966, the first journal '*Computers and the Humanities'* was introduced by Joseph Raben who organized several research conferences and strengthened this field. Under the editorship of Stephen Waite, a newsletter about an amalgamation of classics and computers was published (Schreibman, 2013). In 1966, the TUSTEP programme was launched on text data at the University of Tübingen Wilhelm Ott in Germany.

When large data were stored, information retrieval was an initial step for a minute perusal. So, COCOA concordance program was launched (Russell, 1967), and it counted words and produced concordance of any type of text. In 1969, Generalized Markup Language (GML) was developed (Digital Preservation Management, 2013) to provide supplementary information regarding the text.

Michael Farringdon and Roy Wisbey founded the Association for Literary and Linguistic Computing/ Association for Computers and the Humanities (ALLC/ACH). It organized several conferences in Cambridge University in 1970 to publish standard research works and tools for stylistic analysis, lexicography, teaching and textual analysis (Schreibman, 2013). As a result, text mining of literary text started, and it also contributed to linguistics and lexicography.

2.3.6 Era from 1971 to 1980

In 1970's, Oxford took corpus projects and corpora were built and all paper files were changed into machine readable data. In 1971, Project Gutenberg started to compile literary written works. In the same year, the first email system ARPANET, File Transfer Protocol (FTP), UNIX and floppy disk were introduced. In 1972, C and FORTRAN 66, computer programming languages were introduced, and in the same year, smaller chips and processors were invented by Intel (Digital Preservation Management, 2013).

Then in 1973, the Association for Literary and Linguistic Computing (ALLC) came into being by European digital humanists to promote DH in literary circles. In 1974, Structured Query Language (SQL) was created for programming. In 1975, Kurzweil Reading Machine was invented for blind people to convert text into speech; and in the same year, servers were introduced to monitor other networked systems. Another programming language named BASIC was introduced in 1975. In 1976, Queen Elizabeth II became the first political figure who sent an email (Digital Preservation Management, 2013). During the same year, Oxford Text Archive (OTA) was prepared by Lou Burnard, and it was the stepping stone for the establishment of digital libraries. To generate Old English Dictionary, Old English Corpus was also built (Healey, 1989), therefore, lexicography adopted digital ways for upgradation. Besides, the use of computers for linguistic analysis got fame in the world. Text Encoding Initial (TEI), hypertext, nonlinear text and Extensible Markup Language (XML) were introduced. Even now, no DH course is complete without training in TEI.

To facilitate the growth of computing in humanities, several other funding agencies stepped in this field in the USA. In 1978, The Oxford University Computing Service took COCOA (Corpus of Contemporary American English) to build the Oxford Concordance Program (OCP). In the same year, Dallas Public Library, USA became the pioneer for online library cataloguing. In 1980, FORTRAN 77, a programming language, was developed (Digital Preservation Management, 2013).

In 1980's, an introduction of microcomputers led to text analyser tools for individuals and OCP was improved and changed into Micro-OCP. The new text analysis programme Brigham Young Concordance (BYC) program was launched, and later, it changed into WordCruncher. Individuals used them and devised new changes; hence, an analysis of electronic text was started, and the trend was shifted from printed material to text mining and web mining.

From 1970's to 1980's, network and relational database systems were introduced. Personal computers were used for research, learning, teaching, data generation, data storage and data analysis. Computers and the Humanities, International Conference on Computing in the Humanities (ICCH) and The Association for Computers and the Humanities (ACH) organized several conferences and consolidated beneficial impacts (Hockey, & Marriott, 1979a, 1979b, 1979c, 1980) on DH, data mining and text analytics.

2.3.7 Era from 1981 to 1990

From 1980’s to onwards, digital humanities started to expand worldwide, so database tools served as a foundation for DH. In 1981, a compilation of academic websites named BITNET was introduced (Digital Preservation Management, 2013), and during the same year, an operating system, MSDOS 1.0 was introduced. In 1982, Compact Disk-Digital Audio (CD-DA) and CD player were prepared by Philips and Sony companies (Digital Preservation Management, 2013) to expand digital audio video content. In the same year, Oxford Concordance Program (OCP) was launched, and it got fame worldwide (Hockey, & Marriott, 1979a, 1979b, 1979c, 1980).

In 1985, Microsoft Windows 1.0 was introduced for the first time, and the next year, Standard Generalized Markup Language (SGML) was introduced. In the same year, America established The National Centre for Supercomputing Applications (NCSA) (Digital Preservation Management, 2013). In 1985, Perseus Project was launched at Harvard University, USA to convert traditional literature into digitized text; therefore, various corpora were built for DH. Moreover, Women Writers Project transformed paperbound writings of famous writers into digital text for cross-reference studies; consequently, it promoted digital culture and multidisciplinary studies. DH focused on instance analysis, curation, editing, and modelling projects that interpreted text through computerized quantification (McCarty, 2005).

In 1986, *‘Literary and Linguistic Computing’* (LLC) journal (Kosmos, 2014); and in 1989, *Science Citation Index*® journal were published (Digital Preservation Management, 2013). They promoted digital literacies, and during 1990’s, various projects of DH introduced data visualizations and network studies. DynaText SGML, a publishing tool, was employed in the booking system in 1990. In the 1990's, the use of the internet and websites spread, and this trend revolutionised the world and digital humanities. Moreover, graphic designing, an integration of audio, video and many other features produced new digital content. Furthermore, DH and classrooms have been connected to explore libraries, archives and many other resources as scientists used labs for their scientific studies. DH linked many universities, institutions, students and teachers in virtual universities and distance learning programmes for collaborative learning and teaching in this global village.

Before 1980's, different types of data were analysed under the Statistical Database. In the late 1980's, data mining got the status of a distinct discipline, and it started to expand in biology, medicine and business. From mid-1980, advanced database systems and advanced data analysis were started. In advance, database systems, web databases, text database systems, complex data management, and cloud computing were developed. In advanced data analysis, data warehouse, data mining, knowledge discovery, association, pattern discovery, trends and deviation analysis were done. Moreover, the distribution graph showed the location of the text. Till 1980's, three DOS-based text mining computer programmes, WordCruncher, TAC and Micro-OCP, were introduced (Schreibman, 2013) for text mining. Apart from it, Humanities Computing Yearbook (HCY) was also published (Burnard, 1988) to introduce different projects.

2.3.8 Era from 1991 to 2000

Websites became common as a means of information in 1990's, and new broadband graphics became popular. Consequently, most of the domains of life, goods or services were converted into web appearance. In 1991, Hyper Text Transfer Protocol (HTTP) was introduced, and in 1992, Veronica, a search engine, was introduced in Nevada, USA. In 1994, the Library of Congress was digitized (Digital Preservation Management, 2013) to enhance global readership. Following the same tendency, American university libraries started to keep digital text for academic purposes (Price-Wilkin, 1994). In the same year, TEI textual format was introduced for embedding various textual features.

In 1995, Virtual Reality Modeling Language (VRML) 1.0, Internet Explorer 2.0 and D-Lib Magazine were introduced, and in the next year, World Intellectual Property Organization (WIPO), PNG 1.0 image format and Preserving Access to Digital Information (PADI) were established (Digital Preservation Management, 2013). These developments paved the way for data curation and image mining.

The Universities of Alberta and Guelph compiled History of British Women's Writing, their works and biographies in SGML documents (Brown, Fisher, Clements, Binhammer, Butler, Carter, Grundy, & Hockey, 1997) in order to promote internationalism in research and academia through websites and SGML. In 1998, Roberto Busa got his lifetime achievement award and the prize was named after him. In 1998, XML was coined and Microsoft Windows 98 was developed. In 1999, Google, Bluetooth and HTTP 1.1 were introduced, and Google brought a global revolution in search, search engines and research. Next year, in 2000, PubMed, BioMedCentral, National Archives of Australia, Nordic Web Archive and several other digital archives were built (Digital Preservation Management, 2013), and these digital archives were mined with data mining tools.

2.3.9 Era from 2001 to 2010

In 2000, President Pervez Musharraf digitized Pakistani state documents to become a part of the digitized modern world. In 2001, The Austrian On-Line Archive (AOLA) and The Digital Preservation Coalition, UK were established, and in the next year, Trusted Digital Repositories and Universal Serial Bus 2.0 (USB) were introduced. In 2003, UNESCO issued guidelines for digital heritage, and in the following year, Google started to work in collaboration with Harvard University and Carnegie Mellon University, USA. Furthermore, Digital Curation Centre (DCC), UK started to preserve significant documents (Digital Preservation Management, 2013).

In 2005, National Digital Newspaper Program (NDNP) was established, and in the next year, the foundation of Digital Preservation Europe was laid down (Digital Preservation Management, 2013). To expand research in the domain of DH, a funded programme namely The Digital Humanities Initiative was started in the USA in 2006. In 2009, The World Digital Library was established (Digital Preservation Management, 2013), and in the following year, Google N-grams were shown online for researchers to explore social, cultural and linguistic patterns from 1500 A.D to the contemporary era.

In the 21st century, book reading and book writing also underwent the influence of DH, and it motivated human beings to interpret books with different types of data visualization and knowledge patterns. For this end, Knowledge discovery in the digital environment required human understanding and perspective. Besides, image mining of pictures; audio mining of different podcasts; and video mining of Youtube videos developed new trends of multimedia mining.

With the advent of the 21st century, the first wave of DH spread. Technology got a sudden boost in the early 2000’s and machine learning, data mining and software were introduced on a large scale. As commercialism and technology rooted in all phases of life, Clementine and IBM Intelligent Miner were introduced to serve human beings digitally. Human beings were saved from the laborious and time-consuming task of text mining.

Several scientific, medicine and biological fields got several benefits from Voyant, Lightside, Weka, SPSS, SQL Server, MATLAB and numerous other data mining software. Center for Computer Analysis of Texts (CCAT) in University of Pennsylvania, USA was established; thus, it compiled digitized text from the domain of archaeology and major genres of literature (ETANA Electronic Tools and Ancient Near East Archives, 2019).

There was a transition from digital culture to "screen culture" in the 21st century. Now visuals, audios, videos, animations, and other features have been integrated into the text to enhance visibility and deep understanding for the readers. When more senses were involved in the learning process, learning becomes long-lasting.

2.3.10 Era from 2011 to Present

In 2011, Culturomics work was published by '*Science*'. In the same year, Google Chrome and National Digital Stewardship Alliance were launched. In 2012, Personal Digital Archiving Day Kit was introduced (Digital Preservation Management, 2013), and in October, 2012, data mining was applied in several fields; for instance, NaCTeM provided services regarding text mining of content and books of biology and social sciences for the British community.

In 2013, Digital Public Library of America, International Internet Preservation Consortium and National Digital Stewardship Residency were launched (Digital Preservation Management, 2013). In 2014, Verband Digital Humanities Association expanded worldwide (Kosmos, 2014), and in 2015, Google Books compiled more than 25 million digitized and scanned books (Stephen, 2015) to observe trends and their occurrences at different time intervals.

Writing in humanities and the process of designing in DH gave a new life to text by transforming static pages into interactive designs. Knowledge and critical questions converted a design into an intellectual model. DH dealt with metadata (data about data itself) and metamedium; and DH worked like a polymath to perform multifarious works in multidisciplinary research works.

Authoring visuals and visual arguments was in vogue in many fields, for instance, teaching, learning, natural, social, business and applied sciences. Interactive visuals have superiority over still visuals since interactive visuals elucidated an argument or strengthened the message of text. Visualization of data had been derived from a vast range of data or digitized corpora in a highly systematic way. Roots of mapping lay in visual signposting and relations (Burdick, A., Drucker, Lunenefeld, Presner, & Schnapp, 2012). Following these characteristics, Voyant tools manifest interactivity and multifaceted data visualization to comprehend knowledge patterns.

2.4 Definitions of Digital Humanities

Digital humanities (DH), a multidisciplinary and umbrella term, incorporated humanities, computational linguistics, text analytics, computer science, AI, NLP, ML. Ramsay defined DH thus: “Digital humanities attempts to bring humanistic inquiry and the artefacts of human experience into useful dialogue with digital technology. It is, at once, a practical and a philosophical endeavor…” (Terras, Nyhan, & Vanhoutte, 2016, p. 281). DH was a dire need of contemporary digital age to tackle ever increasing yottabyte (10^{24}) data which were beyond human capacity to process manually. Potential role of DH could be comprehended with Busa’s 22 years of work during the 1950’s and 1960’s, and now that amount of work could have been done in a few minutes with digital tools.

Kim Lacey defined DH, “The digital humanities were awesome! We’re seeing increased cross-pollination between fields like writing and science, literature and computers, history and engineering. In addition to it, there were so many new interactive tools that allow us to (re)discover ideas in a new way!” (Terras, Nyhan, & Vanhoutte, 2016, p. 284).

DH was a miracle in the sense that it enabled a digital scholar to count words, types of collocations/ n-grams, concordance and to draw several accurate data visuals instantly. Meschini elevated the status of DH in the quote: “You know what a miracle was ... another world’s intrusion into this one” (Terras, Nyhan, & Vanhoutte, 2016, p. 281). Sevigny defined DH in these words: “Using digital media to explore, create, analyse and decode meanings in cultural products, current affairs and social life” (Terras, Nyhan, & Vanhoutte, 2016, p. 282). Therefore, text mining decoded multifaceted knowledge patterns from different domains.

Robert Long called DH “a global vehicle” (Terras, Nyhan, & Vanhoutte, 2016, p. 284) which suggested that the entire world was connected and progressed digitally. Machine-readable texts, online content and digital libraries were connected to the world, and knowledge transfer has become easy for academicians. The trend of digitization and the use of technology have played their roles in shrinking global boundaries. In the end, a misconception about DH should be clarified. Although codes and software were basic elements of DH, yet mere use of digital tools or digital media did not lead to DH. DH focused on the extraction of idiosyncratic knowledge patterns and hermeneutic insights with digital tools.

2.4.1 DH as an Intersection of Several Domains

The automatic analysis led to Culturomics with millions of digitized books (Michel, Shen, Aiden, Veres, Gray, Pickett, & Aiden, 2011). After it, DH opened new research dimensions for text mining and corpus studies. Corpus studies paved the way for several other subfields, website corpora, social media corpus, data mining, educational data mining, text mining and sentiment analysis (Pak, & Paroubek, 2010). The analysis of positive, negative and neutral sentiments was an advanced step of text mining of reviews. It resulted in developing Frequently Asked Questions (FAQs), business challenges and shopping trends.

Moreover, Crane (2006) raised an intellectual query, "What to do with a million books" and Moretti (2013) replied to him by introducing the term "distant reading" which explored basic themes by finding frequent patterns from metadata and testified findings of close reading. Moreover, "Distance is … not an obstacle, but a specific form of knowledge… a sharper sense of their overall connections" (Moretti, 2005, p. 1). Since word by word reading of big data was a difficult task, therefore, about 350 novels were voyanted (the use of Voyant tools to mine text, study, teach, conduct research) in Germany, and their cultural patterns were discovered (Vanchena, 2012) in a symposium.

2.4.2 Digital Humanities and Computation

DH included computation as its core; and computational methods were improved with humanist approaches. Some other building blocks of DH were: classification, description, navigation, organization, metadata and above all digitization. Applying various approaches, DH concentrated on linking, interpreting and disseminating knowledge patterns through digital tools.

2.4.3 Three Waves of Digital Humanities

In the earliest stage, the name of humanities computing was given to DH, but "Machine's efficiency as a servant" (McCarty, 2009) was the key idea in text mining. Three waves of DH have been elaborated in the following paragraphs.

The first wave concentrated on text analysis to build the initial infrastructure of text analysis. In this phase, some software research projects were done qualitatively, and different items were retrieved from databases and archives. Furthermore, some research works focused on quantitative and corpus analyses (Digital Humanities Manifesto, 2009 as cited in Rieder, & Rohle, 2012).

The second wave focused on digital analysis of e-literature, web mining and methodological toolkits to study cyberculture and digital culture. Computer codes were the basic parts of digital culture, and DH designed digital tools to expose concealed information. Furthermore, it developed data, metadata and digital archives for various studies. In the second wave, qualitative and interpretive research works were also conducted with several digital toolkits to perform specific functions.

The third wave explored multifaceted computational features. In the contemporary era, studies about software and DH had been separated. New advancements also pointed out anomalies in DH tools and analyses (Liu, 2012). The third wave led to explore computational features from the text. The paradigm of text mining shifted from close reading to distant reading (Moretti, 2005). The current study concentrated on different aspects of distant reading through Voyant tools. The mere use of corpus had become a primitive method, while data visualization tools, for example, Voyant, MONK, MATLAB, Juxta, eMargin, Poemview, Sobek, PRISM were used for distant reading and data visualization.

2.5 Digital Humanities and English Department

Computers and English departments were deeply linked for learning, teaching, publishing and research purposes; for instance, Shakespeare and Jane Austen's complete corpora were studied with Voyant tools. In addition, Shakespeare to second life project was conducted by Humanities Department at the University of Maryland, USA. Moreover, the Digital Humanities Quarterly journal also published content from English literature and linguistics. University of Alberta, Canada, celebrated Day of Digital Humanities on 20th February, 2021, and worked collectively for DH. Furthermore, Twitter and Google Books were also analysed with digital tools.

i. Besides numeric data, text in English or in any other language was more significant than audio, video or images in the computerized data.

ii. Presence and relationship of English text had a long history in computer systems and DH analysis as works of Busa and Lousine have proved earlier.

iii. Linguistics, stylistics, composition, rhetoric and author attribution research works had largely benefitted from DH and text mining.

iv. Digital archives were used for applied theory; hence, Jerome McGann worked exhaustively on the Rossetti Archive.

v. Previously English love letters were analysed to extract key knowledge patterns.

vi. As academia was vastly influenced by e-book and hypertext, so digital culture had emerged as a core study in DH.

vii. Academic publishing through digital tools also paved the way for DH, data mining and data visualization (Kirschenbaum, 2016).

2.6 Digital Humanities in the Global Village

DH addressed emerging challenges of technology and learning through knowledge discovery. The increasing use of personal computers, cell phones, the internet and the advent of social media transformed the world into a digital global village. Again, learning and universities have developed strong ties. According to Kant, knowledge construction was a hallmark of universities (Readings, 1996). Expanding the same notion, DH incorporated the German concept of “Bildung” (Berry, 2012, p. 14), which referred to education and philosophy for personal and cultural maturation. Again, Ryle (1945) defined “digital Bildung” as “dispositional excellence” in computation. In this section, a few developments of DH, data mining and text mining had been discussed briefly in different advanced countries of global village as well as in Pakistan.

2.6.1 Digital Humanities in Canada

In Canada, first of all, Dr. Samuel Cioron founded the Humanities Computing Centre in 1986, and he started to change language material from audiotape to soft material. He developed language modules according to the authoring system of McBookmaster. Later on, media classrooms and language labs were established to meet forthcoming digital challenges. Later computational linguistics was also involved in Humanities Computing (Schreibman, 2013). In 1994, Dr. Geoffrey Rockwell replaced Dr Samuel Cioron. Rockwell started three Humanities Computing courses for humanities students. In addition to it, the Canadian province Ontario started the Access To Opportunities Program (ATOP) for the expansion of technology in humanities. With state funds, the multimedia programme was expanded to teach input and output of information technology to students. It was named multimedia because it merged all types of media, for instance, audio, video, memory storage, graphics, telephone, television, computers, digital tools and software. To conclude, Humanities Computing emerged as an academic and scholarly discipline to exhibit performance, creativity and interpretation (Rockwell, 2016).

2.6.2 Digital Humanities in the USA

There were numerous completed and mid-way DH projects and some of them had been mentioned here. Francophone Digital Humanities (map, timeline); A Literary Tour de France by Robert Darnton (books on French Revolution); Paris Past and Present (with 3D models); Haiti: An Island Luminous (all historical events); The Trans-Atlantic Slave Trade Database (trade of 10 million African slaves); The Early Caribbean Digital Archive (of literature), 19th-Century Caribbean Cholera TimeMap (timeline of cholera) projects were done successfully (University of Florida, 2017a). Various DH labs are functional in MIT, Stanford, Harward, Carnegie Mellon, Pittsburgh to produce high-quality DH projects.

2.6.3 Digital Humanities in Australia

There were numerous ongoing and accomplished DH projects in Australia, for instance, AUSTLANG: Australian Indigenous Languages database; OCCAMS, A research tool to interact and share research; Commonplace Cultures; The Australian Common Reader (Australian National University, 2016). Internationally famous books of the late 18th century were saved through the FBTEE (The French Book Trade in Enlightenment Europe) project (University of Florida, 2017b).

2.6.4 Digital Humanities in Europe

European Association for Digital Humanities prepared *'1914-1918 Online. International Encyclopedia of the First World War'*, ALCIDE to facilitate analysis of humanities; ARIADNE to collaborate archaeologists' work; *'Book of the Dead'* to explore 200-century spellings and images; Burckhardtsource, a semantic digital library; CLiGS (Computational Literary Genre Stylistics); Dantesources (about Dante's works); Eighteenth Century Poetry Archive; German Inscriptions Online; Stylo R Package and many more projects were accomplished from 2012 to 2017 (European Association for

Digital Humanities, 2017). The Baudelaire Song Project, Medieval Francophone Literary Cultures Outside France, Mapping Paris, Medieval Francophone Literary Cultures Outside France (University of Florida, 2017c) projects were in progress in the UK.

2.6.5 Digital Humanities in Pakistan

Pakistan had recently entered the early phase of digital humanities. Integration of learning, digital business and computational linguistics emerged in academia and other private and government departments. Some universities have inculcated DH into their pedagogy. Open University Islamabad used PTV for the first time in 2005 to telecast its academic programmes for distance learning. It was a very meagre beginning of DH for teaching and learning purposes. Nowadays, AIOU had also started its online courses. Furthermore, Virtual University was the first federal government non-profit online university that enrolled students from 2002 onwards. It also had its regional study centres and teachers for replying academic enquiries. It compiled the best and latest digital lectures from renowned university professors (Virtual University of Pakistan, 2015).

The use of digital technology had been started in academic institutions; for instance, several private sector international schools, colleges and universities of Pakistan incorporated digital content, digital interactive board, multimedia, audio-video recording, soft material, and portal system in their curriculum some years ago.

Learn Smart Pakistan (LSP) was the free online platform for Pakistani students and teachers to digitally learn and teach the Pakistani board curriculum. The first digital organization started its pilot project in 2014 to facilitate its students to pass Pakistani board exams with distinctions. To motivate learners, LSP distributed awards as tablets to schools, students and teachers (LSP, 2016).

Now government sector institutes also started to adopt DH, so Muhammad Atif Khan, Minister for Elementary and Secondary Education, Khyber Pakhtunkhwa (KPK), inaugurated interactive whiteboards for 100 KPK schools on 14th December, 2016. It was a pilot project; later, it was extended in the entire KPK (Daily Times, 2016).

In 2017, the Punjab government initiated eLearn Punjab project, a brainchild of the Punjab Information Technology Board (PITB), on 6th January, 2014. It intended to promote digital literacy in Punjab, Pakistan. Digital textbooks of Maths and Science from grades 6 to 10 were prepared, and these digital textbooks followed Punjab Curriculum and Textbook

Board (PCTB). To facilitate learners and teachers, 10,000 videos, 1300 minutes audio, simulations and animations were provided with digital books. These videos were taken from renowned online sources, for instance, Khan Academy. Its prime objective was to promote self-study concepts with facilitation of digital supplementary material. To convey the fruits of this project, 200 multimedia classrooms were started throughout Punjab Government schools (Punjab Information Technology Board Government of the Punjab, 2017, p. ix). Regretfully, all was claimed in Punjab government documents; however, nothing was done practically because the given website did not show any digital content for learning.

Knowledge Platform, an international private organization for the production of digital academic content, opened its functional offices in Pakistan, Singapore, Shanghai, Jakarta and Delhi. It aimed to teach traditional courses with digital methods and support international collaboration. Primarily, it prepared audio, video, ESL games and readers' passages in English (Knowledge Platform, 2017). COVID-19 also played its significant role to transform traditional teaching into digital and hybrid teaching with Zoom, Teams and LMS (Learning Management System) support.

2.7 Journey from Hermeneutics to Digital Hermeneutics

In Greek mythology, Hermes, the son of Zeus and Maia, was a messenger and interpreter to other gods. Keeping in view characteristics of Hermes, hermeneutics was an interpreting theory of revealed, philosophical and literary manuscripts. Its roots also lay in Aristotle's work '*On Interpretation'*. In theology, Talmudic, Vedic, Buddhist, Biblical and Quranic hermeneutics were also composed in different eras. Usually, in hermeneutics, some interpretation rules were devised as a framework; for instance, 13 rules of Rabbi Ishmael were determined for Talmudic hermeneutics (Sion, 2010). Moreover, philosophical, ancient and medieval hermeneutics also opened new interpretive dimensions.

In the contemporary era, traditional hermeneutic frameworks could not manage deep and quantified analysis of big data texts. Modern hermeneutics also interpreted verbal, written and nonverbal expressions. In modern hermeneutics, Schleimacher (1768-1834), Dithey (1833-1911), Heidegger (1889-1976), Gadamer (1900-2002) and Marxist hermeneutics got fame and influenced different disciplines (Malpas, & Gander, 2014). Their theories became the foundation of DH tools.

Hermeneutics had been changed into an interpretive machine in this technological era. Kenny (1992) argued that computers quantified the hermeneutic process, and there was no share of mere intuition. Computers were not just enumerators, but they processed logic and binary principles.

To handle this ever-expanding data, digital hermeneutics required digital tools for text mining. As hermeneutics was interpreted in the light of some specified rules, and their violation was not allowed in that hermeneutic school of thought. Likewise, digital hermeneutics functioned under some algorithm which was a computerized step by step process for the interpretation of text datasets. Moreover, a pseudocode algorithm was used to explore word counts, token words, concordance and comparison of data and distribution of data as per instructions (Rockwell, & Sinclair, 2016). Some algorithms were named as Naïve Bayes tree, Expectation Maximization for finding maximum likelihood), C4.5 (for generating decision trees), CART (for building a decision tree), CN2 (learning for rule induction), Apriori (for finding frequent itemset), k-nearest neighbour (for classification), k-means (for calculating distances), maximal frequent itemset, randomized function K, independent choice logic, PageRank and TwitterRank (Peña Ayala, 2014, p. 68).

2.7.1 Three Hermeneutic Principles

Hermeneutics is based on the following three principles: a. it is supposed to have a textual and coherent unity. b. concordance is generated from paragraphs of the selected text. c. Concordance generation follows a systematic process (Rockwell, 2003). In the current study, concordance has been generated with Contexts tool, and they are arrangeable in length and occurrence.

2.7.2 Computation and Digital Hermeneutics

The discipline of mathematical physics emerged two hundred years before the advent of technology. It envisioned us that maths and physics gave birth to technology, and they were embedded in several academic disciplines. Furthermore, statistics was embedded in machine learning, AI, data mining, data visualization and text analytics tools. Likewise, numeric data clarified and quantified concepts in humanities and sciences.

Unified theory of mathematics functioned in computation. Knobloch (2004) opined that mathesis universalis (science or learning) and universal mathematical science were ideal for the construction of knowledge. There was a distinct discrepancy between hermeneutics and quantification. In the 13th century, Raymand Lullus propounded the concept of ars combinatorial (branch of mathematics for the study of finite and logic). He proposed that a group of propositions should be merged through a logical set of rules, and in their way, knowledge was produced mechanically. This idea was popularized in the seventeenth century, and later on, Leibniz and Descartes were influenced by this idea (Rieder, & Rohle, 2012, p. 78). Furthermore, System theory (an interdisciplinary study of systems) and cybernetics also strengthened this view.

2.7.3 Agile Hermeneutics

Digital tools were adjusted and designed for validation and expedition of the hermeneutic process. In agile hermeneutics, new digital tools were designed, as well as researchers applied those tools for in-depth analysis and textual interpretations. In short, these two processes refined each other for a better hermeneutic approach.

2.7.4 Technology and Digital Hermeneutics

Discussing the nexus between philosophy and technology, the word "technology" had been derived from the Greek word "Technikon" which meant skill. Moreover, the Greek word "techne" had been associated with episteme which referred to craftsmanship, fine art, knowledge and adeptness in skill. "Techne that could do that most human of tasks" (Rockwell, & Sinclair, 2016, p. 25), and it meant that most of the hermeneutic tasks could be performed by digital tools. Essence of technology was called "Enframing" which created order and revealed knowledge. "Technology was ... an instrumentum" (Heidegger, 1954, p. 2) to accomplish academic and non-academic tasks.

Information technology and hermeneutics worked in collaboration to extract knowledge patterns which assisted in the formulation of new hypotheses and unexplored facts. Thus, digital hermeneutics was framed on Heidegger and Vattimo's ontological rules. So, Vattimo introduced "aesthetic pacifism", and he considered "discovery of modern technology as a communication" (translation) (Vattimo, 1996 as cited in Capurro, 2010). In a nutshell, digital tools communicated ideas and patterns from big data; hence, digitized data were termed hyper information (Borgmann, 2000). Digitized life and communication in academic and non-academic settings changed our worldview and interpretive mental makeup. So, DH and cybernetics (automatic communication and control system) collaborated to construct knowledge and ontology; thereupon, digital ontology led to digital reality (Capurro, 2006).

2.7.5 Heidegger’s Positive and Negative Views on Technology

German philosopher Heidegger said, “Language was the house of the truth of Being” and technology was a mode of “ordering revealing”, so “it was the realm of revealing, i.e., of truth” (Heidegger, 1954, p. 18). “Constellation of truth” (Heidegger, 1954, p. 18) and “essence of technology may come to presence in the coming-to-pass of truth” (Heidegger, 1954, p. 19). Again, he argued that technology should be used to reveal the truth from dispersed big data. Likewise, text mining and digital forensics assist in finding the ultimate truth, and the knowledge discovery process is also a means to reveal the truth. Moreover, he challenged the Cartesian ideology and rehabilitated the concept of thingness to the thing (Heidegger, 2010). It was a theory in human-computer interaction that an object transformed into a thing, if it stopped performing its assigned functions.

Viewing the dark side of the picture, “Technology threatens to slip from human control” (Heidegger, 1954). To obstruct the slippery nature of technology, several stops and algorithms were used by Voyant and other digital tools to control the slippery nature of technology and to grasp the desired outcomes.

2.7.6 Screwmeneutics: A Foil to Digital Hermeneutics

Screw is an instrument for manual work and the term “screwmeneutics' ' refers to manual work for the preparation of corpus and its hermeneutic interpretation. When the use of DH increased, the number of its critics also increased, even though they opposed the utilisation of computers for the study of humanities. Ramsay wrote a research paper namely *‘The Hermeneutics of Screwing around...’* (Ramsay, 2014), and it nullified the use of computers as an evaluator of any hypothesis. Instead, active and rational readers were searched to analyse the text. In fact, human beings failed to analyse big data of the complete works of several writers in a few minutes, while digital tools were capable of doing these tasks instantly. A serious drawback of screwmeneutics was the manual preparation of corpus, collocation/ n-grams which were gigantic and time-consuming tasks as Roberto Busa consumed 22 years with screwmeneutic tasks. Currently, drawback of manual file or corpus preparation has been removed by several online tools; for instance, Parsehub tool scrapes big data quickly and correctly (Parsehub, 2019). Voyant tools have the inbuilt potential of online crawlers to scrape large texts automatically.

2.8 Extraction of Knowledge Patterns with Data Mining

Data mining was defined as “the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining was also known as knowledge discovery in databases (KDD)” (Oracle, 2017). Its basic features were: automatic pattern generation, prediction and production of useful information from big databases.

2.8.1 Rationales for Data Mining

There are several rationales for data mining: Firstly, “we were actually living in the data age” (Han, Kamber, & Pei, 2012, p. 1). Secondly, numerous types of data were growing in terabytes from individuals and organisations in every coming second. Thirdly, various sources of information about one thing were limitless in this digital modern age; hence, they were ranging from linear to nonlinear text, websites to social media and audio to video. Fourthly, interdisciplinary, multidisciplinary and transdisciplinary studies of this age drew our attention to the point that data about one topic or organization were present in different disciplines, sources, soft files, hard files and in different types of media. Fifthly, textbooks reflected knowledge sources and to widen students’ academic boundaries, data mining was essential since traditional learners just emphasized on reading comprehension, but data mining led them to corpus summary for the elaboration of stylistic features, collocation patterns/ n-grams for fluency, knowledge graphs for the interrelationship of various themes and characters, and word sense disambiguation for clarity and compilation of information on one topic.

To handle gigantic digital age data, sophisticated data mining tools were required to change raw data into meaningful and interactive knowledge patterns. Their purposes should be analytical, descriptive or predictive according to the needs of individuals or organizations. For these purposes, predictive models were used in 6 out of 10 approaches (Peña-Ayala, 2014a).

2.8.2 Role of Interdisciplinarity in DM

Data mining techniques were mainly derived from statistics, ML, AI, NLP and databases. Statistics dealt with numeric data to organize and to apply to different situations and statistical data mining identified structures with predictive potential. A further development, GenIQ Model solved those problems in which statistical data mining failed (Ratner, 2017). Machine learning designed and applied learning focused algorithms, and NLP concentrated on language processing in DM and tool designing.

2.9 Data Mining Theories

Theoretical evolution in DH contributed on different levels. Some theories were reviewed to link them with DH tools which were outcomes of some theories. Moreover, Knowledge discovery, hermeneutica, ontology and logical reasoning were the outcomes of learning DH. Consequently, ontology as a reality of knowledge paved the way for technology and machine learning.

2.9.1 Bayesian Theory

This theory was derived from Thomas Bayes (1701-1761), who presented an equation that permitted emerging phenomena to update information. Later on, in 19th century, Pierre-Simon Laplace elucidated Bayesian Theorem which employed decision trees in DM research works (Peña-Ayala, 2014a). It was premised on statistics and probability to create knowledge patterns on the basis of previous conditions. Therefore, Bayesian Theory was applied to Bayesian inference and statistical inference, and it became inevitable for probability theory (Jeffreys, 1973). Hierarchical Bayesian Model or Probabilistic Approach was a theoretical approach for data mining processes of classification and clustering. This approach linked all variables in a comprehensible manner because everyone wanted to derive a short and easy text pattern (Gelman, Carlin, Stern, & Rubin, 1995).

Hamalainen, Suhonen, Sutinen and Toivonen (2004) germinated a hybrid model which amalgamated ML and AI to develop a Bayesian Model whose purpose was to provide academic guidance according to their deficient skill. In this model, Bayesian Knowledge Tracing (BKT) was an algorithm for testing learners’ ability and learning in intelligent tutoring systems (Corbett, & Anderson, 1995).

2.9.2 Cognitive Psychology

Gestalt psychologists (since 1890’s) took deep interest in recognition of patterns (Enns, 2005) that were associated with cognitive psychology and information processing. Cognitive psychology had deep-seated ties with computer science and computer graphics.

Psychology harmonized Schema Theory which was postulated on those empirical traces which were drawn on the slate of memory after going through the corpus (Bartlett, 1932). Corpus, an embedded tool in Voyant data mining tools, was the authentic tool to count words, vocabulary density, occurrence and concordance of language, and it laid effects on memory. Barlow (1996) emphasized schema-meaning mapping of the frequent use of language structures. In addition to it, Sinclair’s research on the lexical entities (1996) exhibited form and meaning relationships. Widdowson (1991, 1992, 2000) used corpus for pedagogic purposes. Developing a nexus between corpus and intertextuality, Seidlhofer (2000) posited that intertextuality was essentially required to learn a language from any corpus. Flowerdew (1996) and Tribble (1997) worked on pedagogic requirements, for instance, syllabus designing and development of specialized corpora. Gavioli (2000) suggested that corpus examples should be used for the construction of new sentences on those patterns. This approach quite naturally followed a known to unknown technique for learning and teaching foreign languages.

2.9.3 Information Theory

In 1948, first of all, Claude E. Shannon suggested Information Theory which concentrated on quantification, communication and storage of data. Its major topics were DSL (Digital Subscriber Line), zip files, information and data compression, that is why, it brought revolution in internet, CD (compact disc), mobile technology and linguistics (Shannon, 2009). Cirrus tools provided quantified data and the selected tools extracted information on the basis of quantification. Thus, information theory facilitated the discovery of new knowledge patterns through data mining.

2.9.4 Systems Theory

Bertalanffy, an Austrian biologist, was one of the major contributors of Systems Theory (an interdisciplinary study of systems) which emerged from cybernetics (it is concerned with regulatory and purposive systems) to recognise patterns. It aimed to explore system dynamics and such generalized principles which were applicable to all levels and fields. Its key concepts were system (organized part), homeostasis (maintenance of a system), boundaries (limits), reciprocal transactions (affecting each other), feedback loop (self-correction), microsystem, mesosystem (relation of systems), exosystem (indirect co-influence of two systems on a third system), macrosystem, chronosystem (affecting system). To conclude, any transition in one subsystem affected the entire system (Bertalanffy, 1969).

2.9.5 Knowledge Discovery Theory

Rakesh Agrawal, an internationally recognized computer scientist, was the pioneer of Knowledge Discovery Theory in data mining (Zhu, 2018), and it was premised on databases, statistics and machine learning (Fayyad, Shapiro, Smyth, & Uthurusamy, 1996). “In active data mining paradigm,… rules are discovered, …the history of the statistical parameters associated with the rules is updated… we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct to retrieve rules whose histories exhibit the desired trends” (Agrawal, & Psaila, 1995, p. 1). Summarization involves methods for finding a compact description for a subset of data (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). It was also defined as “the extraction of implicit, previously unknown and potentially useful information from data"” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9; Witten, Frank, & Hall, 2011). It transformed random data into meaningful and valuable information. Knowledge types could be classified into spatial pieces of information, spatial relationship, spatial taxonomy and spatial clustering. During the research process, some characters could be delimited in the data mining process (Rahayana, & Siberschatz, 1998).

Knowledge Discovery Theory had interdisciplinary features; joining fields of computer, computational linguistics, ML, statistics, NLP, AI, DH, systematic information and applicability. This theory was grounded in the rules of association, characteristics, classification, serialized system and prediction. It employed several techniques and theories: evidence theory, cloud model of mathematics, fuzzy sets, rough sets, neural networks, genetic algorithms, SOLAM (data mining with online processes), visualization, exploratory learning and spatial inductive learning (Li, & Wang, 2005).

Extending Knowledge Discovery Theory, Comprehensive Knowledge Discovery Theory discussed spatial objects and spatial relations (Zongyao, & Fuling, 2003).

Framework of DiscoTEX (Discovery from Text Extraction) linked Information Extraction (IE) and standard text mining methods to explore prediction patterns. The prime objective of text mining was "to discover knowledge in unstructured or semi-structured data" (Nahm, 2001). Digital text mining constructed new vistas of knowledge or insights which were valuable for knowledge discovery and "new knowledge" construction. "The unlocked information can lead to new knowledge and improved understanding" (McDonald, 2012). If text mining was useful for society, it produced "positive externalities" (McDonald, 2012) i.e., benefits for third party. The current study explored knowledge patterns, for instance, textual themes in the image of word clouds, neighbouring words as collocations, connectivism of themes and characters, detailed corpus of text and KWIC.

2.9.6 Microeconomic Theory

In DM, optimization was the basis of microeconomic view and optimization evaluated association and clustering processes. This framework facilitated problems of sensitivity analysis, theory of games and segmentation. A pattern was called interesting and beneficial if it transmitted advantages of predictability and decision making (Kleinberg, Papadimitriou, & Raghavan, 1998).

2.9.7 Learning as Research Approach

Johns (1991) and Gavioli (2000) followed Learning as Research Approach which motivated learners to study according to their own desires, interests and remedial activities. As a result, learner empowerment in autonomous learning was promoted. Bernardini (2000, 2002) used this approach in adult students of translation and interpretation at University of Bologna, Italy. She found Learning as a Research Approach which was very useful in discovering various knowledge patterns. They exploited the British National Corpus (BNC) in their class and used WordSmith Tools (Scott, 1996) for the comparison of both corpora. Various tools, for instance, WebCorp, KWICFinder, WebKWIC, Wordnet, Web Concordancer were also frequently used in academia and classrooms (Cobb, Greaves, & Horst, 2000) to prove learning as a research approach.

2.9.8 Pattern Discovery Theory

Pattern Discovery Theory was employed to discover and compare idiosyncratic and innovative knowledge patterns in sequential and time series data. It began its function with minimum structure; and complexity emerged when data required complexity (Shalizi, Shalizi, & Crutchfiel, 2002).

2.9.9 Template Matching and Feature Analysis Theories

Template Matching and Feature Analysis Theories were employed in pattern recognition in the domain of cognitive psychology. Stimulus was juxtaposed with a mental model in Template Matching Theory. Visual stimulus was split into important features which were compared, and a common pattern was derived (Neisser, 2005).

2.9.10 Probability and Statistics Theory

Probability Theory, a key concept of statistics, dealt concepts mathematically by a set of axioms which formulated probability. They had values ranging between 0 and 1, and they were called probability measures. Probability theory was applied in quantitative data analysis to study randomness in any phenomena. Roots of probability theory embedded in Gerolamo Cardano's games of chance in the 16th century. It worked in two ways: Firstly, results of a large number of trials showed regularity. Secondly, interpretations of feature probabilities facilitated to extract a limited, accurate and predictive pattern (Siegmund, 2014). A large number of EDM studies were conducted with probability, statistics and machine learning (Peña-Ayala, 2014a).

2.9.11 TPACK Model

Technological Pedagogical Content Knowledge (TPACK) meant intermingling of technology, pedagogy and content knowledge as shown in figure 5. TPACK Model was an extension of Shulman's idea of Pedagogical Content Knowledge. KDD and TPACK Model were interlinked because KDD extracted knowledge patterns and TPACK Model conveyed content with pedagogy and technology-centred knowledge approach. Emerging trends in academia played their vibrant roles in bringing radical positive changes in technology-focused learning and teaching pedagogies of an academic content (Koehler, Mishra, 2009).

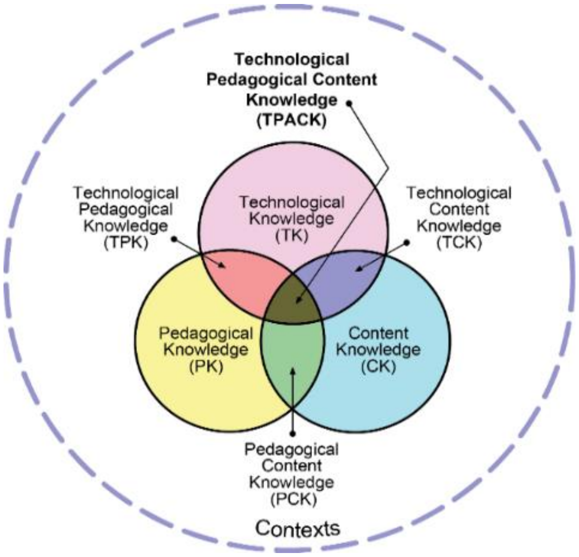

Technological Pedagogical Content Knowledge (TPACK)
Technological Pedagogical Knowledge (TPK)
Technological Knowledge (TK)
Technological Content Knowledge (TCK)
Pedagogical Knowledge (PK)
Content Knowledge (CK)
Pedagogical Content Knowledge (PCK)
Contexts

Figure 5 TPACK features (Koehler, & Mishra, 2009).

Altun and Akyildiz (2017) researched 609 pre-service teachers with certain variables of the TPACK Model. Then, data were analysed with SPSS (Statistical Package for Social Sciences), and it was found that many teachers had these abilities before joining the teaching career, but there was an extensive demand for practical training of pedagogy. In another recent study, Avidov-Ungar and Shamir-Inbal (2017) sought a perception of 130 ICT (information and communication technologies) coordinators from Israeli schools. With a narrative analysis technique, it was explored that technological, pedagogical, organizational leadership and ICT coordinators were change-agents in academia.

2.9.12 Hermeneutica Theory

Stefan Sinclair and Geoffrey Rockwell propounded Hermeneutica Theory along with Voyant tool explanations. They explained some of its features:

i. “Hermeneutica Theory is embedded in a context.”

ii. “It is not like black boxes.” (In the domain of computer programming, it did not examine the actual background programme which was executed.)

iii. “Manipulation is in service of exploration and understanding.”

iv. “It is supplemented by other materials.”

v. “Knowledge bearing tools provoke reflection.”

vi. “Hermeneutic tools fail in interesting ways.”

vii. “They can be extended to expose new things” (Rockwell & Sinclair, 2016, p. 166).

2.10 Seven Ways to Select Data Mining Method

To find the suitability of the data mining method, the following seven methods were suggested.

1. Granularity: It explored the detailed focus of the document.

2. The focus of Algorithm: To analyse words or whole documents, information extraction or clustering should be selected.

3. Available Information: Supervised learning required trained data, while unsupervised learning did not require trained data. The former was more authenticated and beneficial than the latter.

4. Semantics or Syntax: Semantics dealt with meanings, while syntax concentrated on word order, grammar and cognitive effects.

5. Traditional Text or Web Text: Several algorithms were introduced for web or traditional text, but web content had expanded on a large scale due to the internet (Miner, Elder, Fast, Hill, & Nisbet, & Delen, 2012).

6. Information Retrieval: Search engines used information retrieval and text categorization to locate exact words from trillions of sources in a few seconds.

7. NLP: NLP was employed for deeper knowledge owing to brain-like functions of a computer. By the same token, ML became vital for text analysis and NLP processes whose key features were POS tagging, word sense disambiguation, parsing and information extraction.

2.11 Data Mining Techniques for Diversified Data

Clustering, classification and association rule mining techniques were applied on transactional data (a file or table for the transaction) and sequential data (data requires order) mining which were applied on temporal data (values which varied time to time). In addition to them, text mining was also applied on text data (location); multimedia data mining was used for multimedia data (image, video, audio) and web mining was done on web contents (Romero, Romero, & Ventura, 2013). The current research conducted textbook analysis with educational text mining techniques.

2.11.1 Data Mining Approaches

There were two approaches to data mining: hypothesis testing and knowledge discovery (Du, 2010, p. 16). The former checked a hypothesis with data mining to accept or disprove it. On the other hand, the discovery approach commenced with data which drove the investigation and results to innovative ideas. It resulted in supervised or unsupervised learning.

2.11.1.1 Supervised and Unsupervised Learning and Objectives

Supervised learning had a variable in output, while unsupervised learning did not have any particular target variable. Supervised learning used labelled training data, and it required extra parameters. They were used to classify or to predict textual or market trends, for instance, ANNs, KGs, and decision trees. Unlabelled training data were used for unsupervised learning models, and it did not require any extra parameters. Its examples were attribute clustering and association.

As the key objectives of supervised and unsupervised learning are concerned, the data mining process was done for descriptive and predictive purposes (Kantardzic, 2011). Descriptive models were applied on unsupervised learning to generate knowledge patterns which exhibited interrelationship of analysed texts (Peng, Kou, Shi, & Chen, 2008). Predictive models were applied on supervised learning to show futuristic perspectives of dependent variables (Hand, Mannila, & Smyth, 2001). Data from 2010 to 2012 research projects demonstrated that 40% approaches described data patterns, and 60% approaches predicted data patterns.

2.11.2 Goals of Data Mining

Major goals of data mining were “classification, estimation, prediction and data description” (Du, 2010, p. 5). Elaborating these terms, classification meant the compilation of same things in a relevant category; estimation resembled classification of data; prediction referred to a future outcome; data description meant to describe features of data with supporting data visualization, statistical data, tabular data and clustering of data items.

2.11.3 Tasks of Data Mining

There were several tasks of data mining to fulfil the aforementioned goals:

1. Classification: It showed categorization of dispersed elements into some meaningful and interesting patterns which conveyed some comprehensible messages to its viewers.

2. Clustering: Clustering meant a group formation of similar data, but they were different from other objects of another cluster because attribute values differentiated between them. In other words, there was a distribution of objects into subsets. This division was done with clustering algorithms, and it discovered new knowledge patterns. Natural grouping of elements was based on their attributes. Clustering was an unsupervised technique, and usually, clustering algorithms had an option of some stops and iteration to establish certain limits and refinements. Cluster analysis was applied in many fields, and it used partitioning, hierarchical, density, grid and probabilistic models (Han, Kamber, & Pei, 2012).

Single membership clustering quantitatively measured the similarity of documents, while Multiple Membership Clustering divided a document into several categories, for instance, topic modelling (Shaw, 2012, September 20).

SCAN, an algorithm for graph clustering, recognized well-connected elements as clusters. Besides, there were two prominent kinds of clustering: Subspace clustering method which searched clusters in the original vicinity of clusters. Biclustering methods produced new space and position for the search of clusters (Han, Kamber, & Pei, 2012). Clustering, one of the significant data mining methods, was primarily a statistical process, and it had been used in terrorism detection, industry, engineering and marketing (Everitt, 2009). The clustering was generated by click-stream server data that how did learners share academic content and how much time did they spend for various activities and profile generation of learners through OLEs (Object Linking and Embeddings): a system for linking and embedding data, images, and programs from different sources (Antonenko, Toy, & Niederhauser, 2012).

Computing proximity was involved in the analysis process of clusters. Then, the clustering algorithm was applied to attain equidistant items. Clustering development could be hierarchical or non-hierarchical. In the beginning, several clusters were formulated and later, they merged, and the number of clusters decreased. Cluster analysis unveiled automatized cognitive processes and insights about more important factors. Its validity was expressed by automatic information extraction which was almost akin to human information derivation from the large dataset (Antonenko, Toy, & Niederhauser, 2012).

TEXTRISE worked under a learning mechanism which was an amalgamation of rule and instance-based learning system (Antonenko, Toy, & Niederhauser, 2012).

Besides, in Density-Based Clustering Methods, density was measured with the total number of some entities in a specific radius which was named as Eps. Thus, data could be clustered in core point (with Eps), border point (near the border of Eps), and noise point (outside of Eps) (Du, 2010).

3. Graphic Clustering: CHAMELEON was used to measure interrelationship and similarity among various clusters. Eventually, it reached the automatic system in which two clusters should be mixed owing to their extreme similarity. CHAMELEON had the following steps:

4. Construction: A sparse graph was made from the selected data.

5. Segmentation: A sparse graph was divided into sub-graphs.

6. Agglomeration: Subgraphs or sub-clusters were linked for smooth functioning (Du, 2010).

7. Association/ Market Basket Analysis: It was employed to point out common sets of items.

8. Regression: It resembled classification. Furthermore, both logistic regression and linear regression were frequently used for prediction purposes.

9. Forecasting: It took input time-series datasets for the prediction of future trends through data mining.

10. Sequence Analysis: It was used to find patterns in a series of actions, and to break the sequence was also a peculiar act.

11. Deviation Analysis: During the DM process, some rare cases or deviations from sequence were found (Tang, & MacLennan, 2005).

2.11.4 Current Issues and Challenges in DM

The following issues and challenges emerged during DM processes:

i. There were difficulties to transfer data mining results to high tech canonized reporting tools.

ii. Data mining results were derived from statistics.

iii. There was an extreme necessity to enhance user education in data mining.

iv. Algorithms of data mining were limited (Tang, & MacLennan, 2005).

v. EDM lacked particular theories, trustworthy frameworks and field terms.

vi. EDM required valorization of its research contributions and progressive evolution in academia (Peña-Ayala, 2014).

Worldwide DM conferences “ignore the application of DM for education” (Peña-Ayala, 2014), learning, teaching and publishing. This study filled this wide gap and introduced text mining in the domain of ESL, TEFL, ELT and textbook analyses. Data mining was highly recommended only for high tech computer scientists, but it was very productive for all languages and social sciences. Consequently, the DM research trend shifted towards Educational Data Mining to explore knowledge patterns.

2.12 Educational Data Mining (EDM)

Educational Data Mining (EDM) had emerged recently to solve several academic and linguistic issues (Baker, & Yacef, 2009) by taking data from educational settings. Usually, it took unstructured data from traditional classes, textbooks, computerized material, web data, and software to transform them into valuable knowledge patterns. It had been further elaborated in figure 6. ‘‘EDM was both a learning science as well as a rich application for data mining. It enabled data-driven decision making for improving the current educational practice and learning material’’ (Calders, & Pechenizkiy, 2012). The current study provided pedagogical support (Baker, 2010) by unveiling knowledge bearing linguistic patterns with Voyant text mining tools.

Common techniques for EDM were classification, association and clustering through computerized tools. Its major aims were to transform data into knowledge, and to utilise this mined knowledge for constructive decision making, prediction of learning, promotion of learning styles, designing quizzes and evaluating assessments. Besides, EDM provided a pro- learning academic environment maintaining their interest, (Ingram, 1999), and it was used effectively in e-learning, e-commerce (Hanna, 2004), customers and students’ instruction (Romero, Ventura, & De Bra, 2004).

In scientific research, the publication of the first Journal of *'Educational Data Mining'* started in 2009. In 2011, a group of educational data mining conferences started on Learning Analytics and Knowledge. An expansion of knowledge beyond boundaries, research journals and research conferences proceeded side by side in discovering and expanding knowledge of EDM. In 2014, another group of researchers founded ACM Learning, and Pittsburgh Science of Learning Center (PSLC). Besides, DataShop was the largest store of educational data, and it was established in 2004 at Carnegie Mellon University, Pittsburgh, USA (Baker, 2014).

Recently, several technological tools and devices had been introduced in academia to transform static learning material into an interesting learning material (Ha, Bae, & Park, 2000). Consequently, interesting and beneficial knowledge patterns were extracted from large data through KDD (Knowledge Discovery in Databases) (Klosgen, & Zytkow, 2002).

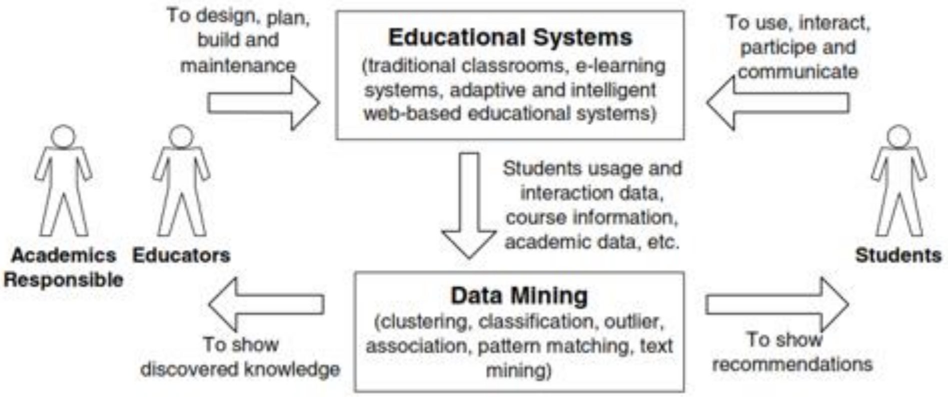

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 6 EDM and students (Romero, & Ventura, 2007, p. 136).

The rationale behind educational data mining was that EDM addressed only educational tools. Secondly, they were easy to use for academia (Zaiane, Xin, & Han, 1998) and its practitioners because educational data were taken from the traditional educational environment, virtual classrooms, web learning and computer-aided learning (Johnson, Arago, Shaik, & Palma-Rivas, 2000). DM found reasons for the outperformance of one section. Usually ELF (executable and linkable format), CLF (Common Loudspeaker File). and server log files were processed for data mining (Koutri, Avouris, & Daskalaki, 2004). DM was also utilised to explore the academic problems, causes (Nilakant, & Mitrovik, 2005) and solutions.

Being a multidisciplinary domain, EDM connected logic programming, ANN, decision tree construction, instance-centred learning, Bayesian learning, rule induction and statistical algorithms (Klosgen, & Zytkow, 2002). DM and EDM incorporated several other disciplines, for instance: probability (Karegar, Isazadeh, Fartash, Saderi, & Navin, 2008), soft computing (Mitra, & Acharya, 2003), machine learning (Witten, Frank, & Hall, 2011), artificial intelligence (Bhattacharyya, & Hazarika, 2006), statistics (Hill, & Lewicki, 2006), and NLP (McCarthy, & Boonthum-Denecke, 2011).

2.12.1 Key Purposes of EDM

EDM provided pedagogical support with software and knowledge engineering in the development of a model. Romero and Ventura (2012) mentioned the following purposes of EDM:

i. EDM facilitated learners' critical thinking about any phenomenon. Moreover, it provided feedback and addressed the academic objectives of students.

ii. It played its part to improve teaching methodology, behaviours and cognition of learners.

iii. EDM led data mining to explore innovative information patterns. Effectiveness of several contents, teaching and learning methods could be tested with EDM.

iv. EDM guided educational administrators to manage human capital and physical setup to promote academic scenarios.

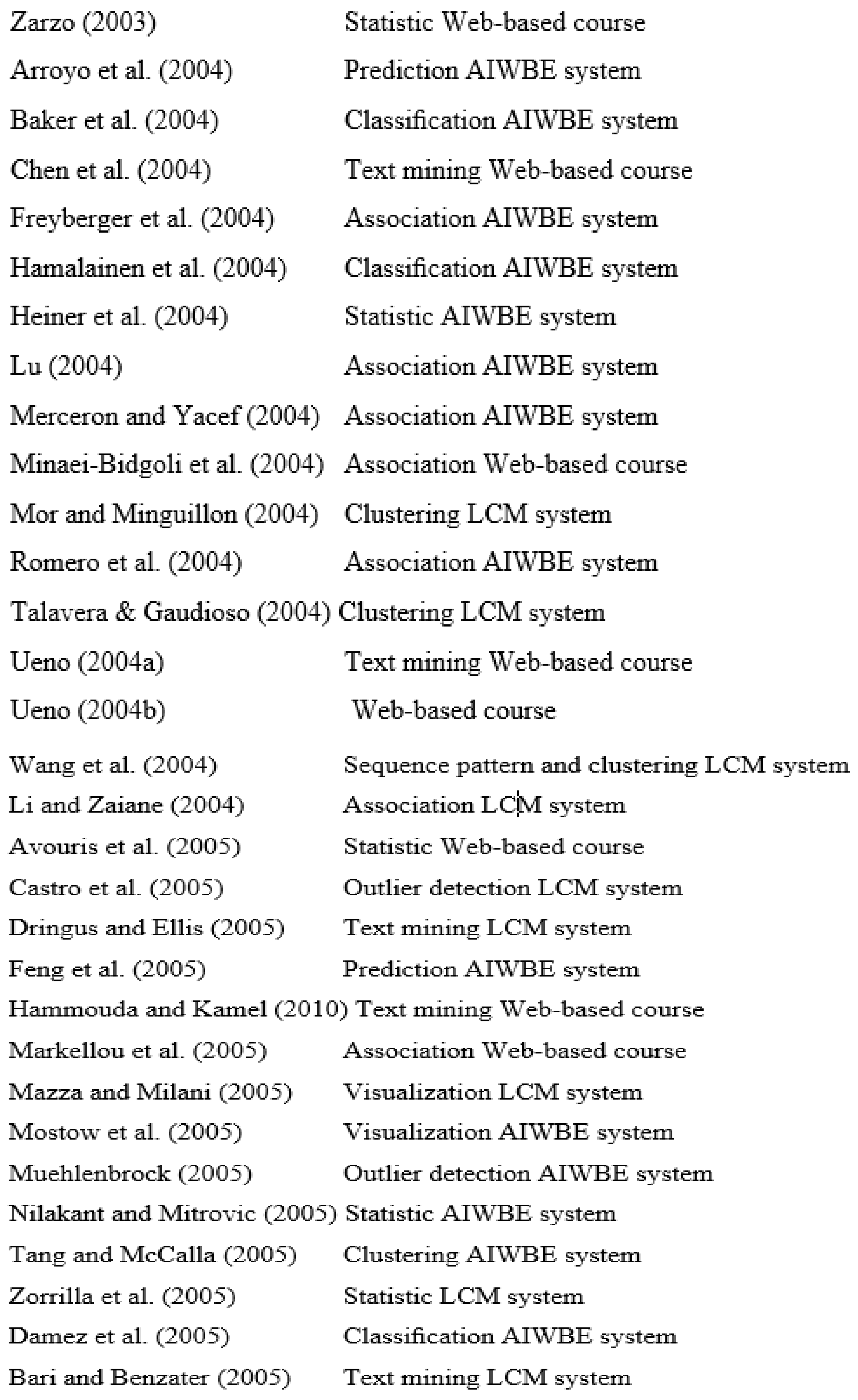
2.12.2 Previous Studies on EDM

Some previous studies widened the horizons of EDM, and explicitly used data mining for academic purposes. Ueno (2004a) suggested the use of DM and TM for learning purposes. Chen, Li, Wang, and Jia (2004) suggested automatic scraping of e-textbooks by deriving target material from websites. Moreover, Tane, Schmitz and Stumme (2004) used clustering and data mining techniques to present similar documents at one place and Hammouda and Kamel (2010) suggested data mining for the study of documents. Therefore, EDM influenced different domains of academia.

Romero and Ventura (2007) studied the period of 1995 to 2005 and reviewed 81 works and among them, seven works were related to text mining. They found data visualization and statistical analyses with DM. Furthermore, web mining was also divided into three parts: a. clustering, classification; b. association rule; c. text mining.

The following list of previous studies showed different works about the application of educational data mining techniques:

Authors	**Mining Task Educational System**
Sanjeev and Zytkow (1995)	Sequence pattern Traditional education
Zaiane et al. (1998)	Statistic and sequence pattern
Beck and Woolf (2000)	LCM systems Prediction AIWBE system
Becker et al. (2000)	Association and classification Traditional education
Chen et al. (2000)	Classification Web-based course
Ha et al. (2000)	Association Web-based course
Ma et al. (2000)	Association Traditional education
Tang et al. (2000)	Text mining AIWBE system
Yu et al. (2001)	Association Web-based course
Zaiane and Luo (2001)	Sequence pattern LCM system
Luan (2002)	Clustering
Shen et al. (2002)	Visualization LCM system
Wang (2002)	Association and LCM system sequence pattern Web-based course
Pahl & Donnellan (2003)	Prediction Traditional education Sequence pattern and statistics
Merceron &Yacef (2003)	Statistic AIWBE system
Minaei-Bidgoli & Punch (2003)	Classification Web-based course
Shen et al. (2003)	Sequence pattern and clustering Web-based course



Zarzo (2003)	Statistic Web-based course
Arroyo et al. (2004)	Prediction AIWBE system
Baker et al. (2004)	Classification AIWBE system
Chen et al. (2004)	Text mining Web-based course
Freyberger et al. (2004)	Association AIWBE system
Hamalainen et al. (2004)	Classification AIWBE system
Heiner et al. (2004)	Statistic AIWBE system
Lu (2004)	Association AIWBE system
Merceron and Yacef (2004)	Association AIWBE system
Minaei-Bidgoli et al. (2004)	Association Web-based course
Mor and Minguillon (2004)	Clustering LCM system
Romero et al. (2004)	Association AIWBE system
Talavera & Gaudioso (2004)	Clustering LCM system
Ueno (2004a)	Text mining Web-based course
Ueno (2004b)	Web-based course
Wang et al. (2004)	Sequence pattern and clustering LCM system
Li and Zaiane (2004)	Association LCM system
Avouris et al. (2005)	Statistic Web-based course
Castro et al. (2005)	Outlier detection LCM system
Dringus and Ellis (2005)	Text mining LCM system
Feng et al. (2005)	Prediction AIWBE system
Hammouda and Kamel (2010)	Text mining Web-based course
Markellou et al. (2005)	Association Web-based course
Mazza and Milani (2005)	Visualization LCM system
Mostow et al. (2005)	Visualization AIWBE system
Muehlenbrock (2005)	Outlier detection AIWBE system
Nilakant and Mitrovic (2005)	Statistic AIWBE system
Tang and McCalla (2005)	Clustering AIWBE system
Zorrilla et al. (2005)	Statistic LCM system
Damez et al. (2005)	Classification AIWBE system
Bari and Benzater (2005)	Text mining LCM system

(Romero, & Ventura, 2007, p. 141)

Baker and Yacef (2009) established an EDM research organization; defined EDM, DM; and evaluated 45 EDM works. They explored outcomes of EDM, for instance, knowledge, effects of learning and student models. In the same year, Pena-Ayala, Dominguez and Medel (2009) reviewed 91 studies regarding EDM, DM and Computer-Based Education System (CBES).

Romero and Ventura (2010) surveyed 235 types of previous research works. Apart from other categories, EDM works were categorized into 11 segments: a. visual data; b. feedback; c. guideline of learners d. learners' performance; modelling for learners; unsuitable behaviours of learners; categorization of students; socializing analysis; mind maps; courseware and planning.

Shu-Hsien, Pei-Hui and Pei-Yuan (2012) reviewed 216 DM works and put them into 9 categories. They suggested three points: a. Social science methods should be integrated into EDM b. Combine some useful tested methods to produce one major EDM methodology c. Formulation of futuristic EDM policy.

2.12.3 Rationale to Use EDM

The question emerged why to use educational data mining with software and computing tools, while human analysis could also be done. The answer to this question was an emergence of current yottabytes (10^{24}) data. Mega tech companies namely Google, Open Content Alliance had digitized millions of books and ProQuest had digitized millions of newspapers (JAH, 2008 as cited in Rieder, & Rohle, 2012). Many databases, repositories and digital archives had been prepared and they cannot be analysed humanly because their reading was beyond human analytical capacity, and many patterns were ignored by human beings. Above all, the whole life was spent to analyse a large corpus even then mistakes were expected. The solution was shifting from close reading to distant reading. So, researchers of different domains had to resort to data mining tools for accurate analyses in the shortest possible time. Previously, 1000 political blogs (Adamic, & Glance, 2005), 54 million Twitter interactions (Cha, Haddadi, Benevenuto, & Gummadi, 2010) and five million Flickr accounts (Prieur et al., 2008) were studied with EDM tools. Big data were precisely encoded in the shape of data visuals, hence, data visualization software could easily handle yottabytes of data in a few seconds. That is why visual data were equal to thousands of words. To conclude, data visualizations had been widely used for the exploration, generation and explanation of big data.

2.13 Text Mining (TM)

Text mining in humanities was introduced after data mining in sciences, that is why it had been presented after EDM. Text mining/ text analytics was named as intelligent text analysis, text data mining or knowledge discovery in a text (KDT) which meant the extraction of interesting and potential knowledge patterns. It explored valuable knowledge patterns and trends which were validated with data evidence. TM incorporated social domains of knowledge, for instance, statistics, maths, NLP, ML and AI.

With manual efforts, those text mining patterns were not easy to locate even after thorough study. New knowledge patterns opened new vistas of research and answered the problematised queries. Text mining incorporated information retrieval, DM, NLP and information extraction (JISC, 2006).

Text data were “sparse and high dimensional” (Aggarwal, & Zhai, 2012, p. 3), and text data were considered as “a bag of words” (Aggarwal, & Zhai, 2012, p. 3). Usually, text mining was restricted to word representations to transform text into numbers and patterns. It showed categorization of dispersed elements into some meaningful and interesting patterns which conveyed some comprehensible and beneficial messages to its viewers.

Supervised learning method trained input data to know about classification. Unsupervised learning methods were used in topic modelling and clustering. Both had close ties since topic modelling formulated clusters. Moreover, transfer learning was used to shift knowledge from one phase to another. Cross-lingual mining of text data addressed information retrieval and cross-comparison of corpora. Different types of data mining studies, for example, text mining, social media mining, opinion mining, multimedia mining (Aggarwal, & Zhai, 2012) sentiment mining, audio mining, video mining, web mining and table mining were also conducted to search knowledge patterns.

2.13.1 Types of Text Mining

Some types of text mining had been discussed below as well as in figure 7.

1. Pre-processing Text: Data should be in free form, and it can be in text or HTML and XML files for an interactive analysis.

2. Text Mining Workflow: Text mining process started from countless unstructured texts and ended on very limited and structured visual patterns. Entity extraction and attribute extractions were done to find various trends and patterns.

3. Text Categorisation: Text categorization required filtering process which selected common features and removed uncommon features.

4. Mining Textified Documents: The process which converted hard file of speech into soft form was named textification. During this process, rhyming words were extracted.

5. Temporal Text Mining (TTM): The time of reading messages was also mined to select the time of digital marketing. Likewise, time of buying some products and coinage of a term for the first time was also searched by TTM method. Through it, important pages could be processed chronologically or reverse chronologically.

6. Distributed Text Mining (DTM): It drew boundaries between different types of data. Sorting out human opinions from the entire world was done through collaborative filtering (Chattamvelli, 2016). Consequently, international research became possible.

2.13.2 Seven Domains of Text Mining

The following seven fields interlinked each other in the process of text mining.

1. Search and information retrieval (IR): It focused on the search of key words from a huge repository. During the search process, Google performed tasks of document matching, search optimization and inverted index.

2. Document clustering: It exhibited document similarity and document clustering tasks from a dataset.

3. Document classification: Document ranking, alert detection and document categorization tasks were also done by forming a group of items.

4. Web mining: As the use of the internet accelerated, the text also changed its forms ranging from text to hypertext, and several other types of media were also integrated into web text. Consequently, a large amount of data was shifted towards websites, web traffic, weblinks, weblogs and they were increasing with every second. Moreover, web text was usually structured with hyperlinks. Web content mining, web structure analysis, web analytics, hypermedia mining, metadata mining and hypertext mining research works were conducted in the domain of web mining. Popular types of web mining were: a. Content of web mining b. Mining of usage and users’ approach or trends c. Mining of layout and design of a website d. multilingual web mining e. semantic web mining. During the process of web mining, innovative, idiosyncratic and meaningful patterns were discovered through clustering, classification, density and trends.

5. Information extraction (IE): It was an initial point of a text mining algorithm. To explore the structured data from unstructured data was the prime tasks of data mining. Entity extraction, relationship extraction and co-reference tasks were conducted in IE. Moreover, it performed a function of term analysis, named entity recognition and fact extraction (JISC, 2006). To conclude, IE explored names and their relationship.

There were some essential features of information extraction. Firstly, entity types and entity words were joined. Secondly, lexical features were also used to extract information. Thirdly, syntactic features of entities or sentences showed relationship, and fourthly, background knowledge was emphasized (Califf, & Mooney, 1999). These relationships have been shown through knowledge graphs generated by Links tool.

In 1995, Named Entity Recognition (NER) recognized named entities from any specialized corpus and classified them. These methods were beneficial for business intelligence, medicine, intelligence agencies and social media analysts. These systems, for instance, TextRunner (Banko et al., 2007) WOE (Wu, & Weld, 2010) and ReVerb (Fader, Soderland, & Etzioni, 2011) were used for knowledge extraction.

6. Natural language processing (NLP): NLP was a mature field in computer science and language technologies to perform certain tasks of tokenization, POS tagging, phrase boundaries and lemmatization.

7. Concept extraction: It inferred semantically homogeneous words or phrases to access the main concept of the content. Collocations/ n-grams, word association and sentiment analysis were used to extract key concepts (Miner, Elder, Fast, Hill, Nisbet, & Delen, 2012, p. 31).

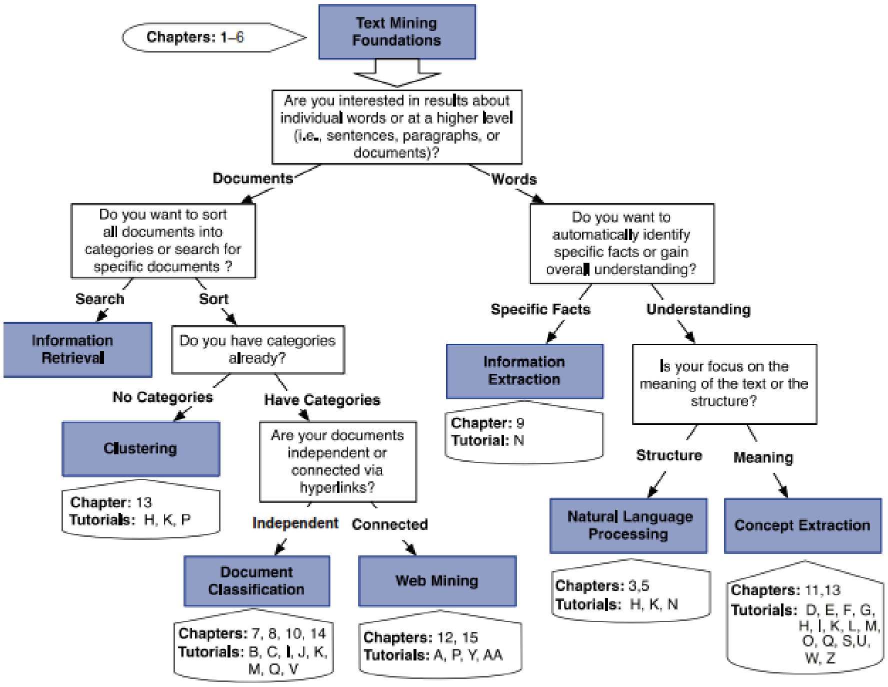

Chapters: 1–6
Text Mining Foundations
Are you interested in results about individual words or at a higher level (i.e., sentences, paragraphs, or documents)?
Documents
Words
Do you want to sort all documents into categories or search for specific documents ?
Search
Sort
Information Retrieval
Do you have categories already?
No Categories
Have Categories
Clustering
Chapter: 13
Tutorials: H, K, P
Are your documents independent or connected via hyperlinks?
Independent
Connected
Document Classification
Chapters: 7, 8, 10, 14
Tutorials: B, C, I, J, K, M, Q, V
Web Mining
Chapters: 12, 15
Tutorials: A, P, Y, AA
Do you want to automatically identify specific facts or gain overall understanding?
Specific Facts
Understanding
Information Extraction
Chapter: 9
Tutorial: N
Is your focus on the meaning of the text or the structure?
Structure
Meaning
Natural Language Processing
Chapters: 3,5
Tutorials: H, K, N
Concept Extraction
Chapters: 11,13
Tutorials: D, E, F, G, H, I, K, L, M, O, Q, S,U, W, Z

Figure 7 Text mining process (Miner, Elder, Fast, Hill, Nisbet, & Delen 2012, p. 33).

2.13.3 Standard Measurement for Text Mining

The prevalent metrics had been given in the following lines.

1. Document Frequency (DF): It meant to count those documents in which the word stem was present, and rare terms were shown in low-frequency columns.

2. Term Variance (TV): It showed the occurrence of a term in a document.

3. Relative Term Variance (RTV): It counted variance of term occurrences.

4. Information Gain (IG): This metric measured absence or presence of an entity in the documents.

5. Mutual Information (MI): It measured sharing and overlapping between two selected documents.

6. X Square Statistics: This metric counted multidimensional categories of data.

7. Term Strength (TS): It counted a term in several documents.

8. FIDF: Frequency meant the occurrence of counting of a term in the selected documents.

9. CNC: This metric analysed link, and it incorporated statistical and linguistic analyses.

2.13.4 Previous Doctoral Research Works on Text Mining

Some previous dissertations on text mining were reviewed in this section. Kongthon (2004) wrote a dissertation titled ‘*A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management’.* A new framework of text mining was suggested to explore valuable hidden patterns in electronic texts. This study elaborated “Technology Opportunities Analysis”. Two algorithms namely “tree-structured networks” and “concept grouping” were applied. The new framework was applied to Thai Science and Technology Abstracts. Moreover, factor analysis and cluster analysis were employed for data analysis. Factor analysis referred to a statistical approach of all variables of the text, and cluster analysis had partitional and hierarchical methods. Two approaches were employed in hierarchical clustering: Agglomerative approach merged clusters, and divisive approach split clusters until a stopping criterion was achieved. This study assisted the decision formation process to find meaningful similar and dissimilar patterns (Kongthon, 2004).

Fiala’s (2007) doctoral research was on ‘*Web Mining Methods for the Detection of Authoritative Sources’* which focused on tightly linked nodes to connect with other web pages. In 1998, Google stepped forth for international connectivity of billions of pages and sites in a few seconds. This dissertation dealt with the modification of the standard PageRank formula which was used for measuring the significance of websites and connectivity of bibliography. New algorithms were applied to Digital Bibliography and Library Project, and two case studies were discussed in it. A major contribution was to discover authoritative research projects from Czech and France through web mining (Fiala, 2007).

In 2015, Gillani’s PhD dissertation was about ‘*From Text Mining to Knowledge Mining: An Integrated Framework of Concept Extraction and Categorization for Domain Ontology’.* Equipping employees with organizational knowledge at the appropriate time and in the most appropriate format was a hallmark of any organization. This dissertation explored the extraction of knowledge patterns, methods of concept extraction and their improvement, ontology enhancement and automatic knowledge extraction for ontological enhancement. This research combined qualitative and quantitative methods and suggested a Promine framework for knowledge extraction and its categorization. Major contributions of this study were bridging process modelling and ontology, concept extraction with Wiktionary, WordNet, concept filtration method with statistics and semantics, concept categorization system and devising of algorithms for word extraction (Gillani, 2015).

2.14 Process Mining and Pre-Processing

Visual information was quick, convenient, synthesising, compressing and informative. Process mining searches tasks, their order and interrelationships with other events. To derive knowledge patterns, advanced level process mining was required. In literature, text mining had been applied on event logs for the extraction and categorization of meanings. Bembenik, Skonieczny, Rybinski, Kryszkiewicz and Niezgodka (2013) also applied TM for the same purpose.

Pre-processing was essential to delimit results corresponding to the required data. Wang (2004) elucidated five approaches: full-text approach, stop words, key word extraction, POS tagging and term extraction. Furthermore, the concept approach extracted semantic concepts. To choose keywords, Hotho, Nürnberger and Paaß (2005) suggested entropy i.e. selection of frequent words from the text. Mathiak and Eckstein (2004) presented text mining steps of tokenization, frequency counting, POS tagging and stemming methods to save time, because a lot of unnecessary data could be kept aside.

Wang (2012) propounded three approaches:

1. Pruned bag of single word approach: It removed common words from the selected documents.

2. Emerging pattern-based bag of single row: Frequency of selected item could be augmented.

3. Bag of frequent item set: A word was considered a single unit, and the most occurring word was called a frequent itemset.

Torunoglu, Cakirman, Ganiz, Akyokus and Gurbuz, (2011) studied the effects of pre-processing on Turkish newspapers. Results showed that the stemming technique was very useful.

1. Tokens: Tokenization took all words, phrases from the textbooks, and its main purpose was to convert the input text into token words, keywords, unique words, occurrence, word boundary and sentence boundary.

2. Stop Words: Stop word filtering was done to delimit a word list, and usually they are function words or thematically useless words. This feature was also present in Voyant tools.

3. POS Tagging: POS tagging was a necessary part of linguistic processing. It was applied in IE, speech synthesis, corpus and term extraction. It usually had 50 to 150 tags with variance in different languages and corpora. Hundreds of POS tags were used in the German language (Voutilainen, 2003). In addition to it, rule-based tagging was also very common. Stochastic approaches (having a random probability) were better than rule-based tagging, and it functioned on the Markov model (Hasan et al., 2007). In addition to them, transformation taggers were combinations of aforementioned both approaches, and they produced reliable results conveniently.

4. Lemmatization: It was a headword in a dictionary, and several graphemes of a word were derived from one lemma. They were filters, and they used statistics or frequency for filtering purposes. This process of finding parts of speech was performed by the implementation of some rules, algorithms and dictionary. The main purpose of lemmatization was to bring the word into its original form by deleting its inflections.

5. Stemming: It cut the last part of a word to find its stem or root word, and it dealt with fixed derivations.

2.15 Knowledge

In philosophy, epistemology meant knowledge of reality; and knowledge meant awareness, acquaintance or understanding of a phenomenon, skill, event or information. Plato defined knowledge in 369 BC in his famous dialogic book named *'Theaetetus'* as a "justified true belief", a true observation or a bona fide judgement with an example event (Plato, 2014). Again, knowledge was a shared part of belief and truth; hence, it demonstrated an evidence through five senses. Likewise, DH laid stress on the collaboration and significance of valid digital knowledge, data visualization and truthful information. The acquired knowledge was equally useful for academic and non-academic purposes (Burdick, Drucker, Lunenefeld, Presner, & Schnapp, 2012).

Knowledge had been classified into four forms:

i. Shallow Knowledge: Information could be searched easily with Structured Query Language (SQL).

ii. Multidimensional Knowledge: Online Analytical Processing (OLAP) was used for searching clusters and information.

iii. Hidden Knowledge: It could be searched through pattern recognition or machine learning algorithms with an economy of words as compared to SQL.

iv. Deep Knowledge: A clue facilitated to search and to dig target knowledge from a database (Adriaans, & Zantinge, 2009).

2.15.1 Difference among Data, Information and Knowledge

Data were organized stores or transcriptions of some known facts. Palace (1996) defined data as the embodiment of texts, digits and facts that could be processed with computers. Its major types were metadata, operational and non-operational data, and in this way, various types of data were accumulated to develop a database. The database was an organized repository of data for speedy and ubiquitous access. The data warehouse was a more refined stage in which a data repository was cleansed, summarized, organized and formatted to curb redundancies (Chattamvelli, 2016). Its three types were operational data, non-operational data and metadata. Operational data were consisting of accounts and bills of enterprises; non-operational data were unchangeable, for instance, student demographics; and metadata were data about data.

Several entities produced data which constructed meaningful information patterns which were also attained with the summarization of data. Extracting useful, meaningful, purposeful and authentic pieces of information was the sole purpose of knowledge discovery in data mining. So, information could be called semantics of collected data, and knowledge was the outcome of verified information. “It takes the form of heuristics, assumptions, associations and models” (Du, 2010, p. 1) which were derived implicitly or explicitly from the data.

2.15.2 Knowledge Discovery Process in DM

Knowledge discovery process underwent six steps: data selection, data cleaning, enrichment, coding, data mining and reporting. The significant point was that these six steps moved to and fro to accomplish data mining tasks. In the first phase, workable data were selected. Secondly, some cleaning processes were applied beforehand, while some processes were applicable only in the case of pollution. Sometimes data were provided at various places, or customers changed their place without informing the stores, or name spellings were misspelt. Therefore, pattern recognition algorithms were applied to cleanse the data. The third phase was of enrichment and behaviour of clients could be assessed with given data. It should be considered how new information was attached with existing information. The fourth phase was coding which was a creative and repeated procedure.

Some common coding steps were: “address to region”, “birth date to age” (Adriaans, & Zantinge, 2009, p. 57). The fifth step was data mining which incorporated query tools, data visualization, case-based learning, association rules, genetic algorithms, statistical techniques, OLAP, decision trees and KGs. Almost 80% of knowledge patterns were obtained from SQL, whereas 20% of knowledge patterns were taken by other advanced tools. Interesting patterns emerged with comparisons of subgroups. The sixth step was writing about findings of data mining with communicative words and interactive visuals (Adriaans, & Zantinge, 2009).

2.15.3 Structures and Benefits of Knowledge Patterns

The rule meant a pattern which occurred with regular intervals, and rule-based methods were learnt manually or automatically. Then the text was juxtaposed with rules. Manual rules were so much laborious, while automatic rules were classified as top-down (Soderland, 1999) and bottom-up (Califf, & Mooney, 1999).

Common structures of knowledge patterns were called “Bateson/Volk meta patterns” (Dixon, 2012, p. 196), and they were found in the shape of spheres, binary and sheets. They were usually used in pedagogy, video games, software and mobile software (Dixon, 2012). Besides, the major benefit of patterns was solidifying data analysis with patterns. They showed links among generated data, data analysis and different signposts in the metadata. Furthermore, analysis of patterns was “an empirical process” (Dixon, 2012, p. 198), since it gave a valid and visual proof to strengthen any underpinning.

2.15.4 Ontology Learning and Knowledge Patterns

Ontology, a branch of metaphysics, was a reality-based concept of being and interrelationship of priori arguments (Blackburn, 2016, p. 116). Ontology was defined as “a formal specification of a conceptualization” (Gruber, 1993). Ontology learning meant the generation of an automatic process to derive knowledge patterns from multiple input sources. George, Vangelis, Anastasia, Georgios and Constantine (2009) pointed out six steps of ontology learning: synonym identification, taxonomic relation identification, non-taxonomic relation identification, term identification, concept identification and rule acquisition.

Maedche and Staab (2004) explicated salient KAON Text-To-Onto system for ontology learning. For ontology learning, a four stepped cyclic OntoLancs model (Gacitua, Sawyer, & Rayson, 2008) was presented: POS tagging and semantic annotation; concept extraction; construction of ontology and dictionary and extracted concepts were linked with bootstrapping ontology which was employed for semi-supervised relation extraction. Nie and Zhou (2008) classified ontology learning into three segments: concept extraction, interrelationship extraction and axioms extraction.

2.15.5 Transformation of a Pattern into a Knowledge Pattern

If a digital or manual pattern was drawn without exact statistics, it was a mere pattern, but an amalgamation of visual and data produced a knowledge pattern. When Voyant or any other tool generated a word cloud/ Cirrus, putting the cursor on any word showed its correct occurrence in the data, it was called a knowledge pattern. Consequently, statistical knowledge changed a simple pattern into a knowledge pattern.

Implication and terseness of a pattern with correct information was a means of knowledge. Knowledge patterns were results of text mining, and they referred to correct "rules, models, statements" (Du, 2010, p. 3). Knowledge patterns assisted to understand information with its language (Alexander, 1979) to transmit knowledge. Furthermore, "Patterns implicitly required narrative" (Berry, 2012, p. 14), and in literature, these patterns built a narrative for comprehension. Literature was written to spread its narratology among masses. Furthermore, data visualizations elaborated informative links among items of a subset. These knowledge patterns facilitated individuals and organizations for formulation of policies and prediction of results.

In this technological era, pattern-focused knowledge pursuit was subservient to Alexandrian study. Alexander (1979) opined that different patterns were required to analyse different research settings, so the Alexandrian pattern catalogued them, and structuralism searched for links in the language.

Peirce propounded abductive reasoning for knowledge generation, thereupon, knowledge was constructed by guesses and rational intuitions in social sciences. Then inductive and deductive reasoning proved the utility of knowledge patterns (Burch, 2010). Being authentic parts of research, knowledge patterns were epistemological. Its validity was determined with its utility and correctness, since patterns were used "as a justifiable knowledge generation and validation technique" (Dixon, 2012, p. 192). To conclude, validity and reasoning strengthened knowledge patterns.

2.15.6 Knowledge Patterns in Data Mining

Mining referred to find precious nuggets from the scattered and large amount of data (Han, Kamber, & Pei, 2012, p. 6). Similarly, data mining was "knowledge mining from data". Any type of data could be mined through data mining techniques. Knowledge discovery process followed 7 steps: i. Data cleaning 2. Data integration 3. Data selection 4. Data transformation 5. Data mining 6. Pattern evaluation 7. Knowledge presentation. In the data mining process, several domains, for instance, statistics, machine learning, information retrieval, visualization, algorithm and pattern recognition were also used (Han, Kamber, & Pei, 2012).

Data mining produced numerous patterns, but a few of them were potentially significant to represent knowledge. A pattern became interesting, if it was easily understandable, potentially useful, novel, valid and statistical to test a hypothesis (Han, Kamber, & Pei, 2012, p. 21). Computers searched various patterns from big data as Google searched different results instantly. Precisely, regex, a regular expression or a sequence of characters, was a pattern which found thesaurus-based synonyms of the target word from a large database. Thus, patterns became comprehensive, systematic and extensive for knowledge discovery.

2.16 Data Visualization for DH

Visualizations were better than mere numeric data because complex patterns and data became easily understandable. Data visualization presented big data in the form of a condensed image. Data were visualized with a simulation which employed a mathematical system or a computer program. A computer model consisted of algorithms and equations to observe the response of a system. By running a programme, a simulation changed into computerized visuals (The Editors of Encyclopaedia Britannica, 2017). Python, its libraries and R language also changed textual data into data visualization. Voyant text mining tools also took help from embedded computer programming to extract useful knowledge patterns.

Text mining communicated the implicit messages from a huge database. Visuals facilitated communication of both meaning and data in interesting interactive visuals. Moreover, data visualization was also generated to serve the purposes or to find the answers of research questions.

Likewise, in EDM, a distillation of data for human judgment used “information visualization methods” (Barahate, 2012, p. 13).

Moretti (2005) used a visual learning technique for the study of historical novels. Moreover, charts, word clouds, timelines, graphs were also used (Jänicke, Franzini, Cheema, & Scheuermann, 2015) for long lasting learning and teaching purposes. Similarly, topic modelling was applied in print media, scientific research work, social media and digital libraries (Blei, Ng, & Jordan, 2003), and search of key topics with clustering was also used for classification of similar themes (Scrivner, & Davis, 2017).

Breadth-First Search algorithm was used to locate the briefest path between different characters, places, titles, organizations, thus interrelationship was studied in the datasets (Barabási, & Lászlo, 2002). Besides, the interrelationship of 5 Victorian novels was analysed with Mandala browser (Brown, Ruecker, Radzikowska, Milena, Patey, Sinclair, & Antoniuk, 2009). In conclusion, Voyant tools exhibited different types of data visualizations to serve different purposes.

2.16.1 Various Patterns of Visual Data

Visualization revealed different types of patterns, for instance, word clouds and KGs, but it is essential to select the most appropriate visual to serve a specific purpose. Line charts manifested trends; multiline charts showed different categories of data; step charts exhibited granularity; scatter plots showed the co-relationship; heat grid displayed relationship; bubble charts presented the existence of different categories in the data; and treemaps exhibited sharing of hierarchical data. Besides, colour coding was also used for categorization, divergence and sequence.

2.16.2 Data Visualization Categories

Tables and graphics were also parts of data visuals.

1. Tables: Tables could be drawn in structured forms with rows and columns to show numeric or non-numeric cell value data, for instance, corpus and collocations/ n-grams were shown in tabular form in the current study.

2. Graphics: Graphics presented data in a picture form. Development in parent data changed graphics automatically, hence, interactivity enhanced the value of data visualization. An effective graph must be vivid, clear and terse for quick comprehension of knowledge pattern. Furthermore, Graphical Processing Units (GPU) were those chips which visualized data efficiently.

2.16.3 One Variable Diagrams

One variable diagram communicated a univariate piece of information, and they were drawn digitally.

1. Line Charts: They exhibited data with consecutive intervals, but variations and graphs could also be shown. They facilitated the search for trends and changes in a certain time frame.

2. Bar Charts: X and Y axes of bar charts were drawn to show their certain categories and data. A segmented bar chart was subdivided to show other variables.

3. Histograms: It showed counting of frequency in a pictorial form, so a reasonable number of classes, equal width of the class and continuous intervals were qualities of a histogram.

4. Pictogram: If the frequency was converted into a picture or an icon, it was called pictogram.

5. Time Charts: Time of occurrence of a phenomenon was shown with time charts.

6. Temporal Histograms: One axis showed time, and the other axis displayed entities.

7. Spatial Histograms: These histograms showed space in the world as an international publishing company showed the location of its customers on the world map.

8. Pareto Diagrams: Spikes of Pareto diagrams were arranged in a descending sequence. Thus, the shortest and longest spikes were placed on the right and left side.

9. Pie Charts: It showed a proportion of the data in a 360-degree circular form.

10. Radar/Polar/ Spider Chart: Analysis of two or more categories started from the centre to the outer side.

11. Frequency Polygons and Frequency Curves: Frequency polygon was made by linking top points of histogram bars with a straight line. Frequency curves focused on the division of sample points to evaluate the normality or abnormality of the parent population.

12. Stem and Leaf Plots: They dealt with those small numeric sets which had common attributes. Leading digit formed a group to keep all parts of a set together, so the stem was placed in the left column, while leaf was positioned on the right side.

13. Overlay Charts: One chart overlaid the other chart, for instance, the line chart was easily overlaid on bar charts.

2.16.4 Multi-Variable Diagrams

They were made with a combination of two or more variables and they had been discussed in the following lines.

1. Scatterplot/ Scattergram: Two or more variables were graphically represented through scatterplot. Star, dot or x symbols were used for each type of data to show the presence of data. These symbols showed the strength of variables in a scatterplot.

2. Bubble Chart: Multi-coloured bubbles represented different variables, and their quantity and presence had been shown with bubbles.

3. Contour Plots: They used colour-coded shades or lines to display heterogeneous data, for instance, epidemics in different regions of the world had been visualized through contour plots.

4. Quantile- Quantile (Q-Q) Plots: Before analysis, parent data were normalised. Q-Q plots visually checked normalcy of the parent population and displayed it on X and Y axes.

5. Chernoff Plots: They visualized emotional and facial attributes in different data trends.

6. Box and Whisker Plots: They had a box and whiskers to show different data segments.

7. Stem Plots: Stem plots showed outliers from the data. Outlier detection was the initial step for any statistical analysis, because the outlier detection algorithm searched anomalies from datasets.

8. Hierarchical Charts: Data were arranged according to a hierarchy of values, for example, family tree, decision tree, organizational hierarchy setup.

9. Polar Trees: They were drawn with tree nodes of circles, while the root was kept in the centre.

10. Cause and Effect Diagrams: They were named as fish-bone diagrams, whereas the spine was a central arrow of the diagram. They demonstrated multifaceted causes of an issue. It had 4M (methods, materials, machines, people) and 4P (policies, procedures, plans, people) approaches (Chattamvelli, 2016).

2.17 Textbook Analyses with Human Computing

Textbook analysis started from the 1980s but usually, they followed manual content analysis and different criteria were compiled for textbook analyses (Chambers, 1997; Cunningsworth, 1984, 1995; Sheldon, 1988; Williams, 1983). Pedagogical methods were derived from the study of textbooks (O'Neill, 1993; Ranalli, 2003; Swales, 1995). Textbook analysis concentrated on academic activities for the classroom (Jacobs & Ball, 1996). Major drawbacks of these studies were human counting of token words, concordance. Therefore, several researchers were reluctant to perform hectic manual tasks of word counting. To facilitate human beings, corpus and text mining tools were introduced.

2.18 Previous Studies on PTB Intermediate English Books

Various theses had been written on analysis of intermediate English textbooks in Pakistani universities, and they had been reviewed here. Old syllabus of intermediate English books has been analysed and five lesson plans have been designed to teach those lessons (Basharat, 2004). Moreover, the novel '*Goodbye Mr. Chips'* as well as Book 2 of intermediate have been evaluated with a user study approach. Five lesson plans have been prepared for classroom teaching (Iqbal, 2011). Another PTB intermediate level study was conducted to evaluate the effectiveness of poetry while teaching English grammar. In this user study, five lesson plans from five different poems were designed to teach parts of speech from poetic lines (Naser, 2012). This study had severe drawbacks, because poets have poetic license to violate grammatical rules, and sometimes, several poetic lines make one sentence. Another comprehensive study was conducted to explore gender related themes and characters through the FAWE framework of Kabira Masinjila. It was a manual counting and content analysis of themes and characters, then they were presented in comparative tables. It found that male characters and roles were more dominating than feminine characters and roles in intermediate English textbooks (Hussain, 2009). Its major shortcoming was that its themes were not counted. Secondly, the lessons which were not having both genders, they were totally ignored.

To conclude, all works were evaluated manually, and none of them used digital humanities tools to extract knowledge patterns, therefore, this gap was prominent and the current study filled this niche with Voyant text mining study. Another drawback was that most of the studies restricted their scope to five lesson plans, but no work covered all intermediate syllabus of 1st and 2nd year. The current study covered all lessons in detail, therefore, it was the most comprehensive which used five text mining tools.

2.19 Earlier Corpus-Based Studies of EFL Textbooks

Johns (1991) worked on data-driven learning which pioneered the use of corpus in language teaching and learning. As a linguist explored language variations and data from the corpus, similarly a student should also interact with data for exploring the original language corpora. Johns (1991) focused on this trend of learning linguistic patterns through corpus principles: "identify, classify, generalise". Later Leech (1997, p. 10) "invites the student to obtain, organize and study real-language data according to individual choice" and it enhanced autonomous learning.

The use of corpus had been proved a very beneficial source for learning and teaching any language. "Corpus was a useful tool for language learners and teachers" (Meunier, & Gouverneur, 2007, p. 153). The occurrence of words and collocations/ n-grams were vital in corpus studies. Biber, Johansson, Leech, Conrad and Finegan (1999, pp. 992-993) noted that five, four- and three-word lexical bundles were used with a ratio of 1:10:100 in one million words textbooks corpus.

Burnard and McEnery (2000), Sinclair (2004), Connor and Upton (2004) worked extensively on the use of corpora in TEFL. Botley, McEnery and Wilson (2000) emphasized the need of multilingual corpora for research and academic potential of learners. Granger, Hung and Petch-Tyson (2002) developed a deep-seated association between second language learning and corpora. Mukherjee and Rohrbach (2006); and O'Keeffe, McCarthy and Carter (2007) compiled native and second language learners' corpora, and used them for language learning and teaching.

Later, the research trends shifted to the corpus-based study of textbooks. Some other textbook studies dealt with grammar aspects and lexical bundles (Biber, Conrad. & Cortes, 2004; Gabrielatos, 1994; Koprowski, 2005; Meunier, & Gouverneur, 2007; Romer, 2004a, 2004b). Some textbooks studies highlighted the specific register and their use in English for Academic Purposes (Biber, Conrad, Reppen, Byrd, & Helt, 2002; Paltridge, 2002; Swales, 2002), and some studies dealt with phraseology in the textbooks (Biber, Conrad. & Cortes, 2004; Gouverneur, 2008; Koprowski, 2005; Meunier, & Gouverneur, 2007). These six types of research applied automated methods for textbook analyses (Anping, 2005; Biber et al., 2004; Chujo, 2004; Gouverneur, 2008;

Meunier, & Gouverneur, 2007; Romer, 2004b, 2006), and they used already published corpora in their textbook studies. The current study introduced an emerging trend of digital text mining of intermediate English textbooks with corpus, text mining and data visualization.

Biber, Conrad, Reppen, Byrd and Helt (2002) built the first textbook corpora named the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus). It had 27 million words of spoken and written discourses in American universities. They worked on lexical bundles to show the relationship between classroom discourse and textbook discourse. Second textbook corpus in the world was named German English as a Foreign Language Textbook Corpus (GEFL TC) which was compiled in Germany by Romer (2004a). It had 100,000 words. She conducted two studies, and found that spoken English and content of textbooks varied in the use of progressives and modal auxiliaries. Later, it resulted in the incorporation of real linguistic expressions in textbooks to introduce natural language.

Chujo (2004) took lemmatized word lists from English and ESP textbooks to compare them with the wordlist from the British National Corpus (BNC). Anping (2005) built a corpus of 100,000 words from foreign and Chinese TEFL books. He conducted corpus-driven studies with the comparison of lexis and semantic tagging.

Meunier and Gouverneur (2007) and Gouverneur (2008) built textbooks corpora named TeMa Corpus of 700, 0000 words to study phraseology of two adjectives. Gabrielatos (2005, p. 5) named textbook corpora “pedagogic” or “pedagogical”. Willis (2003) coined the term pedagogic corpus for academic needs, and Hunston (2002, p. 16) defined it thus: “a corpus consisting of all the language a learner had been exposed to. It could consist of all the course books, readers etc a learner had used, plus any tapes etc they have heard.” Delimiting this extensive definition of Hunston (2002), “The examples of textbook corpora could be referred to as pedagogic corpora as they all consisted of representative samples of textbook data intended for the teaching of EFL” (Huntston, 2010, p. 186).

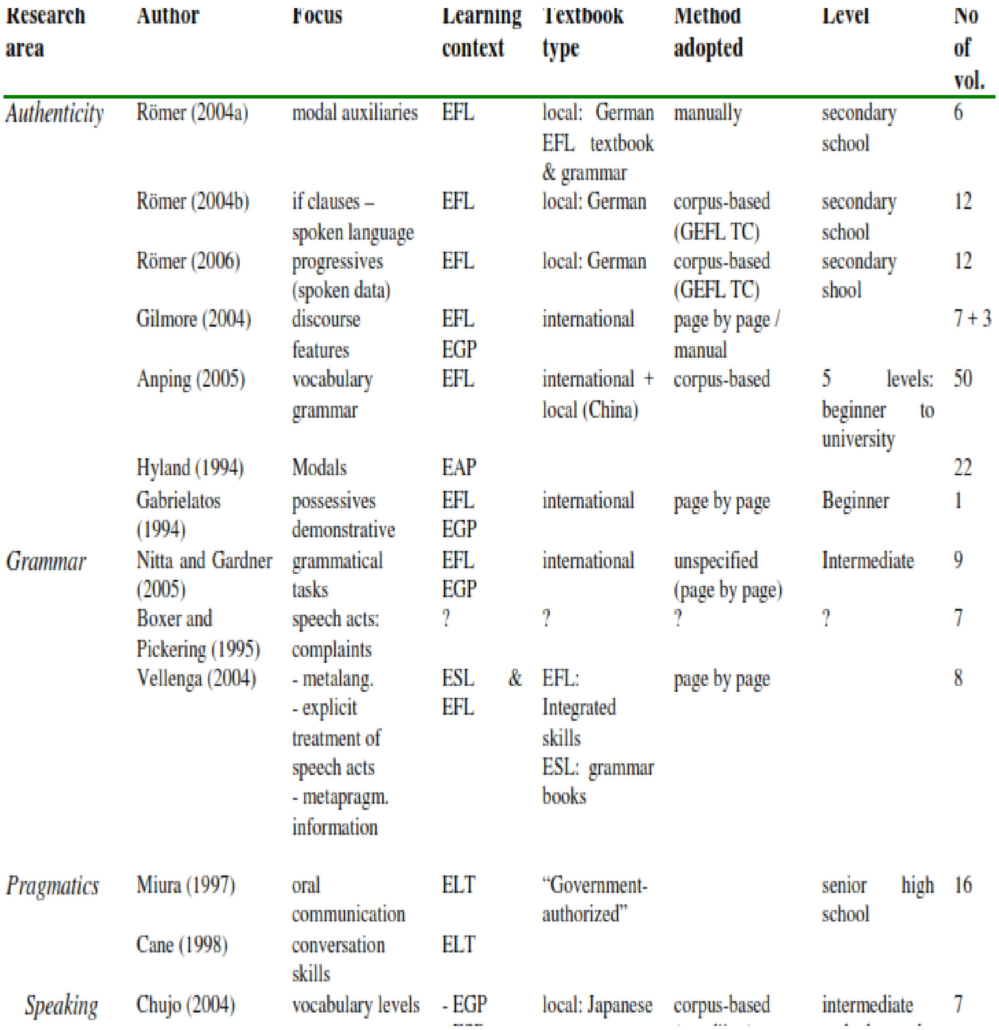
Hunston (2002, p. 16) directed to two uses of pedagogic corpora: firstly, to raise awareness about textual contents, words and their usages as Biber et al. (2004) did. Secondly, the newly built corpora of textbooks could be compared with other authentic textbook corpora as Romer (2004a, 2006) did.

Biber et al. (2004) worked on the university register through textbooks as well as teaching, and he presented lexical bundles with a frequency-driven technique. His major finding was that classroom teaching procedures employed more lexical bundles as compared to textbooks. Previous research works used several other words for lexical bundles, for instance: formulae, routines, prefabricated patterns, lexical phrases and fixed expressions. Weinert (1995) had pointed out two ways of research: one was to specify lexical bundles, and the second was to analyse the discourse function of the selected lexical bundles. To follow these ways, Biber et al., (2004) explored functions of lexical bundles, and he found 84 lexical bundles during teaching, 43 lexical bundles in the conversation of teachers and students, 27 lexical bundles in textbooks, and 19 lexical bundles in the classroom academic prose. Lexical bundles exhibited certain beliefs of discourse articulators (epistemic stance bundles, attitudinal stance bundles), frequent references (identification bundles, imprecision bundles, bundles specifying attributes, deixis bundles) and discourse formation (focus bundles/ topic introduction bundles, topic elaboration bundles). Among them, referential bundles were used more in teaching than textbooks.

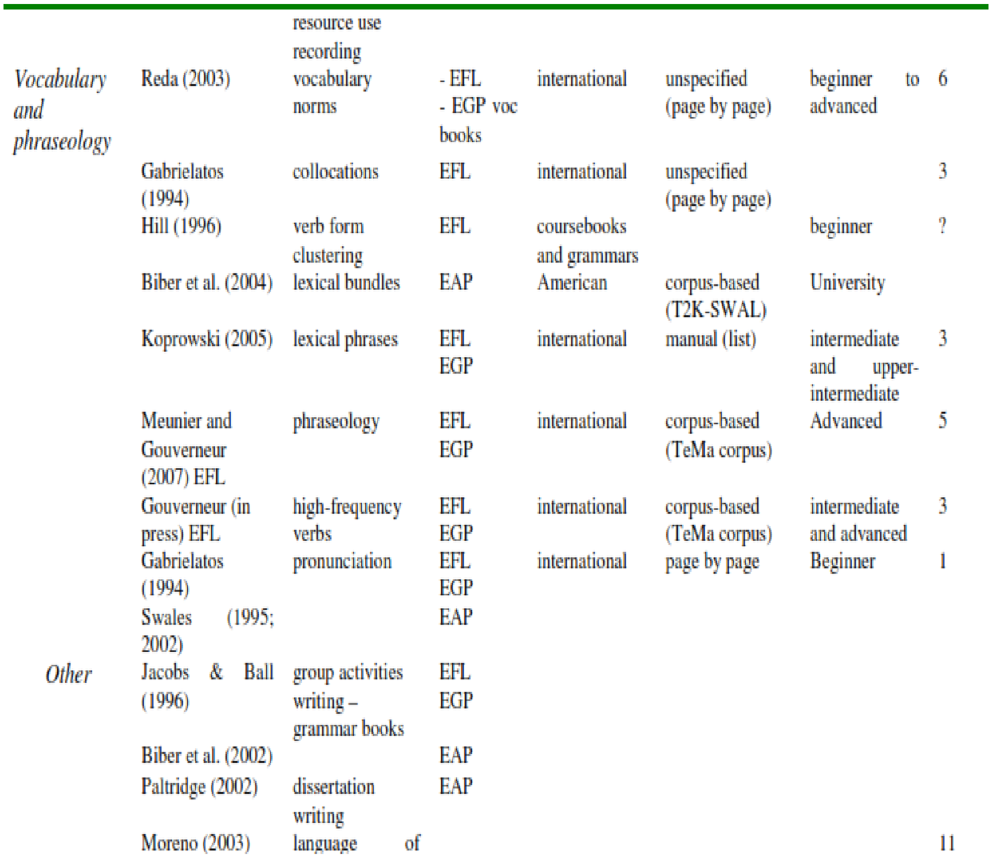
Romer (2006) took 100 high-frequency progressive forms from spoken British English Corpus, since Germans felt difficulty in learning progressive forms of English. The inquiry was how progressive forms (to be+looking) were dealt differently in English textbooks and in daily life. A comprehensive functional and contextual study was carried out with a data-driven corpus approach. For this study, three major corpora, for instance, British National Corpus of 10 million words, Spoken Bank of 20 million English words, and GEFL TC of 100,000 words were exploited by Romer. Analysis of this study covered present progressive, past progressive, present perfect progressive, past perfect progressive, progressive infinitive and modal auxiliaries, therefore, they were compared manually in the study. The prime aim was to explore significant linguistic patterns around the progressive form. It found that textbooks were not utterly matching the naturally occurring spoken corpus. Furthermore, EFL textbooks did not keep common lexical entities of daily life in them. It was strongly recommended that EFL textbook writers must use frequently occurring authentic linguistic items in textbooks, so that the real scenario of language learning and textbooks should support each other. Similarly, Beaugrande (2001) emphasized the same notion that textbooks could become more beneficial, if they covered authentic patterns of native English.

Gouverneur (2008) studied current phraseological uses of two verbs “make” and “take” with qualitative and quantitative research methods in the TeMa corpus of three English textbooks. Two major findings were: There were clear differences in the phraseology of verbs “make” and “take”. Three selected textbooks had similarities in their pedagogical choices; usually, delexicalized verbs ("get" in the examples, get out, get married) were not taught through English textbooks (Sinclair, & Renouf, 1988). This study also found that both verbs were present in three textbooks. Moreover, several exercises were devoted to teaching collocation patterns/ n-grams regarding verbs "make" and "take". Only a few collocation patterns/ n-grams of one textbook resembled the other two textbooks.

In the last two decades, several researchers analysed ESL textbooks through corpus studies and they had been mentioned in figure 8.



Research area	Author	Focus	Learning context	Textbook type	Method adopted	Level	No of vol.
Authenticity	Römer (2004a)	modal auxiliaries	EFL	local: German EFL textbook & grammar	manually	secondary school	6
	Römer (2004b)	if clauses – spoken language	EFL	local: German	corpus-based (GEFL TC)	secondary school	12
	Römer (2006)	progressives (spoken data)	EFL	local: German	corpus-based (GEFL TC)	secondary shool	12
	Gilmore (2004)	discourse features	EFL EGP	international	page by page / manual		7 + 3
	Anping (2005)	vocabulary grammar	EFL	international + local (China)	corpus-based	5 levels: beginner to university	50
	Hyland (1994)	Modals	EAP				22
	Gabrielatos (1994)	possessives demonstrative	EFL EGP	international	page by page	Beginner	1
Grammar	Nitta and Gardner (2005)	grammatical tasks	EFL EGP	international	unspecified (page by page)	Intermediate	9
	Boxer and Pickering (1995)	speech acts: complaints	?	?	?	?	7
	Vellenga (2004)	- metalang. - explicit treatment of speech acts - metapragm. information	ESL & EFL	EFL: Integrated skills ESL: grammar books	page by page		8
Pragmatics	Miura (1997)	oral communication	ELT	"Government-authorized"		senior high school	16
	Cane (1998)	conversation skills	ELT				
Speaking	Chujo (2004)	vocabulary levels	- EGP	local: Japanese	corpus-based	intermediate	7



		resource use					
Vocabulary and phraseology	Reda (2003)	recording vocabulary norms	- EFL - EGP voc books	international	unspecified (page by page)	beginner to advanced	6
	Gabrielatos (1994)	collocations	EFL	international	unspecified (page by page)		3
	Hill (1996)	verb form clustering	EFL	coursebooks and grammars		beginner	?
	Biber et al. (2004)	lexical bundles	EAP	American	corpus-based (T2K-SWAL)	University	
	Koprowski (2005)	lexical phrases	EFL EGP	international	manual (list)	intermediate and upper-intermediate	3
	Meunier and Gouverneur (2007) EFL	phraseology	EFL EGP	international	corpus-based (TeMa corpus)	Advanced	5
	Gouverneur (in press) EFL	high-frequency verbs	EFL EGP	international	corpus-based (TeMa corpus)	intermediate and advanced	3
	Gabrielatos (1994)	pronunciation	EFL EGP	international	page by page	Beginner	1
	Swales (1995; 2002)		EAP				
Other	Jacobs & Ball (1996)	group activities writing – grammar books	EFL EGP				
	Biber et al. (2002)		EAP				
	Paltridge (2002)	dissertation writing	EAP				
	Moreno (2003)	language of					11

Figure 8 Corpus Analysis of Textbooks (Meunier, & Gouverneur, 2010, pp. 3-4).

2.20 Former Corpus Driven Research Works in Textbooks

Corpus tool in Voyant produced total words, unique words, vocabulary density, average words per sentence from the text, so it was necessary to discuss some previous studies in the domain of corpus.

The selected textbooks were structure-centred (Chalker, 1994) to teach some linguistic structures. Corpus facilitated educationists to comprehend the existing gap between textbooks and real-life language. If there were wide gaps, textbooks should be adapted to meet learners' academic, linguistic, and social requirements. Future course books should be influenced by corpus (Hunston, & Francis, 1998) for upgradation and matching natural language structures.

The following research works found incompatibility of textbooks with corpora; for instance, Sinclair and Renouf (1988) worked on “delexicalized” verbs for example, “do, make, take” which were not taught in English textbooks. Going beyond limited verb study, Willis (1990) worked on discourse structure and complexity of the noun phrase, and he found these features very frequent in corpora, but very least in English textbooks.

Including grammatical categories, Willis (1994) researched indirect speech, passive voice and second conditionals to explore that textbooks should also incorporate these grammatical aspects in the textbooks. In the next year, Carter and McCarthy (1995a, 1995b) studied ellipsis, dislocation and topicalization from CANCODE corpus. They found a list of common features in spoken English.

Fox (1998) studied vague language in phrases, but such phrases were not found in classroom discourse. It showed deviation from real language situations. After a couple of years, Conrad (2000), as well as Conrad and Biber (2000), explored register and adverbial stance markers. They found that grammatical patterns differed from other genres of academic writing and journalism.

2.20.1 Results of Corpus Driven Research Works

The influences of the modelling approach and such corpus-driven studies had started to appear in academia that famous English textbook series *'Headway'* started to include exercises of delexicalized verb exercises. *'Cutting Edge'*, a course of Longman series, introduced reported speech (Cunningham, & Moore, 1999 as cited in Ranalli, 2003). Five volume *'Innovations'* Natural English series by Hugh Deller incorporated vague words in it (Dellar, & Hocking, 2000). Furthermore, lexicon upgradation started by taking help from various corpus studies. These were the practical contributions of the corpus, concordance and text mining of ESL books. There was rationality to adapt books according to real exposure to language. On the other hand, Widdowson, Cook and Owen were against this corpus-led changes in ESL textbooks (Widdowson, 2000), but textbooks should have presented natural language which had been mentioned in different corpora.

2.21 Data Mining of Textbooks

Indian textbooks have been studied diagnostically to improve linguistic and content quality with data mining techniques. Stanford POS Tagger was used to assign a part of speech to each sentence. Errors were corrected with the use of WordNet. Microsoft Web NGram service was also used to prune phrases. This study pointed out weakly written segments of textbooks for further refinement (Agrawal, Gollapudi, Kannan, & Kenthapadi, 2012). This study was quite technical with algorithms; therefore, social science students could not conduct it, because the proposed algorithm detected weak areas of the textbooks.

Indian textbooks had been augmented and improved with data mining techniques. The book corpus was prepared with 17 online available textbooks to enrich textbooks with algorithms. This study aimed to enrich textbooks for dispersion of key ideas, syntactic complexity of writing and image mining (Agrawal, 2013). The main gap in this study was that the selected books were not specifically English textbooks; rather there were mixed books including social science, commerce, science. Another deficiency from a linguistic point of view was that separate language issues had not been discussed, but an algorithm had been proposed that was a good contribution from the perspective of computer science.

2.22 Texts and Poetry Analyses with DH Tools

Different textbooks, pieces of prose and poems had been analysed with various DH tools. They had been reviewed in this section.

2.22.1 Japanese EFL Textbook Analysis with Posit Tool

The Posit Text Profiling Toolset (Weir, 2007) analysed several features of textbooks, including Parts of Speech (POS) tagging, token words, number of all parts of speech, number of sentences, frequency, N-Grams and concordance. Three Japanese EFL textbook series had been analysed and compared with one another through Posit tool (Weir, & Ozasa, 2010), and some of its features showed similarity with Voyant tools. One deficiency was that Voyant tools were unable to do POS tagging of the text.

2.22.2 Text Analysis with Wordseer Tool

WordSeer was a digital tool for literary text analysis. Muralidharan and Hearst (2012) claimed that it was better than the MONK and Voyant tools. WordSeer, a sense-making tool, searched grammatical items, their context, visuals, comparison and contrast of different sources and different utterances of male or female characters. Shakespeare’s language was analysed with WordSeer tool, and it was found that he used physical language for feminine characters and sentimental language for male characters during the discussion of love (Muralidharan, & Hearst, 2012). Thus, WordSeer tool explored feelings, cognitive abilities and implicit knowledge patterns.

2.22.3 Analyses of Poems with Digital Tools

Poemage, a data visualization tool, exhibited sonic typology of a poem. It facilitated readers to create new poems. It had three segments: on the left side, group of words related with sonic; in the middle, the connection of words with text based on sonic; and sonic typology had been visualized in the extreme right side (McCurdy, Lein, Coles, & Mayer, 2016). Another tool, ProseVis (Clement, 2012) also facilitated its users to find sonic patterns of the poem.

Other tools, for instance, GistIcons (DeCamp, Frid-Jimenez, Guiness, & Roy, 2005), Docuburst (Collins, 2006), Compus (Fekete, & Dufournaud, 2000), and Galaxies (Wise, Thomas, Pennock, Lantrip Pottier, Schur, & Crow, 1995) were also used to create a visualization, synopsis, semantic study and extraction of thematic patterns from poems.

Word clouds summarized the themes of the text. To generate word clouds, Wordle tool (Viegas, Wattenberg, & Feinberg, 2009) and TextArc tool (Paley, 2002) were utilised. Tag Clouds (Collins, Viegas, & Wattenberg, 2009) visualized data with a frequency of words. Some other tools FeatureLens (Don, Plaisant, Zheleva, Gregory, Tarkan, Auvil, Clement, & Shneiderman, 2007), Phrase nets (Ham, Wattenberg, & Viegas, 2009), Arc Diagrams (Wattenberg, 2002), The Word Tree (Wattenberg, & Viegas, 2008), WordSeer (Muralidharan, & Hearst, 2012) tools exposed interesting knowledge patterns from the text. PoemViewer tool (Abdal-Rahman et al., 2013), exhibited assonance, rhyme, internal rhyme and the alliteration of a poem. Furthermore, Myopia tool (Chaturvedi, Gannod, Mandell, Armstrong, & Hodgson, 2012) decoded meter, emotion and personification from poems.

2.22.4 Interactive Text Mining Suite (ITMS)

Interactive Text Mining Suite (ITMS), a corpus and its visualization toolkit, was used in literary analyses, and it linked close reading and distant reading. A specialized corpus was made by compiling 1000 source languages and translated lines from *'Romance of Fleminca'* in an experimental study. The word cloud showed that verbs and pronouns dominated in the English translation, while common nouns and proper nouns were least used. The word "not" was found frequently, and it referred to jealousy in Fleminca's husband. It also unveiled that the original text used more punctuation marks than the translated text (Scrivner, & Davis, 2017). These features had been included in Voyant tools for conducting such comparative studies.

2.23 Voyant Tools

Voyant was the most effective text mining tool which was specially designed to create ease even for social scientists who were unaware of different file formats, Python and R programming languages. Voyant open-access tools eased uploading of data in any form, for instance, weblinks, pdf and word file. Furthermore, 25 tools of visual, grid and corpus were available for presenting data visualization interactively.

2.23.1 Principles of Voyant Tools

Voyant tools present interactive data as a piece of evidence in order to interpret any type of text hermeneutically. HyperPo and Taporware data mining software analyse small data, but Voyant tools process big data too. Thus, their capacity and speed are far better than other tools. Voyant accumulate all the following characteristics in it.

1. Modularity: Digital tools must be adaptable according to different configurations.

2. Generalization: They should have the ability to process all types of texts. Voyant tools support several languages including Urdu, Punjabi etc.

3. Domain sensitivity: They fulfil specific intellectual requirements of text analytics.

4. Flexibility: The tools are capable of performing various functions in different formats.

5. Internationalization: They support significant languages of the world.

6. Performance: They perform the text mining process efficiently and accurately. Moreover, Voyant tools do not require any crawler or any file transfer requirement as Antconc tool does.

7. Separation of concerns: In one glance, it shows five different skins or panels for different types of analyses, for instance, Summary, Cirrus, Phrases, Links and Contexts.

8. Extensibility: It must accept new challenges either in the improvement of existing tools or the creation of new tools, and Voyant possesses the quality of extensibility.

9. Interoperability: They should be able to connect to other users' APIs.

10. Skinnability: Digital tools must be divided into different skins to accomplish researchers' different demands simultaneously, and Voyant showed five skins at a time.

11. Scalability: Digital tools must be equally adept for small or large corpora.

12. Simplicity: Text analysis tools must be simple, since their users are from humanities and unaware of different technicalities of programming and files.

13. Ubiquity: They are able to be embedded anywhere on the web.

14. Referenceability: These tools and their results can be presented with authentic evidence (Sinclair, & Rockwell, 2015a).

2.23.2 NLP Tools Vs Voyant Tools

To address emerging challenges of NLP, GATE, Sketchengine, R language, Python language, Python libraries, Wordsmith tools, TAPoR, WebNLP, Python NLTK and countless other sophisticated tools had been introduced. IT experts felt ease to operate these tools easily, but many new entrants from social sciences faced difficulties in data entry, but Voyant tool was very flexible and user friendly, because it accepted all types of data in pdf, word, copy-paste contents, pasting links of websites for automatic data crawling and single entry or multiple entries at the same time. Voyant tools were the most user-friendly tools, because they exhibited all qualities of easy processing. Besides, Python NLTK (Bird, 2006) analysed text and produced visuals, but its major failing was that it demanded practical knowledge of Python which social scientists did not have. Likewise, Commandline caused difficulties, whereas Treetagger tool (Schmid, 1994) did not visualize the data. Voyant showed 25 interactive visuals, but it lacked lemmatization and POS tagging (Radzikowska, Ruecker, & Sinclair, 2011). Addition of thesaurus to consult meanings for each word could enhance its usage for language learners in the future. Another off the shelf tool, WebNLP was very user-friendly for social science data miners. It displayed the option of wordclouds, bubblelines, frequency lists, collocations/ n-grams and scatter plot for data visualization.

2.24 Previous Research Works with Voyant and Similar Tools

The use of five Voyant tools had been mentioned in the following subheadings.

2.24.1 Previous Literature on Summary/ Stylometry

Corpus played a central role in Summary tool, and its significance was evident with the quote that “Interpretive analysis could learn from and build on corpus linguistics, stylistic analysis” (Rockwell, & Sinclair, 2016, p. 19). Stylometry was a computational stylistic analysis with corpus, and it presented quantified data to be discussed. Total words, unique words, vocabulary density, average words per sentence, and the most frequent words were the results of corpus and stylometric analysis.

Some other stylistic qualities were also declared as a benchmark, for instance, an average length of sentences, unique words, total word count and vocabulary density by dividing unique words with total word. To incorporate these features in stylistic qualities, Summary tool was used for the extraction of stylometry.

One dimension was that human efforts of quantification were very laborious and time-consuming, on the other hand, computer usage in academia saved human beings from this tiresome act of word counting. Therefore, computers were used to quantify text (Stamatatos et al., 1999) for text mining. Now, the work of Roberto Busa’s whole life could be done in a few minutes with computers. Therefore, supercomputers, different problem-solving algorithms and trained models could mine text quickly and accurately (Argamon et al., 2003; Zhang et al., 2002).

Some other studies also concentrated on content analysis to reveal stylistic characteristics (Krippendorff, 2003), still it showed a deficiency of fluidity. The age had shifted to precise, mathematical and quantitative findings which were more preferable than qualitative approaches. Consequently, quantification of stylometry was incorporated in the current study.

There were different literary controversies about the authorship of different literary and non-literary works. It could be the case of the Shakespeare-Marlowe authorship controversy or Ayesha Gulalai’s blame of obscene text messages in Pakistan. These controversies could be fixed by analysing the stylistic qualities of both texts through the Summary tool, as some earlier studies had already done (Stamatatos et al., 2000; Van Halteren et al., 2005).

Many studies were conducted to assess the stylometric features of different literary and non-literary works with DH tools. In the beginning, word lengths were determined as a stylistic quality of a writer, and they were shown with histograms (Malyutov, 2006).

Another aspect was finding the age of the text to reveal stylistic qualities of a particular age, and this study fell in the domain of “stylochronometry” (Stamou, 2008) as the carbon 14 test explored the age of any object through rays. The summary tool also informed about the date of uploading and creation of corpus. Another method of knowing stylochronometry was to compile prominent themes of the text and to search their etymology to determine the period of selected vocabulary items; for instance, Bacon used "hath" (has/have) to denote Renaissance age. Likewise, the etymology of words "Aprill" (April) and "soote" (sweet) informed that Chaucer used these words in the 14th century. In conclusion, archaic or modern linguistic entities reveal their physical age.

Afterwards, SVM (Support Vector Machine), a supervised machine learning model, was implemented in computational stylistics (Stamatatos, 2009). Moreover, the N-gram approach was brought into action in the domain of stylometry. That is why, these technological studies, models and tools were followed in the current study.

Previous studies were concerned with literary language, but one study focused on stylistic characteristics of tourism web-based English. Then the Tourism English Corpus (TEC) was built, and it was compared with Freiburg-LOB Corpus of British English (FLOB). Both of them showed equal lexical density; however, content words of TEC were more in number than FLOB. TEC length of sentences was smaller than FLOB; more nouns and adjectives were found in TEC than FLOB; and TEC employed more verbs than FLOB (Kang, & Yu, 2011, p. 129). Besides, both features of vocabulary density and the average length of sentences matched the current study. As its gaps were concerned, corpus sizes were not equal; consequently, they created unequal results. TEC was a reduced corpus as compared to FLOB.

A stylometric study was conducted in Bangladesh. Some other qualities were added along with KGs. A corpus was built by selecting 150 short stories from each writer, including Rabindranath Tagore, Sarat Chandra and other writers to explore stylistic qualities. KG, SVM and decision tree algorithms were used to discriminate the style of the aforementioned writers. In this study, SVM was verified as the most authentic algorithm for extracting stylometric features (Chakraborty, 2012).

Dutch university students' writings and their styles were analysed with the perspective of the author's age, sentiments, genre, personality, gender. For this purpose, CLiPS consisted of 1126 students' documents and 305,000 token words. It discovered personality, sentiments, author, age, gender and genre. The SVM algorithm was applied in this study to automatically classify text in the given categories (Verhoeven, & Daelemans, 2014). This Dutch research ignored the average length of sentences, vocabulary density and the most frequent words; therefore, this gap was filled by Summary tool.

Statistics had become an integral part of machine learning; text mining and data mining research works; thereupon, the most occurring words had been derived with statistical algorithms. Statistics facilitated in determining attributes of an author and gender recognition. Then Brown corpus was applied, and stop words filtered the noise of data to find lemmas (Amancio, 2015). There were difficulties for social science students in comprehending formulae of statistics, complex diagrams; therefore, the current Voyant study solved these issues and saved social scientists from programming and technical issues.

Some pseudo writers claimed authorship of a renowned novel, *'Go Set a Watchman'*. On this blame, its computational stylistic analysis was conducted with clusters and frequencies. Its cluster tree was compared with Harper Lee's corpus of 28 novels. Stylometry of the novel *'Go Set a Watchman'* matched Lee's 28 novel corpus; therefore, the reality dawned upon all that the controversial novel also belonged to Harper Lee (Gamerman, 2015). The current research also studied the stylistic features of each lesson from intermediate English textbooks.

Stylometry with R presented quantifiable qualities without complexities of coding, programming and embedding of other tools. Furthermore, French and English literary works were studied and interpreted with R packages, and they generated N-grams, corpus and enthralling diagrams (Eder, Rybicki, & Kestemont, 2016). R package qualities produced results like the Summary tool, but R package missed the frequently occurring corpus words. Summary tool extracted the most recurrent words from the corpus, and they denoted hedges and narratology too.

Digital tools, for example, Stylo and Cluto facilitated the extraction of computational stylistics features from the Polish language. It explored average word length, frequent punctuation marks, average sentence length, style markers and word frequencies to differentiate various stylistic qualities without any programming expertise (Eder, Piasecki, & Walkowiak, 2017). If Stylo and Cluto tools were juxtaposed with Voyant tools, Stylo and Cluto tools showed some additional features of exploring the average length of words, style markers, punctuation marks and POS tagging.

Li, Ji and Xu (2017) prepared a literary corpus of Mo Yan and Zhang Wei's works to discriminate their measured stylometric characteristics. Statistical results exposed that their literary styles diverged in the construction of sentences, length of diction and description of the societal system. Zhang Wei's language fluctuated and varied recurrently, while Mo Yan's language style remained consistent throughout his works.

Nobel laureate Ernest Hemingway's literary work had been re-examined in Sweden to assess common postulation and stylometric characteristics in his delineation of feminine characters. Therefore, nouns, token numbers and adjectives were enumerated to characterize stylistic structures (Sundberg, & Nilsson, 2018).

Recently, a study was conducted to explore similar features between James Joyce and O'Brien's literary works. Therefore, their stylistic qualities were juxtaposed, and it was found that O'Brien was following his teacher James Joyce's stylistic qualities. This Joycean study manifested data with a delta analysis tool, stylistic clusters and average length of sentences with graphics (O'Sullivan, Bazarnik, Eder, & Rybicki, 2018). This research expressed graphical and numeric data to demonstrate common stylistic features. Thus, stylometric features were extended; and quantified criticism was introduced to discuss literary aspects precisely. Generalised literary criticism received less response as compared to numeric and quantified criticism; hence, numeracy gave authenticity to data analysis.

Summary tool in Voyant suite was used to extract quantitative and qualitative sylometric features from 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*). The most recurrent style of these stories was dialogic, and their vocabulary density ranged from 0.307 to 0.362 which was counted by unique and total words. Furthermore, the average length of sentences ranging from 9.9 words to 12 words, hence, it referred to simplicity and shortness of sentences. In short, Summary tool also extracted key motifs of each short story separately to give an overview of the story (Ullah, & Mahmood, 2019, pp. 1-17).

2.24.2 Foregoing Publications on Cirrus/ Word Cloud

Cirrus displayed major key themes in the big font size, and minor themes occupied the small place in small font size. In addition to colourful themes of different sizes, the statistical weight of each theme was shown by putting the cursor on any theme, and this statistical weight transformed an artistic word cloud into an interesting knowledge pattern.

Voyant tools were applied by Jannidis to study 350 German novels and he extracted knowledge patterns, visuals, statistical knowledge and interactive data (Burrows, 2002, p. 267) in 21st St Louis Symposium on German Literature and Culture. Likewise, the application of word clouds shifted from academia to politics for instance the gist of political speeches was derived through word clouds during Obama and McCain's presidential election campaign (GitHub, 2014). The current dissertation also extracted themes of one novel and other intermediate English textbooks.

If a word occurred frequently in the corpus, its size and font would be bigger to show its frequent occurrence. There was one shortcoming that the context of each word was not present here; therefore, Contexts tool in Voyant suite provided an interactive context of the keyword. Thus, word sense disambiguation (WSD) could be resolved with the simultaneous use of Cirrus and Contexts tool (Ramsden, & Bate, 2008). In addition to it, Python and the NLP tool kit also resolved the issue of WSD by finding the context of certain ambiguous words (Bird, Klein, & Loper, 2014). With the help of these tools, previewing techniques became beneficial, quick and accurate by combining key ideas and their context.

Clement, Plaisant and Vuillemot (2008) described that PosViz tool created visuals to comprehend word usage. Various colours were associated with certain parts of speech, and an integration of POS tagging was a peculiar quality of PosViz visual. It was deprived of interactivity and statistical weight, while the current research extracted 25 to 500 themes in one colourful word cloud with their statistical weight.

Word clouds were aptly applied in the academic fields as a visual summary of any textual data. Summary of interviews (McNaught, & Lam, 2010) was also exhibited through word clouds. This method was equally useful for the summarization of any linear or nonlinear texts.

Some other uses of word clouds were also found in academia for instance, word clouds facilitated teaching and learning processes for all language skills (Hayes, 2008). Thus, four learning styles had been presented with word clouds: accommodation, convergence, assimilation and divergence. As a result, they increased the interest, inspiration and information level of students. Therefore, word clouds were found useful in teaching and learning roles (Miley, & Read, 2011). Word clouds simultaneously chiselled knowledge and sense of appreciation. They were suitable for previewing, vocabulary learning, extraction of key ideas and brainstorming of ideas.

Semantic word clouds possessed greater importance than ordinary word clouds. In this study, the Latent Dirichlet Allocation (LDA) algorithm model explored semantic shades with contexts. They were equally beneficial in media studies and business intelligence. Likewise, TopicVec tool was applied to analyse showbiz industry content with word clouds (Nguyen, Chang, & Hui, 2011).

Some other research projects expanded the use of word clouds and applied them to emotion mining. For example, Shakespearean plays, '*As You Like It'*, *'Hamlet'*, *'Frankenstein'* and some other stories were mined to discover emotion-related knowledge patterns (Muhammad, 2012). Cirrus extracted emotional themes to discover psychoanalytical conditions of characters, because the dramatic genre exposed more mental conflicts and emotional state of mind than other genres.

One more study was conducted with the collective function of Cirrus and Contexts tool. In one study, Google Books Corpus N-grams data were exhibited in the form of Cirrus as well as in other data visualizations (Muhammad, 2012). Sometimes, only word cloud could mislead semantically and contextually for instance the use of word "like" was confusing whether it was an action verb or just a simile. When KWIC was visible, ambiguity was resolved, and the same had been argued by Ramsden and Bate (2008).

In the unsupervised learning, topic modelling also extracted key motifs through text mining process and Latent Dirichlet Allocation (LDA) algorithm. Therefore, knowledgeable and fascinating main ideas and epistemological themes were derived from French dramas (Schoch, & Schoech, 2012).

Word clouds were also used to extract topics from the corpus. This topic modelling technique was applied to 3346 Irish, British and American novels. Function words were removed to highlight key ideas of the text, but Voyant stop words had automatic, editable and disabled qualities of stop words. The main ideas with the word "enemy" were separated from novels of Jokers (Jockers, & Mimno, 2013).

Yeates (2013) produced word clouds and key themes from the big data of 1500 apocalyptic novels. Owing to the utility of Voyant tools, several types of research projects were conducted to explore useful knowledge patterns. Next year, Voyant tools were applied to magazines to explore their main themes (Grier, 2014).

Contexts tool showed the interactive context of key words to disambiguate word sense, but ReCloud, a semantic word cloud, had incorporated context in the word cloud for word sense disambiguation. In the business intelligence domain, the satisfaction level of customers was determined with the semantic cloud. So, reviews about Amazon and Yelp were mined with mixed methods (Wang, Zhao, Guo, North, & Ramakrishnan, 2014). To conclude, review mining had been used extensively in research and business intelligence.

Some other digital tools also created different types of word clouds; for instance, Wordstorms tool kit produced Cirrus to extract motifs of the selected data. Ideally, a good word cloud should not skip any significant theme or character during the filtering process of stop words (Castella, & Sutton, 2014). Therefore, Cirrus tool was empowered with the aforementioned two qualities to produce standardized and refined results.

Several other text mining research projects were steered on big data of complete literary works of Shakespeare, Jane Austen, J. K. Rowling's Harry Potter, Sir Arthur Conan Doyle's detective novels namely '*Sherlock Holmes*' and George Eliot's one novel, '*Mill on the Floss*' with Voyant text mining tools (Sinclair, & Rockwell, 2015b).

Some other digital tools exhibited main motifs in a circular shape. Lohmann, Heimerl, Bopp, Burch and Ertl (2015) argued that RadCloud and ConcentriCloud tools were better than other word cloud tools because of their three-layered thematic circles. The most inward circle displayed the most common themes from the selected documents. The central circle remained empty if no word matched other documents. In the experimental research, 7 Harry Potter novels were investigated with CorecentriCloud with adaptable colour and size. Comparing it with Cirrus tool, which had the interactive ability to increase or decrease the number of themes, however, CorecentriCloud was deprived of interactivity. Cirrus tool showed its statistical weight which transformed the word cloud into a knowledge pattern. Furthermore, digital topic extraction ability had been embedded in some other tools for instance, in Parallel Tag Clouds, ManyEyes and DocuBurst exhibited key text concepts.

Clustering and classification procedures had been applied for the topic modelling process which "identifies clusters of words that could be the major topics" (Rockwell, & Sinclair, 2016, p. 129). Word clouds/ Cirrus exposed most occurring characters and themes which could be studied in depth. Different digital tools produced multi-designed Wordle, Tagxedo, Word Cloud Explorer and particular algorithms performed their roles to produce and refine them.

Kemman, a PhD student in the University of Luxembourg, used analysis of 1000 WikiLeaks emails through Voyant tools in his classroom activities, and found answers to the points related with who, what, when, where (Kemman, 2016).

Over time, new software and hardware brought drastic changes in DH. Previous images were transformed into interactive knowledge bearing visuals. Wordle tool created still multi-shaped word clouds but later, interactive word clouds were in vogue, for example, ManiWordle (Koh et al., 2010 as cited in Seyfert, & Viola, 2017) exhibited flexibly rotating and moving word clouds. Afterwards, SparkClouds (Lee et al., 2010 as cited in Seyfert, & Viola, 2017) manifested sparkline below each word. Later WordlePlus (Jo et al., 2015 as cited in Seyfert, & Viola, 2017) showed words pop and word classification in an interactive word cloud. Then some word clouds modified themselves with the input of different types of data and word occurrence (Seyfert, & Viola, 2017).

A toolkit of Interactive Text Mining Suite was employed to extract major themes from the text of the novel '*Romance of Flamenca'*, and its English translated version. So, in this study, data visualization, corpus and DH were merged to discover knowledge patterns (Scrivener, & Davis, 2017) that were innovative, idiosyncratic and interesting for learners.

Interactive word clouds were generated by Cirrus tool in Voyant suite, and it discovered an epigrammatic digital interactive visual to show main characters "Tommy" (496), "Ruth" (455), "I" (Kathy) (355); key nostalgic themes of "remember" (143), "thought" (126); and some other motifs about "Hailsham (203), "carer" (74), "sex" (80) and sex "lectures" were mentioned (8) (Ullah, Uzair, & Mahmood, 2019, pp. 83-98). Consequently, Cirrus tool presented a preview of the Nobel prize-winning novel with distant reading technique presented by Moretti (2013).

2.24.3 Previous Studies on Phrases/ Collocations/ N-grams

Different terms, such as standard phraseology, bigrams, trigrams, multiword expressions, formulaic language, lexical bundles, language clusters or collocations/ n-grams, were used to describe the umbrella term of collocations, and they also showed slight variations in their meanings. So, some previous studies had been reviewed to get an insight into collocations.

Different studies elaborated various collocation patterns/ n-grams and 37 standard phrase patterns were found (Benson, Benson, & Ilson, 1986); nevertheless, the current dissertation extracted 167 types of collocation patterns/ n-grams. Later, it was argued that the study of phraseology eased comprehension of the systematic coinage of idioms (Sinclair, 1987) and collocation patterns/ n-grams.

Four types of factor analysis were done to investigate different senses and usage of the word “right”. The first factor showed concreteness, the second factor expressed themes, the third factor exhibited a general sense of all right, and the fourth factor showed the occurrence of the word “right” at the end of clauses. These factors showed that the grammatical category was not mentioned along with these collocations/ n-grams (Biber, 1993), but the current study added grammatical patterns. Thus, the gap of the grammatical category with collocation patterns/ n-grams was filled in the current study.

The contribution of research was to bring positive changes in academia. So, the usefulness of collocation/ n-grams for language learning had been argued with three perspectives: ubiquitous nature of chunks, chunks as a unit of proficiency, challenging role of chunks for second language learning (Nesselhauf, 2003). Likewise, language acquisition also progressed from unigram to bigram, then trigram and sentence. Therefore, acquiring standard phraseology was the second step of language learning.

Collocations/ n-grams facilitated in language learning, teaching and translation processes. Multiword expressions from multilingual documents, including newspaper articles and other soft files were analysed visually with Fips syntactic parser tool which supported English and French languages. Afterwards, the log-likelihood ratio statistical test was applied to generate collocations/ n-grams of French and English which could assist in language learning and translation process with the fastest pace (Seretan, Nerima, & Wehrli, 2004).

Later, the trend shifted to cross-comparison of different genres and texts to find their similarities and dissimilarities. Biber, Conrad and Cortes (2004) analysed lexical bundles from university textbooks and spoken discourse on the criterion of their frequency. Classroom teaching and textbook lexical bundles were juxtaposed with prose lexical bundles and conversational lexical bundles. The paper found that stance bundles and discourse organizing bundles were used more than conversational lexical bundles in the classroom teaching process. Classroom teaching employed more referential bundles than educational prose (Biber, et al., 2004).

Koprowski (2005) pointed out the problem that academic material writers intended to include collocations/ n-grams in the syllabus, but they did not have any criterion or instructions for their selection, that is why content writers of textbooks subjectively chose and included collocations/ n-grams in textbooks. Moreover, commercial book publishers included multiword expressions in their ESL books. In this continuation, Wang and Good (2007) pointed out shortcomings in textbooks that resulted in flaws of pedagogy and learning. That is why non-native learners faced difficulties in learning the English language. Responding to this criterion- problem of collocations/ n-grams, six criteria were adopted to shortlist collocations/ n-grams from forty terms concerning collocations/ n-grams. So, 1000 most occurring collocations/ n-grams were investigated from ten million BNC (British National Corpus), and eventually, 2000 most appropriate phrases were selected for the spoken English syllabus (Shin, & Nation, 2007). This paper argued to include phraseology for teaching and learning the English language.

Hsu (2008) studied lexical units in ESL books published from 2003 to 2005, and he found inconsistency of collocation patterns/ n-grams in these ESL books. He emphasised that teachers must know about the significance of collocation/ n-grams input through textbooks. The current study proposed that before teaching and thorough reading of any textbook, its collocation patterns/ n-grams should be extracted through Phrases tools to familiarize learners with specific phraseology.

Cheng, Greaves, Sinclair and Warren (2008) identified discrepancies in the phrases through Concgram software. Moreover, phraseology occupied a central position in the understanding and construction of language use. Their significance became more evident when Sinclair (2007) named phrases as Meaning Shift Unit (MSU), and he explored that to what extent collocations could accept linguistic modifications. There was yet unexplored capacity to probe interphrasability, intertextuality and intercollocability. On the other hand, one research on school textbooks presented a weakening trend of incorporation of collocations/ n-grams (Wray, 2012). The current study aimed to implicate collocations/ n-grams in textbooks.

There were also discrepancies in the use of collocations/ n-grams between native and non-native language users. Korean word list, curriculum and corpora of non-native speakers were compared with native English speakers (Choi, & Chon, 2012). Non-native speakers did not utilise corpora and native language; hence, this deficiency led to their slow progression in learning and teaching English.

Another study presented themes of collocations/ n-grams which could be selected as criteria for the selection of collocations. Formulaic language showed six themes: 1. Theory: grammar, lexis 2. Clinical: linguistic disorders 3. Development: mother tongue acquisition 4. Learning and teaching 5. Culture: cultural discrepancies and trends 6. Text: corpus or corpora (Wray, 2013). In the same year, the text of poetry was mined through Voyant tools to produce collocations/ n-grams, word clusters and key word lists (Mohs, 2013).

Modern lexicographers utilised corpora for the inclusion of new words in new editions of dictionaries. In 2014, during the Pakistani sit-in of PAT (Pakistan Awami Tehreek), a new Proper noun GULLU BUTT was uttered many times and due to its frequent occurrence, Oxford Advanced Learners Dictionary included gulluism, gulluish, gullunise words.

Another Korean study made a corpus of 9 Korean Middle School textbooks published by three different publishers. The Simple Concordance Program was used in this study to extract collocations/ n-grams. Their lexical chunks were repeated 3.6 times, but common collocations/ n-grams were found in a very small number. It revealed that there was no harmony or fixed policy about the inclusion of collocations/ n-grams in English textbooks (Lee, 2015). The current study also argued that collocations/ n-grams should have been included in textbooks and the Punjab Textbook Board should approve a policy to include collocations/ n-grams in English textbooks as foreign books had done.

Following the same trend of the selection of collocations/ n-grams from textbooks, the formulaic language was investigated by comparing primary school English textbooks taught in Hong Kong, and English textbooks taught in the United Kingdom. Thus, native and non-native differences in formulaic language were investigated with the corpus. Results of the study showed that Hong Kong English textbooks used less formulaic expressions than UK English textbooks. Moreover, lexical structures were diversified in the form and use in Hong Kong English textbooks, when juxtaposed with UK English textbooks (Russell, 2017). The use of native speakers' corpus, for example, BNC was capable of alleviating this linguistic discrepancy.

Through corpus, 100 academic quadgrams were extracted from eight domains of academia. To accomplish this task, a corpus of 120 million words was built. A set of WordSmith Tools was applied to extract collocation patterns/ n-grams automatically. Later the selected collocations/ n-grams were evaluated with Wilcoxon rank-sum test (a = 0.05) to search their frequent use in academic discourse. Its major findings were the existence of the most frequent collocations/ n-grams for instance, 'as a result of', 'the case of', and 'at the end of' (Da Silva, Orenha-Ottaiano, & Babini, 2017). One difficulty for social science researchers was the application of statistical tools and tests, while the current dissertation did not apply any different programming or statistical test because statistical formulae were inbuilt in Voyant tools, and they extracted collocations/ n-grams in terms of their length and occurrence.

Business English collocations/ n-grams were taught for 8 weeks to 23 undergraduate students with COCA (Corpus of Contemporary American English) and Wikipedia corpora in Chinese academic institutions. Furthermore, Antconc tool was also used to study these corpora. Research data were collected from tools of questionnaires, pre-test, post-test and reflective journals. This study explored the utilisation of data to increase collocation/ n-grams awareness and autonomous learning through corpus (Chen, 2017). Likewise, register of business English and exposure of target language could be learnt best from Cambridge Business English Corpus.

In multi-word expressions (MWEs), 33% were figurative collocations/ n-grams and 33% were idioms. There were phonological MWEs whose construction was based on alliteration and assonance. This study found that language differed due to various MWEs and interdisciplinary themes (Siyanova-Chanturia, Conklin, Caffarra, Kaan, & van Heuven, 2017). This study searched different types of collocations n-grams and differentiated between collocations n-grams and idioms, while the classification of the standard collocations/ n-grams (phrases and idioms) based on their grammatical categories.

2.24.4 Former Research Works on Links/ KGs

A knowledge graph was defined as a graphical descriptor of the interrelationship of real things and their classes in a schema (Paulheim, 2017). KG and ontology relation (Ehrlinger, & Wöß, 2016) was the foundation of KG. Initially, Google graphs were introduced and popularized in 2012 to represent knowledge patterns. Moreover, they had been studied with relationship mining techniques which discovered linear correlationship among relevant variables (Barahate, 2012, p. 13).

One segment of the current study also focused on finding connections among characters and key themes of any text. Some other terms, for instance, KG systems, neural networks, connectionist systems and parallel distributed processing systems denoted KG which showed nodes and some connections. Another term, "computational neuroscience" (Churchland, & Sejnowski, 1992) employed mathematics, models and abstraction in the study of neuroscience. KG was generated with the formula of multiplication of inputs with weight. To conclude, KG acted as interconnections of different entities.

NAGA, a new semantic engine, was structured to generate interrelationship of millions of web KGs, and they were extendable with increasing size of data. Hermeneutica Theory proposes that "They can be extended to expose new things" (Rockwell, & Sinclair, 2016, p. 166). This paper proposed the concept of searching knowledge patterns and key terms in place of browsing web pages. Moreover, devising a new scoring model was based on a generative language model (Kasneci, Suchanek, Ifrim, Ramanath, & Weikum, 2008).

In one study, the correlation between a query and an entity (Blanco, & Zaragoza, 2010) was searched and ranked, while Voskarides, Meij, Tsagkias, De Rijke and Weerkamp (2015) found interrelationship of two entities. Another pertinent study was conducted by Fang, Sarma, Yu and Bohannon (2011) who searched a list of interconnections of an entity coupling. Next year, Baalen (2012) used links to find interrelationship of first-person pronouns and second person pronouns, collocation/ n-grams of lemmatised first-person pronouns per song with Links tool, though its visual was a bit different than the current version of Links tool.

Automated knowledge aggregation from knowledge repositories was the goal of this research. Sar-graph was constructed on the theme of “marriage” to show the interrelationship of texts from BabelNet, WordNet, Verb-Net, UWN (University World News) and covered them with semantic knowledge graphs. Thus, Sar-graph identified arguments and the relation of argumentative entities (Uszkoreit, & Xu, 2013).

Explaining the relationship of KG is a big challenge, and in this paper, KG and text from corpora were compared and scored. It showed how well a KG was designed (Voskarides, Meij, Tsagkias, De Rijke, & Weerkamp, 2015). The current dissertation generated KGs with Links tool to explain the interrelationship of various nodes to derive a meaningful message from the text. Moreover, KGs and textual quotes were compared to approve or refute KGs. In the same fashion, Hermeneutica Theory was “supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166) for cross-checking and verifying the discovered knowledge.

KGs were also used to discover knowledge from large documents. Two problems sustained: first was about the relatedness of themes and their explanations; second was about showing pairs and their relatedness. To address these issues, the RECAP tool was designed for constructing KGs (Pirrò, 2015).

The role of a character named Spock from the Star Trek science-fiction movie was drawn in a multipronged KG. Statistical Relational Learning (SRL) methods were applied to extract large knowledge graphs whose major shortcoming was a lack of common sense (Nickel, Murphy, Tresp, & Gabrilovich, 2016).

KGs were constructed with Microsoft Satori, Cyc, Freebase and Wikidata. Web KGs were also structured with DBpedia and Yago. NELL, PROSPERA, or KnowledgeVault to extract knowledge patterns from semi-structured data, but no approach could claim the production of absolute correct results from KGs. To correct them, heuristic methods were applied. This paper suggested various heuristic ways to attain and refine KGs (Bordes, & Gabrilovich, 2014 as cited in Paulheim, 2017).

KGs were used to determine stylistic qualities of literature (Tweedie, Singh,. & Holmes, 1996). Following the same trend, another stylistic research was carried on works of three literary personalities. Middleton’s 90 examples, Jonson’s 164 specimens and Shakespeare’s 168 samples were analysed with Cascade-Correlation network architecture. It claimed that Shakespeare was the true author of Thomas Kyd’s *‘The Spanish Tragedy’,* and Madison was the person who originally wrote twelve disputed Federalist papers (Waugh, Adams, Tweedie, & Waugh, 2000). To summarise, KG was also studied for author identification. There was a niche of quantification in KGs and without statistical information, the statistical comparison was ambiguous. To present a stylometric summary, total words, unique words, vocabulary density, the average length of sentences and key themes with occurrence were extracted to understand and compare the two literary works.

2.24.5 Preceding Scholarly Works on Contexts/KWIC/WSD

In the 1940s, WSD research projects were started when machine translation found errors. Word sense disambiguation (WSD) meant to explore the exact sensible meaning of a word according to its usage in the text. Meaning, semantic field, orthographic pattern, pronunciation, parts of speech, register and context determined sense of any word. According to the type of ambiguity, syntactic, semantic, pragmatic, lexical or semantic solutions (Turdakov, 2010) were applied to resolve WSD. Such ambiguities might have appeared in homonyms (same spelling and pronunciation and different meaning), homographs (same spelling, meaning and pronunciation), homophones (different meaning, same pronunciation and spelling), heteronym (same spelling, different meaning and pronunciation), heterograph (same pronunciation, different meaning and spelling), polysemy (different but related meaning, same spelling and pronunciation) and capitonyms (different meaning in capitalization, same spelling and pronunciation) (Hutchins, 1999). The exploration of bidirectional context facilitated in word sense disambiguation.

Changing the trend from computer programmers to linguists and linguistics which solved problems of WSD. Key Words in Context (KWIC) were incorporated under Roberto Busa’s seminal work on concordance. In the 1960s, Peter Luhn at IBM introduced KWIC which disambiguated difficulties in semantic shades (Fischer, 1971).

Productive works on WSD were done by the collaboration of linguists and computer programmers. Statistics, word frequency, algorithms, training of algorithms, different types of corpora, thesauri and lexicons were used to differentiate between wrong and true sense of words. In the 1960s, it was strived to produce such programmes which could work on cognitive frameworks of human beings. Furthermore, different dictionaries were used to disambiguate the word sense of problematized words (Collins, & Loftus, 1975).

Mostly computer scientists worked on these aspects. WSD algorithms premised on sense definition, sources of knowledge, context display, concept demonstration and algorithms to discriminate word sense. Different types of applications were required to address different types of ambiguities; for instance, psycholinguistic and polysemic disambiguities were required to solve ambiguities in the written language (Kilgarriff, 1997).

Hockey (2001) opined qualities of the perfect KWIC tool which could present data in an alphabetic sequence and numbers of occurrence. Therefore, Contexts tool possessed the aforementioned qualities of perfection; for this purpose, KWIC facilitated readers to disambiguate the sense of words. Therefore, WSD and KWIC had deep-seated ties, since the former was a problem, and the latter was its remedy. Some more studies on WSD were reviewed, and Bowker and Pearson (2002) explained various uses of KWIC that could find lemmas and word family from the corpus.

A hybrid approach was applied to measure the relatedness of word sense on WSD. Set algebra was applied to WordNet to formulate regular rules. Moreover, Boolean operators were applied to lexical instances. This experimental research established a quantified system of semantic relatedness with statistics to exhibit productive results on Semcor, a small selection of Brown corpus (Yang, 2008).

Word Sense Disambiguation (WSD) had two aspects: all senses of each word must be determined. Then a particular sense should be based on its context. WSD was studied in the domain of NLP and computational linguistics. WSD was solved by Key Word In Context (KWIC). The key word was centralized, and interactive bidirectional context disambiguated word sense with human reflection. Quantification of statistical data led to reliability, validity and knowledge patterns since numerics transformed a visual or a word into a knowledge pattern.

In multilingual texts, some languages were written right to left while some were composed from left to right. This study discussed bidirectional presentation of code mixed texts (presence of Arabic and Persian texts in English texts) that which order should be followed. Following one order caused trouble for the other languages which were written in the opposite order. This problem was solved with an algorithm and corpus files of those languages (Rychlý, & Kovár, 2007). Oriental and Occidental languages were disambiguated with bidirectional text presentations. Contexts tool was capable of disambiguating bidirectional context.

A large variety of algorithms ranging from AI to data mining were applied for WSD. These algorithms were classified according to the categories of unsupervised, supervised and bootstrapping (Zhou, & Han, 2005), and among them, the supervised learning algorithm was the most common. Ensemble algorithms, for example, LazyBoost, as well as AdaBoost, were proved accurate and graph-based algorithms disambiguated All-Words dataset (Ponzetto, & Navigli, 2010). To conclude the discussion on WSD algorithms, one algorithm was not sufficient for all sorts of data, because one algorithm dealt with only one sense in one discourse, so the application of various WSD algorithms produced the best results.

In the 1980s, the embedding of memory into computers became easy and affordable, so corpora, thesauri and dictionaries were utilised to solve WSD. Then sense annotation algorithms were introduced. Afterwards, WordNet like lexicons classified semantic and lexical ambiguities. Later, the trend of computation resorted to human-like sense comprehension (Hwang, Choi, & Kim, 2011). Afterwards, semantic networking was emphasised. To explore ambiguity in the written text was more painstaking as compared to spoken language. Moreover, classification algorithms were used for sense identification. Later, automatic sense taggers were used to disambiguate word sense (Bhala, & Abirami, 2014).

Digital dictionaries, ontologies and corpora were also used for sense detection. Semcor was an extensively used corpus for sense annotation. Moreover, Wikipedia encyclopaedia was also used for sense annotation (Mihalcea, 2007). Detecting part of speech was done with FrameNet and VerbNet (Bhala, & Abirami, 2014). When a part of speech was finalised, the issue of semantic ambiguity was resolved since each part of speech carried its particular sense.

From the last fifty years, corpus and KWIC could not get their true status to explore true language in the classroom, and still learners were deprived of corpus-driven language learning techniques. KWIC was a “linguistic detective” to find application and uses of language in academic settings (Green, 2018).

A corpus of academic English was prepared, and it comprised 895 Elsevier journal articles from 2011 to 2015. The total words of this corpus were 5686428. Furthermore, this specialised corpus was classified into health, physical, life, and social sciences. Since Elsevier was the biggest database of journals, conference proceedings and books, it was very difficult to search the relevant material from a big database. So, a concordance or KWIC program was initiated on the corpus website, www.kwary.net, to facilitate researchers in the exploration of the relevant content in the shortest possible time (Kwary, 2018). To conclude, Information retrieval, Named Entity Recognition and KWIC facilitated resolving various ambiguities.

2.25 Conclusion

For the construction of new knowledge, previous knowledge had been mentioned, because new knowledge was established on the foundation of the previous studies, and this approach took readers from known to unknown dimensions. The current section presented a chronological evolution of various domains closely pertinent to the current research. Moreover, the data analysis section also resorted to this section for cross-checking and discussion whether findings of the current study matched the previous studies or not; therefore, this act strengthened the current study, and showed its close relationship with previous literature. Therefore, this chapter encapsulated major developments as well as research projects in the field of DH, data mining, major data mining theories, educational data mining, text mining, web mining, knowledge patterns, data visuals and previous research works on the five Voyant tools. Thus, it covered an entire academic and research paradigm from an umbrella term of DH to the contemporary Voyant text mining tools.

The coming chapter explains the research methodology and theoretical framework.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Research design set out by defining key concepts regarding the method, methodology, mixed methods and combining two wide-ranging research strategies; for instance, Saunders, Lewis, & Thornhill's (2012) research onion which comprised research philosophy, research approach, research strategies, time horizons, data collection methods; and Durant (2004)'s seven-pointed typology of research for instance i. Framework for Research ii. Data Generation iii. Research Method Approach and Rationale iv. Data Quality and Data Mining v. Data Handling and Data Analysis vi. Research Management and Application of Research vii. Research Skills. Both research methodologies were triangulated to chalk out the blueprint of the research design of this study, as shown in figure 9. After delineating them, reliability, validity and limitations of the study were also discussed.

Table 1 Application of the Research Onion to Current Study

Features	**Application**
i. Research philosophy	Post positivism (post empiricism)
ii. Research approach	Inductive
iii. Research strategies	The experimental strategy of text mining; triangulation of Saunders, Lewis, & Thornhill's (2012) research onion; Durant's (2004) seven-pointed typology of research
iv. Time horizons	Research duration between 2017-2018
v. Data collection methods	Digitization of intermediate English textbooks

Table 2 Durant (2004)'s Seven-Pointed Typology of Research

Features	**Application**
i. Framework for Research	Triangulation of Knowledge Discovery Theory and Hermeneutica Theory
ii. Data Generation	Digitization of Intermediate English Textbooks
iii. Research Method Approach and Rationale	Mixed-Methods Approach
iv. Data Quality and Data Mining	DM Rules and Voyant Tools for DM
v. Data Handling and Data Analysis	Text, Visual and Quantitative Data Analysis
vi. Research Management and Application of Research	Ethical Considerations and implications
vii. Research Skills	Data Presentation Architecture

A research method is a rational and systematic pursuit of knowledge to address the research problem with particular methods and theoretical frameworks. Research is primarily meant for a journey from known to unknown and to produce a knowledgeable perspective on canonized and scientific paradigms.

Differentiating between a method and methodology, research methods are individual parts of research, while methodology or the scale of the methodology is vaster than a method. A method is a research technique or tool, while a methodology is a complete system of various methods or a set of principles to acquire authentic deep knowledge. Furthermore, research methodology makes the research endeavour a compact whole to respond all research questions and to accomplish all research objectives (Surbhi, 2016). Moreover, the methodology is a justification to opt for one or more accurate research methods which must be tested by canons of reliability and validity. In short, the method is a part and methodology is a whole.

This research aimed to find answers to research questions by applying all methods of required data, data generation and data analysis. Mixed methods approach was applied to acquire qualitative and quantitative data which cross-validated each other. The generated data in the design of visuals revealed unexplored knowledge patterns from data with Knowledge Discovery Theory (Agrawal, & Psaila, 1995, p. 1; Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996; Cabena, Hadjinian, Stadler, Verhees, Zanasi 1998, p. 9) and Hermeneutica Theory (Rockwell, & Sinclair, 2016, p. 166).

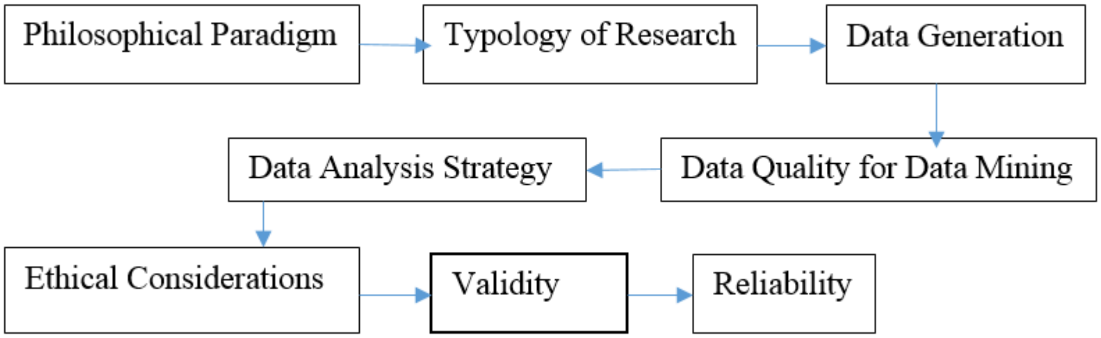


Figure 9 Flowchart of Research Methodology

Generally, scientific experiments produced the same results every time, and they became law. In social sciences, text mining/ text analytics and its created interactive knowledge patterns made it a scientific subject because the same results appeared every time. Furthermore, data mining in digital humanities had been declared as an experimental scientific research.

3.2 The Research Onion

Saunders, Lewis, & Thornhill (2012) discussed the research onion to write a research methodology. This section discussed its five points in the following sections.

3.2.1 Research Philosophy

Philosophy was the underpinning for establishing a research paradigm. The current study adopted post-positivism (post empiricism) philosophy, reflecting scientific, empirical, and inductive research methods. This study was scientific because its results remained the same, valid and reliable. It was empirical because digital tools were applied, and new digital experiences were conducted on intermediate English textbooks. The current research was inductive because inductive reasoning moved from specific observations to broad generalizations; therefore, it aimed to develop a theory. In addition to it, it chose a relativistic perspective in which nothing was considered an absolute reality since the truth was a relative term. Its data were collected through mixed methods (combination of qualitative and quantitative data). Its ontological ground was objective, but the knowledge was presented through personal experience (The Writepass Journal, 2012). This personal experience could be humanistic or machine-generated with the help of digital tools.

3.2.2 Inductive Research Approach

The inductive approach is a logical reasoning process to proceed from data and examples to a theory development. Data generation and data analysis processes confirmed or rejected research propositions clearly (Saunders et al., 2012). In this research approach, the effects of hypothetical inquiries, theories and tools were analysed and tested to confirm their authentic implications during all phases of research.

Five hypothetical research inquiries had been raised in the current study: Cirrus tool showed themes and characters; Phrases tool searched collocations/ n-grams; Links tool explored the interrelationship of various themes and characters; Summary tool generated corpus summary; and Contexts tool exhibited word sense disambiguation. The aforementioned investigations were tested and confirmed to find the desired results. Therefore, the results of Voyant tools were tested, and in most cases, they confirmed their textual outputs. In addition to it, the triangulation of Knowledge Discovery Theory and Hermeneutica Theory was applied to test and interpret the capability of Voyant tools to search for several unknown knowledge patterns. Consequently, the current study verified its productivity and authenticity.

3.2.3 Research Strategies

This study followed the triangulation of two strategies, namely Saunders, Lewis, & Thornhill’s (2012) research onion and Durant’s (2004) seven-pointed typology of research; therefore, their detail had been mentioned in 3.1 in a tabular form.

3.2.4 Time Horizons

Initially, the current study was conducted during 2017-2018. Later, it was revised in 2021.

3.2.5 Data Collection Methods

Intermediate English textbooks taught in Punjab boards of intermediate and secondary education colleges in Pakistan were digitized for text mining. Its detail had been delineated in the 3.3.2 section.

3.3 Typology of Research

Durant (2004) delineated the following seven-pointed typology of research:

3.3.1 Theoretical Framework

The framework was a guiding and supporting structure in which a research design performed its entire function. So, the theoretical framework of this research had been discussed in this section. The current study triangulated Knowledge Discovery Theory and Hermeneutica Theory to address five research queries, because the former generalised classical foundation of the current study, and the latter discussed Voyant tools and their roles in text mining.

3.3.1.1 Knowledge Discovery Theory (KDD)

KDD was introduced in 1989, and knowledge was declared as the output of the manipulation of data. In the 1990's Rakesh Aggrawal, the top most Indo-American computer scientist at Microsoft became the pioneer of Knowledge Discovery Theory (Zhu, 2018) which comprised machine learning, databases and statistics (Fayyad, Shapiro, Smyth, & Uthurusamy, 1996). Moreover, all knowledge extracting techniques and tools belonged to the knowledge discovery process. "In active data mining paradigm,… rules are discovered, …the history of the statistical parameters associated with the rules is updated… we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct to retrieve rules whose histories exhibit the desired trends" (Agrawal, & Psaila, 1995, p. 1). Summarization involves methods for finding a compact description for a subset of data (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). KDD process was a "set of various activities for making sense of data" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 82).

Another definition of KDD was "the extraction of implicit, previously unknown and potentially useful information from data" (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9; Witten, Frank, & Hall, 2011). Therefore, raw data were transformed into meaningful and useful knowledge patterns, for instance, Summary tool exposed corpus, Cirrus tool showed themes and characters; Links tool revealed interconnectivity of key ideas in knowledge graphs; unique words, words per sentence to quantify stylometric features of any piece of writing; and Contexts tool disambiguated word sense.

This theory was premised on the rules of the association, characteristics, classification, serialized system and prediction. A set of theories for instance, cloud model of mathematics, visualization, evidence theory, neural network, fuzzy sets, rough sets, genetic algorithm, exploratory learning and spatial inductive learning were at work in the process of unveiling knowledge patterns (Li, & Wang, 2005).

Text mining process aimed "to discover knowledge in unstructured or semi-structured data" (Nahm, 2001), as this study worked on intermediate English textbooks taught in Punjab education boards, Pakistan. Those interesting and beneficial knowledge patterns had "positive externalities" (McDonald, 2012) for learners and teachers. This research extracted knowledge patterns and interactive visuals for better understanding and interpretation of the text during agile hermeneutics. So, corpus summary, word clouds, collocation patterns/ n-grams, KG and key words in context (KWIC) were visual knowledge patterns. To interpret these knowledge patterns, triangulation of Knowledge Discovery Theory and Hermeneutica Theory was applied on generated data; and data analysis was interpreted in the harmony of research questions and objectives. Theoretical framework and philosophy strengthened each other; hence, post-positivism, KDD and Hermeneutica Theory were correlated with the bond of ontology. Focusing on conceptual taxonomy, it interlinked micro and macro ideologies and concepts (Thompson, 2016).

The main approaches of KDD and Hermeneutica Theory were a distillation of data for human judgment, prediction, relationship mining, clustering and discovery models. The prediction had been classified into three types: classification, regression and density estimation (Baker, 2010). Other DM methods included classification, clustering, summarization and deviation detection (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, pp. 44-45). So, text mining predicted and facilitated rational decision-making on the basis of data.

3.3.1.2 Theoretical Interface of KDD, Hermeneutica Theory and EDA

The exploratory Data Analysis (EDA) approach summarized statistical data in visual forms to identify knowledge patterns. It informed more than explicit data information or hypothesis testing; hence, this approach was applied to "small volume data" (Chattamvelli, 2016, p. 13); for instance, short stories or poems from the selected textbooks. Similarly, KDD and Hermeneutica Theory also extracted hidden data patterns, and presented them as interactive data visuals for instance, Cirrus (word cloud) and Links (knowledge graph). The emergence of big data necessitated and rationalized integrating EDA, KDD and Hermeneutica Theory to extract knowledge patterns. Mainly, Knowledge Discovery Theory emphasized on the extraction of knowledge patterns, and then these knowledge patterns were applied to many other domains. Secondly, Hermeneutica Theory was propounded by Voyant tool designers, and its application had been done by its tool designers in the book *Hermeneutica: Computer-assisted interpretation in the humanities*. So, the amalgamation of both theories was most appropriate to discover knowledge patterns. Another justification was that human computing could not handle big data with the economy of time, while text analytics dealt with big data to find accurate, useful and informative knowledge patterns. Furthermore, pedagogical support and knowledge engineering processes (Baker, 2010) were performed in EDA, while potential idiosyncratic patterns were explored in KDD, so the similarity of their objectives justified their triangulation.

3.3.1.3 Knowledge Discovery Theory and Mixed Methods

The research framework discussed mixed methods or "multidisciplinary research" or "combination of qualitative and quantitative approaches" (Durant, 2004, p. 10). This study interblended linguistics and computer sciences, including digital humanities, English language, applied linguistics, educational data mining, text analytics, Machine Learning (ML), Natural Language Processing (NLP), (AI) and statistics. Therefore, it became a multidisciplinary research.

3.3.1.4 Hermeneutica Theory

Sinclair and Rockwell (2016) propounded Hermeneutica Theory along with Voyant tool explanations. They explained its following features:

i. "Hermeneutica Theory is embedded in a context."

ii. “It is not like black boxes.” (In the domain of computer programming, it did not examine the actual background programme which was executed.)

iii. “Manipulation is in service of exploration and understanding.”

iv. “It is supplemented by other materials.”

v. “Knowledge-bearing tools provoke reflection.”

vi. “Hermeneutic tools fail in interesting ways.”

vii. “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166).

3.3.1.5 Interface between Hermeneutica Theory and Voyant tools

Hermeneutica Theory was propounded by designers of Voyant tools to explain the working of Voyant tools. So, Voyant tools and their generated visuals can be analysed best with Hermeneutica Theory. Another rationale was that mixed-methods approach required an amalgamation of two theories, so Knowledge Discovery Theory and Hermeneutica Theory were amalgamated and applied in the data analysis chapter.

Different features of Hermeneutica Theory had been visualized in the interface of Voyant suite. One theoretical aspect, “It is supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166) was linked with similar features of different tools, for example Summary tool also extracted key themes and Cirrus tool also extracted key themes. Furthermore, they can be read in Reader tool. Another feature, “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166), and Links tool generated knowledge graphs which required reflection and hermeneutic abilities to interpret different nodes at multidimensional levels. One more theoretical underpinning, “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166), and this interactive feature is the hallmark of all tools; for instance, themes in Cirrus, number of phrases, nodes of the knowledge graph, number of themes and bidirectional context can be extended to explore new linguistic entities and new pieces of information. One theoretical postulate, “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166) elucidated Contexts tool which showed the bidirectional context of any key word for word sense disambiguation.

3.3.2 Data Generation and Rationale

Research questions and objectives play the crucial role to lead any research. They determine the need of qualitative, quantitative or mixed methods research. Then machine-readable data were generated from intermediate English textbooks.

Intermediate English textbooks as research data (population) were taught in different educational systems. So, the purposive sampling technique was applied in this study. The major purpose was to choose those English textbooks which were studied by the largest number of students in Pakistan. Secondly, those books covered most of the literary genres.

The first rationale behind these sample books was that majority of students from Pakistan and 9 Boards of Intermediate and Secondary Education Punjab, Punjab Technical Education Board students studied these books for their annual exams. The current study delimited intermediate English textbooks taught and examined in nine Punjab Boards (Rawalpindi, Sargodha, Faisalabad, D.G. Khan, Lahore, Gujranwala, Sahiwal, Multan, Bahawalpur), and Punjab Board of Technical Education, Lahore. Five educational boards quoted the total number of their students thus: BISE Multan 74,491 students, BISE Bahawalpur 57, 339, BISE Faisalabad 97,528 students, BISE Sargodha 53,246 and Gujranwala 141, 726 students. The number of intermediate students enrolled in five boards was 4224, 330 (The Dawn, 2019). If all Punjab board students were counted, approximately more than 1 million students studied these books every year. So, the intermediate was an essential academic gateway for a large majority of Pakistani students before starting any professional studies or career.

Four intermediate English textbooks comprised six genres. In 1st year, Book I included 15 short stories, and Book III covered three one-act plays and 20 poems. 2nd year English syllabus included Book II which covered ten literary essays and five biographical essays and James Hilton's one novel titled *'Good Bye Mr Chips'*. Furthermore, educational data mining specifically required educational content or textbooks.

Research data were available online in the form of scanned images, so they were composed and preserved digitally; thus, this process was named digital reformatting. Its corpus consisted of 82,487 words, and it was named Pakistani Intermediate English Textbook Corpus Zafar Ullah (PIE TCZU). The novel was downloaded and tallied with the original paper textbook because there were some changes; thus, the data were cleansed. The rationale behind the digitization of texts was to enable them for machine readability, interchangeability and transmission ability to exhibit its interactive visual data, numeric values and visual trends. Besides, this study was a secondary research, since it digitally analysed secondary sources of textbooks. This study was justifiable, because many critics considered learning or teaching as a "neglected stepchild" in digital humanities (Gold, 2012). One more justification was that TEFL textbooks were considered as the most influential source of learning a foreign language (Littlejohn, 1998, p. 190). Five research questions showed the necessity and justification of the required data. The research questions presented various tools to produce textual, visual and tabular data.

3.3.2.1 Combination of Research Questions, Variables and Methodology

The current study addressed the following research questions for each unit of intermediate English textbooks:

i. How does text mining summary discover stylometric features from intermediate English textbooks?

ii. How does an interactive word cloud/ Cirrus reveal major themes and characters from intermediate English textbooks?

iii. What types of collocation patterns/ n-grams have been unveiled to extract the standard phraseology with its parts of speech?

iv. How do knowledge graphs present the interrelationship of various key themes and characters for digital hermeneutics?

v. How does the context of certain problematized words disambiguate the word sense by showing interactive bidirectional context?

In question i, text mining summary was an independent variable, but stylometric features were dependent variables. The text analytics summary was created by the Summary tool to produce quantitative and qualitative data.

In question ii, a word cloud was an independent variable and themes or characters were dependent variables. Word cloud was generated with Cirrus tool to display key themes of the data. Clicking on each theme in Cirrus showed statistical information too. The qualitative data were shown as themes, and quantitative data were displayed with statistics.

In question iii, collocation patterns/ n-grams were independent variables and the extraction of standard phraseology was a dependent variable. Phrases tool was required to generate collocations/ n-grams of each lesson for learning a language. Word count and length were in the quantitative form, while collocations/ n-grams were in the qualitative form.

In question iv, the KG term was an independent variable and the interrelationship of various themes and characters was a dependent variable. Links tool created a colour-coded KG which showed the interrelationship of certain words. Several words and their relationships could be added or decreased interactively.

In question v, the context of selected words was an independent variable, while semantic comprehension was a dependent variable. To explore the certain context, Contexts tool was applied to find a key word and its bidirectional context for word sense disambiguation.

3.3.2.2 Research Questions as a Hermeneutic Activity.

"Start with questions" (Rockwell, & Sinclair, 2016, p. 169), and the raised questions also altered as the researcher went deeper into the text and digital tools. The researcher explored anomalies in the results of Voyant tools, but they led to new points of inquiries and solutions. "New questions will come faster than answers" (Rockwell, & Sinclair, 2016, p. 169). Therefore, any type of answer, whether affirmative or negative, could lead to new hermeneutic dimensions. Besides, five research questions were harmonious with five research objectives.

3.3.2.3 Interface between Research Objectives and Tools.

This study endeavoured to accomplish the following objectives:

i. To produce a summary of text mining to extricate quantified information about stylometry, vocabulary density, the average length of sentences and the most frequent words in the corpus.

ii. To generate Cirrus/ word clouds to unveil the prominent motifs and characters.

iii. To point out collocation patterns/ n-grams to extract the most frequent standard phraseology.

iv. To create knowledge graphs to explain the interconnectivity of various themes and characters in digital hermeneutics.

v. To explore the bidirectional context of ambiguous words to comprehend the contextual word sense.

Cirrus, Phrases, Links, Summary and Contexts tools in Voyant open-access suite facilitated in accomplishing the aforementioned five objectives respectively. These tools provided fast and accurate quantification along with data visualisation. Machines just support the quantification and data visualisation process. "It is you who decide … what the result mean" (Rockwell, & Sinclair, 2016, p. 55). It meant that the researcher interpreted new insights and meanings from data visualization and knowledge patterns. In addition, these objectives were harmonized with the five research questions. In this section, visual, qualitative and quantitative data were generated with Voyant text mining tools.

Table 3 Intersection of Research Questions, Objectives, Visuals and Tools

RQ No.	**Obj No.**	**Tool from Voyant and Visual Name**	**Terms**
1	1	Summary	Corpus, Total words, Unique words, Vocabulary density, Stylometry
2	2	Cirrus	Cirrus, word cloud
3	3	Phrases	Phraseology, Collocations/ n-grams, Phrases
4	4	Links	Knowledge Graphs
5	5	Contexts	Word sense disambiguation, KWIC (Key Word In Context)

3.3.3 Mixed Methods Approach and Rationale

The mixed methods approach combined qualitative and quantitative methods in data collection and analysis. Their triangulation facilitated in refined, comprehensive and authentic comprehension of research issues, as it was compared with one approach or research method (Creswell, & Clark, 2017). Some qualitative content, for instance, themes in word clouds, collocations/ n-grams, knowledge graphs, most occurring words and bidirectional context were present in data analysis. Some quantitative content, for example, statistical frequency, number of collocations, total words, unique words, vocabulary density, average words per sentence and counting of most occurring words.

Voyant text mining tools use inbuilt statistical models, algorithms, topic modelling, probability models and Python libraries to quantify texts. These automated machine learning models are far advanced, efficient and accurate than traditional statistical methods and packages. Furthermore, this study exhibits the use of latest text mining methods which intersect NLP, AI, Statistics and ML, that is why it does not require the use of SPSS, descriptive or inferential statistics. Descriptive statistics summarizes the given quantities. This study does not present quantitative data from population or samples. Input is given in the form of texts, and the same has been done in other text mining studies. Inferential statistics is used for hypothesis testing and this study does not present any hypothesis, so the use of inferential statistics is inapplicable in this study. To conclude, this study utilises statistics used by Python libraries and statistical models which have been embedded in Voyant tools.

The rationale for using mixed methods for the current study was due to its nature because data analysis was done with digital tools. Being cryptographic machines, computers produced “numbers, color, letters and so on” (Rockwell, & Sinclair, 2016, p. 26). Furthermore, it was a fact that text or qualitative data were an integral part of text mining. Besides, computers “replace intuition with quantification” (Kenny, 1992 as cited in Rockwell, & Sinclair, 2016, p. 26) to explore innovative knowledge patterns to support hermeneutics. Consequently, mixed methods were most appropriate for data generation and data analysis.

The mixed methods approach was designed in the following four methods:

1. **Triangulation Design:** It is comprised of convergence and data transformation models.
2. **Embedded Design:** It contained experimental and co-relational methods.
3. **Explanatory Design:** It incorporated follow-up and participant selection models.
4. **Exploratory Design:** It included instrument development and taxonomy development models (Creswell, & Clark, 2017). From these four types, the current research mainly opted

for embedded design since this experiment was conducted with five tools voyantly (the use of Voyant tools to mine text, study, teach, and conduct research). Moreover, the co-relation of two variables had been studied in this dissertation as they had been mentioned in each research question and in 3.3.2.2.

3.3.3.1 Comparison of Mixed Methods, Multi Methods and Triangulation.

Due to the interchangeability of multi-methods and mixed methods terms, in multi-methods, "multiple types of qualitative or quantitative data were collected" (Creswell, & Clark, 2017, p. 273). It combined two or more methods of the same type either qualitative or quantitative. Contrary to it, mixed methods "incorporate collecting both qualitative and quantitative data" (Creswell, & Clark, 2017, p. 273). It essentially combined at least one qualitative and one quantitative method. Consequently, mixed methods is more robust and diverse. Campbell and Fiskel (1959) introduced the triangulation research method, and later Webb et al. (1966) and Denzin (1970) expanded it. Denzin (1970) mentioned four types of triangulation: data, investigator, theoretical, and methodological triangulation. The current study followed three features, namely data triangulation, theoretical triangulation and methodological triangulation.

3.3.3.2 Justification of Triangulation

Knowledge Discovery Theory and Hermeneutica Theory have been triangulated for presenting a solid theoretical framework. Firstly, both theories share some common features, that is why they have been triangulated. Their common features are: Secondly, digital humanities is an intersection of humanities and computer science, so it was quite rational to amalgamate one pure computer science theory and the other one from humanities. Thus, KDD represented computer science theory while Hermeneutica Theory represented humanities. As Hermes, the Greek messenger god, used to explain god's messages to other gods and goddesses. Thirdly, Hermeneutica theory was propounded by Voyant tool designers and they explained the theory in the light of their tools and its hermeneutic dimensions. The aforementioned features justified the triangulation of the two theoretical frameworks. The triangulation had been justified in the light of these points: Olsen and Wendy (2004) emphasized that triangulation enhanced deep perception and comprehension of the research. Webb et al. (1966), Smith and Kleine (1986) and Denzin (1978) posited that triangulation was very suitable for enhancing the accuracy of the research. To sum up, triangulation was employed to explain statistical tests and their accuracy with other measuring tools (Kadushin, Hecht, Sasson, & Saxe, 2008), while mixed-methods opted for both qualitative and quantitative methods in one research.

3.3.3.3 Eight Phases of Mixed Methods

The mixed methods model performed eight phases:

(1) Research questions led to the research methods; (Five research questions guided to the use of mixed methods.)

(2) The decision of suitability of mixed methods approach; (Visuals, statistics and texts were involved in this research, so mixed methods approach was most appropriate to analyze visuals, statistics and texts.)

(3) Selection of mixed methods design in the research; (Presence of qualitative and quantitative data led to selection of mixed methods.)

(4) Data collection: (Hard copies and scanned texts were present but this study digitized these books.)

(5) Data analysis: (Data had been analysed in the light of Knowledge Discovery Theory and Hermeneutica Theory.)

(6) Data interpretation: (During data interpretation, various pedagogical, curriculum designing, insights had been derived. By the same token, interpretations led to implications for this research for learners, teachers and publishers.)

(7) Legitimization of data: (It is legal and permissible to digitize and analyze textbooks.)

(8) Drawing appropriate conclusions: (Conclusions had been delimited in major findings.) (Johnson, & Onwuegbuzie, 2004). The current study followed the proposed research strategy of mixed methods and its all phases.

3.3.3.4 Purposes of Mixed Methods

The purpose behind applying the mixed methods approach was to minimise the intrinsic biases of the researcher, measure biases, sampling biases, and procedural biases. Furthermore, it was employed for cross-checking of methods and to increase the credibility and validity of the research. It was a verification process of several methods for a single point of inquiry. Moreover, it was used for completeness and confirmability. When several methods of inquiry and theories were applied on a single point, their results validated and authenticated one another's findings. Consequently, its results were better than a single tool or a single method of inquiry.

3.3.3.5 Rationale for Mixed Methods

The rationale for the use of mixed methods was that one method was not sufficient for the current research design. Firstly, pragmatism supported mixed methods. Secondly, both qualitative and quantitative type of visuals and data were generated by Voyant text mining tools; hence, mixed methods were required in both data collection and data analysis phases. Thirdly, this study tested the functionality of Voyant tools for qualitative and quantitative knowledge patterns. Fourthly, the research questions of the current study necessitated mixed methods in analysis. The new insights had been derived to benefit linguistics, pedagogy and academia. Fifthly, hybrid learning (Adriaans, & Zantinge, 2009, p.103) was supported through data mining tools, and this hybridity required an application of mixed methods.

3.3.4 Data Quality for Data Mining

A standard and good quality data produced the best knowledge patterns, so the emphasis was laid on data cleaning. Ten rules for reliable DM had been presented here:

3.3.4.1 Ten Golden Rules for Reliable DM Environment.

The current study followed most of the features from ten data mining rules.

1. Supporting Big Data: Big data or billions of entities or items were mined in the DM process, and the current study covered 84,400 words data.

2. Supporting Hybrid Learning: Learning tasks had been classified into classification, knowledge engineering and problem-solving tasks. To detect different patterns, different learning algorithms were required. The current study promoted hybrid learning with different types of data.

3. Establishment of Data Warehouse: Without a data warehouse, no DM was possible. The corpus was also a data warehouse for EDM.

4. Data Cleaning: Whenever a polluted element was found, it was removed in no time to acquire the desired results. Advanced tools and de-duplication processes were applied to clean data. The current textual data were cleansed and adapted for getting better results.

5. Dynamic Coding: Creative coding was the essence of data mining to unveil new knowledge patterns. The application of codes was based on time or attributes, and the current study coded time of data visualization in the figures of Summary tool in Ch. 4.

6. Integration with DSS: Knowledge discovery procedure started with DSS (Decision Support System) activities. The current study also integrated Hermeneutica Theory into Knowledge Discovery Theory to interpret various knowledge patterns.

7. Addition of Extendible Architecture: Wide-ranging capacity to integrate new tools was enabled. The current study extended uses of Voyant tools for analysing code-mixed texts and analysis of new oriental languages, as mentioned in 5.7 and 5.10

8. Capacity to Integrate Heterogeneous Database: Complete data were not available in a single uniformed format. So, systems and tools must support various files, formats and data warehouses. Voyant tools can support web links, MS word, pdf, text and other files.

9. Server Architecture: The server was a flexible machine that transferred the burden of visuals to the local machine; as a result, DM processes became faster.

10. Introduction of Cache Optimization: Data were stored in separate files in internal memory for quick and frequent access. To maximize the function of the learning algorithm, a low-level integration was added with a database for better performance (Adriaans, & Zantinge, 2009, pp.102-106).

Data archiving saved older data that was occasionally used or accessed. The main advantage was that it decreased primary storage. It could be stored online or offline in a file or object form. Archive transparent was used to save exact data (Komprise, 2017) to enhance data quality and speed.

3.3.4.2 Voyant: Text Mining Tools and Rationale.

Computers were used as an instrument and “heuristic tools” (Berry, 2012, p. 69). On the other hand, Knorr-Cetina (1992) criticized making humanities a laboratory science subject. He posited that humanities must remain very close to humanism and not with machines. Countering the argument, Rockwell and Sinclair (2016) argued about the use of Voyant tools for “thinking through tools and experimenting with texts” (p. 6). Concluding the debate, technology-facilitated in the accurate analysis and thematic access from voluminous texts. Traditional close reading took decades to study the big data, while technological study/ distant reading/ hyper reading took a few minutes to read large data precisely. Thus “reading and playing can be combined in tools that have the ability to generate live panels” (Rockwell, & Sinclair, 2016, p. 63).

Open access Voyant tools were developed by Stefan Sinclair and Geoffrey Rockwell in Canada in 2003, and later, they were upgraded in 2013. The important tools had been described in the following lines: Bubblelines (the size and number of bubbles which showed presence of a linguistic item in a corpus); Corpus Summary (it presented density, unique words); Cirrus (word cloud along with its frequency); Corpus Grid (it showed corpus data with token words, unique words); Corpus Type Frequency Grid (it showed frequencies from the highest to the lowest level); Document Type Collocate Frequencies Grid (ordered frequencies with words); KWIC Grid (key words were shown in their context); Knots (colourful corpus generated threads which were twisted and they showed interrelationship of different terms while different colours showed their linguistic codes); Links (it showed interconnectivity of words as KG in a corpus); Mandala (it showed frequency and linkage); Reader (it showed full text in its original form); RezoViz (a visualization tool to show ties among people, places and organizations);

Scatterplot (it visually displayed the usage of words in a corpus); Termometer (it showed variations in frequency in different time spans); Termsradio (it showed frequency on scrolling line graph in different time spans); Type Frequencies Chart (it showed word occurrence on a line graph); Word Count Fountains (it showed frequency of words in the visual of fountains) (Sinclair, & Rockwell, 2017).

As a justification of Voyant tool for this research was concerned, “Voyant and other text analysis tools fit into the tradition of research practices of the humanities” (Rockwell, & Sinclair, 2016, p. 195). These tools were visual, analytical and repleted with hermeneutic knowledge patterns. Furthermore, these tools were used for deriving knowledge patterns in twenty-two renowned universities, six libraries in top-ranking universities, and University of Melbourne, Australia used them in digital humanities classes (Sinclair, & Rockwell, 2017). Digital tools just assisted the researcher who derived insight from visual data and tools. The same was conveyed by Voyant designer “it is you who decide … what the result mean” (Rockwell, & Sinclair, 2016, p. 55). It meant that the researcher who interpreted new insights and meanings from data visualization and knowledge patterns.

3.3.4.3 Delimitation in Voyant Tools

From 25 Voyant tools, five tools, namely Summary, Cirrus, Phrases, Links and Contexts**,** were delimited to accomplish the current research.

i. In Summary tool, the top 10 most occurring words were specified for each lesson except poems whose only five most occurring words were specified because of their terseness.

ii. To refine results, Stopwords (function words) were automatically excluded from the results and visuals.

iii. Voyant tools were interactive, but to export them on a piece of page, there was a requirement of a still image. Besides, there was also a need to specify a certain number of themes to create harmony and uniformity in the analysis section. Research data consisted of short stories, poems, one-act plays, essays, biographies, so 25 words Cirrus was selected for each of them. The novel was a big document, so its 95 words Cirrus was generated.

iv. Phrases tool generated a large number of collocations/ n-grams from the corpus. So, to delimit them, the top 15 phrases concerning their length were chosen in descending order to learn standard phraseology from each lesson of three textbooks. Further delimitation was done by selecting only standard phraseology, and substandard collocations/ n-grams were omitted.

v. With Contexts tool, the searching context of any one word from each lesson was delimited to comprehend their contexts and disambiguate word sense. Since the novel was a big document, the context of two words was selected for this study.

3.3.4.4 Rationale for the Use of Voyant in Text Mining.

In the data mining process, Voyant text mining tools were used. Firstly, the major reason to use Voyant tools was the quick generation of the corpus, statistics, tabular data and multiple visuals; hence, they saved time and labour. The second justification was the search for accurate statistical data and corpus for small or big data. Thirdly, Voyant tools were the most user-friendly tools for technophobes. Uploading data was an easy task for social scientists. Fourthly, Voyant was "an aid" (Rockwell, & Sinclair, 2016) for agile hermeneutics and interpretive thinking. To conclude, "Voyant is ideal" for teaching purposes in the classrooms (Graham, Milligan, & Weingart, 2013).

3.3.4.5 Data Mining Procedure

After composing of data, each unit file was uploaded on Voyant tools for individualized text analytics. Voyant tool, an aid to thinking, facilitated the process of interpretation and hermeneutics. Furthermore, its generated data were visible on five skins, "an environment for interpretation" (Rockwell, & Sinclair, 2016, p. 72) of Voyant tools. There were options to utilise more tools for the selected data. Then, analysed data of each skin was exported to a Portable Network Graphics (PNG) file.

3.3.4.6 Data Mining Processes

Data mining had three processes.

1. Data Preparation/ Training Data: This preparatory phase selected, collected, formatted, cleaned and processed data.

2. Real Process of Data Mining/ Testing Data: In the actual data mining phase, a suitable data mining solution was chosen according to the selected data. The selection of suitable parameters led to the creation of useful information patterns.

3. Post Data Mining Process/ Result Validating Data: This phase selected, evaluated, and interpreted patterns; hence, only informative and interesting patterns/visuals were selected with a uniform criterion.

3.3.4.7 Seven Data Mining Phases

There were seven phases of data mining in principle.

1. Data Mining: A warehouse was created by collecting data.

2. Data Filtering: It extracted the required data records from operational data.

3. Data Standardisation: Usually, variables were codified in a standard format for instance, 00, 01. On the other hand, Voyant data did not require standardization.

4. Data Cleansing: It meant to load data warehouse and harmonize it according to standards while all mismatched data were abandoned.

5. Data Summarisation: Data warehouses were summarized and named data cubes.

6. Data Security: When many users accessed the same data warehouse, there were risks of data jamming, hijacking or overloading. Fool proof data security and privacy systems must protect data and their functions.

7. Data Visualization and Analysis: Most of the data mining or text analytics tools exhibited data in the form of various interactive visuals. After exporting data on word files, those visuals were analysed according to research questions (Zhong, & Skowron, 1999). Following them, the current study also analysed data visualization in Ch.4.

3.3.5 Data Handling and Data Analysis

The current research followed mixed-methods approach for data analysis, since it showed qualitative, quantitative and visual data. Moreover, three types of data and their analyses validated one another’s results. Qualitative data elucidated data mining patterns in the display of Cirrus, Phrases and Contexts. Thus, unstructured or semi-structured data were transformed into structured data. Besides, textual quotes were presented in inverted commas to differentiate between exact data words and interpretation. Previously, descriptive and predictive approaches were applied to text mining, and the current study also applied descriptive and interpretive approaches to knowledge patterns. Quantitative data showed numeric data of word count, corpus, statistics and frequency of collocation patterns/ n-grams, while visual data were manifested in the form of word clouds and KGs.

3.3.5.1 Data Analysis and Interpretation.

John Tukey, a renowned statistician of last century, informed some decades ago that “Data analysis must progress by approximate answers” (Tukey, 1962, p. 14). Powers (1996) used computation for the interpretation of the text. The current study concentrated on statistics which led to produce multiple visuals and their interpretations.

Text mining covered three steps: demarcation, digitization and encoding (Rockwell, & Sinclair, 2016, p. 102). Demarcation or fixing boundaries to show the text as a string for fast and easy processing. The digitized text could be taken from the internet or composed manually as it was done in the current study. Voyant tools did not necessitate any encoding with UTF-8.

The question arose that how a computer-processed the text in the analysis phase. Digital tools could not read semantic shades, but they performed some machine learning rules and processes of homogeneous or heterogeneous characters to save human labour. Mechanical data had been purged from human subjectivity. Machine-generated analysis and visuals were epistemologically stronger than human analyses (Rieder, & Rohle, 2012), because they were free from human biases. Above all, they were interactive to be modified according to data and queries. Hao Wang, a famous mathematician, said that machines processed many functions faster and more accurate than human beings; that is why machines were “persistent plodders” (Wang, 1963, p. 93). Even technology gave “an exact answer for the wrong question” (Tukey, 1962, p. 14).

Visuals had compressed argumentation power to represent some significant epistemological information, data and universalism. Images and visuals elaborated scientific phenomena and now linguistics had also changed into a scientific study. In research analysis, visuals, numeric data and texts were explained as a "literary technique" (Shapin, & Schaffer, 1985). Graphical visuals informed about the substance of data, however, graphics did not guide about methodology or design of the visuals (Tufte, 2001). The critical researcher derived knowledge from interactive visuals. Furthermore, the same data could have been revealed with various interactive visuals because "rhetorical potential of hermeneutica lies in their interactivity" (Rockwell, & Sinclair, 2016, p. 193). Same visuals were also called knowledge patterns, since they were replete with accurate and quantified information. "Patterns as a method of analysis" (Dixon, 2012, p. 192) were a means of knowledge discovery in Voyant tools.

3.3.5.2 Rationale for Data Analysis.

The raised questions were answered with the evidence of visuals, qualitative data, quantitative data, corpus, interpretive ability and hermeneutic canons to make data analysis reliable. When tools were involved in text mining, data were quantified. Secondly, nobody could falsify corpus, statistics and visuals, because they were built on statistical data. Human computing and machine-generated results were found identical. Thirdly, quantitative and qualitative analyses matched, and they counter checked each other to enhance reliability and validity.

3.3.6. Research Management and Application of Research

This segment dealt with managerial issues involved in the research process. Fulfilment of ethical considerations transformed this research work into a legal and academic activity.

3.3.6.1 Ethical Considerations.

No human participant was involved in this research; therefore, there was no need to seek any permission from human participants. Research data consisted of intermediate English textbooks which were open-source in scanned form and accessible to all and sundry in hard copies. Therefore, no permission was required to mine textbooks. Secondly, Voyant tools were also open access which meant anybody could use it without prior permission or payment. I emailed Geoffrey Rockwell and Stefan Sinclair, designers of Voyant tools, to seek permission and guidance, and they gladly approved to conduct this study. They replied to all my queries in detail.

3.3.6.2 Setting.

This academic inquiry was conducted in Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA and Faculty of English Studies at National University of Modern Languages (NUML), Islamabad, Pakistan.

3.3.7. Research Skills

The current study belonged to digital humanities since this research used web technology of Voyant tools for the generation of visuals and their analysis. So, it required competence in Voyant text analytics tools. Furthermore, visual data were one segment of Data Presentation Architecture (DPA). They were analysed with DPA skills that sought to recognize, search and display data to convey meanings and to proffer knowledge effectively, efficiently and accurately (DMK Force, 2008). The researcher acquired computer expertise; studied Applied Machine Learning course; and attended many lectures and workshops about DH, AI, ML, big data, language technology at Carnegie Mellon University, USA (Appendix H) to mine textual data successfully.

3.4 Validity

Validity means believability and rational feasibility of the research process and research findings. It had internal validity and external validity. Internal validity referred to internal procedures and tools to authenticate the entire process, while external validity showed the capability to generalize findings of this study for other phenomena (University of California, Davis, 2009).

3.4.1 Evaluation Parameters for Data Mining Results

Following parameters validated the interestingness of discovered patterns in text mining.

1. Conciseness: Knowledge patterns must be limited since brevity is the soul of wit.

2. Coverage: It showed generalizability that results were generally based on association if data were comprehensive.

3. Reliability: Its reliability laid in accordance with association rules and classification rules. The derived patterns were reliable, if they were applied to the majority of the text.

4. Peculiarity: Knowledge patterns must be unique and different from norms.

5. Diversity: Knowledge patterns were not identical to other patterns.

6. Novelty: Novelty was the basic goal of data mining.

7. Surprising: Surprising patterns were interesting, and they differed from existing patterns.

8. Utility: Data mining processes were meant for the accomplishment of some objectives. If these patterns provided some utility, they were interesting for research and knowledge discovery.

9. Applicability: If these knowledge patterns were applicable for prediction, exploration, interpretation and problem solution, they were very valuable (Du, 2010, p. 26). Knowledge patterns of the current study showed all the aforementioned parameters.

3.4.2 Concurrent Validity

Criterion validity had been divided into two further branches: Concurrent validity and predictive validity. "Concurrent validity focuses on the extent to which scores on a new measure are related to scores from a criterion measure administered at the same time" (Salkind, 2010, p. 209). Through this validity method, findings and results were measured and compared with an already established measure. In the current study, Cirrus tool generated key themes and characters, and they were compared with humanly extracted characters and themes mentioned by Hussain (2019)..

3.4.3 Interpretive Validity

Corpus Summary, Cirrus, Phrases/ Collocations/ n-grams, KGs, KWIC with visual, quantitative and textual data were interpreted in the current research. Consequently, they harmonized and cross-validated the interpretation.

3.4.4 Validity of Methodology

Mixed methods approach was applied to methodology and results to strengthen their findings. Thus, cross-validation reinforced and proved the findings of the current research.

3.4.5 Theoretical Validity

Triangulation of Knowledge Discovery Theory and Hermenutica Theory was applied for text mining. The extracted knowledge patterns were based on facts (Cohen, Manion, & Morrison, 2002), and they were validated with textual evidence and other Voyant tools.

3.5 Reliability

Reliability means repeatability of the same research results after conducting the same specific research process. Such tested findings were reliable, scientific, and true in their pragmatism because the same results and data visualizations were found every time. Moreover, the reliability of Voyant tools was expressed with this function, that Voyant tools were better able to perform tedious and complex tasks “without over-interpreting them” (Rockwell, & Sinclair, 2016, p. 43).

3.5.1 Reliability of Voyant Tools

As the reliability of Voyant tools was concerned, they were applied in research papers, teaching and symposiums. Some examples were: 22 famous universities of the world, including University of North Texas, USA, Stanford University, USA, Michigan State University, USA and others used Voyant tools for teaching to their graduate students. Moreover, University of Melbourne, Australia taught Voyant tools to their PhD digital humanities students. seven international digital humanities and data mining conferences employed Voyant tools. Six universities, including University of California, Indiana University, Western Michigan University, USA mentioned Voyant tools on their library pages. Twenty-eight literary research works and sites employed Voyant tools. 13 blogs introduced and discussed Voyant tools (Sinclair, & Rockwell, 2017). Voyant based research projects were published in impact factor journals, and they were cited hundreds of times in other impact factor journals and books. The aforementioned uses and citations of Voyant tool proved its reliability for knowledge patterns, research, academia, pedagogy and library study.

3.6 Conclusion

Concluding this section, the researcher followed the triangulation of the research onion (Saunders' et al., 2012) and Durant (2004)'s seven-pointed typology of research. In the research onion, the current study chose post-positivism philosophy, inductive approach, experimental strategy, 2017–2018-time horizon and digitization of ESL textbooks for data collection. Mixed methods approach was applied for data analysis. Besides, Durant (2004)'s the seven-pointed typology of research: i. Framework for Research ii. Data Generation iii. Research Method Approach and Rationale iv. Data Quality and Data Mining v. Data Handling and Data Analysis vi. Research Management and Application of Research vii. Research Skills were also followed in the research design of the current study.

The subsequent section mines intermediate English textbooks with five Voyant tools.

CHAPTER 4

DATA ANALYSIS

4.1 Introduction

Carly Fiorina, CEO of HP said, “The goal is to turn data into information, and information into insight” (Lloyd, 2016). Data information and analysis leads to deriving insights for better understanding and intellectual depth. To materialize the same idea, text mining insight theory has been presented in contribution section. Therefore, textual and statistical data are present in Summary phase; this section examines data visualization in the form of Cirrus (word cloud); tabular data as collocations/ n-grams; Knowledge Graph (KG); and tabular data in Contexts panel. In 1st year of intermediate, Book I and Book III are taught; and in 2nd year, Book II and a novel *‘Good Bye Mr Chips’* are taught; hence, they have been analysed in the sequence of class wise teaching and not in the sequence of book numbers, and the same has been shown in figure 10.

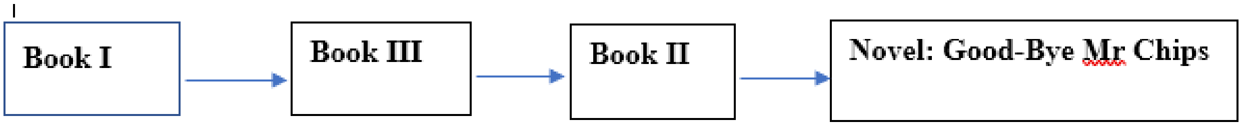


Figure 10 Flowchart of Research Data during Data Analysis

Digitized text or corpus of each lesson has been changed into five types of visual data with Voyant text mining tools, and then previously known knowledge about themes has been confirmed with the work of Hussain (2019), but this is just a partial and very small aspect of this study. Major work revolves around discovering new knowledge patterns, linguistic aspects and data visualization. The same has been buttressed through Knowledge Discovery Theory by Rakesh Aggrawl and Hermeneutica Theory by Geoffrey Rockwell and Stefan Sinclair. Consequently, new knowledgeable patterns have been discovered to explore new insights in the light of Knowledge Discovery Theory and Hermeneutica Theory. Primarily, data mining aims to describe and predict data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Research questions, objectives, data generation tools and data visualization correspond to one another in this study, and their description has been presented in table 4.

Table 4 Interrelationship of Data Analysis Elements

RQ No.	**Obj No.**	**Tools and Visual Names**	**Terms**
i	i	Summary	Corpus, stylometry, vocabulary density,
ii	ii	Cirrus	Cirrus, word cloud
iii	iii	Phrases	Phraseology, Collocations/ N-Gram, Phrases
iv	iv	Links	Knowledge Graphs
v	v	Contexts	KWIC, Word sense disambiguation

4.2 Data Analysis of Book I

Book I contains 15 short stories (Appendix A), and almost 12 of them have been written by American short story writers; 2 indigenous short stories have been translated from Urdu to English, and 3 Persian tales of Sheikh Sadi (as one short story) have been translated into English.

4.3 Text Mining of Short Stories

Each short story has elaborated its potentially useful knowledge patterns for instance, key themes and characters, collocation patterns/ n-grams, knowledge graphs, stylometric qualities and word sense disambiguation of problematized words.

1. *Button, Button* by Richard Matheson

i. 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); think (11); asked (10);

Figure 11 Summary, Button, Button

Stylistics is the study of the literary style of any writer, and computational stylistics extracts total words, unique words, vocabulary density, average words in a sentence and the most occurring words/ themes or characters. Summarization shows a report of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45); so, Summary tool exposes Richard Matheson's computational stylistic qualities/ stylometry. In this short story, total vocabulary is 2152 words, but unique words are 660. Vocabulary density has been derived by keeping unique words as the numerator and total words as a denominator (unique words/ total words; 660/2152= 0.306, and then their answer is round figured to find vocabulary density. The foregoing discussion implies that after how many words, a new word appears in the text. Mathematically, it is named Inverse Absolute, and it is taken with the division of total words by unique words (Simpson, 2000). Higher vocabulary density expresses less repetitive usage, which increases the vocabulary difficulty level for intermediate-level readers. Lower vocabulary density informs about the use of frequently repeated words in the text; consequently, it leads to textual ease for readers. In this short story, 9.9 average words per sentence discover the knowledge pattern that most of the sentences are small and simple; hence, basic language learners can easily read them. The most frequent words present key themes and characters of the short story.

Corpus and statistics strengthen each other, and their collaboration derives stylistic patterns, since statistics has been used for the extraction of stylistic features (Amancio, 2015). Moreover, findings of the current study verify these literary stylometric studies (Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O'Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018). In Hermeneutica Theory, "exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166) are hallmarks of hermeneutica, so Summary tool explores the corpus summary which makes us understand stylometric features of any text and its writer.

ii. Cirrus

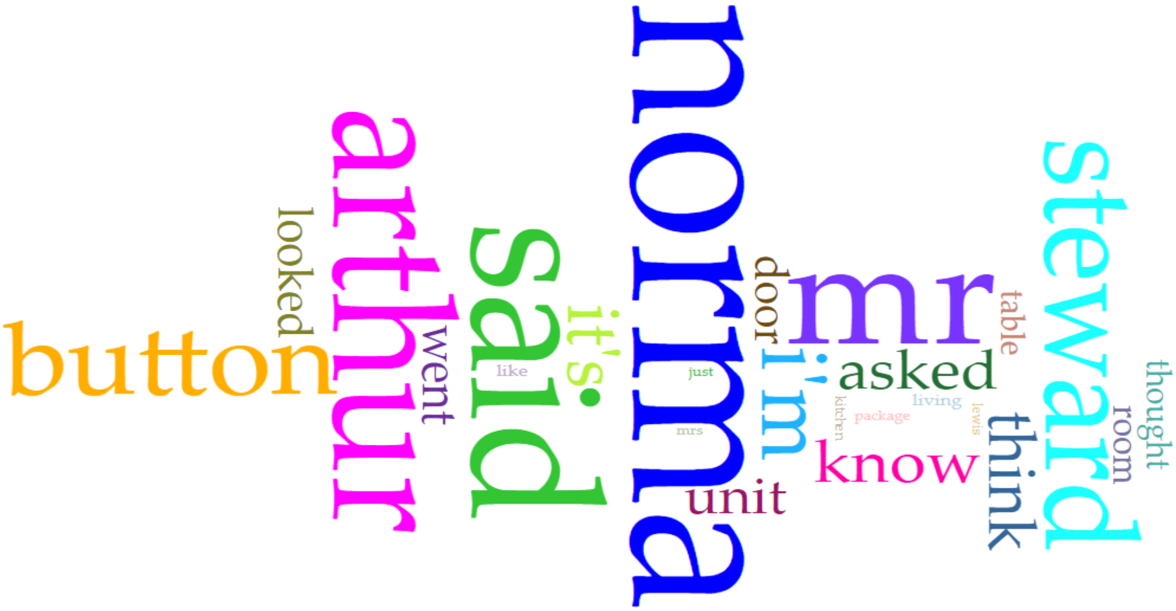


Figure 12 Cirrus, Button, Button

Cirrus, a Latin word for word cloud, is used for topic modelling which generates topics based on statistics within the domain of Machine Learning (ML). In Cirrus tool, word cloud manifests thematic and statistical information. Theme means a major subject of discussion in the text, and these themes and characters predominate in the Cirrus. The word “said (31)” refers to a dialogic style of the short story, and characters exchange ideas directly during their interaction. Another interesting discovery of knowledge pattern is “say (5)”, “saying (4)”, “said (31)” and “asked (10)” which are 50 words, while “think (11)”, thought (8)” are 19 words. The foregoing discussion implies that an act of saying occurs almost three times more than the thinking process. Therefore, the thinking process is used less, and talking is more in this short story; hence, this attitude leads to the destruction and tragedy of Norma.

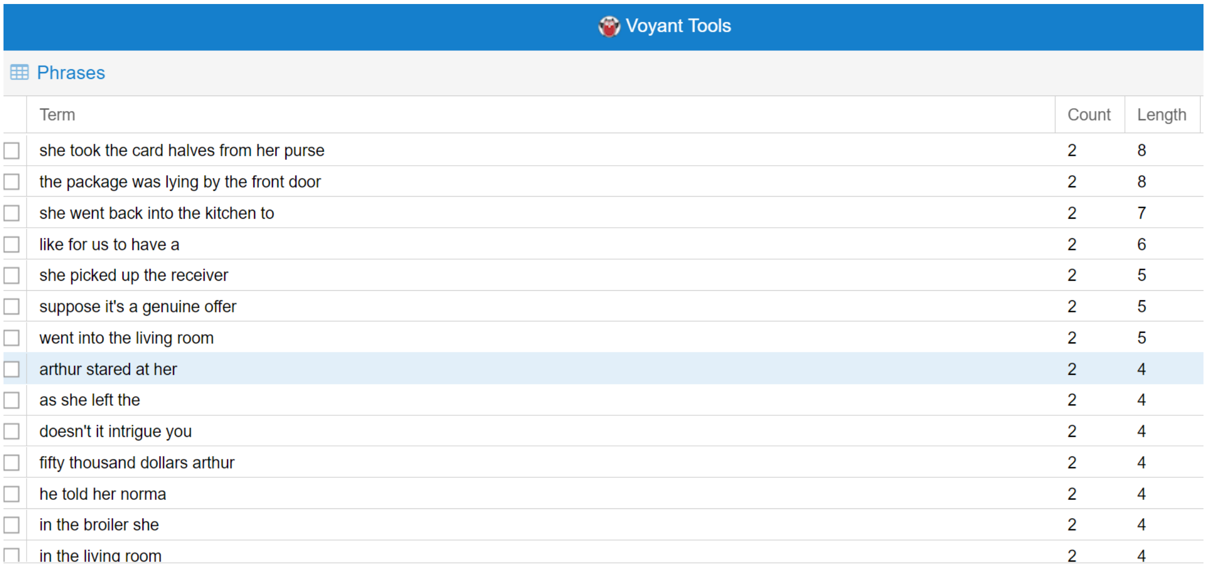
In non-living things, “button (17)” is the central object in the title of the short story, while in human beings, “Norma (41)”, the protagonist of the short story and real perpetrator of pushing the button, discusses Arthur and Steward most of the time. Another major character is her husband, “Arthur (29)” who studies most of the time and advises her to shun this button-pushing act. On the other hand, “Steward (19)” is an accomplice and persuader to push the button. Eventually, Norma pushes the button, and invites her tragic consequences to be a widow.

Hermeneutica Theory is “supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166), that is why Cirrus themes have been explicated by textual references. Moreover, there can be a difficulty to understand a theme from Cirrus, so Contexts tool supplements Cirrus tool by providing textual context and disambiguates its word sense. Furthermore, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). Cirrus tool quantifies key themes and presents them in the compact word cloud form.

The current study confirms findings of the earlier studies of Burrows, 2002; Burch, & Ertl, 2015; Hussain, 2009; Jockers, & Mimno, 2013; Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015; Sinclair, & Rockwell, 2015b; Scrivener, & Davis, 2017; Yeates, 2013 mentioned in chapter 2.

The human analysis discusses three characters, namely “Norma, Steward and Arthur” who have been discussed here. Human analysis extracts certain key themes which are “selfish woman, greed, resisting against temptation, reading, wisdom, shrieking, impulsively, smiled, shrugged” (Hussain, 2009, pp. 38-40). Comparing human and Voyant extracted themes and characters, their character extraction is same, but their themes are different because human beings give names to those themes differently according to their cognition, inclination, understanding and range of vocabulary, while Cirrus text mining tools concentrate on statistical weight and frequent repetition of any theme.

iii. Phrases



	Term	Count	Length
□	she took the card halves from her purse	2	8
□	the package was lying by the front door	2	8
□	she went back into the kitchen to	2	7
□	like for us to have a	2	6
□	she picked up the receiver	2	5
□	suppose it's a genuine offer	2	5
□	went into the living room	2	5
□	arthur stared at her	2	4
□	as she left the	2	4
□	doesn't it intrigue you	2	4
□	fifty thousand dollars arthur	2	4
□	he told her norma	2	4
□	in the broiler she	2	4
□	in the living room	2	4

Figure 13 Phrases, Button, Button

Some collocations/ n-grams are selected on the criterion of their frequent occurrence for example, “she took the card halves from her purse” (Prn+V+Art+N+N+Prep+Prn+N), “the package was lying by the front door” (Art+N+Aux+V+Prep+Art+Adj+N), “she went back into the kitchen” (Prn+V+Prep+Prep+Art+N), “like for us to have a” (V+Prep+Prn+Inf V+Art), “may be some eccentric millionaire is” (Aux+Adj+N+Aux), “she picked up the receiver” (Prn+V+Prep+Art+N), “suppose it’s a genuine offer” (V+Prn+Aux+Art+Adj+N), “went into the living room” (V+Prep+Art+N), “Arthur stared at her” (N+V+Prep+Prn), “as she left” (Adv+Prn+V), “doesn’t intrigue you” (Aux+V+Prn), “fifty thousand dollars” (Adj+N+N), “he told her” (Prn+V+Prn) and “in the broiler” (Prep+Art+N). Overall, the occurrence of each phrase is 2, and their length ranges from 4 to 8 words. Statistics counts the length of each phrase, and transform qualitative data into a knowledge pattern. They are standard collocation patterns/ n-grams to learn and teach because some previous studies on collocations/ n-grams have been used to learn and teach English (Nesselhauf, 2003; Seretan, Nerima, & Wehrli, 2004; Shin, & Nation, 2007). Another significant advantage of collocations/ n-grams is that they enhance fluency in using four language skills.

iv. Links

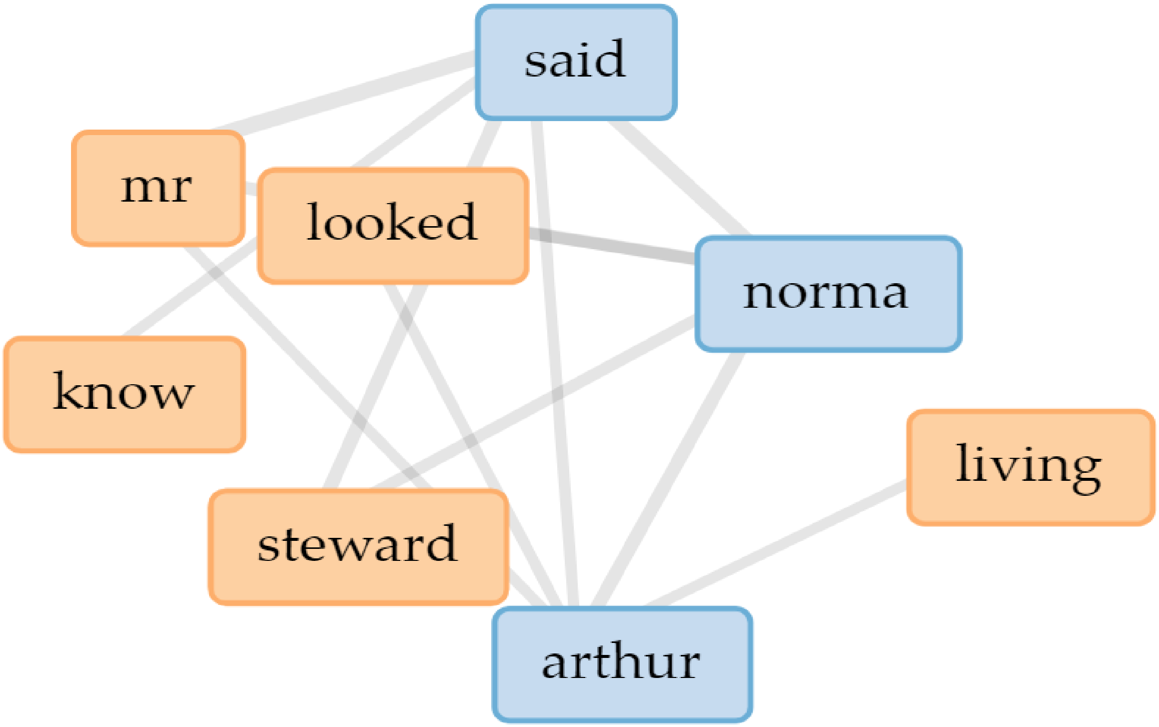


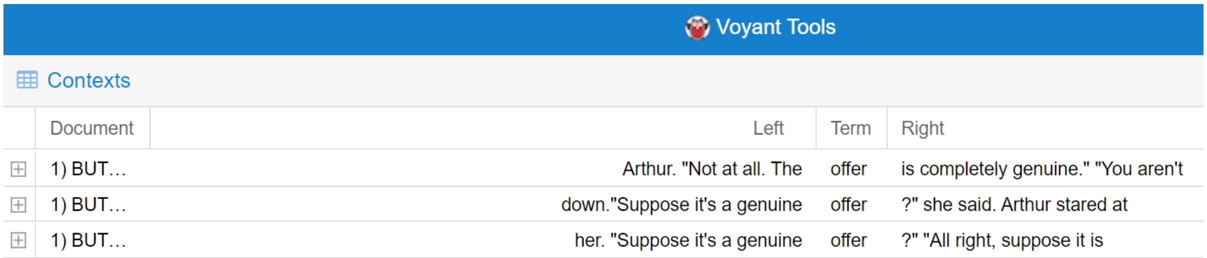
Figure 14 Links, Button, Button

A knowledge Graph (KG) shows the interrelationship of various themes and characters to explore an interesting knowledge pattern. Moreover, some nodes are zoomed in to concentrate on them, and some nodes are filtered to refine certain key themes. “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166), so different nodes of knowledge graphs have been refined for exploration and understanding of multidimensional hermeneutic patterns. One KG shows “said, Norma, Steward, Arthur”, and it shows that these characters converse in the story. It also reveals the interrelationship of Arthur and Norma because they discuss a lot about pushing the button unit to win the reward of $50,000. Moreover, the word “said” indicates a dialogic and direct conversational style for the progression of story.

In another knowledge pattern, “Steward, Norma” reveal that Steward and Norma discuss button pushing issue frequently, but Steward and Arthur are not interlinked because they rarely talk to each other. The background situational reality shows that Norma and Steward converse with each other several times, but Arthur just unwillingly exchanges one unpleasant talk with Steward. Another KG of “Arthur, living” reveals that he frequently stays in his living room, and remains busy reading some book. It seems that he manifests more bookish knowledge as compared to the pragmatic and socio-cultural knowledge. Arthur dislikes the offer of pushing the button, and to become rich on account of taking the life of some innocent person in any part of the world. Arthur is against money-making and materialism at the cost of the precious life of any human being. One knowledge graph,,“Arthur, said, Norma, know”, refers to a deficiency of shared knowledge and mutual understanding between the couple. Lack of knowing each other is the most problematised quality, and it causes their tragedy. The very thought-provoking and concluding statement of Steward for Norma is, “Do you really think you knew your husband?”, and it verifies the evidence of the KG. The word “looked” refers to the appearance of characters during dialogues, because nonverbal expressions are stronger than verbal expressions. As this short story has been written in 1970 for TV, so keeping in mind its viewers, more visual words have been used in it.

To conclude, Hermeneutica Theory guides that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166), especially in the interpretation of knowledge graphs, because different nodes have been linked with certain themes and characters, for instance, Steward and Norma are linked because of their frequent discussion about pushing the button and winning the amount to materialise her materialistic desires.

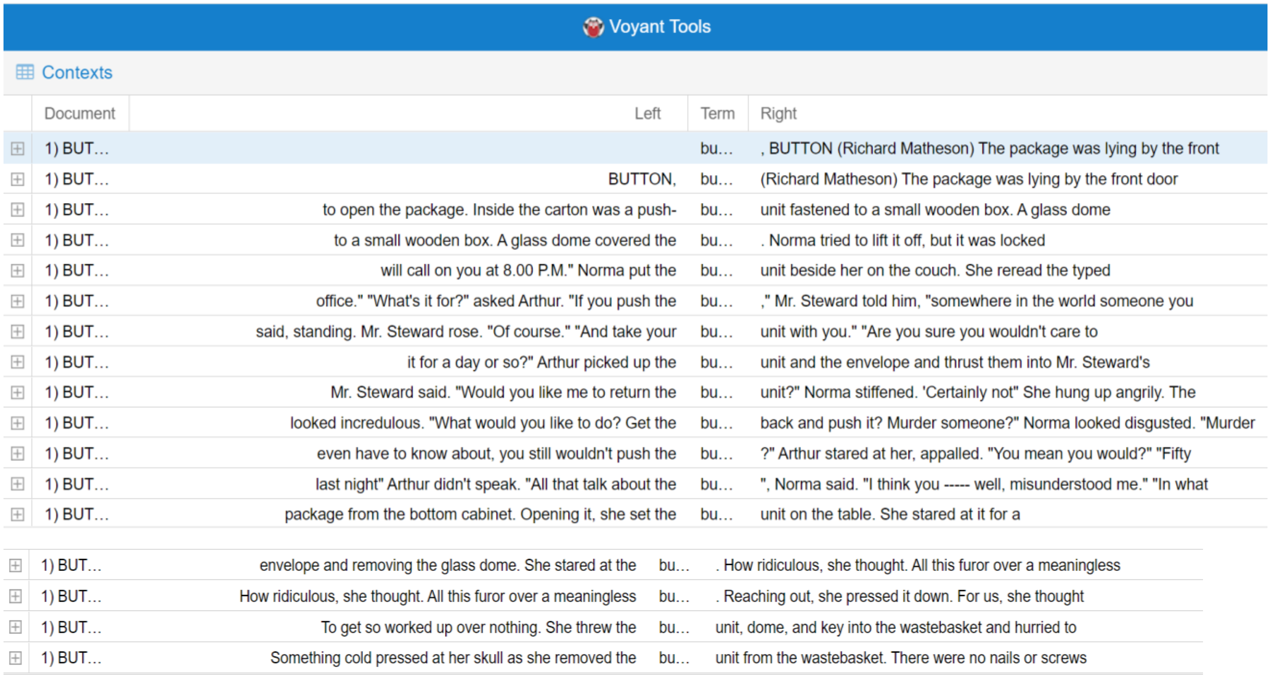
v. Contexts



Document	Left	Term	Right
1) BUT…	Arthur. "Not at all. The	offer	is completely genuine." "You aren't
1) BUT…	down."Suppose it's a genuine	offer	?" she said. Arthur stared at
1) BUT…	her. "Suppose it's a genuine	offer	?" "All right, suppose it is

Figure 15 Contexts, Button, Button

Hermeneutica Theory is embedded in a context (Rockwell, & Sinclair, 2016, p. 166). So, the use of context for finding exact word sense is the first step of hermeneutica. Word sense disambiguation (WSD) clarifies the sense of words with interactive context. Noun and verb are the same lexically, but they differ grammatically; for instance, the word "offer" can be used as an offer (verb), offer prayer (verb), accept an offer (noun), offer letter (adjective). Information retrieval (IR) searches the required word "offer", and clarifies the word sense, part of speech and semantic shade of any word. The same findings have been validated in the 2nd chapter of this dissertation by Kwary (2018). In the literature review chapter, part of speech recognition disambiguates the word sense (Bhala, & Abirami, 2014). Figure 15 shows that the word "offer" has been used three times in this short story, and every time it is used as a noun. Thus, lexical ambiguity has been resolved in figure 15.



Document	Left	Term	Right
1) BUT…		bu…	, BUTTON (Richard Matheson) The package was lying by the front
1) BUT…	BUTTON,	bu…	(Richard Matheson) The package was lying by the front door
1) BUT…	to open the package. Inside the carton was a push-	bu…	unit fastened to a small wooden box. A glass dome
1) BUT…	to a small wooden box. A glass dome covered the	bu…	. Norma tried to lift it off, but it was locked
1) BUT…	will call on you at 8.00 P.M." Norma put the	bu…	unit beside her on the couch. She reread the typed
1) BUT…	office." "What's it for?" asked Arthur. "If you push the	bu…	," Mr. Steward told him, "somewhere in the world someone you
1) BUT…	said, standing. Mr. Steward rose. "Of course." "And take your	bu…	unit with you." "Are you sure you wouldn't care to
1) BUT…	it for a day or so?" Arthur picked up the	bu…	unit and the envelope and thrust them into Mr. Steward's
1) BUT…	Mr. Steward said. "Would you like me to return the	bu…	unit?" Norma stiffened. 'Certainly not" She hung up angrily. The
1) BUT…	looked incredulous. "What would you like to do? Get the	bu…	back and push it? Murder someone?" Norma looked disgusted. "Murder
1) BUT…	even have to know about, you still wouldn't push the	bu…	?" Arthur stared at her, appalled. "You mean you would?" "Fifty
1) BUT…	last night" Arthur didn't speak. "All that talk about the	bu…	", Norma said. "I think you ----- well, misunderstood me." "In what
1) BUT…	package from the bottom cabinet. Opening it, she set the	bu…	unit on the table. She stared at it for a
1) BUT…	envelope and removing the glass dome. She stared at the	bu…	. How ridiculous, she thought. All this furor over a meaningless
1) BUT…	How ridiculous, she thought. All this furor over a meaningless	bu…	. Reaching out, she pressed it down. For us, she thought
1) BUT…	To get so worked up over nothing. She threw the	bu…	unit, dome, and key into the wastebasket and hurried to
1) BUT…	Something cold pressed at her skull as she removed the	bu…	unit from the wastebasket. There were no nails or screws

Figure 16 Contexts, Button, Button

Key Word In Context (KWIC) solves semantic issues too by showing the context of any problematized word. Moreover, interactive context disambiguates the semantic shade of a word. The word “button” is ambiguous, whether it is a verb or a noun. If it is a noun, it may refer to a shirt button or an electric button. Besides, students feel ambiguity in exams, when a comprehensive note on the title is asked in exams, and students feel difficulty in compiling and presenting all relevant information in the textual sequence. All development regarding the “button” unit has been contextualised in figure 16 to give a detailed analysis of the required information. The word “button” has been used 17 times, and it is used as a noun every time. To conclude, KWIC resolves grammatical ambiguity, semantic ambiguity and presents a comprehensive note on the key word.

2. *Clearing in the Sky* by Jesse Stuart

i. Summary

This corpus has 1 document with 2,228 total words and 599 unique word forms. Created about 10 minutes ago (on 16th October 2017).

Vocabulary Density: 0.269

Average Words Per Sentence: 12.8

Most frequent words in the corpus: i (88); he (53); said (17); land (15); mountain (14); father (13); path (12); years(12); jess (9); asked (8);

Figure 17 Summary, Clearing in the Sky

Computers quantify big data of texts (Stamatatos et al., 1999), therefore, several algorithms (Argamon et al., 2003; Zhang et al., 2002) have been designed and trained to extract features of computational stylistics. Figure 20 presents the stylistic characteristics of the short story in a quantifiable form. Total words in this corpus are 2228, and almost its 1/4th words are unique and numbered as 599. It means that the writer utilizes the same vocabulary items almost four times in this short story. Consequently, this vocabulary repetition creates ease for readers. Its vocabulary density is 0.269 which is quite appropriate for beginner level readers. The average words per sentence are 12.8 which indicate the stylistic quality of longer sentences in this short story.

ii. Cirrus

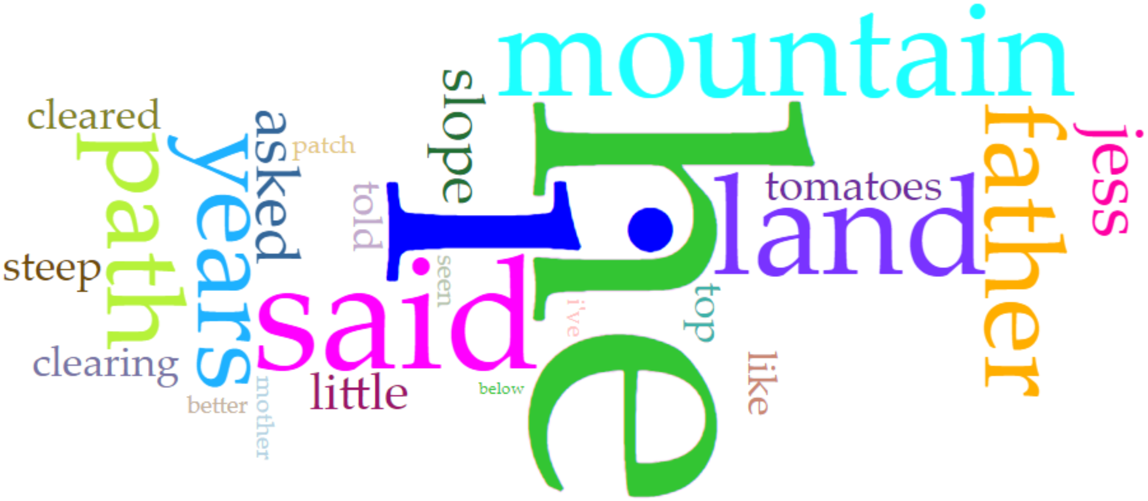


Figure 18 Cirrus, Clearing in the Sky

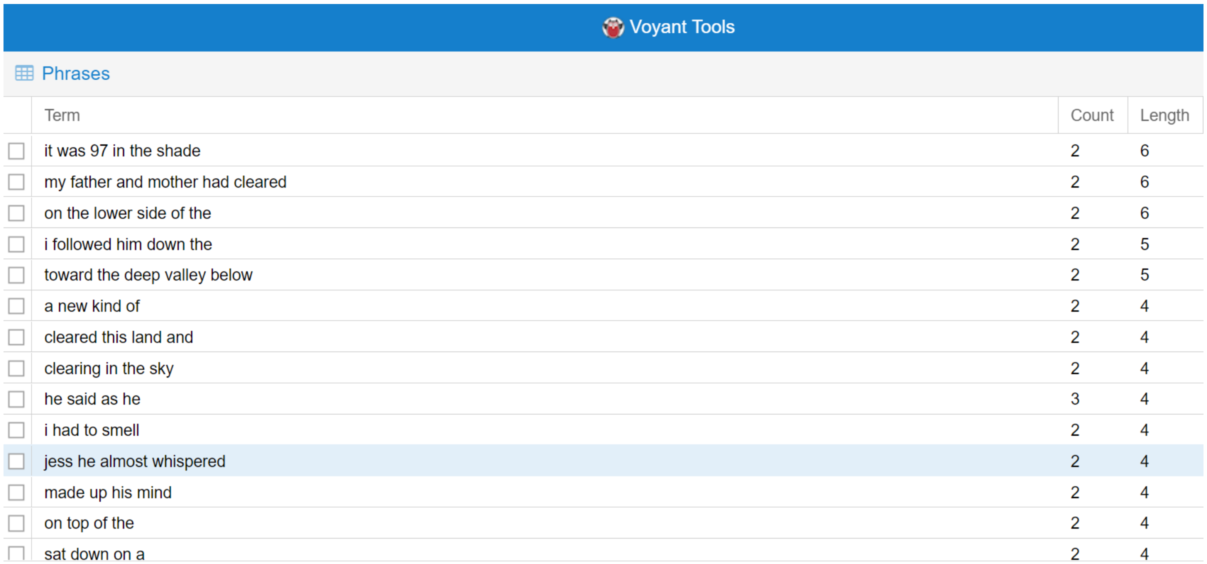
KDD "describes the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). This Cirrus shape describes various themes as the query is raised. Major themes of this short story are pertaining to botanical and zoological nature as some words "corn (4)", "tomatoes (7)", "farm (4)", "land (15)", "alfalfa (4)", "fox (4)", "mountain (14)", "top (8)", "slope (8)", "path (12)", "trees (8)", "yams (4)" and "clearing (7)" indicate. Cirrus aptly performs the task of topic modelling to present them as a cluster. Besides, this Cirrus informs about main characters of the short story, for instance, "I (88)", "he (53)", "Jess (9)" and "father (13)". There are two main characters: "I and Jess" and they refer to the same person who is the author too, while words "he and father" refer to one person who is the author's father.

These two characters use words "said (17)" and "asked (8)"; therefore, these themes reveal the dialogic nature of plot expansion of the story. The theme of "years (12)" is also very dominating because it discusses all memories of previous years as well as a Biblical reference about the average human age which is 70 years.

The theme of "mountain (14)" is also dominating in the short story, and the old man clears some patches of "land (15)" from the mountain top. The theme of "path (12)" dominates because it serves as a parameter to measure the old man's health. As his health improves gradually, he chooses steeper paths to prove his willpower and stamina. Moreover, he leads his son from those paths to the mountain top to show his fields of "corn (4)", "tomatoes (7)", "farm (4)", "alfalfa (4)" and "yams (4)". Moreover, he shows how he hunts fox squirrels to feed his family.

The human analysis discusses "old man, his son Jesse and their home" as the characters of this story. Prominent themes are "physically unfit, climbing up the mountain, willpower, hard work, vigor, strength and determination" (Hussain, 2009, pp. 41-42). Comparing Cirrus tool with human analysis, two human characters are common in both studies, but "home" as a character has not been extracted by Cirrus tool. Human and Cirrus generated different themes. One limitation is present in machine learning that a human mind can choose any theme with synonyms, but this text mining study specifies only twenty five themes. In fact, "They can be extended to expose new things" (Rockwell, & Sinclair, 2016, p. 166). The extension of themes in Cirrus tool can facilitate finding more themes because text mining tools are interactive.

iii. Phrases



	Term	Count	Length
□	it was 97 in the shade	2	6
□	my father and mother had cleared	2	6
□	on the lower side of the	2	6
□	i followed him down the	2	5
□	toward the deep valley below	2	5
□	a new kind of	2	4
□	cleared this land and	2	4
□	clearing in the sky	2	4
□	he said as he	3	4
□	i had to smell	2	4
□	jess he almost whispered	2	4
□	made up his mind	2	4
□	on top of the	2	4
□	sat down on a	2	4

Figure 19 Phrases, Clearing in the Sky

The most occurred 15 collocations have been extracted, for instance, "it was 97 in the shade" (Prn+Aux+N+Prep+Art+N), "my father and mother had cleared" (Prn+N+Conj+N+Aux+V), "on the lower side" (Prep+Art+Adj+N), "I followed him" (Prn+V+Prn), "toward the deep valley" (Prep+Art+Adj+N), "a new kind of" (Art+Adj+N+Prep), "cleared this land" (V+Prn+N), "clearing in the sky" (N+Prep+Art+N), "he said as he" (Prn+V+Adv+Prn), "I had to smell" (Prn+Aux+Prep+V), "he almost whispered" (Prn+Adv+V), "made up his mind" (V+Prep+Prn+N), "on top of the" (Prep+N+Prep+Art) and "sat down on a"

(V+Prep+Prep+Art). These phrases occur 2 or 3 times in this short story, and the length of these phrases ranges from 4 to 6 words. Apart from enhancing language fluency, they also teach the correct use of prepositions through the examples of "made up his mind", "on the lower side".

iv. Links

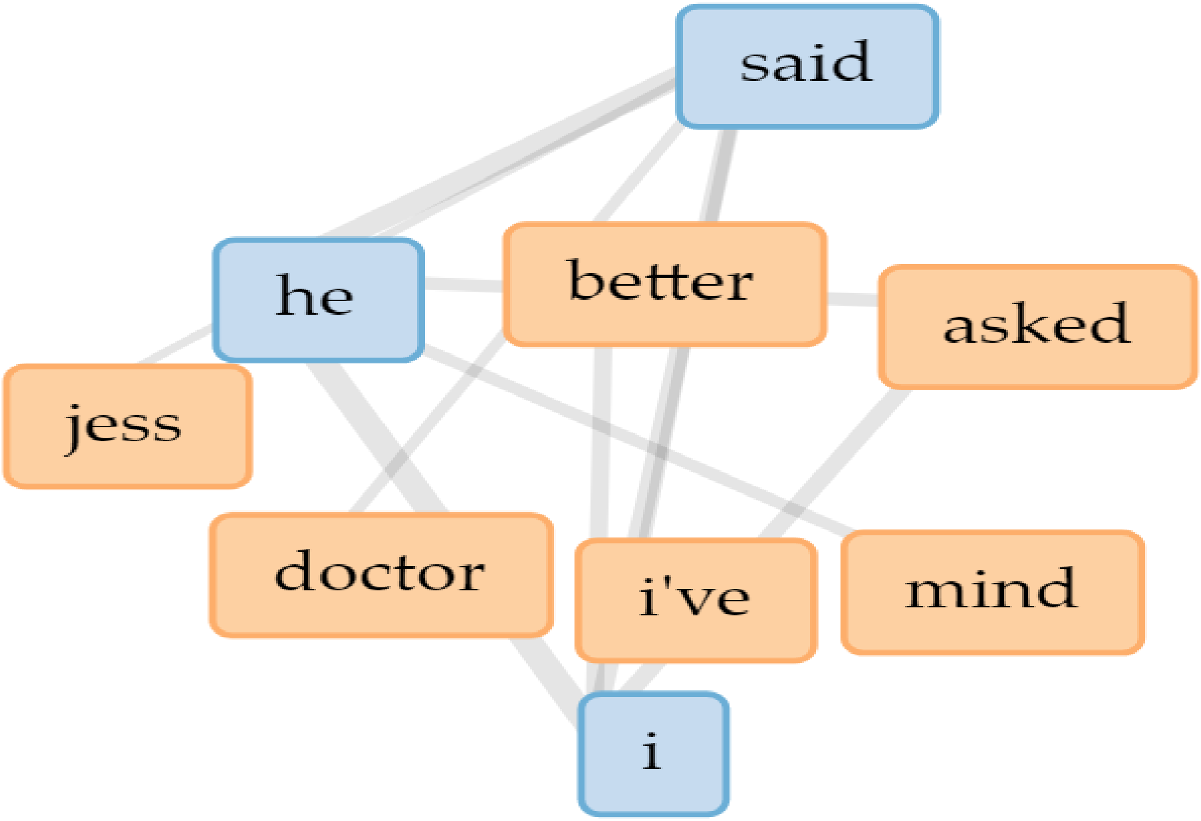
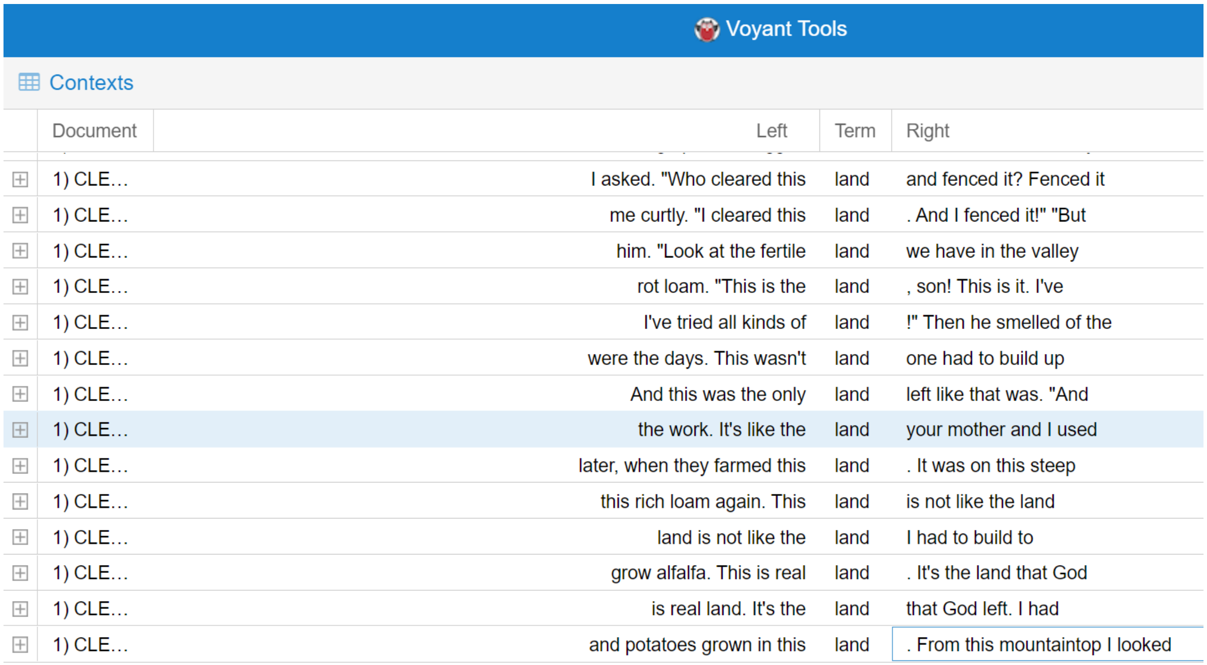


Figure 20 Links, Clearing in the Sky

In this KG, "I, Jess, he, said" words are linked, and they denote that there are dialogues between two major characters "he" and "I", and this short story is developed with conversations between "he" and "I", that is why the word "said" is used 17 times in this corpus. Apart from it, the words "I, asked, he" are interlinked since both characters ask different questions from each other, revive memories and exchange views. Another connection of "doctor, he, said" is evident that a doctor informs him about the old man's short span of life, while the old man falsifies the doctor's prediction by dint of his willpower and endurance. Another KG "I've, better" shows that I have never seen better alfalfa and other vegetation than the production of my cleared patch. Another KG of "he, mind" refers to the textual phrase "he made up his mind". It means that when he determines to do something, he does it at any cost. He recovers from lethal disease by dint of his adamant determination and healthy natural activities.

v. Contexts



Document	Left	Term	Right
1) CLE…	I asked. "Who cleared this	land	and fenced it? Fenced it
1) CLE…	me curtly. "I cleared this	land	. And I fenced it!" "But
1) CLE…	him. "Look at the fertile	land	we have in the valley
1) CLE…	rot loam. "This is the	land	, son! This is it. I've
1) CLE…	I've tried all kinds of	land	!" Then he smelled of the
1) CLE…	were the days. This wasn't	land	one had to build up
1) CLE…	And this was the only	land	left like that was. "And
1) CLE…	the work. It's like the	land	your mother and I used
1) CLE…	later, when they farmed this	land	. It was on this steep
1) CLE…	this rich loam again. This	land	is not like the land
1) CLE…	land is not like the	land	I had to build to
1) CLE…	grow alfalfa. This is real	land	. It's the land that God
1) CLE…	is real land. It's the	land	that God left. I had
1) CLE…	and potatoes grown in this	land	. From this mountaintop I looked

Figure 21 Contexts, Clearing in the Sky

The word “land” as a verb refers to the landing of a plane, and as a noun, it refers to a piece of geographical land. To disambiguate word sense, Contexts tool is used. The old man frequently praises his cleared land at the top of the mountain, and 15 times, the word “land” is used as a noun, and it means rich fertile ground. Furthermore, if one wants to find adjectives for the noun “land”, Contexts data show “rugged land”, “fertile land” and “real land”. Contexts tool not only disambiguates lexical and semantic disambiguity but also finds adjectives of the key word “land”.

3. *Dark They Were, and Gold Eyed* by Ray Bradbury

i. Summary

This corpus has 1 document with 1,858 total words and 672 unique word forms. Created 18 seconds ago (on 16th October 2017).

Vocabulary Density: 0.362

Average Words Per Sentence: 7.5

Most frequent words in the corpus: said (25); harry (24); rocket (13); bittering (12); earth (10); looked (10); wife(9); away (8); children (8); sam (8)

Figure 22 Summary, Dark They Were, and Gold Eyed

Summarization shows a condensed report of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45) in the qualitative and quantitative forms. So, the Summary tool generates data that facilitate readers to point out the stylistic characteristics of each writer and his/her literary work. This story has 672 unique words which are used almost three times in this story; hence, 1858 total words are present in this corpus. By dividing unique words with total words, its vocabulary density reaches 0.362. Furthermore, 7.5 average words have been written in a sentence because short sentences enhance beginner-level readers' readability and fluency. Analysing the first ten most occurring words, five words consist of human characters, and 5 words are related with "said, rocket, earth, looked, away".

ii. Cirrus

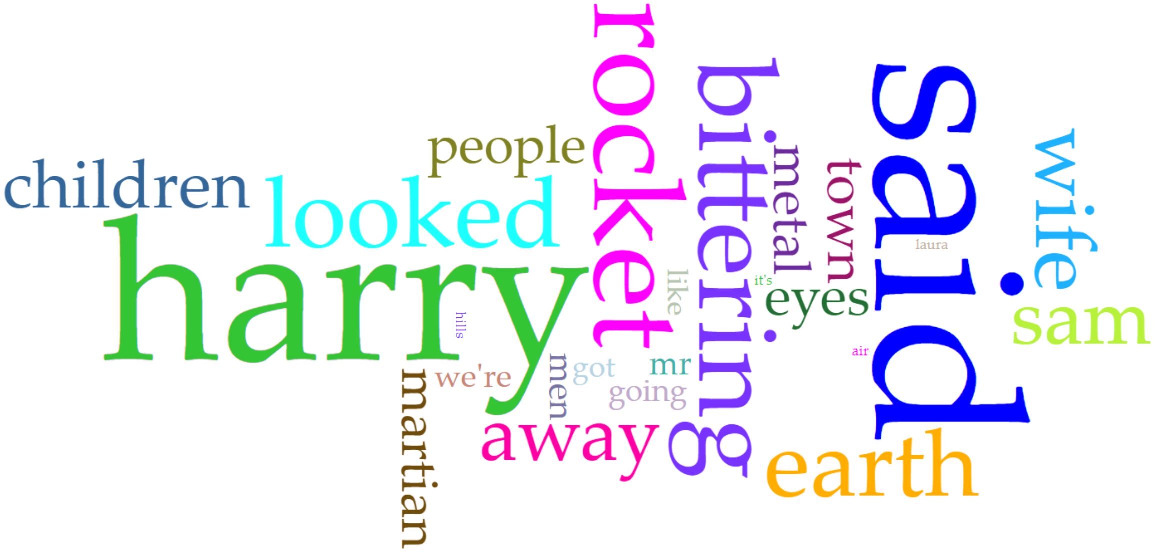
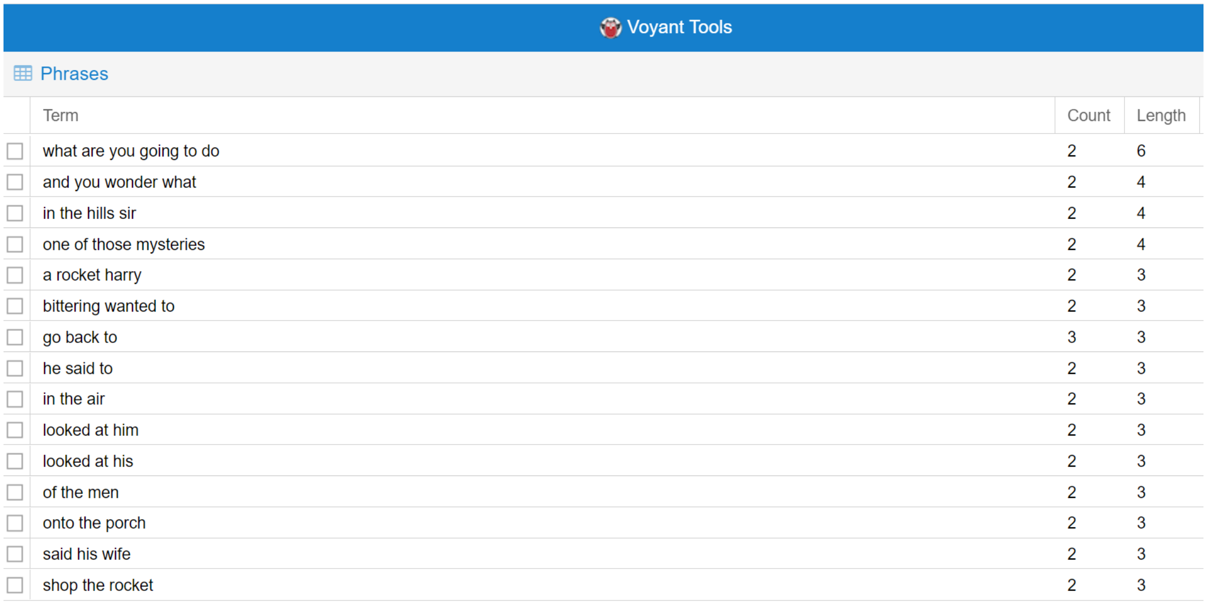


Figure 23 Cirrus, Dark They Were, and Gold Eyed

This Cirrus reveals major quantified themes and characters of this short story by topic modelling. Words "rocket (13)", "rockets (5)", "build (4)", "built (4)", "metal (7)" and "Martians (7)" express the idea of a visit to Mars through rockets, and their interaction with Martians who are earth people. Then there is another theme of "settlement (4)" which is inhabited by "earth (10)" and "people (7)". Then some other climatic and geographical themes are also highlighted with words of “dark (4)”, “air (5)”, “yellow (4)”, “valley (4)” and “hills (5)”. Besides, major characters of the short story are “Harry (24)”, “Bittering (12)”, “wife (9)”, “children (8)”, “Sam (8)”, “Cora (4)” and “Laura (5)” who have gone to Mars to get refuge during war circumstances. Most of the situations revolve around “Harry”, so his name occurs 24 times in this Cirrus. Besides, the word “said” is spoken 25 times, and it clearly discovers knowledge patterns that there is a progression of the story through dialogues; and characters directly address each other.

Human analysis reveals that the main characters of this story are “Harry, his wife Cora and children”. Key themes of this short story are “Mars, build rocket, go back to earth, wove tapestries, played songs, war on earth, changing shapes, lose identities and foreseeing nature of husband” (Hussain, 2009, pp. 42-43). The comparison of human analysis with Cirrus analysis shows that both have extracted the same characters. Furthermore, Cirrus shows the occurrence of each theme and character, but the human analysis does not count theme occurrence. Both human analysis and tool analysis have mentioned common themes of “Mars” and “build rockets”, while other themes are different.

iii. Phrases



	Term	Count	Length
□	what are you going to do	2	6
□	and you wonder what	2	4
□	in the hills sir	2	4
□	one of those mysteries	2	4
□	a rocket harry	2	3
□	bittering wanted to	2	3
□	go back to	3	3
□	he said to	2	3
□	in the air	2	3
□	looked at him	2	3
□	looked at his	2	3
□	of the men	2	3
□	onto the porch	2	3
□	said his wife	2	3
□	shop the rocket	2	3

Figure 24 Phrases, Dark They Were, and Gold Eyed

From figure 23, some standard phrases have been extracted “what are you going to do” (Prn+Aux+Prn+V+Inf V), “and you wonder that” (Conj+Prn+V+Prn), “in the hills sir” (Prep+Art+N+N), “one of those mysteries” (Prn+Prep+Prn+N), “go back to” (V+Adv+Prep), “in the air” (Prep+Art+N) and “looked at him” (V+Prep+Prn). These collocations occur two to three times, and their length ranges from 3 to 6 words. They not only exhibit the ideology and narratology of the short story but also facilitate in constructing new collocation patterns/ n-grams based on the old ones. The main narratology of mystery, wonder-struck conditions and a desire to go back is apparent in these collocations. Again, narratology construction requires reflection and thinking of human beings, as Hermeneutica theory emphasises "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166).

iv. Links

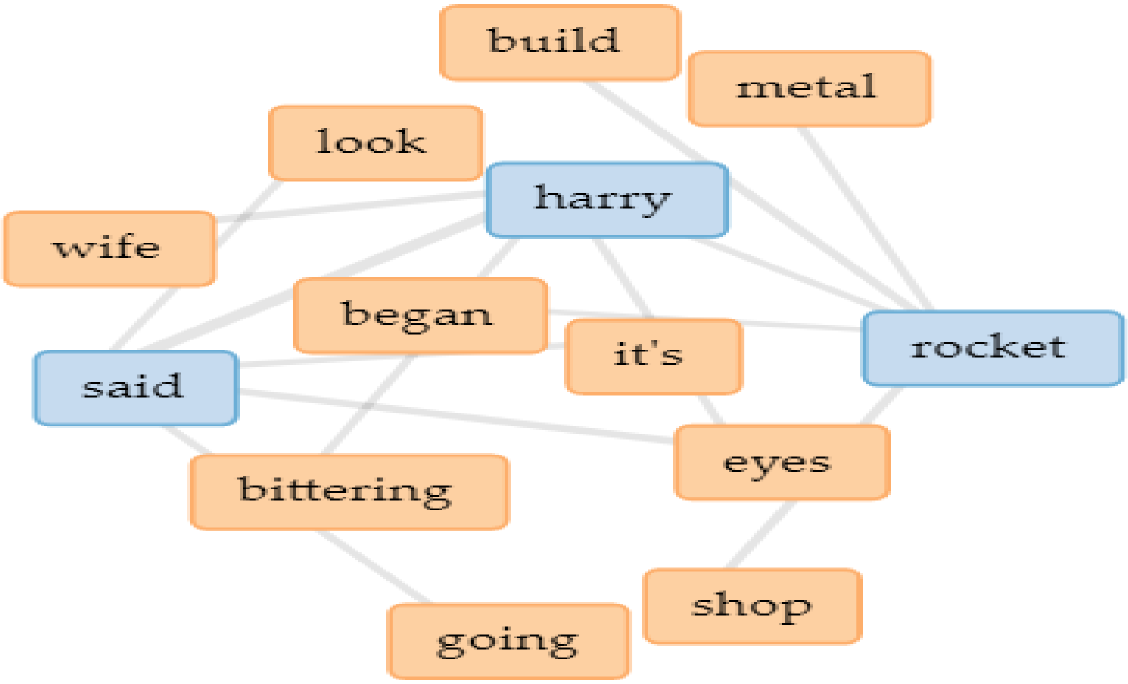
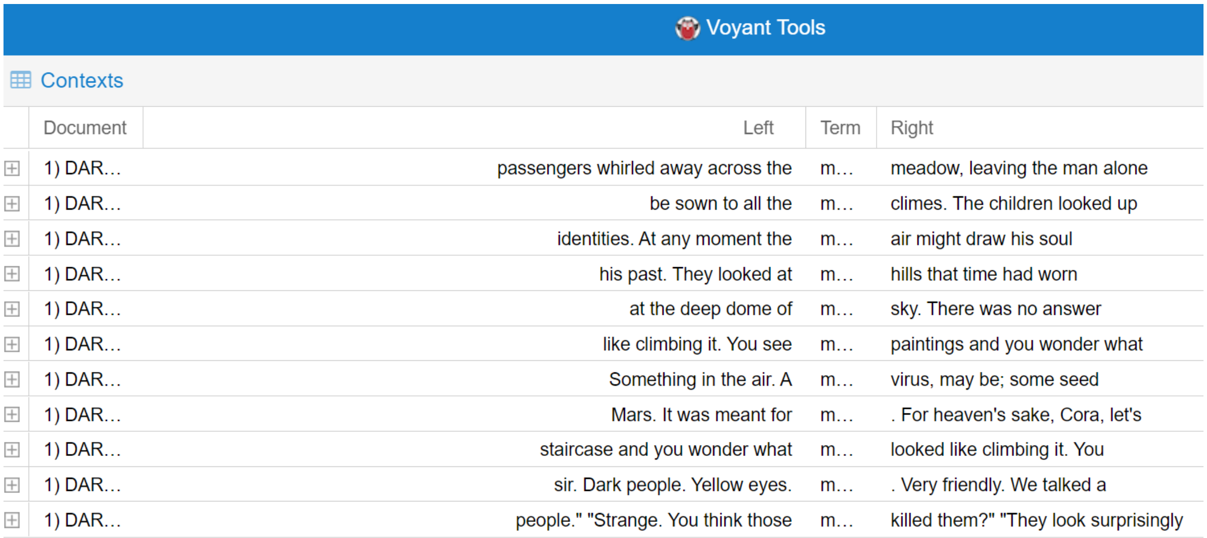


Figure 25 Links, Dark They Were, and Gold Eyed

The KG interlinks the characters and themes of the short story aptly. Links tool draws node linkages among "build, rocket, Harry" which suggest that Harry Bittering tries to build a rocket to go back to the earth. Another KG of "rocket, Harry, shop, metal" denotes the act of buying metal to build a rocket in the iron shop. Then, KG of "Harry, rocket, began" indicates that he starts to work on metal though he cannot accomplish his rocket-building task because of his incompetence in this very scientific task. The KG of "Bittering, Harry, wife" shows that their family name "Bittering" is attached with their names, for example, Harry Bittering and Mrs Bittering. Moreover, the KG of "Harry, eyes" refers to the transformation of the colour of his eyes into golden under the influence of Martian climate. To conclude, KGs interlink themes and characters of the text. Likewise, relationship mining among relevant variables (Barahate, 2012, p. 13) has been done by various KGs.

v. Contexts



	Document	Left	Term	Right
⊞	1) DAR…	passengers whirled away across the	m…	meadow, leaving the man alone
⊞	1) DAR…	be sown to all the	m…	climes. The children looked up
⊞	1) DAR…	identities. At any moment the	m…	air might draw his soul
⊞	1) DAR…	his past. They looked at	m…	hills that time had worn
⊞	1) DAR…	at the deep dome of	m…	sky. There was no answer
⊞	1) DAR…	like climbing it. You see	m…	paintings and you wonder what
⊞	1) DAR…	Something in the air. A	m…	virus, may be; some seed
⊞	1) DAR…	Mars. It was meant for	m…	. For heaven's sake, Cora, let's
⊞	1) DAR…	staircase and you wonder what	m…	looked like climbing it. You
⊞	1) DAR…	sir. Dark people. Yellow eyes.	m…	. Very friendly. We talked a
⊞	1) DAR…	people." "Strange. You think those	m…	killed them?" "They look surprisingly

Figure 26 Contexts, Dark They Were, and Gold Eyed

Readers face lexical ambiguity of adjective or noun for word “Martians”. So, Contexts tool expresses that the word “Martians” as an adjective have been used in the first seven sentences and from 8th to 11th sentences “Martians” as a noun has been used. On the next level, semantic ambiguity confuses readers who want to know about the descriptions of “Martians”. Just finding context of the key word “Martians” shows the overall presence of the word in the corpus that the word “Martians” appears 11 times in this short story, for instance, Martian meadows, Martian climes, Martian air, Martian hills, Martian sky, Martian paintings, Martian virus, Martian look, Martian friendly attitude and probability of Martians as killers. To summarize, such extensive WSD cannot be delineated without the utilization of Contexts tool, and the same has been emphasized in the theory that “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166).

4. *Thank You, M'am* by Langston Hughes

i. Summary

This corpus has 1 document with 1,361 total words and 426 unique word forms. Created 9 seconds ago (on 17th August 2017).

Vocabulary Density: 0.313

Average Words Per Sentence: 12.5

Most frequent words in the corpus: said (28); boy (27); woman (23); got (10); face (9); door (8); run (8); going (7); large (7); behind (6)

Figure 27 Summary, Thank You, M'am

Stylometry means the extraction of any writer's linguistic style by analyzing the literary work's language. Unique words of this corpus are 426, and they are repeated almost three times, so their total words are 1361. Vocabulary density is calculated using Inverse Absolute (Simpson, 2000) (a division of total words by unique words), so the vocabulary density of this short story is 0.313. This corpus suggests that Langston Hughes writes sentences with 12.5 words on average, and they are longer than Ray Bradbury's '*Dark They Were And Golden Eyed'*. The most frequent words discuss the main characters of "boy (27)" and "woman (28)". The most occurring word "said (28)" informs about the stylistic qualities of dialogues in this short story. Thus, the quantification of word occurrence (Amancio, 2015) displays key themes and characters of the story.

ii. Cirrus

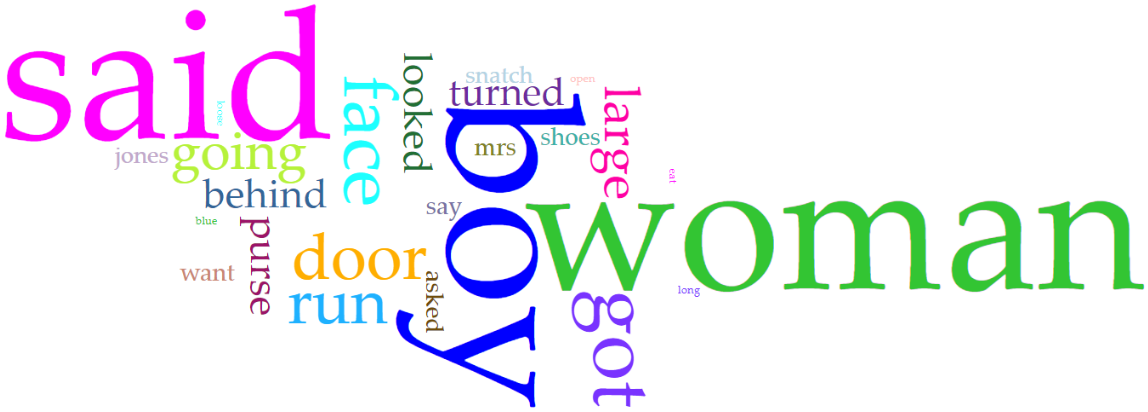
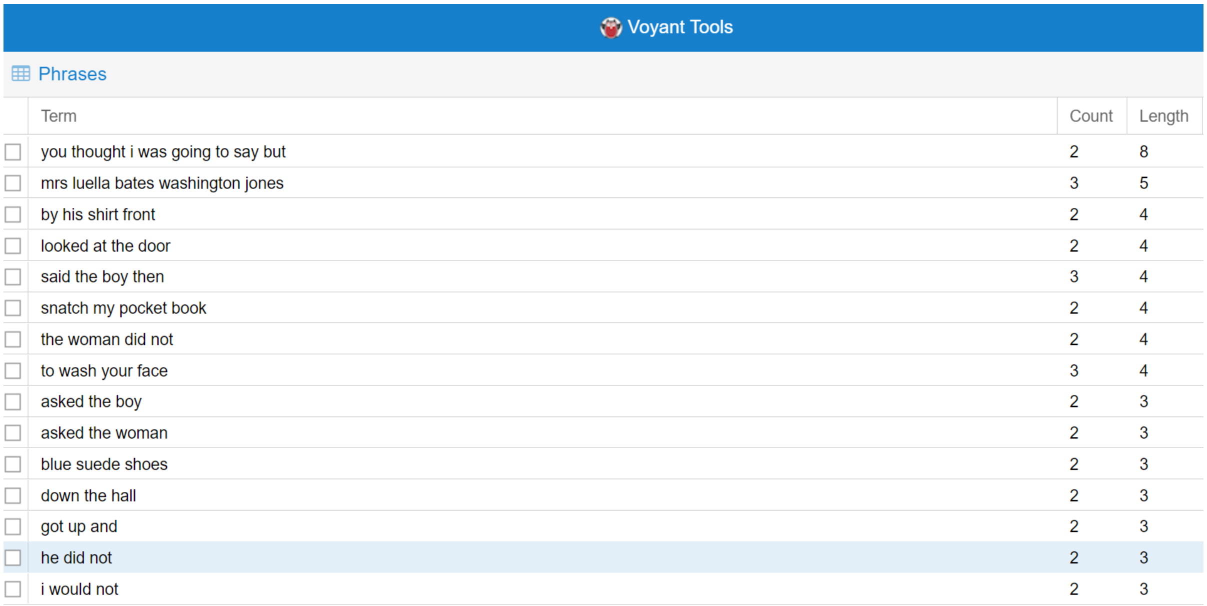


Figure 28 Cirrus, Thank You, M'am

The key characters of this short story are "boy (27)" and "large (7)" "woman (23)", and she is the protagonist of this short story. The most occurring theme of this short story is "said (28)". "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166). After human reflection, it reveals that this story is dialogic in nature, and it progresses with different dialogues of characters. Textual evidence also proves that the boy named Roger and the woman named Mrs Luella Bates Washington Jones talk directly in the story, and the plot develops, till it reaches its climax and resolution stage. Key events of this short story are to “snatch (5)” the “purse (6)”; and after it, the boy tries to “run (8)”, but the heavy “woman (23)” drags the “boy (27)” to her home to wash his “face (9)” and to reform him perpetually. She provides the “hungry (3)” boy with delicious food. She keeps the “door (8)” open, but the boy does not run from her house because both of them have developed a sense of mutual trust. Then she asks him about the cause of purse snatching. Having known the fact, she gives him ten dollars to buy blue suede shoes, and advises him that “Shoes got by devilish ways will burn your feet.”

Human analysis reveals that two human characters, “woman” and “boy” have been portrayed, and their key motifs are “sympathy, give money, generous, good behaviour, large, powerful, frail and willow-wild” (Hussain, 2009, pp. 44-45). Cirrus also presents both characters and elaborates the woman’s good acts to feed the hungry boy, but human analysis expresses it just as a “good behaviour”. To conclude, Cirrus analysis is more detailed and quantified than human analysis.

iii. Phrases



	Term	Count	Length
□	you thought i was going to say but	2	8
□	mrs luella bates washington jones	3	5
□	by his shirt front	2	4
□	looked at the door	2	4
□	said the boy then	3	4
□	snatch my pocket book	2	4
□	the woman did not	2	4
□	to wash your face	3	4
□	asked the boy	2	3
□	asked the woman	2	3
□	blue suede shoes	2	3
□	down the hall	2	3
□	got up and	2	3
□	he did not	2	3
□	i would not	2	3

Figure 29 Phrases, Thank You, M'am

In this story, standard phraseology is “you thought I was going to say” (Prn+V+Prn+Aux+V+Inf V), “by his shirt front” (Prep+Prn+N+Adv), “looked at the door” (V+Prep+Art+N), “snatch my pocket book” (V+Prn+N), “to wash your face” (Inf V+Prn+N), "blue suede shoes" (Adj+Adj+N), "down the hall" (Prep+Art+N), got up" (V+Prep), "on the daybed" (Prep+Art+N) which enrich fluency in speaking and writing. Statistically, one phrase occurs a maximum of three times, and the length of phrases ranges between 3 to 8 words.

iv. Links

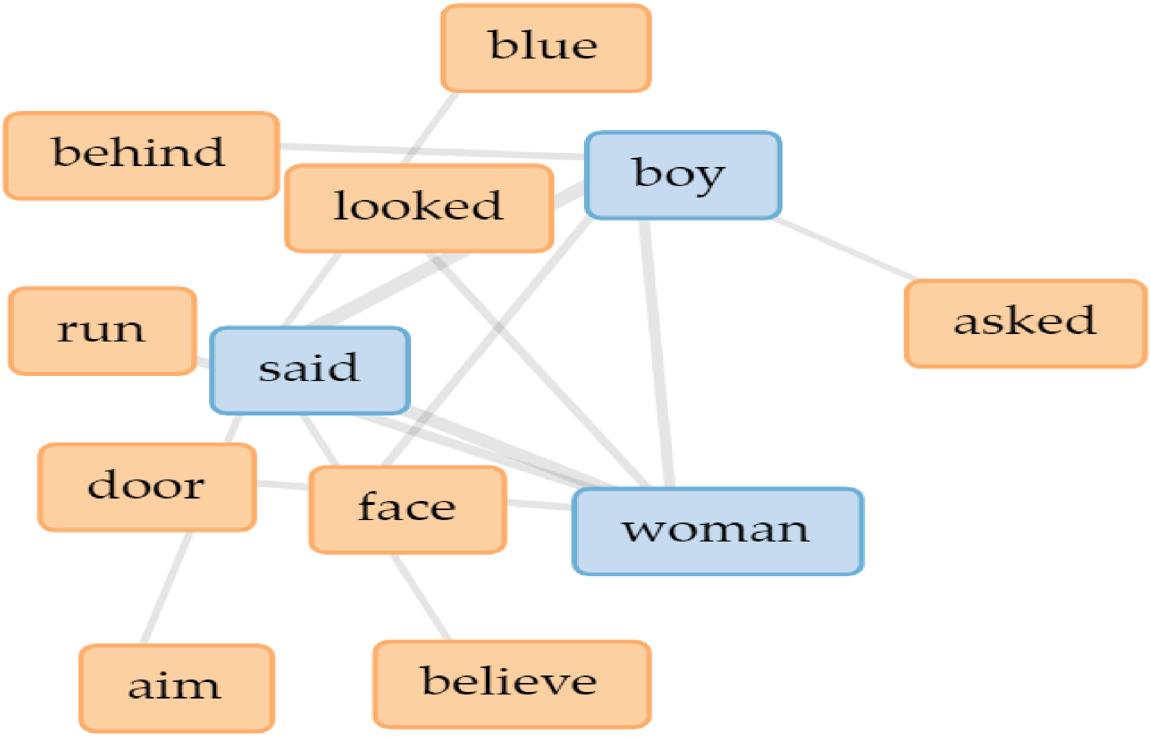
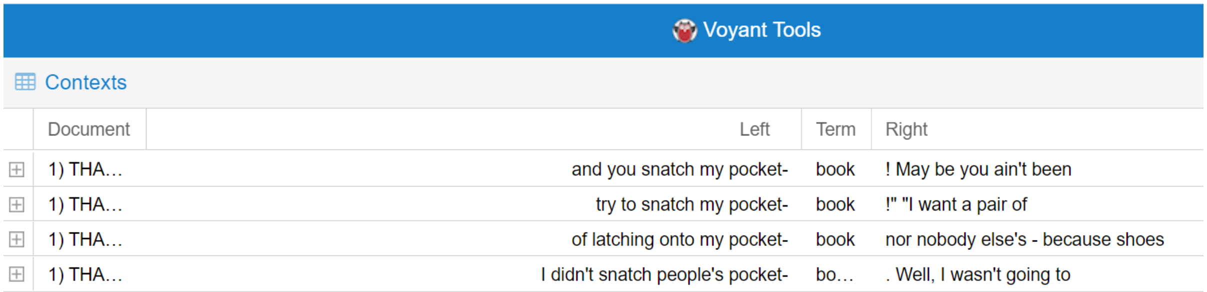


Figure 30 Links, Thank You, M'am

"In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). Two active conversant characters of the short story are woman and boy, and the same has been visualized in the KG of "woman, boy, said". The nodes of "woman, boy, asked" refer to the incident when the woman asks the boy some questions about his life, family, intentions behind purse snatching. Another KG of "woman, boy, door, face" shows that the woman asks the boy to wash his face, and keeps the door open because now she trusts in him. The KG of "boy, said, blue" conveys the idea that the boy yearns to buy blue suede shoes with the stolen money.

v. Contexts



Document	Left	Term	Right
1) THA...	and you snatch my pocket-	book	! May be you ain't been
1) THA...	try to snatch my pocket-	book	!" "I want a pair of
1) THA...	of latching onto my pocket-	book	nor nobody else's - because shoes
1) THA...	I didn't snatch people's pocket-	bo...	. Well, I wasn't going to

Figure 31 Contexts, Thank You, M'am

There is a word sense ambiguity that the word "book" as a noun refers to some reading material, while as a verb, it refers to the reservation of some seat. To resolve this semantic ambiguity, its context has been checked. The word "pocket book" has been used four times, and it refers to a purse, and it has no connection with the above-mentioned meanings. Besides, it unveils that Voyant tool deals with the hyphenated word "pocket-book" as two words, whereas the English language considers this compound noun as a single meaningful word.

5. *The Piece of String* by Guy De Maupassant

i. 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), day (4); great (4)

Figure 32 Summary, The Piece of String

Computational stylistics means the derivation of the style of any writer through the application of the digital tool. Moreover, "summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). This compact description has been presented in figure 35. Unique words in this corpus are 413, and they are repeated almost 2.5 times in this short story; that total words are counted 1007. Vocabulary density is 0.410 which has been derived with Inverse Absolute (Simpson, 2000) (a division of total words by unique words). The average words per sentence are 12.9 which indicate the medium length of sentences. Vocabulary density and the average length of sentences guide us to select reading material for different levels of readers. Besides, the most occurring words reveal the name of the central character and themes, for instance, “Mr (20)”, “Hubert (17)”, stolen thing “pocket (12)” and “book (10)”. In reality, Mr Hubert picks the “string (7)” from the ground, while villagers keep on blaming him as a liar. The word “said” reveals the use of dialogues for the plot development.

ii. Cirrus

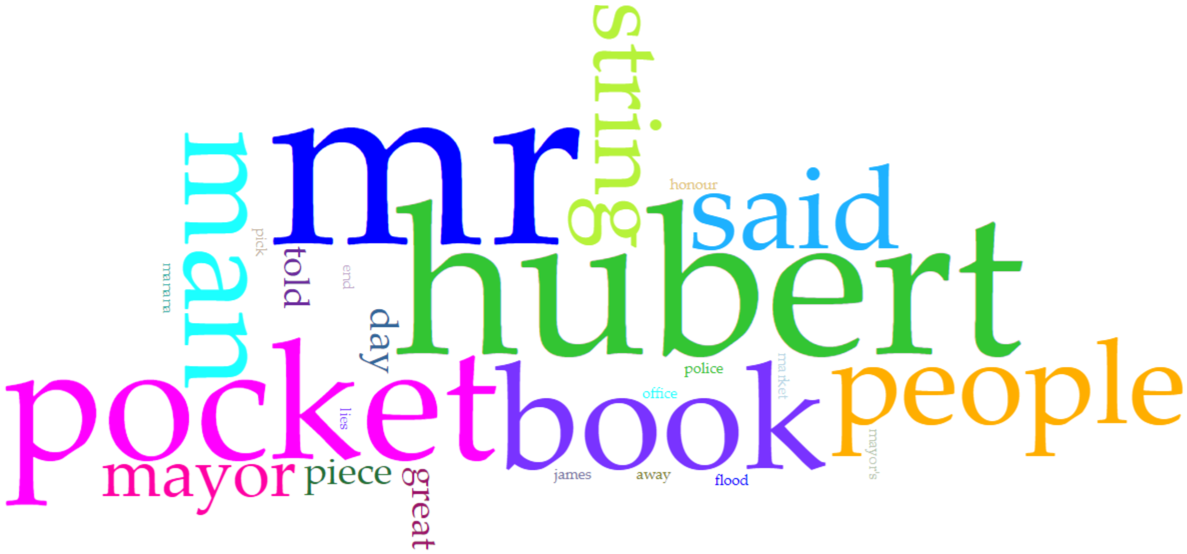


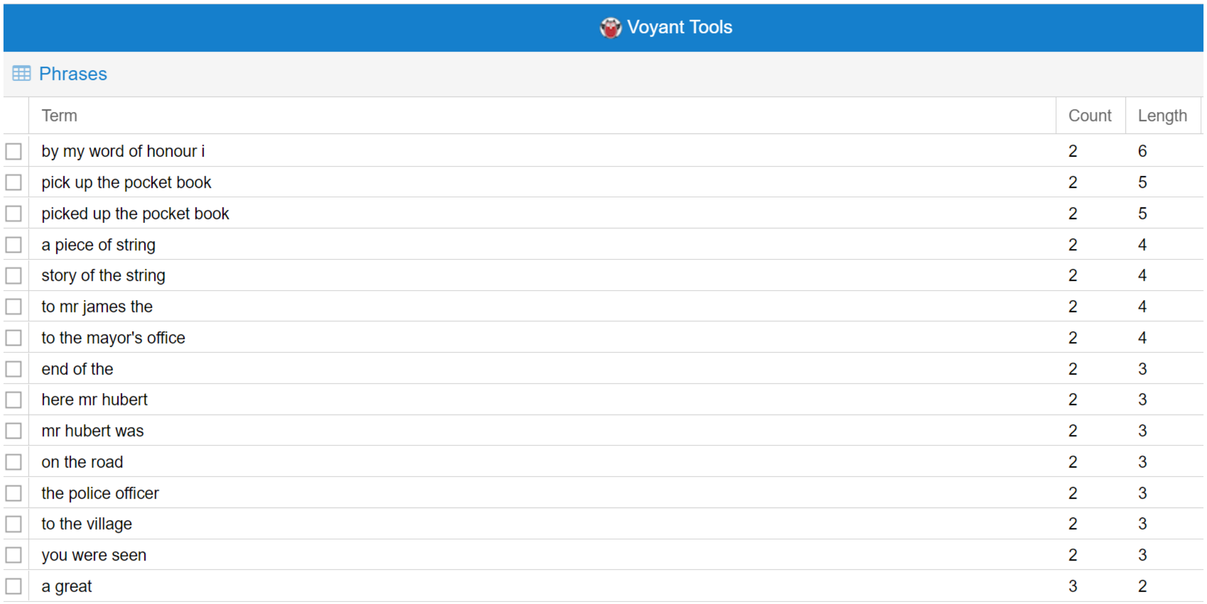
Figure 33 Cirrus, The Piece of String

The central character and tragic hero of this short story is “Mr Hubert (17)”, and he remains the most discussed topic throughout the story. Another important character is “James (3)” who is the owner of the lost “pocket book (10)”. Then this loss is announced with an incentive of a prize by the drum beater. Afterwards, the “police (3)” officer takes Hubert from the market feast to the “mayor’s office (3)” for scolding, threatening and further inquiry. “Manana (3)” blames Hubert for stealing the pocketbook from the mud. Hubert shows a piece of “string (7)” and takes an oath on holy “book (10)” to prove his “innocence (2)”, but nobody believes in his oath and statement. Later on, when “George (2)” returns the pocket book to its owner, even then villagers continue to call Hubert a liar and thief. “People (8)” of that locality play their inquisitive and teasing role in spreading rumours against Hubert, and they enjoy the story of string, whereas Hubert painfully claims his innocence and truthfulness before them.

This Cirrus discovers knowledge patterns about prominent themes of “truth (3)” and “lies (3)” which are found in equal numbers as foils. “Innocence” of Hubert cannot be proved in his life, but his hypersensitivity and grief over the blame of falsehood lead him to imminent death. During his life, he informs about the possible cause of his imminent death in this proverbial sentence, “There is nothing so shameful as to be called a liar.” Besides, the word “said (8)” reveals that all characters talk directly, and the plot is developed through direct conversations and dialogues. Another study resembles this short story on emotional grounds, so in the literature review section, emotion mining of Shakespearean plays (Muhammad, 2012) has been done with Cirrus tool, and findings of both studies are similar.

Human analysis reveals that ten male characters have been portrayed, but no female character is present in the story. (Hussain, 2009, p. 45). Here computer-generated analysis performs better than human analysis because Cirrus tool extracts “Hubert” as a central character. Moreover, Cirrus discusses quantified themes of “pocket book”, “string”, but the human analysis does not mention them.

iii. Phrases



	Term	Count	Length
□	by my word of honour i	2	6
□	pick up the pocket book	2	5
□	picked up the pocket book	2	5
□	a piece of string	2	4
□	story of the string	2	4
□	to mr james the	2	4
□	to the mayor's office	2	4
□	end of the	2	3
□	here mr hubert	2	3
□	mr hubert was	2	3
□	on the road	2	3
□	the police officer	2	3
□	to the village	2	3
□	you were seen	2	3
□	a great	3	2

Figure 34 Phrases, The Piece of String

This corpus shows long frequent collocation/ n-gram phrases, for instance, “by my word of honour” (Prn+N+Prep+N), “pick up the pocket book” (V+Prep+Art+N), “a piece of string” (Art+N+Prep+N), “story of the string” (N+Prep+Art+N), “on the road” (Prep+Art+N), “to the village” (Prep+Art+N), “you were seen” (Prn+Aux+V). Some substandard or very common phrases, such as “a man” and “about his”, are excluded for attaining standard phraseology. Collocations/ n-grams occur twice or three times, but their length ranges from 2 to 6 words.

iv. Links

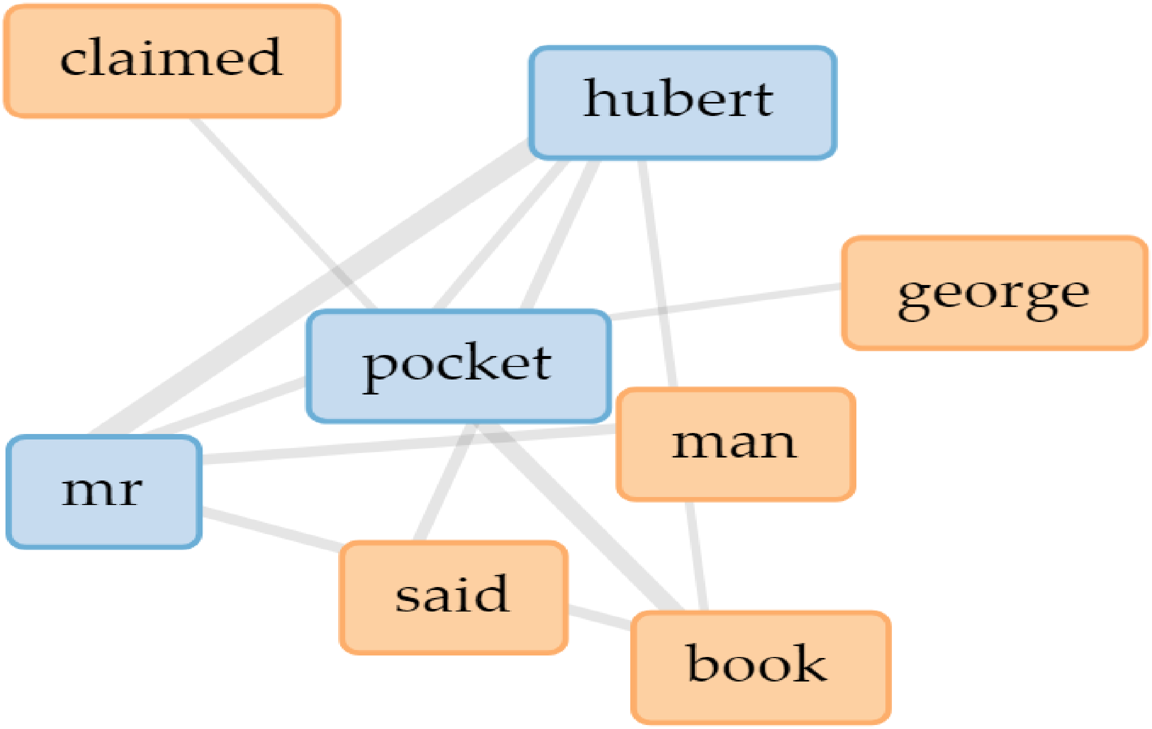
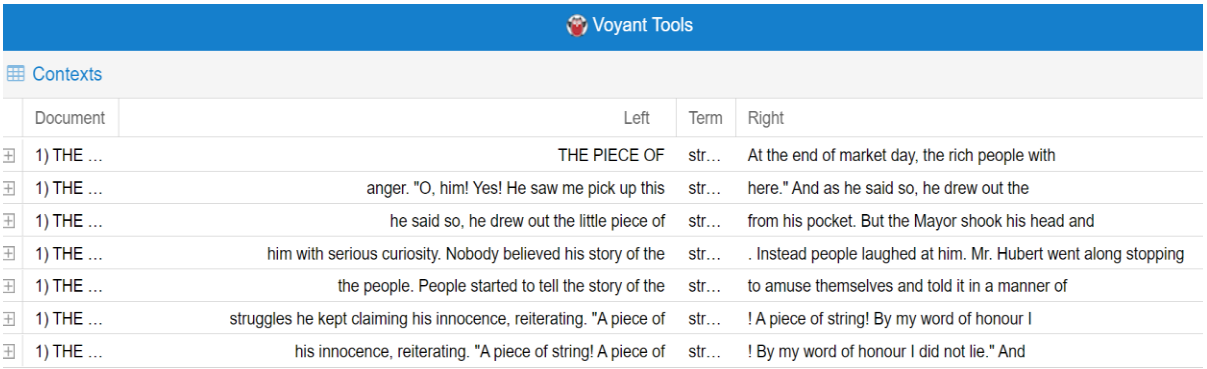


Figure 35 Links, The Piece of String

The KG of “pocket, book, Mr, Hubert” reveals the main plot of the story that Manana and other villagers blame that Mr Hubert has stolen the lost pocket book. Though Hubert nullifies this blame, yet nobody trusts in his clarification and swearing. Another KG “Hubert, claimed, pocket, book” shows that Hubert exonerates himself from the blame of stealing the pocket book. Afterwards, the KG “George, said, pocket, book” exposes that George finds the lost pocket book on his way to the market. Even then, nobody trusts in Hubert’s truthfulness.

v. Contexts



Document	Left	Term	Right
1) THE ...	THE PIECE OF	str...	At the end of market day, the rich people with
1) THE ...	anger. "O, him! Yes! He saw me pick up this	str...	here." And as he said so, he drew out the
1) THE ...	he said so, he drew out the little piece of	str...	from his pocket. But the Mayor shook his head and
1) THE ...	him with serious curiosity. Nobody believed his story of the	str...	. Instead people laughed at him. Mr. Hubert went along stopping
1) THE ...	the people. People started to tell the story of the	str...	to amuse themselves and told it in a manner of
1) THE ...	struggles he kept claiming his innocence, reiterating. "A piece of	str...	! A piece of string! By my word of honour I
1) THE ...	his innocence, reiterating. "A piece of string! A piece of	str...	! By my word of honour I did not lie." And

Figure 36 Contexts, The Piece of String

"Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166) because true word sense is revealed by reading the context of the problematized word. In this short story, the word "string" conveys different semantic shades as a noun, adjective or verb, such as thin rope, text in computer science, a string of musical instruments, and a series of events. To resolve this semantic and lexical ambiguity, KWIC is searched. In this corpus, the word "string" as a noun is used seven times, and every time it means a small piece of rope.

6. *The Reward* by Lord Dunsany

i. Summary

This corpus has 1 document with 1,255 total words and 446 unique word forms. Created about a minute ago (on 20th October 2017).

Vocabulary Density: 0.355

Average Words Per Sentence: 17.4

Most frequent words in the corpus: jorkens (12); said (11); long (10); terbut (9); court (8); acrobat (7); came (7); ambition (6); asked (6); country (6)

Figure 37 Summary, The Reward

Computational stylistics automatically analyses any text and exhibits literary style accurately with digital humanities tools. This corpus comprises 446 unique words, and they have been used almost three times in this short story until their total words reach the number of 1255. Vocabulary density is calculated using Inverse Absolute (Simpson, 2000) (a division of total words by unique words), so its vocabulary density is calculated at 0.355. Lord Dunsany writes longer sentences which have an average of 17.4 words in a sentence. Its most frequent words have shown three central characters for illustration: “Jorkens (12)”, “Terbut (9)” and “court (8)” “acrobat (7)”. In addition to it, the whole short story concentrates on the fulfilment of the vaulting “ambition” to become a court acrobat in the royal court of a country where no such post exists.

ii. Cirrus

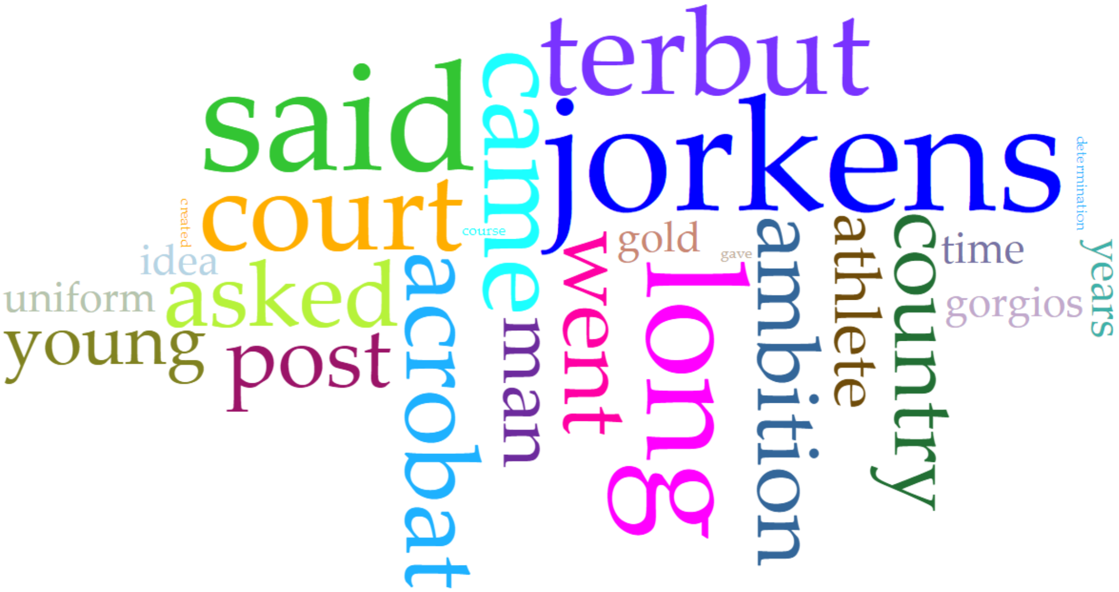
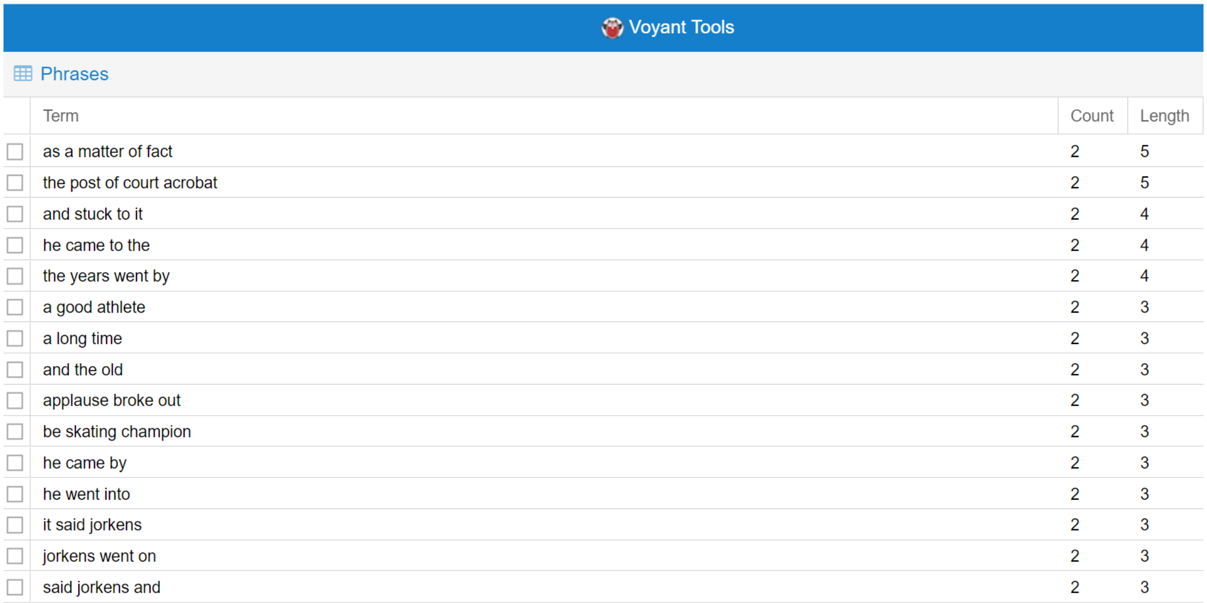


Figure 38 Cirrus, The Reward

Key characters of the short story are “Jorkens (12)”, “Terbut (9)” and “acrobat (7)” whose name is “Gorgios (4)”. Jorkens propagates the significance of determination, and the story of a court acrobat has been narrated to exemplify determination. Moreover, this Cirrus shows key themes of the short story, and one of the key themes is “ambition (6)”. Two debatable points are “opportunity (3)” or “determination (3)”, so both persons argue in favour of their notions. Another theme of “politics (3)” suggests that the athlete temporarily takes part in political activism to become a court acrobat in the “royal (3)” palace. Meanwhile, he continues his athletic activities. During the inaugural ceremony, he wears a tight-fitting “red (3)” “uniform (4)” as a court acrobat. He tries to jump on “hurdles (3)”, but he is too weak to perform any acrobatic activity in front of the audience. Another theme of “applause (2)” is also shown to commemorate and crown his lifelong efforts and achievement. So, ambitions come true, if one “stuck (3)” to it, and devotes all energies and time to the real ambition. Besides, the word “said (11)” informs about direct conversations of characters, and their dialogues develop the plot of the story.

Human analysis shows that this short story discusses "male centred, ladies, brilliant dresses" themes and characters (Hussain, 2009, p. 46), while Cirrus tool presents more comprehensive themes and named characters., for example, "Jorkens (12)", "Terbut (9)" and "acrobat (7)" whose name is "Gorgios (4)" in the story. Cirrus presents themes of "politics, royal, hurdles, uniform, said" etc.

iii. Phrases



Phrases

	Term	Count	Length
□	as a matter of fact	2	5
□	the post of court acrobat	2	5
□	and stuck to it	2	4
□	he came to the	2	4
□	the years went by	2	4
□	a good athlete	2	3
□	a long time	2	3
□	and the old	2	3
□	applause broke out	2	3
□	be skating champion	2	3
□	he came by	2	3
□	he went into	2	3
□	it said jorkens	2	3
□	jorkens went on	2	3
□	said jorkens and	2	3

Figure 39 Phrases, The Reward

This corpus reveals standard phraseology, for instance, "as a matter of fact" (Adv+Art+N+Prep+N), "the post of court acrobat" (Art+N+Prep+Adj+N), "the years went by" (Art+N+V), "a long time" (Art+Adj+N), "applause broke out" (N+V+Prep), "he came by" (Prn+V+Prep), "he went into" (Prn+ V+Prep) and "sticks to it" (V+Prep+Prn). Substandard phrases, for instance, "said Jorkens and" are left out. All selected multiword expressions occur twice, and their length ranges from 3 to 5 words.

iv. Links

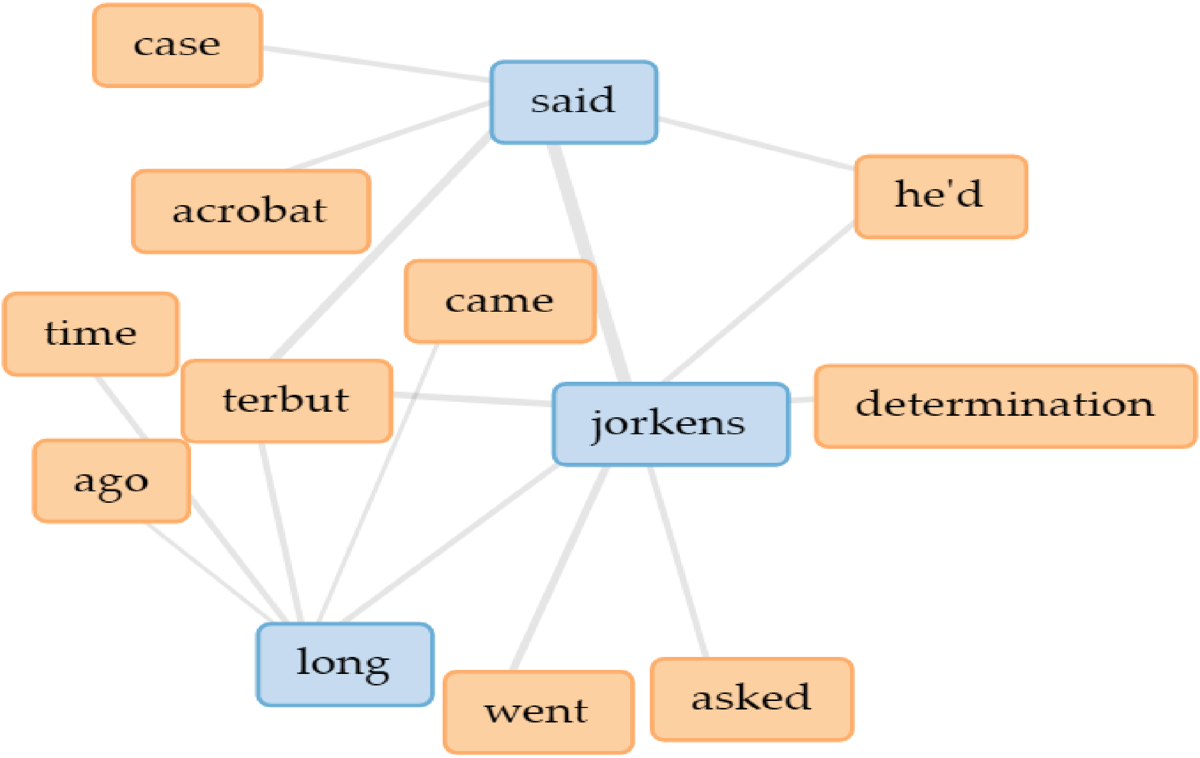
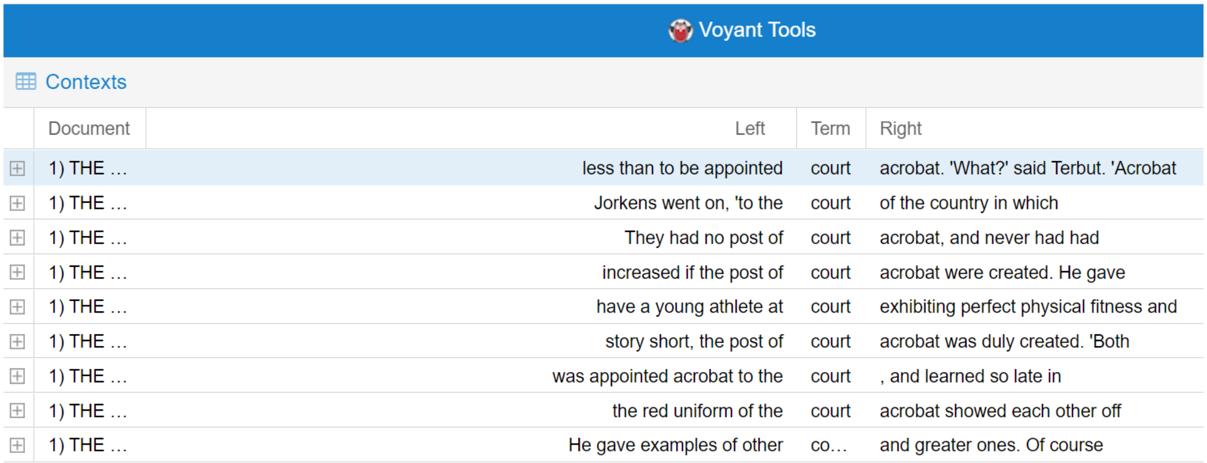


Figure 40 Links, The Reward

“Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166), so different nodes of knowledge graphs have been explored for the exploration and understanding of multidimensional hermeneutic patterns. Links tool establishes several knowledgeable connections among different characters, qualities and issues; for example, both major characters Terbut and Jorkens negotiate about determination; hence, the KG of “Jorkens, determination, Terbut, said, asked” is drawn digitally. Moreover, the KG “said, asked” refers to characters’ direct talks in the short stories. It discovers knowledge patterns that this story has direct conversations and dialogues along with narrative technique. Another KG of “long, time, came” refers to the “case” of court “acrobat” who toils for a long time to create the post of court acrobat and to become the most appropriate person to fill the acrobat post in the royal court. To conclude, these KGs verify textual information as concurrent validity.

v. Contexts



Voyant Tools

Contexts

	Document	Left	Term	Right
⊞	1) THE …	less than to be appointed	court	acrobat. 'What?' said Terbut. 'Acrobat
⊞	1) THE …	Jorkens went on, 'to the	court	of the country in which
⊞	1) THE …	They had no post of	court	acrobat, and never had had
⊞	1) THE …	increased if the post of	court	acrobat were created. He gave
⊞	1) THE …	have a young athlete at	court	exhibiting perfect physical fitness and
⊞	1) THE …	story short, the post of	court	acrobat was duly created. 'Both
⊞	1) THE …	was appointed acrobat to the	court	, and learned so late in
⊞	1) THE …	the red uniform of the	court	acrobat showed each other off
⊞	1) THE …	He gave examples of other	co…	and greater ones. Of course

Figure 41 Contexts, The Reward

The word “court” has several semantic shades, for instance, a prestigious place for queen or king, a decision place about disputes, a small road, a playground, a start of an amorous relationship, an encounter of risk and act of pleasure. To resolve this semantic ambiguity, its context is explored, the word “acrobat” has been found nine times in this short story, and it collocates as “court acrobat” five times (in 1st, 3rd, 4th, 6th, 8th sentences). The key word is found four times in 2nd, 5th, 7th, 9th sentences referring to a royal seat where the king and royal family live, and important government issues are decided.

7. *The Use of Force* by William Carlos Williams

i. Summary

This corpus has 1 document with 1,268 total words and 467 unique word forms. Created 13 seconds ago (30th October 2017).

Vocabulary Density: 0.368

Average Words Per Sentence: 12.2

Most **frequent** **words** in the corpus: throat (12); said (11); look (10); mother (10); child (9); come (6); let (6); doctor(5); open (5); teeth (5);

Figure 42 Summary, The Use of Force

Figure 45 exhibits the literary style of William Carlos Williams, and those stylistic qualities have been named stylometry. This short story uses 467 unique words, and they are reused almost three times; hence, their total words are 1268. Its vocabulary density is 0.368 which indicates its easy and repeated vocabulary. In every sentence, an average of 12.2 words have been used, and they refer to the average length of sentences. Its most frequent words reveal that this short story is related to the throat problem of a child, and the doctor forcibly examines the child's throat. Validation is an important part of each research, so the extracted theme of Summary tool matches the findings of Cirrus tool. Thus, "It is supplemented by other materials" (Rockwell, & Sinclair, 2016, p. 166). This concurrent validity enhances the reliability and validity of both Cirrus and Summary tool.

ii. Cirrus

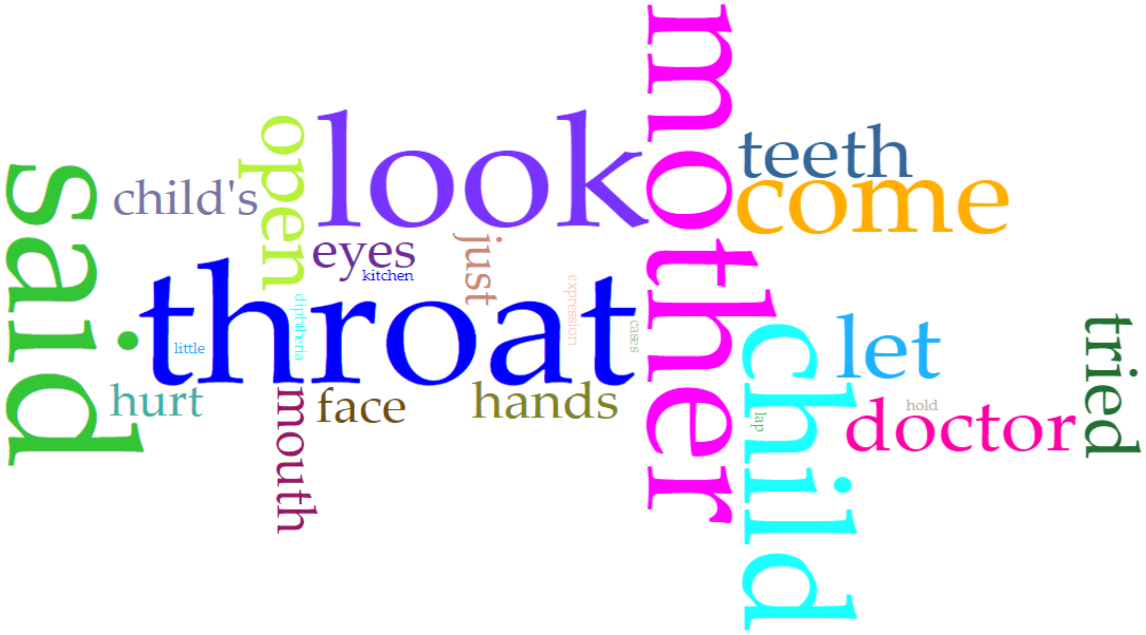


Figure 43 Cirrus, The Use of Force

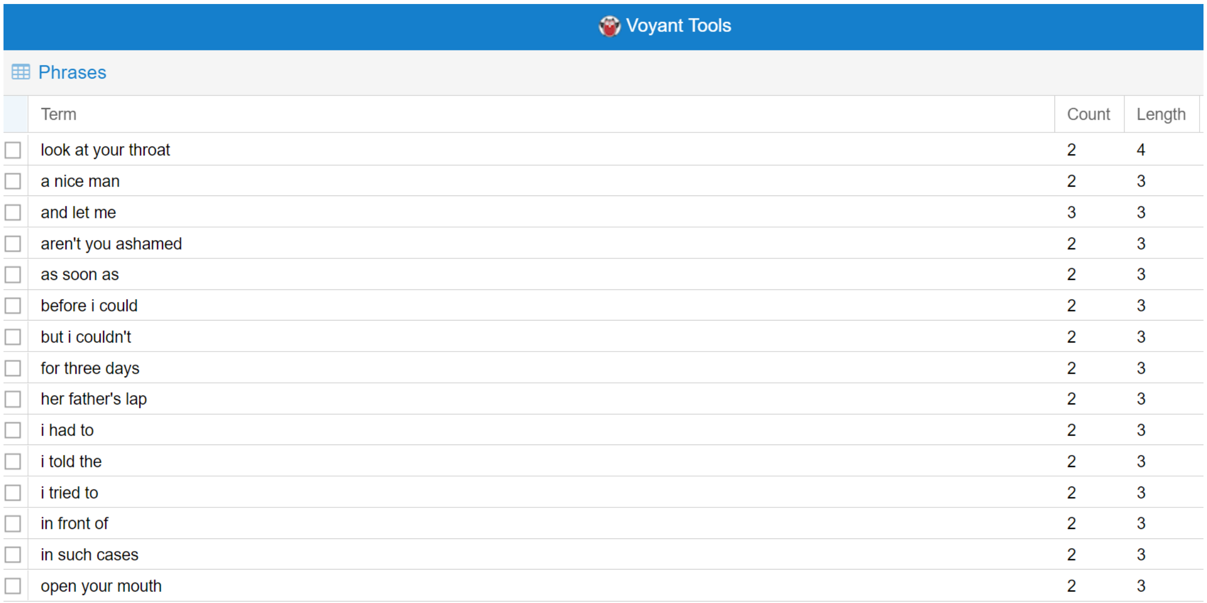
One proper noun, "Mathilda", a young patient of diphtheria, dominates in this short story. Firstly, most of the major characters of this short story are unnamed, for instance, two pronouns "I (46)", "she (24)" and some common nouns, for example "doctor (5)", "child (4)", "mother (10)" and "father (4)". In addition to it, the pronoun "I (46)" refers to the doctor, and the doctor narrates the entire story as his first-hand experience in the first-person narrative. This story-telling technique emphasises the writer's point of view, the depiction of personal feelings and unveiling of the narrator's mind. Consequently, the dialogues of various characters evolve the plot of the story. Moreover, the word “said (11)” also guides the dialogic nature of this short story because the doctor tries to convince the child and her parents during the forced diagnosis of diphtheria.

Very interesting knowledge aspect has been discovered through Cirrus that some facial organs have been mentioned for example “face (4)”, “teeth (5)”, “throat (12)”, “mouth (4)” and “eyes (4)” because these body organs are usually examined by most of the doctors. The doctor tries to examine Mathilda’s throat, while she crushes his spatula into pieces. Then “hands (4)” are mentioned because the doctor and parents hold her hands to examine her throat by force, while she claws at the doctor with her hands.

Ailing conditions of the child have been exhibited with Cirrus themes of “diphtheria (3)”, a throat disease of the “child (9)”, “fever (2)” and “hurt (4)”. The “doctor (5)” tries to give proper treatment to save the patient’s life because he has already seen some victims of diphtheria, and this situation has been informed through the theme of “tried (5)”. Themes of “fought (2)”, “attack (1)” “furious (2)”, “hurt (4)” and “bleeding (1)” reveal a resisting situation between the doctor and the patient. This Cirrus reveals several topics and knowledge patterns with their statistical weight. Findings of the current study match McNaught and Lam’s (2010) studies about finding themes from interviews.

Human analysis shows themes as “his commanding behaviour on his wife, male domination, female subordination, snubbing nature of husband, sympathetic nature of wife, furious, assault, attack” (Hussain, 2009, pp. 46-48). Human analysis and Cirrus tool extract the same characters, but some themes, for instance, “attack, furious” are the same, whereas the rest of the themes are different.

iii. Phrases



	Term	Count	Length
□	look at your throat	2	4
□	a nice man	2	3
□	and let me	3	3
□	aren't you ashamed	2	3
□	as soon as	2	3
□	before i could	2	3
□	but i couldn't	2	3
□	for three days	2	3
□	her father's lap	2	3
□	i had to	2	3
□	i told the	2	3
□	i tried to	2	3
□	in front of	2	3
□	in such cases	2	3
□	open your mouth	2	3

Figure 44 Phrases, The Use of Force

Frequent standard phrases are “look at your throat” (V+Prep+Prn+N), “a nice man” (Art+Adj+N), “aren’t you ashamed” (Aux+ Prn+V), “as soon as” (Subordinate conjunction), “I had to” (Prn+Aux+Prep), “in front of” (PP), “in such cases” (Prep+Det+N) and “open your mouth” (V+Prn+N). First knowledge discovery is an extraction of frequent phrases, and second knowledge discovery is the ideology of the speaker, which is reflected in collocation/ n-gram; for instance, “I had to” phrase suggests a link with the title of the short story *‘Use of Force’* in which the doctor is compelled to use force to save the patient’s life. Moreover, some collocations/ n-grams, “look at your throat” (V+Prep+Prn+N) and “open your mouth” (V+Prn+N) also suggest imperative sentence structure form. The order also expresses power relationship, utter power of discourse producer and complete powerlessness of discourse consumer. These collocations/ n-grams are repeated twice, and their length varies from 3 to 4 words.

iv. Links

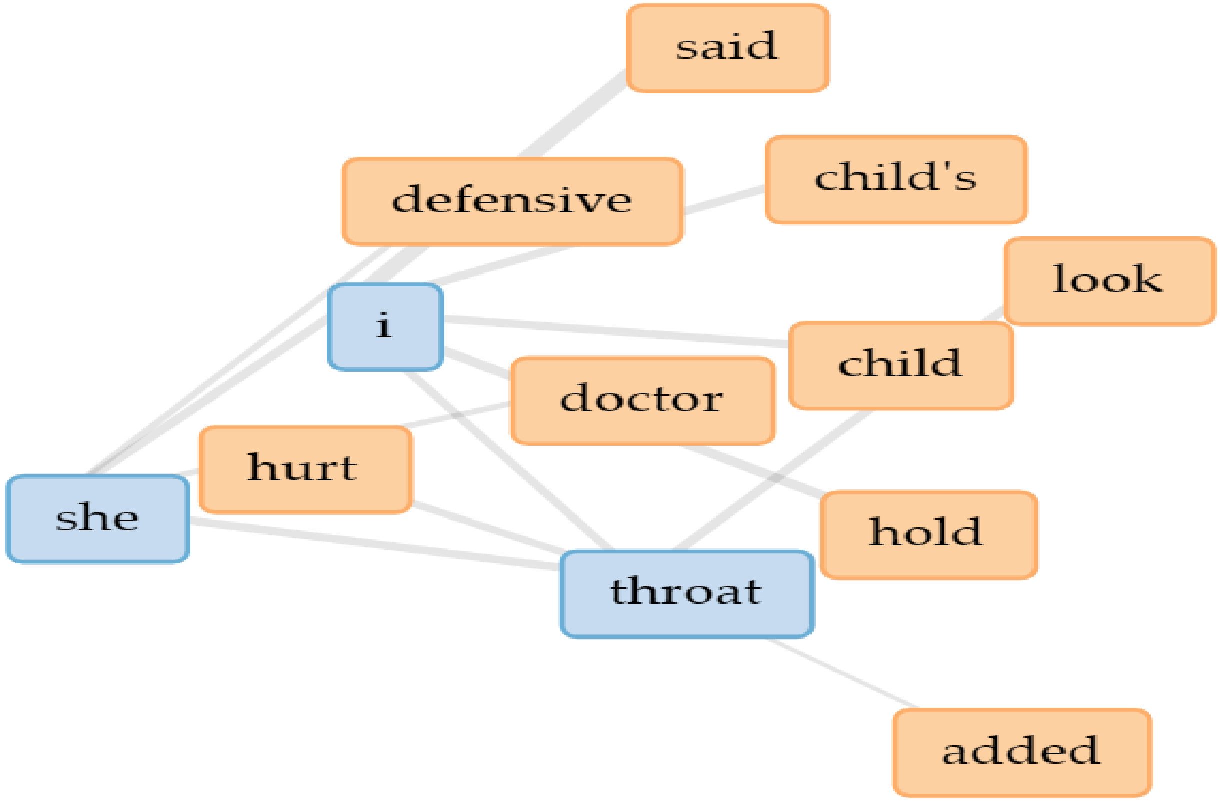
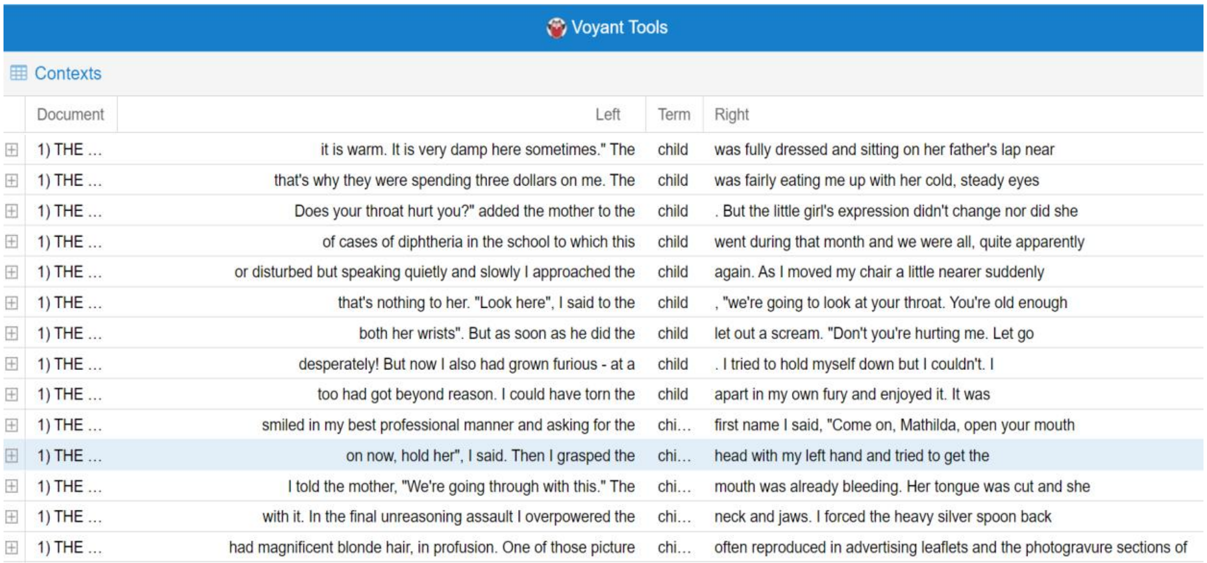


Figure 45 Links, The Use of Force

Links tool generates a set of linkages with nodes, and they are similar to 100 billion neurons in a human brain. So, figure 44 is named a set of various KGs. Major bonds are among “I, she, throat”, and it reveals the knowledge pattern that “I” (doctor) treat a diphtheria patient’s “throat”. Another KG “Throat, she, hurt” shows painful throat and swollen glands when the doctor examines her throat by dint of his authority and force. Another KG of “I, defensively, child” describes the initial situation when the doctor requests the patient and tries to convince the child with care and consideration. After the use of force, the doctor has to behave defensively in the wake of her attacks. Later on, the same doctor becomes more aggressive after she claws into the doctor’s eyes, and throws his glasses away. One KG of “hold, doctor, I, she” suggests that the doctor firmly grasps her arms and inserts a wooden spatula inside her throat to diagnose her disease. In a fit of fury, she chews the wooden spatula with her teeth. To summarise, all KGs testify and validate the text of the story; hence, it serves the purpose of concurrent validity.

v. Contexts



	Document	Left	Term	Right
⊞	1) THE …	it is warm. It is very damp here sometimes." The	child	was fully dressed and sitting on her father's lap near
⊞	1) THE …	that's why they were spending three dollars on me. The	child	was fairly eating me up with her cold, steady eyes
⊞	1) THE …	Does your throat hurt you?" added the mother to the	child	. But the little girl's expression didn't change nor did she
⊞	1) THE …	of cases of diphtheria in the school to which this	child	went during that month and we were all, quite apparently
⊞	1) THE …	or disturbed but speaking quietly and slowly I approached the	child	again. As I moved my chair a little nearer suddenly
⊞	1) THE …	that's nothing to her. "Look here", I said to the	child	, "we're going to look at your throat. You're old enough
⊞	1) THE …	both her wrists". But as soon as he did the	child	let out a scream. "Don't you're hurting me. Let go
⊞	1) THE …	desperately! But now I also had grown furious - at a	child	. I tried to hold myself down but I couldn't. I
⊞	1) THE …	too had got beyond reason. I could have torn the	child	apart in my own fury and enjoyed it. It was
⊞	1) THE …	smiled in my best professional manner and asking for the	chi…	first name I said, "Come on, Mathilda, open your mouth
⊞	1) THE …	on now, hold her", I said. Then I grasped the	chi…	head with my left hand and tried to get the
⊞	1) THE …	I told the mother, "We're going through with this." The	chi…	mouth was already bleeding. Her tongue was cut and she
⊞	1) THE …	with it. In the final unreasoning assault I overpowered the	chi…	neck and jaws. I forced the heavy silver spoon back
⊞	1) THE …	had magnificent blonde hair, in profusion. One of those picture	chi…	often reproduced in advertising leaflets and the photogravure sections of

Figure 46 Contexts, The Use of Force

When a common noun is used, there is ambiguity as to whom it is being referred to. In this short story, the word "child" created reference ambiguity. After finding the context of the "child" which is used 13 times for the patient child named Mathilda, and it is used for other patient children in the last 2 sentences. The user of Contexts tool can easily locate exact and the most pertinent quotes, as it has been emphasized in Hermeneutica Theory which is "embedded in a context" (Rockwell, & Sinclair, 2016, p. 166).

8. *The Gulistan of Sadi* by Sheikh Sadi

There are three tales of Sheikh Sadi in this short story, so each of them has been visualized and analysed separately because of their different themes and stories. These stories have been taken from Sheikh Sadi's English translation of the Persian tale book '*Gulistan-e-Sadi'*.

i. Summary

This corpus has 1 document with 851 total words and 365 unique word forms. Created about 2 minutes ago (1st November 2017).

Vocabulary Density: 0.429

Average Words Per Sentence: 15.5

Most frequent words in the corpus: king (17); boy (10); said (9); boat (5); asked (4); gulistan (4); sergeant (4); slave (4); world (4); became (3);

Figure 47 Summary, The Gulistan of Sadi

This corpus consists of 365 unique words, and they have been repeated almost a bit more than twice. Total words have been calculated at 851. Another fact is that there are three tales in this section and each tale has a different setting and diverse characters. The most common character in the three tales is king, and no other character has been repeated commonly in the three tales. that is why its vocabulary density is 0.429. Only function words, for instance, prepositions, articles have been repeated. Sheikh Sadi's translated tales have 15.5 average words in a sentence.

ii. Cirrus of Tale I

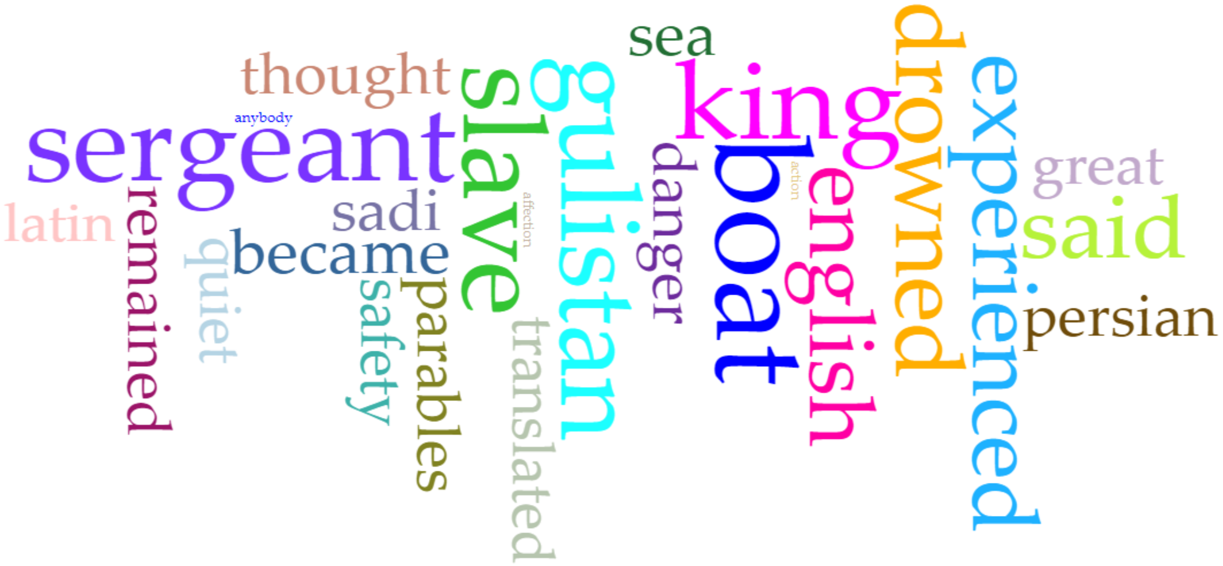


Figure 48 Cirrus, The Gulistan of Sadi

This Cirrus highlights main themes pertaining to the 1st tale about a king and a slave. The theme of "Gulistan (4)" indicates that the story has been taken from Sheikh "Sadi (2)" 's world-renowned book '*Gulistan-e-Sadi'* originally written in "Persian (2)". The most occurring theme is "boat (5)", a setting of the tale. In this story, the king "(4)" has been perturbed by the behaviour of a "slave" who is also voyaging in the same boat. The "sergeant (4)" offers himself voluntarily to quieten the disturbing slave. Having got permission from the king, he throws the slave into the “sea (2)” water. After facing the true “danger (2)” of drowning, he becomes “quiet (2)” and peaceful for the rest of the voyage.

ii. Cirrus of Tale II

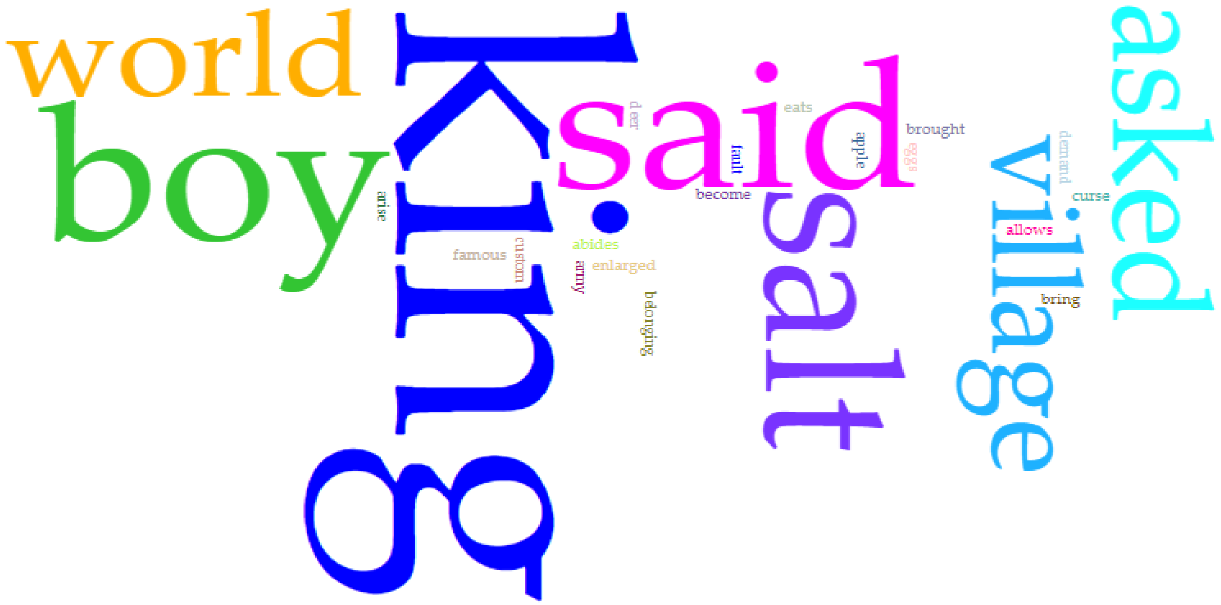


Figure 49 Cirrus, The Gulistan of Sadi

Figure 48 Cirrus also reveals the story related to the “king (6)”. The theme of “salt (3)” refers to an event when the king sends his workers to the “village (2)” to bring “salt (3)” for the hunted “deer (1)”. He asks whether you pay the price of salt or not. Then he explains that the foundation of evil is always small, and subordinates of the king always accelerate the intensity of the king's wrongdoings. The theme of “said (3)” suggests the use of dialogues for the development of the plot in the story.

ii. Cirrus of Tale III

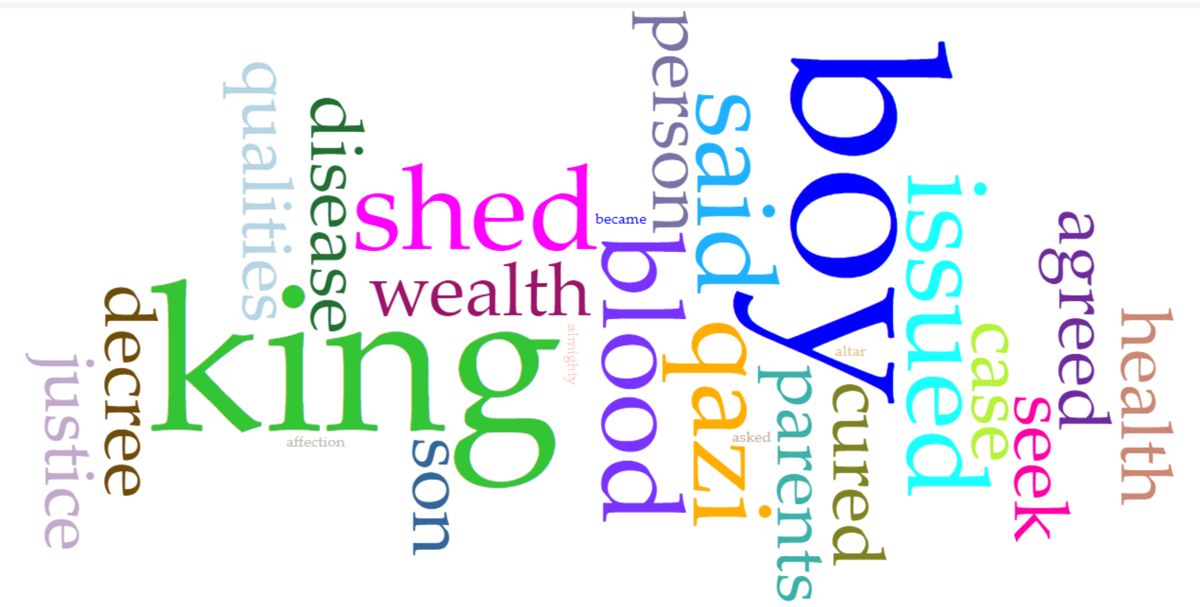
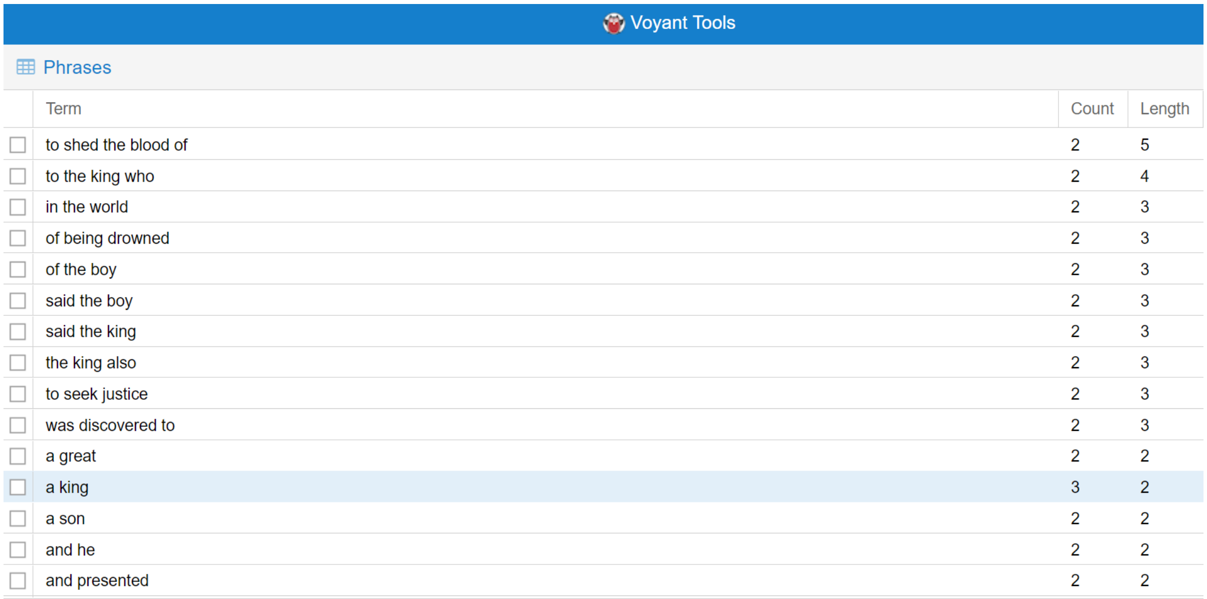


Figure 50 Cirrus, The Gulistan of Sadi

Figure 49 shows the most repeated characters, namely a “boy (7)” and a “king” (7). In this story, the king becomes sick, and his doctor has advised him to eat the bile of a boy who has some particular remedial characteristics. So, Qazi issues decree to “shed (3)” the “blood (3)” of such boy, and parents are also “agreed (2)” to slaughter their “son (2)” for heaps of “wealth (2)” from the “king (7)”. Then the boy looks towards the sky to seek “justice (2)” from God. The king questions about the boy’s body language, then the boy answers that first of all, he hopes for justice from parents, afterwards from Qazi, and then from the king, but all of them disappoint him, so lastly, the boy looks towards the sky seeking only God’s justice. “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). Text mining tools extract certain knowledge patterns and data visualizations, but human deep reflection and cognition draw meaningful results and interpretations from data visualization. The theme of “said" has been used three times to demonstrate the direct exchange of ideas among different characters. This tale shows the conversational style of Sheikh Sadi. Motifs of the current Cirrus match the Cirrus of 1500 apocalyptic novels (Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015) that both have highlighted key characters and themes.

Human analysis shows that male characters are twelve, whereas only one female character, “mother” (Hussain, 2009, p. 48). Cirrus tool extracts the detail of the characters of the king, qazi, boy. Moreover, Cirrus extracts key themes, for instance, justice, wealth, blood, sea. Consequently, Cirrus tool covers almost all key aspects which have been validated by textual evidence.

iii. Phrases



	Term	Count	Length
□	to shed the blood of	2	5
□	to the king who	2	4
□	in the world	2	3
□	of being drowned	2	3
□	of the boy	2	3
□	said the boy	2	3
□	said the king	2	3
□	the king also	2	3
□	to seek justice	2	3
□	was discovered to	2	3
□	a great	2	2
□	a king	3	2
□	a son	2	2
□	and he	2	2
□	and presented	2	2

Figure 51 Phrases, The Gulistan of Sadi

Substandard phrases, for example “and the” have been excluded from figure 50, while standard phrases have been included in this section: “to shed the blood of” (Inf V+Art+N+Prep), “in the world” (Prep+Art+N), “to seek justice” (Inf V+N) and “of being drowned” (Prep+Aux+V). Background ideology in the 3rd tale is about the boy who looks towards the sky to seek justice from God, and he complains that Qazi issues an unjust decree to shed the blood of an innocent soul. These collocations/ n-grams occur 2 to 3 times, and their length ranges from 2 to 5 words.

iv. Links of Tale I

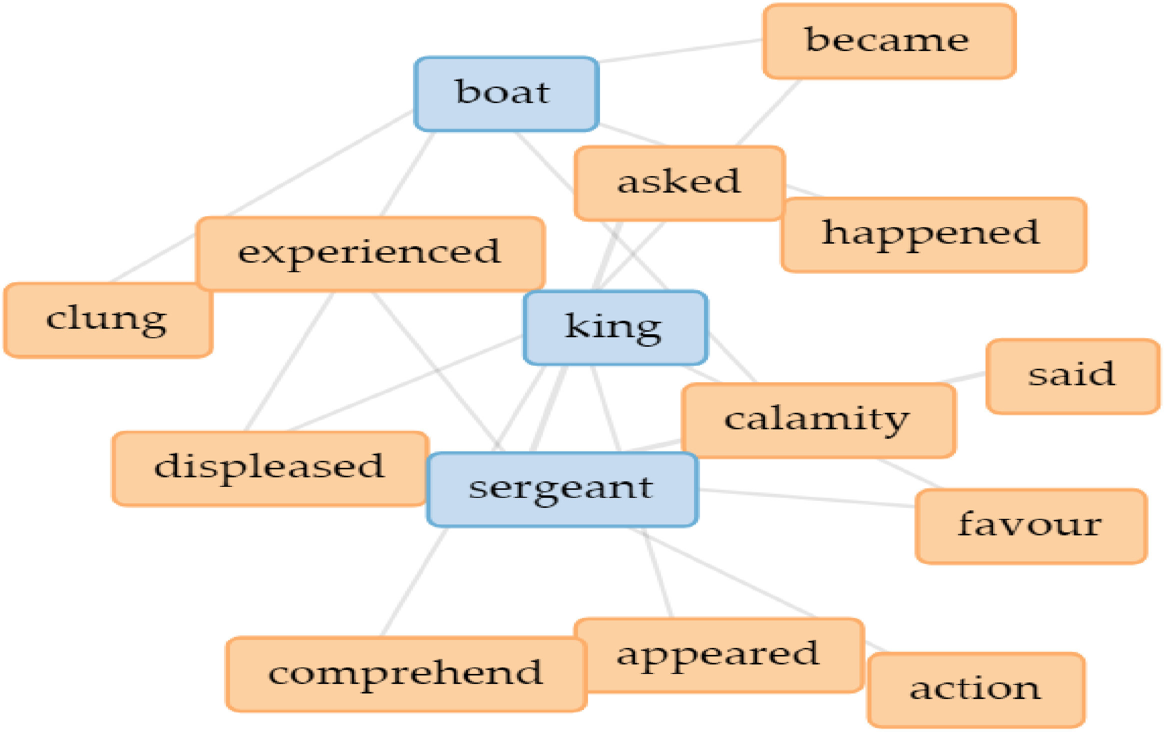


Figure 52 Links, The Gulistan of Sadi

Most of the time, the king and sergeant interact with each other; that is why their names have been shown in blue colour KG "king, sergeant" which suggests their dominant roles in the short story. The KG of "became, king, displeased" expresses an unpleasant mood of the king, while he is voyaging in the boat. The reason for displeasure is creating fuss by the slave who has boarded the boat for the first time. The KG about "asked, king, sergeant, action" denotes that the sergeant seeks royal permission to silence the slave, and the king permits him to take an action. Another KG of "clung, boat, calamity" shows the situation of the slave who does not realise the significance of safety in the boat. After undergoing an experience of the calamity of falling into the river, the slave clings to the backside of the boat and sits calmly.

iv. Links of Tale II

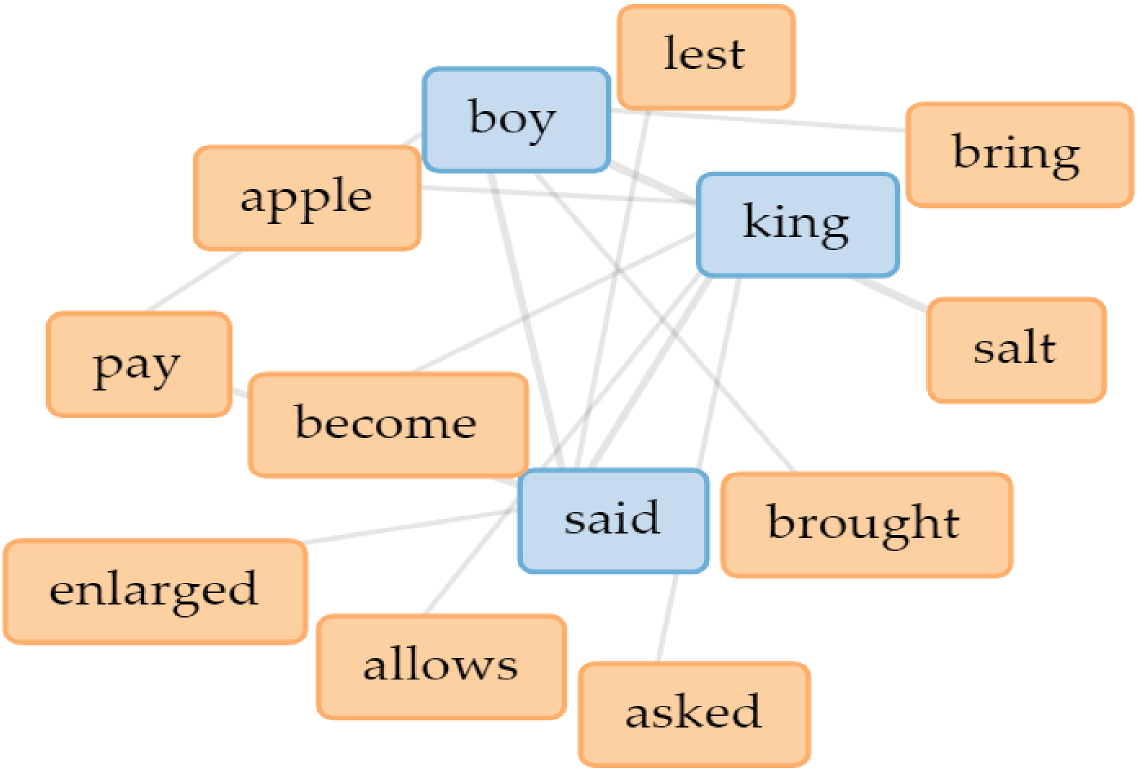


Figure 53 Links, The Gulistan of Sadi

Knowledge Discovery Theory is defined as “the extraction of implicit, previously unknown and potentially useful information from data” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). Primarily, there are two main characters conversing in the KG, “boy, said, king”. Furthermore, the word, “said” shows that their dialogues proceed the plot of the 2nd tale. The KG of “said, king, salt” informs that the king orders to get salt from the villagers. The king advises to pay money for the salt.

iv. Links of Tale III

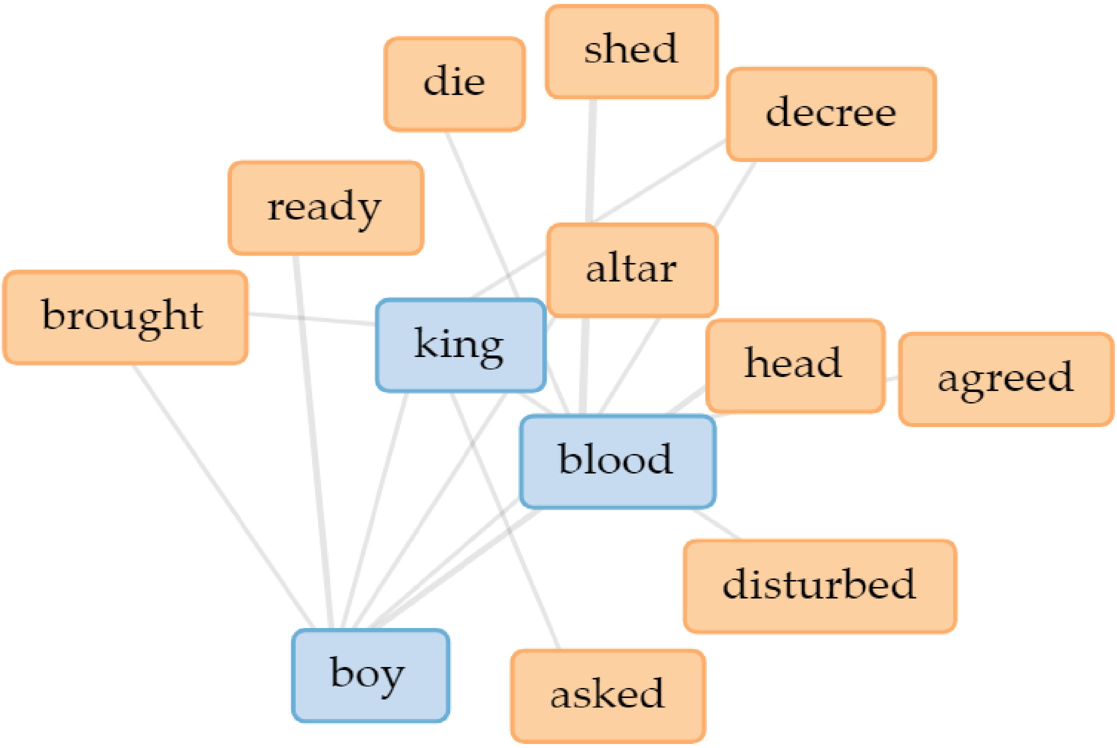
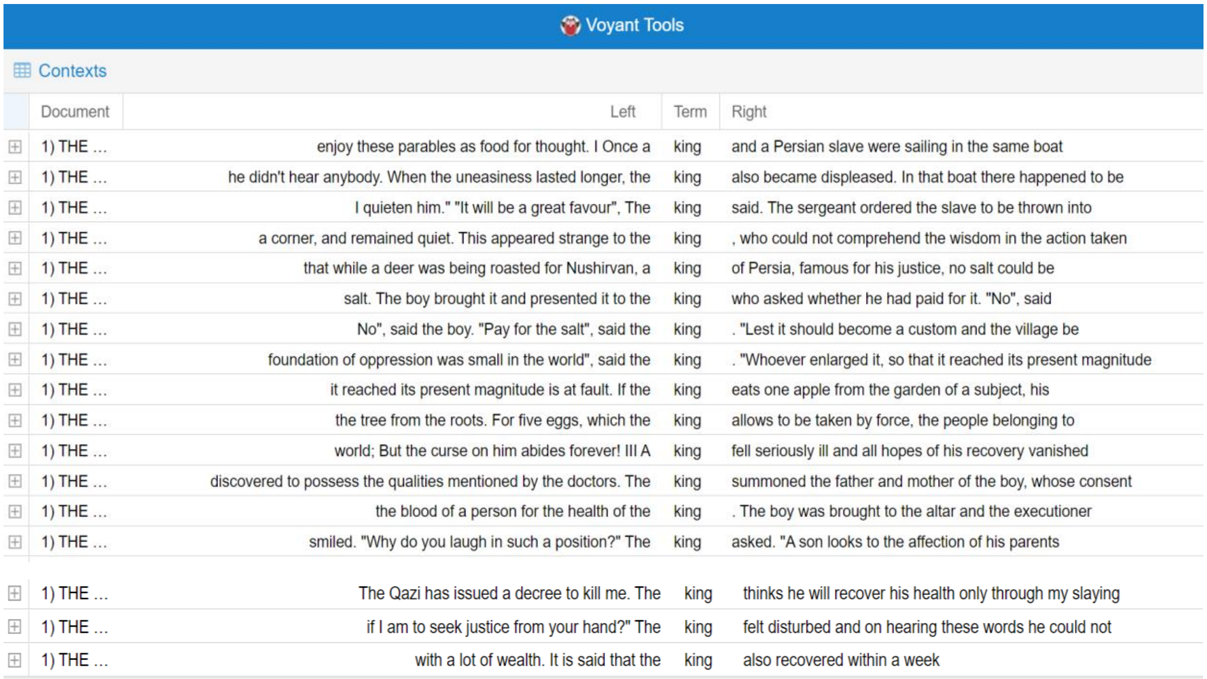


Figure 54 Links, The Gulistan of Sadi

The boy has been brought to the altar to be slaughtered for the king's health recovery, and this situation has been exhibited in the KG of “brought, boy, altar”. Apart from two human conversationalists, the word “blood” is very prominent in the KG because it has been discussed repeatedly and has deep-seated ties with other characters. The king searches a boy’s bile for his health, and the Qazi issues the decree for the bloodshed of an innocent boy to treat the king. The KG of “king, blood, shed, decree” expresses that Qazi issues a decree to shed the blood of a certain boy to cure the king of lethal disease. Afterwards, parents also agree upon the bloodshed of their son for the sake of abundant money, and to please the Shadow of God (king). Eventually, the boy looks towards the sky, and considers God his sole redeemer from the intended atrocities of worldly people. Another KG of “king, blood, disturbed” reveals that after listening to the boy’s judicious answer, the king is deeply grieved and disturbed due to boy’s bloodshed order, so he kisses the head of the boy with fatherly affection, and releases him.

v. Contexts



Document	Left	Term	Right
1) THE ...	enjoy these parables as food for thought. I Once a	king	and a Persian slave were sailing in the same boat
1) THE ...	he didn't hear anybody. When the uneasiness lasted longer, the	king	also became displeased. In that boat there happened to be
1) THE ...	I quieten him." "It will be a great favour", The	king	said. The sergeant ordered the slave to be thrown into
1) THE ...	a corner, and remained quiet. This appeared strange to the	king	, who could not comprehend the wisdom in the action taken
1) THE ...	that while a deer was being roasted for Nushirvan, a	king	of Persia, famous for his justice, no salt could be
1) THE ...	salt. The boy brought it and presented it to the	king	who asked whether he had paid for it. "No", said
1) THE ...	No", said the boy. "Pay for the salt", said the	king	. "Lest it should become a custom and the village be
1) THE ...	foundation of oppression was small in the world", said the	king	. "Whoever enlarged it, so that it reached its present magnitude
1) THE ...	it reached its present magnitude is at fault. If the	king	eats one apple from the garden of a subject, his
1) THE ...	the tree from the roots. For five eggs, which the	king	allows to be taken by force, the people belonging to
1) THE ...	world; But the curse on him abides forever! III A	king	fell seriously ill and all hopes of his recovery vanished
1) THE ...	discovered to possess the qualities mentioned by the doctors. The	king	summoned the father and mother of the boy, whose consent
1) THE ...	the blood of a person for the health of the	king	. The boy was brought to the altar and the executioner
1) THE ...	smiled. "Why do you laugh in such a position?" The	king	asked. "A son looks to the affection of his parents
1) THE ...	The Qazi has issued a decree to kill me. The	king	thinks he will recover his health only through my slaying
1) THE ...	if I am to seek justice from your hand?" The	king	felt disturbed and on hearing these words he could not
1) THE ...	with a lot of wealth. It is said that the	king	also recovered within a week

Figure 55 Contexts, The Gulistan of Sadi

There are three tales and the character of the king is common in all of them. So, to disambiguate his diverse roles in three different tales, it is deemed to retrieve information for the key word “king”. Figure 55 unveils that the word “king” has been used 17 times. The word “king” is used four times in the 1st tale, six times in the 2nd tale, and last seven times in the 3rd tale.

9. *The Foolish Quack* (Folk Tale)

i. Summary

This corpus has 1 document with 833 total words and 316 unique word forms. Created 36 seconds ago (on 1st November 2017).

Vocabulary Density: 0.379

Average Words Per Sentence: 16.3

Most frequent words in the corpus: camel (9); man (9); cure (7); old (7); doctor (6); men (6); said (6); answered(5); struck (5); throat (5);

Figure 56 Summary, The Foolish Quack

It has 316 unique words, and almost more than double words have been used in the entire short story, and they become 833 words. By dividing unique words with total words, 0.379 vocabulary density has been calculated. Average 16.3 words per sentence have been written, and it indicates longer sentences than previous stories. Stylometric analysis shows that its sentence length is suitable for advanced level learners, but its vocabulary density is at the level of intermediate learners. Writing style is dialogic in nature

ii. Cirrus

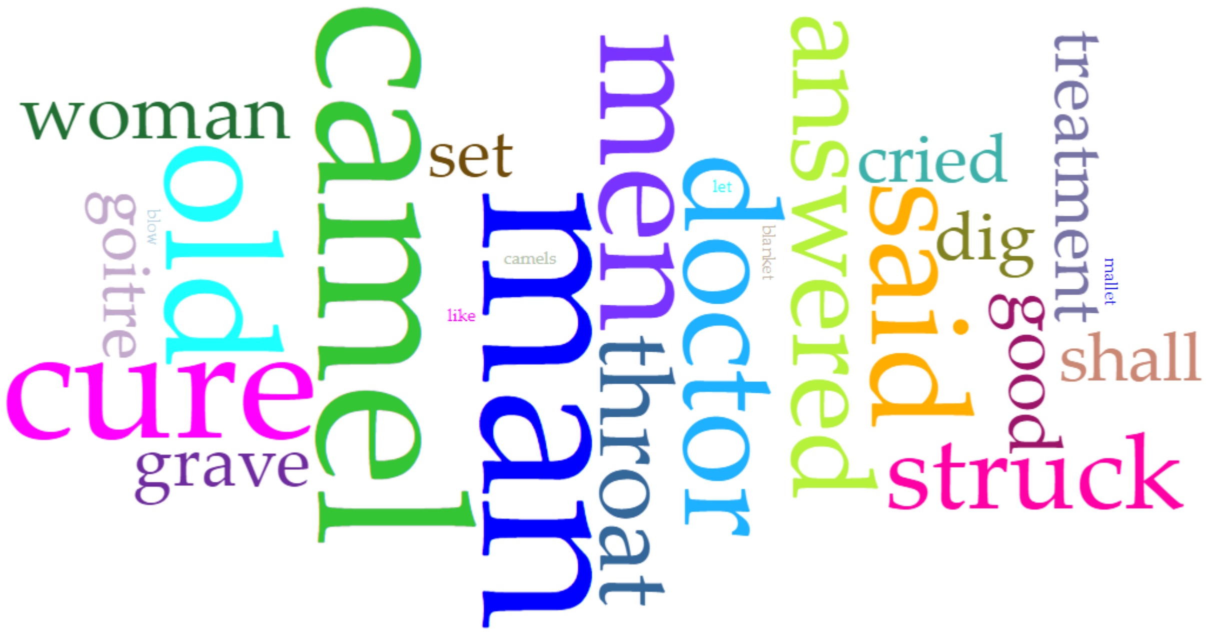


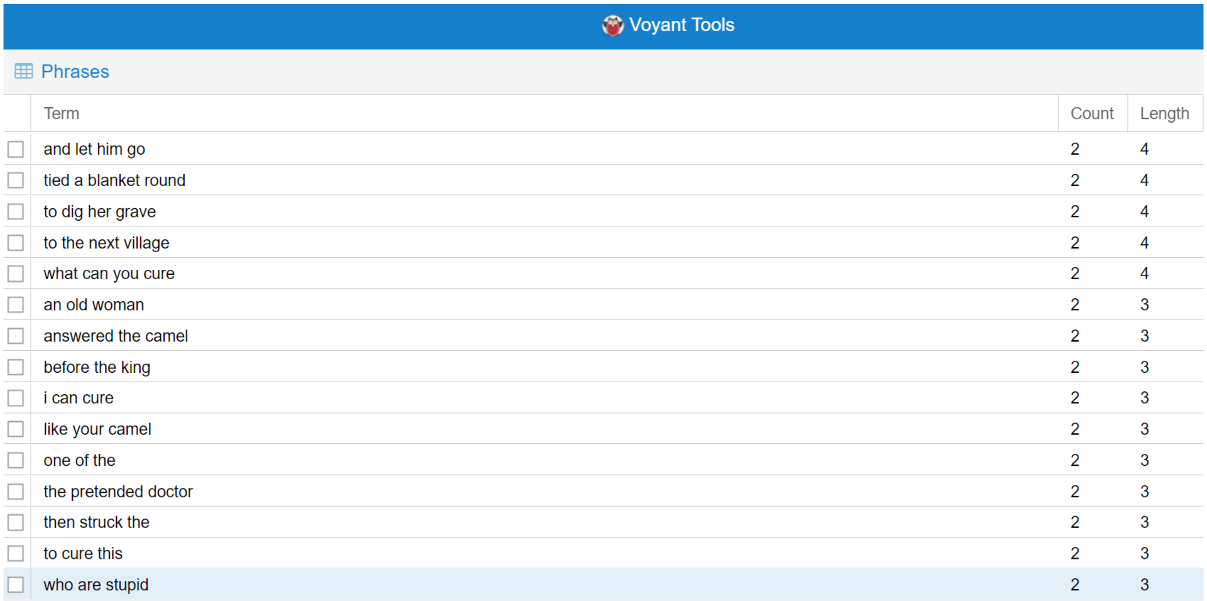
Figure 57 Cirrus, The Foolish Quack

"In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). Figure 56 Cirrus shows the key themes of "pretended (3)", "foolish (2)", "doctor (6)" to elaborate the title '*Foolish Quack'*. Themes of "camel (9)", "man (9)" and "men (6)" suggest that the foolish quack learns the art of curing throat swelling from camel men, and he applies the same animal-related treatment method to human beings. This knowledge pattern has been shown in the Cirrus with themes of "melon (3)", "blanket (3)", "throat (5)", "goitre (4)", "woman (4)", "struck (5)" and "cried (4)". At the end, the same camel man hits the quack with a heavy stick to teach him the obvious differences between a man and an animal's physique, stamina, endurance and treatment style.

The theme of "died" has been found twice: once the poor old feeble woman passes away by the first hit of the mallet, while the second time, he disowns the responsibility of digging the grave of the sick person. Consequently, villagers refrain from his brutal treatment. The theme of "beat (2)" has been found twice, and the quack has been beaten twice, once by the villagers and the second time by the camel man. The words "said (6)" and "answered (5)" have been found in the Cirrus, and they indicate the use of dialogues among foolish quack and villagers; and foolish quack and camel man.

The human analysis discusses one feminine character, an old lady ("a poor old creature"), and remaining male characters, for instance, quack, camel men and villagers. Key themes of this story are "foolish, villain, wretch" (Hussain, 2009, pp. 48-49). Cirrus-generated characters and humanly analysed characters are the same, but their themes are different except for one common theme, namely "foolish".

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	and let him go	2	4
□	tied a blanket round	2	4
□	to dig her grave	2	4
□	to the next village	2	4
□	what can you cure	2	4
□	an old woman	2	3
□	answered the camel	2	3
□	before the king	2	3
□	i can cure	2	3
□	like your camel	2	3
□	one of the	2	3
□	the pretended doctor	2	3
□	then struck the	2	3
□	to cure this	2	3
□	who are stupid	2	3

Figure 58 Phrases, The Foolish Quack

As standard phraseology is concerned, the following collocation patterns have been selected from the first 15 most occurring phrases: "and let him go" (Conj+V+Prn+V), "to dig her grave" (Inf V+Prn+N), "to the next village" (Prep+Art+Adj+N), "before the king" (Adv+Art+N), "the pretended doctor" (Art+Adj+N), "I can cure" (Prn+Mod+V), "one of the" (N+Prep+Art) and "who are stupid" (Int Prn+Aux+Adj). These co-occurrences show the ideology of people who intend to let the foolish quack go and not take him to court to punish him. Moreover, standard phraseology facilitates learning correct English; for instance, the phrase “before the king” is right, while the phrase “opposite the king” is wrong. So, phraseology establishes correct linguistic norms for language learners and especially for second language learners. All collocations/ n-grams occur twice, and their length ranges from 2 to 4 words.

iv. Links

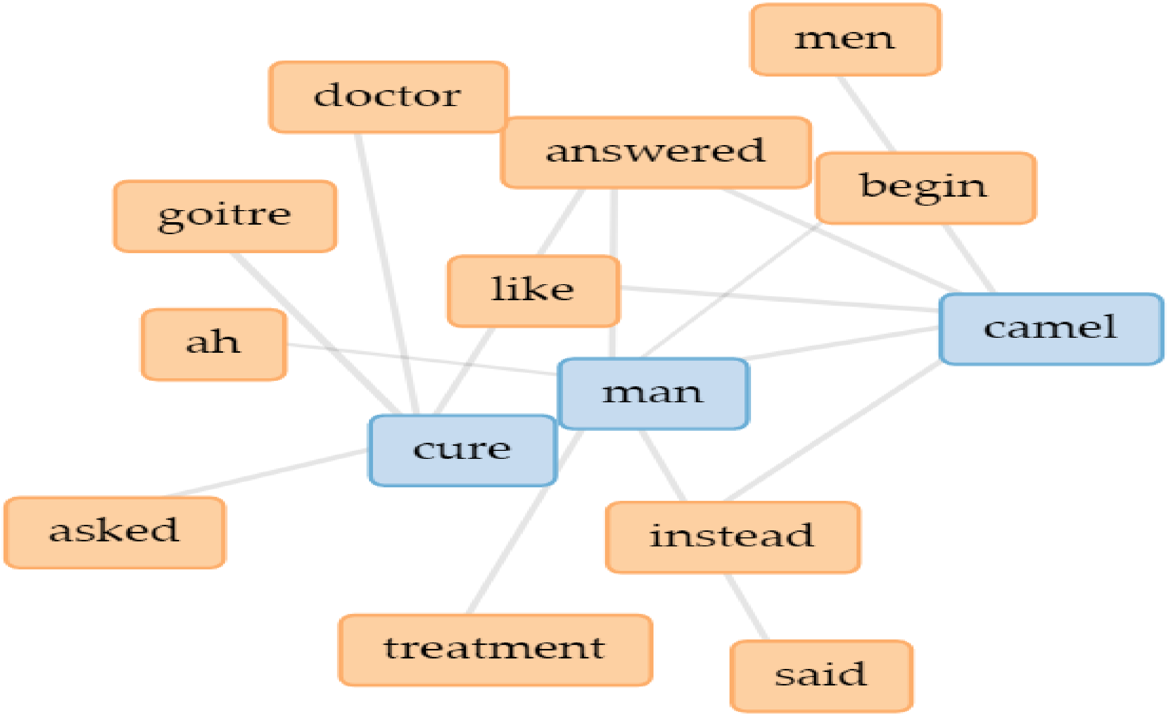
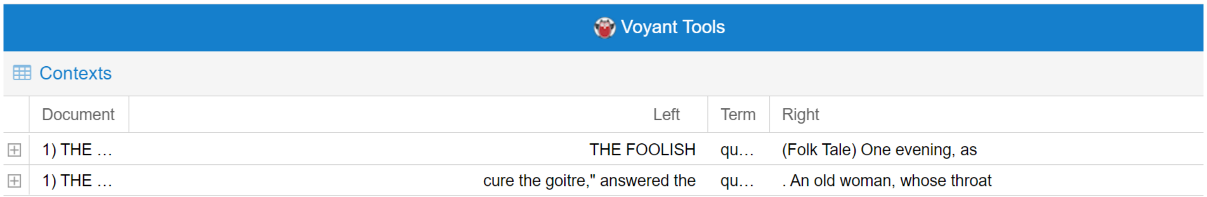


Figure 59 Links, The Foolish Quack

The KG of “cure, doctor, goitre” refers to the fake doctor’s start of goitre treatment. The KG of “treatment, camel, man, begin,” suggests two events: firstly, he cures the camel throat when it eats a melon and it stucks in its throat. Secondly, he beats the foolish quack with an iron stick to teach him a practical lesson that he should not apply animal treatment methods to fragile human beings. Another KG of “ah, man” suggests the concluding scene when the wretched man (quack) realises his mistake that he wrongly applies camel treatment technique to the feeble poor old woman, and in this self-revealing situation, he utters the interjection of “ah”.

v. Contexts

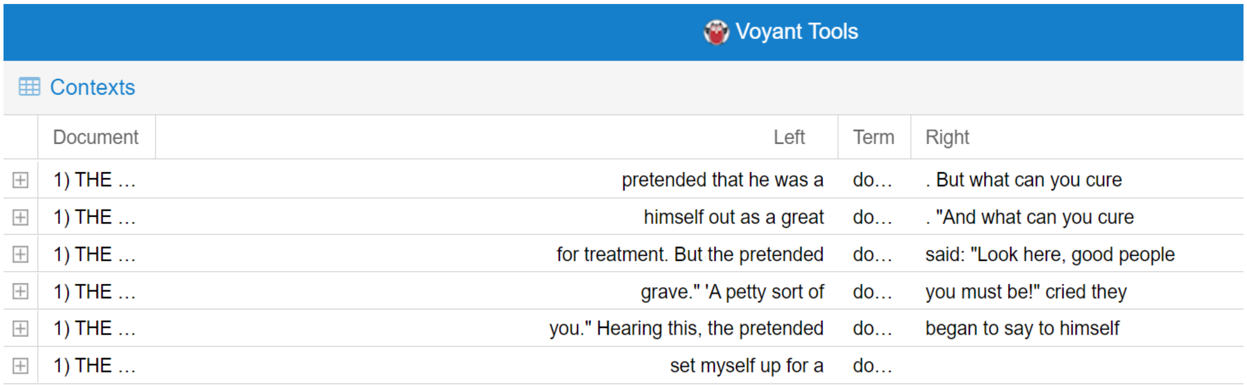


Voyant Tools

Contexts

Document	Left	Term	Right
1) THE …	THE FOOLISH	qu…	(Folk Tale) One evening, as
1) THE …	cure the goitre," answered the	qu…	. An old woman, whose throat

Figure 60 Contexts, The Foolish Quack



Voyant Tools

Contexts

Document	Left	Term	Right
1) THE …	pretended that he was a	do…	. But what can you cure
1) THE …	himself out as a great	do…	. "And what can you cure
1) THE …	for treatment. But the pretended	do…	said: "Look here, good people
1) THE …	grave." 'A petty sort of	do…	you must be!" cried they
1) THE …	you." Hearing this, the pretended	do…	began to say to himself
1) THE …	set myself up for a	do…	

Figure 61 Contexts, The Foolish Quack

The title of the short story is *'The Foolish Quack'*, but the key word "quack" has been used only twice in figure 60, once in the title of the short story, and the second time in the text, whereas the word "doctor" has been used 6 times in the text (figure 61). There is a confusion whether any real "doctor" is present or not in the story. This confusion is resolved by applying Hermeneutica Theory which is "embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). By finding the word "doctor" through Contexts tool, this semantic ambiguity can be resolved easily. The adjective "pretended" is used only three times with the noun "doctor" in 3rd and 5th sentence. In the 1st sentence, the word "pretended" as a verb has been mentioned. In 2nd sentence "himself out as a great doctor", in 4th sentence "a petty sort of doctor" and in 6th sentence "set myself up for a doctor" indicate that he is a "pretended doctor" or "quack" who is equal in terms of their meaning and role. G.B. Shaw writes an essay titled, *'Are Doctors men of Science?'* and declares that there is only one difference between a doctor and a quack. The former can sign a death certificate, while the latter cannot do so, but both drag a patient to the grave (Shaw, 1930).

10. *A Mild Attack of Locusts* by Doris Lessing

i. Summary

This corpus has 1 document with 990 total words and 407 unique word forms. Created 6 seconds ago (on 2nd November 2017).

Vocabulary Density: 0.411

Average Words Per Sentence: 12.9

Most frequent words in the corpus: margaret (11); locusts (8); old (7); swarm (7); came (5); farm (5); smoke (5); thick (5); air (4); come (4);

Figure 62 Summary, A Mild Attack of Locusts

"Summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). Here the description of a subset has been done in a quantified manner. This quantified summary of the corpus elaborates stylometric patterns. Its unique words are 407, and they are repeated more than two times, so total words become 990, and vocabulary density is calculated at 0.411. Doris Lessing uses 12.9 average words per sentence, and it suggests her writing style with a medium length of sentences. The most frequent words present major characters as "Margaret (11)" and significant events, for instance, the attack of "swarm (7)" of "locusts (8)" and efforts to resist the "swarm (7)" attack with noise, fire and "smoke (5)".

ii. Cirrus

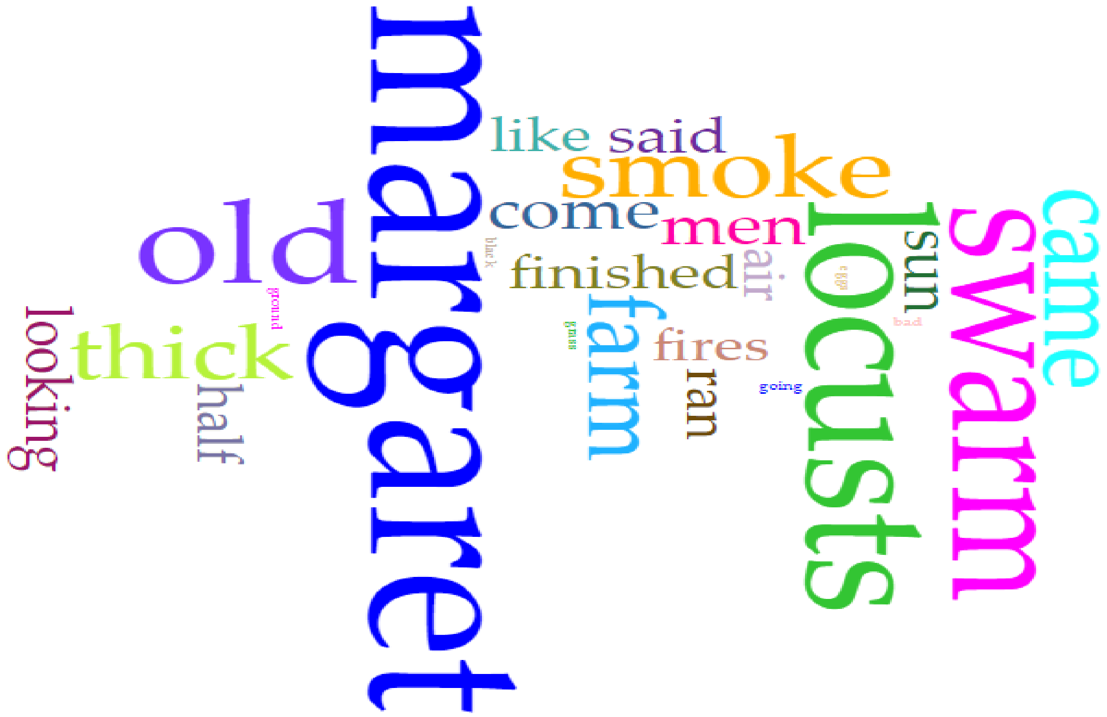


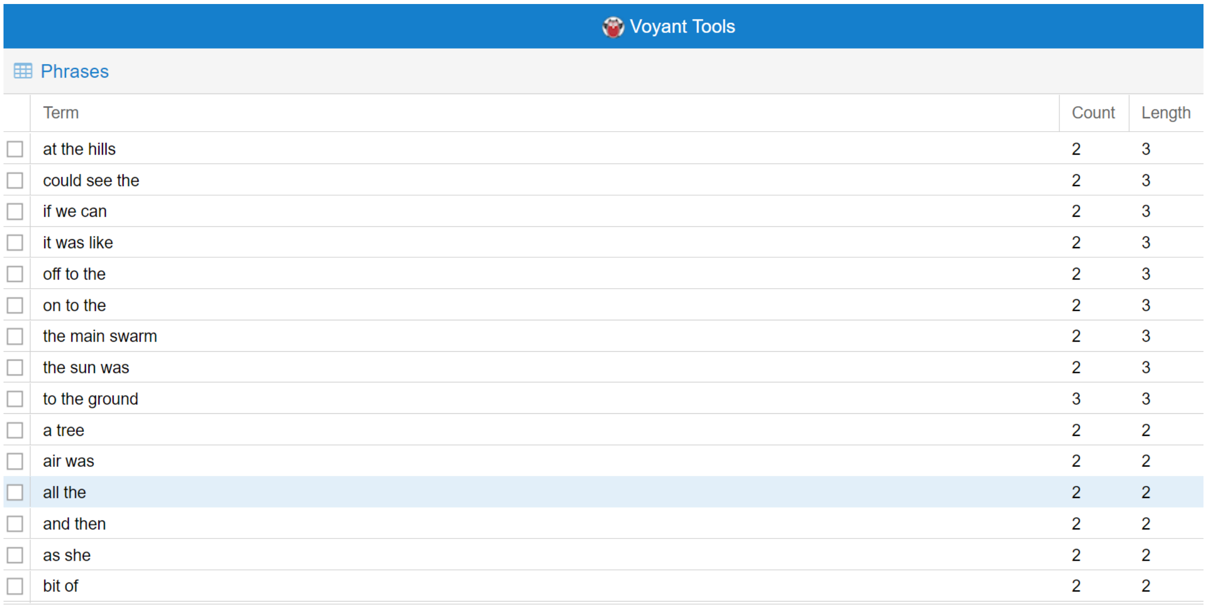
Figure 63 Cirrus, A Mild Attack of Locusts

The most occurring character is “Margaret (11)” who overwhelms and dominates the plot of the short story from beginning to the end. This Cirrus shows some other characters, for instance, “Old Stephen (3)”, “houseboy (2)”, “men (4)”, “cookboy (2)”. The theme of “hurry (3)” and “going (3)” refer to the speedy arrival of workers on the farmland to perform precautionary measures while saving crops from locust attack. Themes of “run (1)”, “come (4)”, “came (5)”, “coming (2)”, “going (3)”, “hurry (3)” indicate that all characters come in a hurry to produce smoke, fire, and bell sounds to get rid of the imminent arrival of a swarm of locusts.

Just one glance towards Cirrus shows some significant themes, namely “Margaret (11)”, “locusts (8)”, “swarm (7)”, “farm (5)”, “smoke (5)”, “fires (4)”, and some minor themes. Then a swarm of locusts attacks fields, and farmers burn the fire to produce smoke because locusts do not land on the smoky and noisy fields. Themes of “insects (3)”, “eggs (3)”, “eaten (2)”, “bad (3)”, “finished (4)”, “ruin (2)” show effects of locusts on the fields since they ruin crops by eating and laying their eggs. The theme of “said (4)” and “asked (2)” show the inclusion of dialogues and direct conversations among characters.

Human analysis shows characters of Margaret and labourers. Key themes of the story are "beating the ploughshare, collecting the cans, throwing wet leaves on fire, male domination, female character is powerless, shouted, yelled, pouring out of compounds, shouting, giving orders" (Hussain, 2009, pp. 49-50). Comparing Cirrus and human analysis, Cirrus extracts all characters precisely by naming them Old Stephen, houseboy, men, cookboy, while human analysis has not done so. Furthermore, themes of "smoke, fire" are common in both studies; however, the rest of the themes are different.

iii. Phrases



	Term	Count	Length
□	at the hills	2	3
□	could see the	2	3
□	if we can	2	3
□	it was like	2	3
□	off to the	2	3
□	on to the	2	3
□	the main swarm	2	3
□	the sun was	2	3
□	to the ground	3	3
□	a tree	2	2
□	air was	2	2
□	all the	2	2
□	and then	2	2
□	as she	2	2
□	bit of	2	2

Figure 64 Phrases, A Mild Attack of Locusts

From the first 15 phrases, certain standard phrases have been chosen, for instance, "at the hills" (Prep+Art+N), "if we can" (Conj+Prn+Mod), "it was like" (Prn+Aux+Adj), "the main swarm" (Art+Adj+N) and "to the ground" (Prep+Art+N). The phrase "the main swarm" denotes the group attack of locusts because they always attack in swarms. Farmers are not afraid of individual crop-eating insects, but they are worried about the landing of the main swarm in which millions of locusts attack a crop, and they not only eat the whole crop in a very short time but also lay eggs and moths for future destruction of the agricultural land. Besides, some collocations/ n-grams occur 2 to 4 times, and their length ranges from 2 to 3 words.

iv. Links

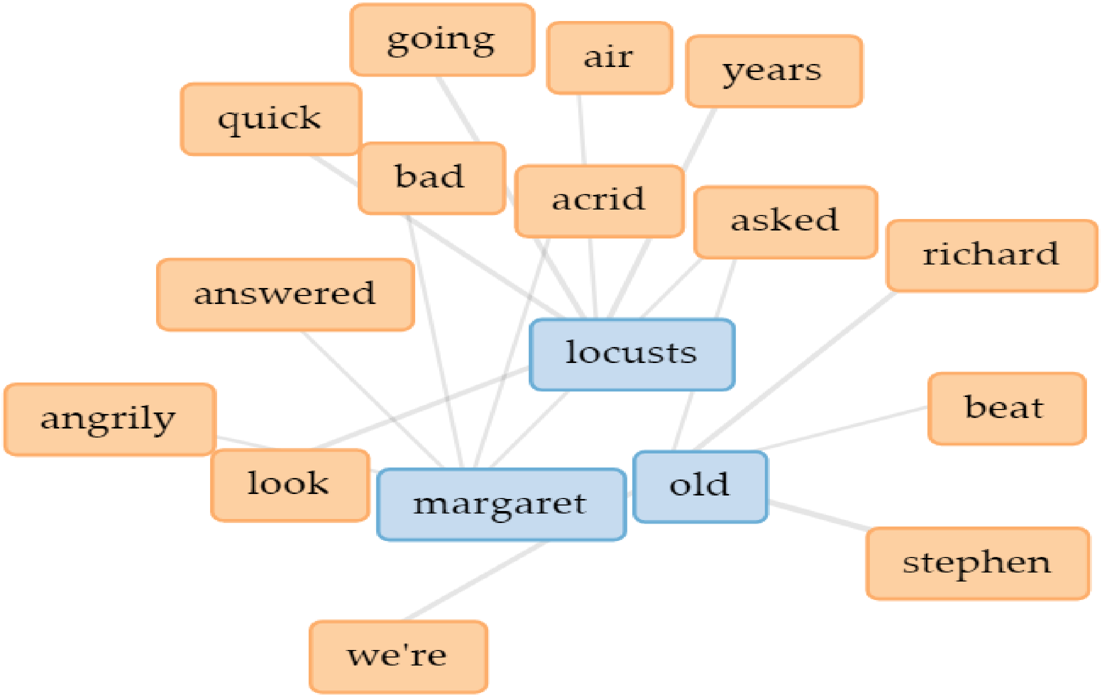
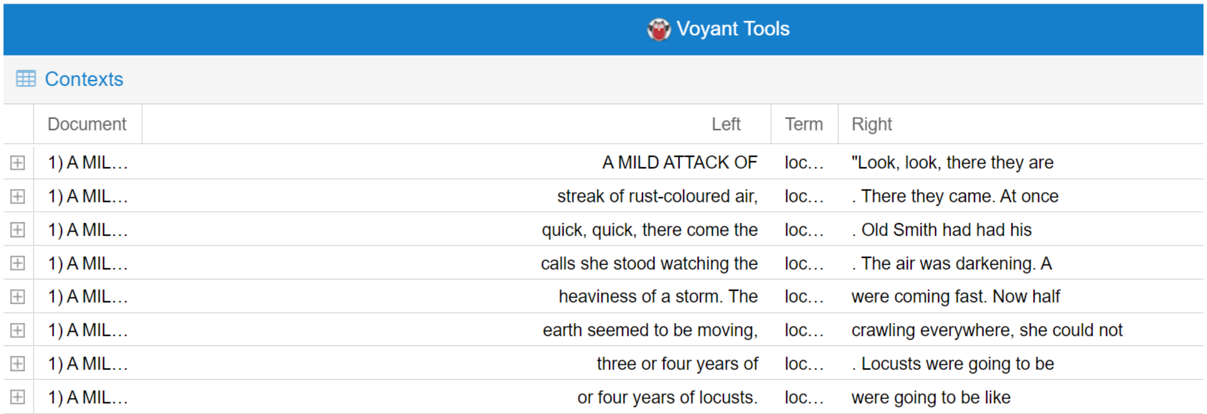


Figure 65 Links, A Mild Attack of Locusts

"Manipulation is in service of exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166). Links tool also shows the relationship of themes, but it is the task of humans to interpret various nodes of knowledge graphs by establishing their hermeneutic interrelationship. Thus, human reflection also supports the data visualization of a tool. The KG of "quick, locusts" refers to the event when people warn one another with the words "quick, quick". Another KG of "bad, locusts" discloses that locusts are prevailing on the farms like bad weather. The KG of "locusts, air, quick" shows that locusts come quickly with fast air flow. The KG of "old, Margaret, beat, Stephen" builds the relationship of asking to beat the bell with full sound to stop locusts and burn the fire while making it "acrid and black". Moreover, these KGs also discover a knowledge pattern that Margaret is ordering workers angrily and hastily because swarms are about to attack the fields.

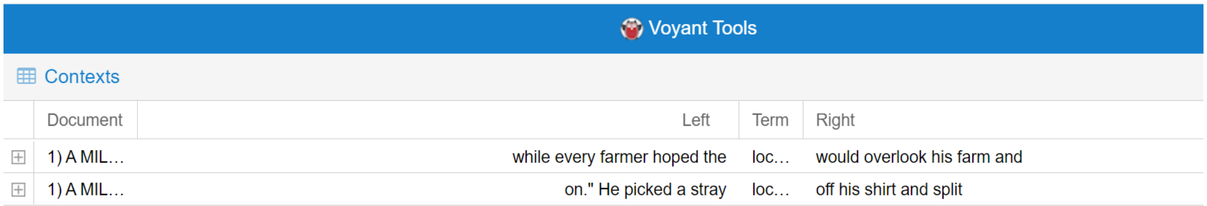
The KG of "old, Stephen, Richard" suggests that Richard and Old Stephen are issuing orders to servants on how to burn the fire and produce smoke in this combat against the swarm of locusts. Moreover, an adjective "old" has been added with Stephen and Smith, but Smith is not present in this KG. This is a partial deficiency of Links tool in the generation of the KG. Hermeneutic tools "fail in interesting ways" (Rockwell, & Sinclair, 2016, p. 166), and they find new dimensions with an improvement of tools. To solve such deficiencies, the KG should be extended since a big KG encompasses more themes, whereas a small KG misses some important pieces of information.

v. Contexts



	Document	Left	Term	Right
⊞	1) A MIL…	A MILD ATTACK OF	loc…	"Look, look, there they are
⊞	1) A MIL…	streak of rust-coloured air,	loc…	. There they came. At once
⊞	1) A MIL…	quick, quick, there come the	loc…	. Old Smith had had his
⊞	1) A MIL…	calls she stood watching the	loc…	. The air was darkening. A
⊞	1) A MIL…	heaviness of a storm. The	loc…	were coming fast. Now half
⊞	1) A MIL…	earth seemed to be moving,	loc…	crawling everywhere, she could not
⊞	1) A MIL…	three or four years of	loc…	. Locusts were going to be
⊞	1) A MIL…	or four years of locusts.	loc…	were going to be like

Figure 66 Contexts, A Mild Attack of Locusts



	Document	Left	Term	Right
⊞	1) A MIL…	while every farmer hoped the	loc…	would overlook his farm and
⊞	1) A MIL…	on." He picked a stray	loc…	off his shirt and split

Figure 67 Contexts, A Mild Attack of Locusts

Lexical ambiguity between singular and plural usage exists in this short story. In figure 66, the word “locusts” has been found eight times, and in figure 67, a “locust” usage has been shown twice. Locusts refer to a disastrous event of a swarm attack, and all characters prepare themselves to stop this attack, whereas, a locust betokens an event that happens with an individual. Singular and plural words unveil different knowledge patterns and situations.

11. *I Have a Dream* by Martin Luther King

i. Summary

This corpus has 1 document with 771 total words and 255 unique word forms. Created 8 seconds ago (2nd November 2017).

Vocabulary Density: 0.331

Average Words Per Sentence: 18.4

Most frequent words in the corpus: freedom (15); let (14); ring (13); dream (12); day (9); able (8); land (6); mountain (6); faith (5); free (4);

Figure 68 Summary, I Have a Dream

The unique vocabulary of Martin Luther King's speech comprises 255 words, and the total words are 771, which are about three times more than unique words. Vocabulary density is 0.331 which has been calculated using Inverse Absolute (Simpson, 2000) (a division of total words by unique words). The language of Martin Luther King Jr's speech is simple and repetitive to convey the message to common protesters. This repetition serves the purpose of rhetoric. The average words per sentence are 18.4, which betoken the use of long sentences and connected ideas during the speech.

ii. Cirrus

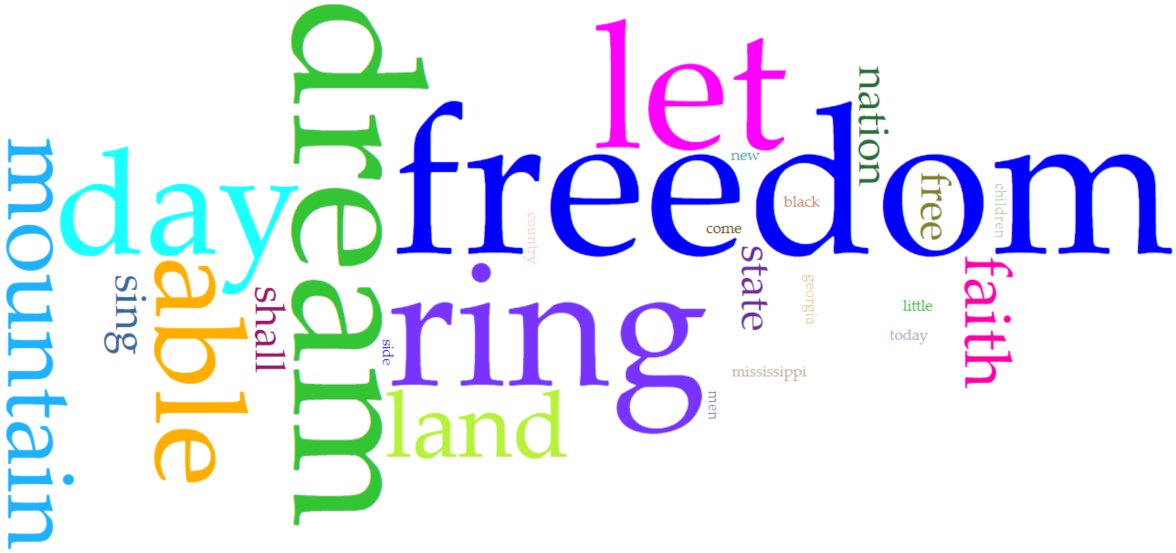


Figure 69 Cirrus, I Have a Dream

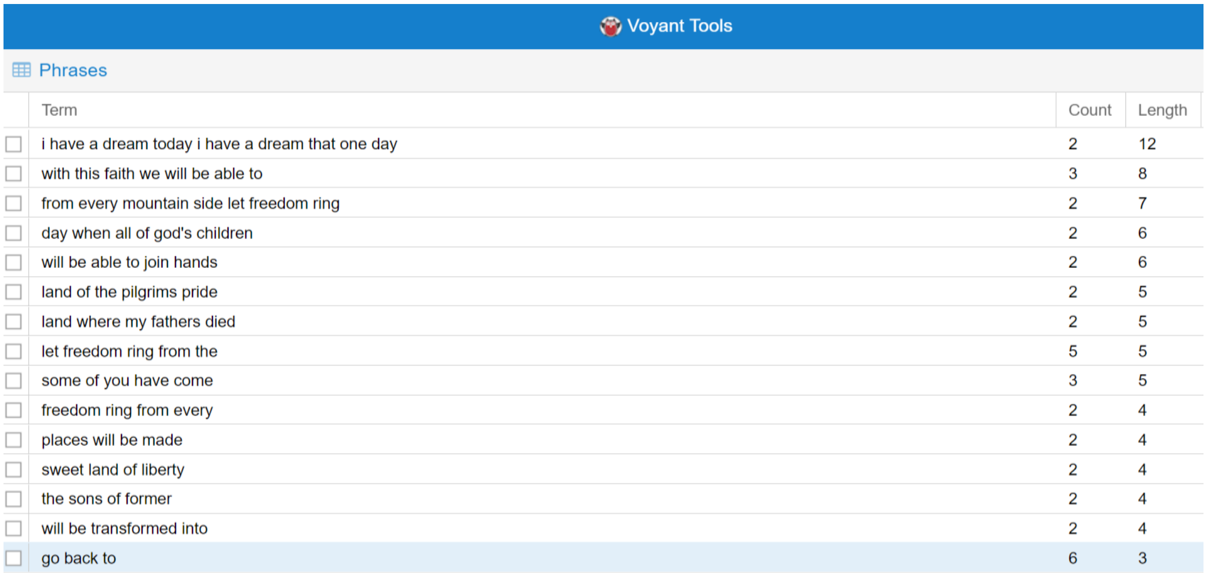
Cirrus is generated by following the topic modelling, and this Cirrus displays the theme of "freedom (15)" of Afro-Americans. Moreover, the word "dream" has been used 12 times, and he emphasizes his vision and prospect of equal rights for black and white children of gods. In addition to it, they would hold hands as a bond of unity and reconciliation. Extending this dream, the themes of "brotherhood (2)", "liberty (2)" and "free (4)" are its ingredients.

No movement can progress and expand without a certain vision and faith, so “faith (5)”, “free (4)” and “state (4)” themes are dominating in this speech. Another motif is “able (8)” and each leader motivates his people by giving them the realization that they are able to revolutionize.

Themes of “let (14)”, “ring (13)” and “sing (4)” denote that the Whites of the USA are hampering ways of social freedom of black people, and now they have stood up for their equal rights; hence, they claim equality without any discrimination of caste, colour and creed. Preference should not be given owing to skin colour, but on the basis of the content of character. They demand to change the yardstick of nobility. Another political study of Obama and McCain’s speeches (GitHub, 2014) also explores themes with Cirrus, and its findings align with the current study.

Human analysis proves it a male-dominated speech (Hussain, 2009, p. 51). Cirrus tool performs better in extracting key themes, for instance, “freedom, free, liberty, brotherhood”. Textual evidence and statistical weight proved that Cirrus analysis is accurate, quantified and authentic, as compared to human analysis.

iii. Phrases



	Term	Count	Length
□	i have a dream today i have a dream that one day	2	12
□	with this faith we will be able to	3	8
□	from every mountain side let freedom ring	2	7
□	day when all of god's children	2	6
□	will be able to join hands	2	6
□	land of the pilgrims pride	2	5
□	land where my fathers died	2	5
□	let freedom ring from the	5	5
□	some of you have come	3	5
□	freedom ring from every	2	4
□	places will be made	2	4
□	sweet land of liberty	2	4
□	the sons of former	2	4
□	will be transformed into	2	4
□	go back to	6	3

Figure 70 Phrases, I Have a Dream

There are some repeated sentences in this speech, for instance, “I have a dream today I have a dream that one day”, “land where my fathers died”, “let freedom ring” and “from every mountainside let freedom ring”, therefore, they have been excluded from the category of phrases. In this speech, standard and repetitive collocation patterns/ n-grams are, “with this faith we will be able to” (Prep+Prn+N+Prn+Aux+V+Prep), “when all god’s children” (Adv+Adj+N), “will be able to join hands" (Aux+V+Inf V+N), "land of the pilgrims' pride" (N+Prep+Art+N+N), "sweet land of liberty" (Adj+N+Prep+N), "the sons of former" (Art+N+Prep+Adj) and "go back to" (V+Adv+Prep). These phrases express Martin Luther's vision that all black and white people are the children of the same god, so there should not be any discrimination based on caste, colour and creed. He motivates all races to collaborate for a noble cause of equal rights and to transform this sweltering world into a "sweet land of liberty". Furthermore, he strives to make it as sacred as the "land of the pilgrims" for all races. These collocations/ n-grams are repeated two to six times, and their length ranges from 3 to 12 words.

iv. Links

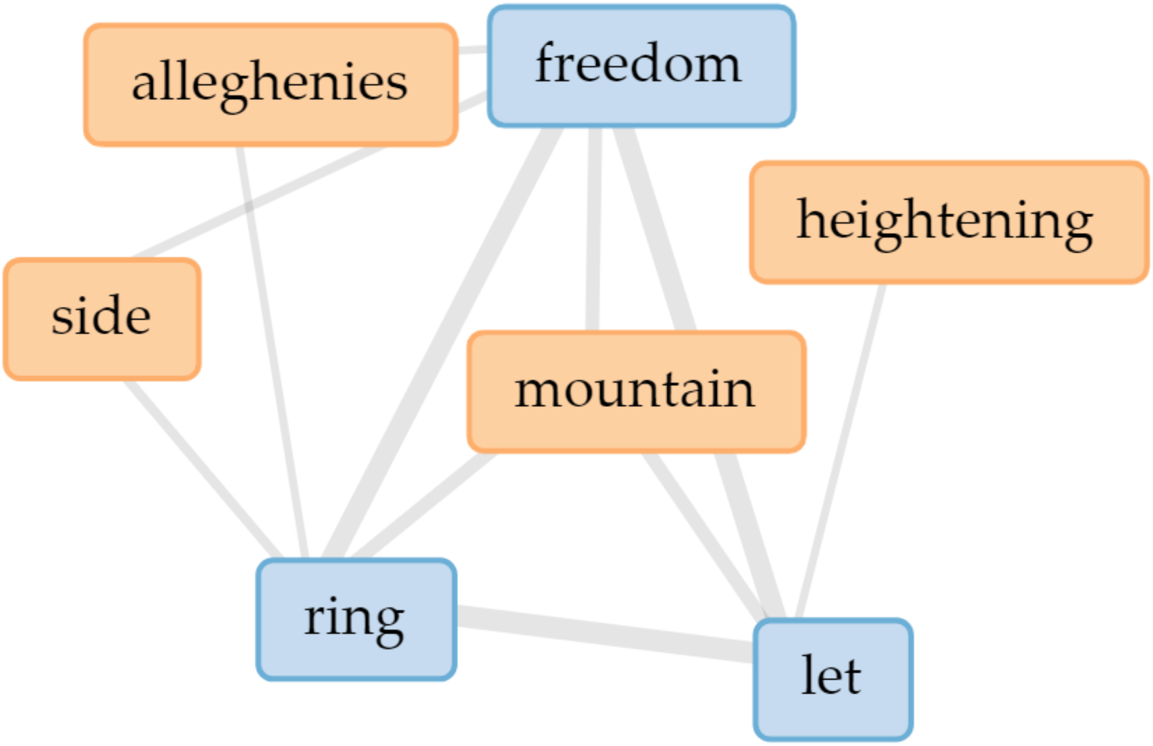
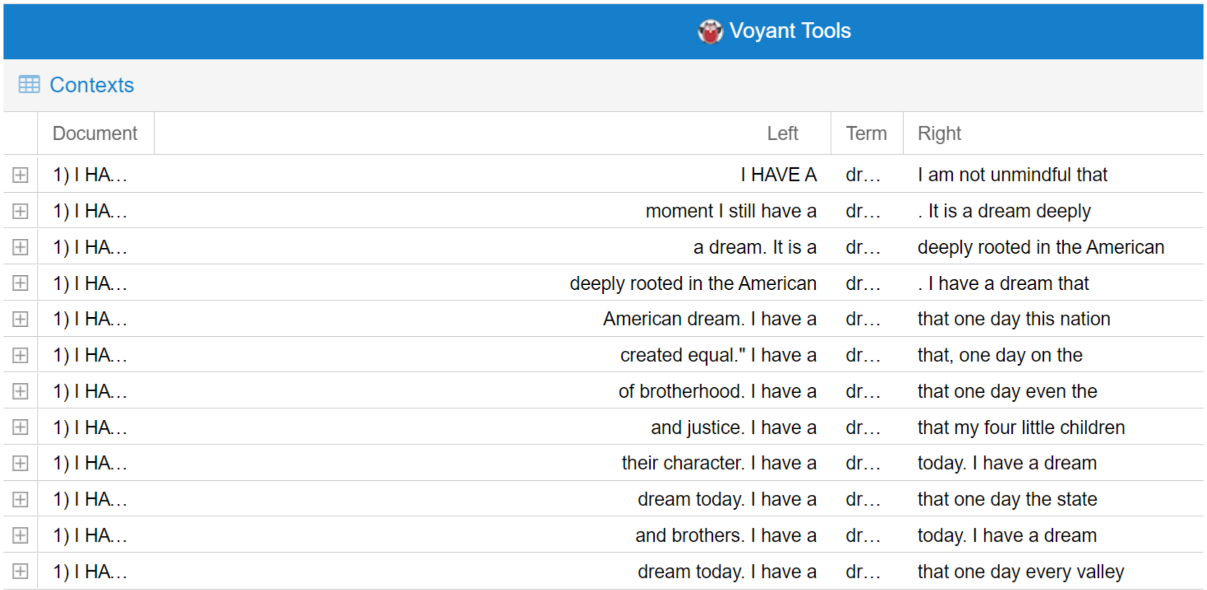


Figure 71 Links, I Have a Dream

Triangular KG of "let, freedom, ring" expresses that the song is based on freedom and equality, and it would be chanted on the mountains of Tennessee. That is why this triangle has been linked with the word "mountain". Then the same triangle has been connected with the word "Alleghenies" because the freedom ring would be audible on Alleghenies of Pennsylvania. "Heightening" is an adjective of the noun "Alleghenies", so they are linked with each other.

v. Contexts



	Document	Left	Term	Right
⊞	1) I HA…	I HAVE A	dr…	I am not unmindful that
⊞	1) I HA…	moment I still have a	dr…	. It is a dream deeply
⊞	1) I HA…	a dream. It is a	dr…	deeply rooted in the American
⊞	1) I HA…	deeply rooted in the American	dr…	. I have a dream that
⊞	1) I HA…	American dream. I have a	dr…	that one day this nation
⊞	1) I HA…	created equal." I have a	dr…	that, one day on the
⊞	1) I HA…	of brotherhood. I have a	dr…	that one day even the
⊞	1) I HA…	and justice. I have a	dr…	that my four little children
⊞	1) I HA…	their character. I have a	dr…	today. I have a dream
⊞	1) I HA…	dream today. I have a	dr…	that one day the state
⊞	1) I HA…	and brothers. I have a	dr…	today. I have a dream
⊞	1) I HA…	dream today. I have a	dr…	that one day every valley

Figure 72 Contexts, I Have a Dream

Semantic disambiguity prevails in this speech whether the word "dream" refers to denotative or connotative semantic shades, or it is a dream during sleep or a vision for a great cause. He utters the phrase, "I have a dream" 12 times during his speech, so it is significant to know about his dreams for an Afro-American audience. The underpinning of Hermeneutica Theory lays stress on the search of context, as it "is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). To solve this WSD, Contexts tool produces complete contextual data regarding the key word of "dream" in Martin Luther King Jr's speech. This visionary dream covers dogmas of brotherhood, rising up of the nation, social justice, equality, the exaltation of valleys or downtrodden people. Contexts tool also reveals the knowledge pattern that these dreams have connotative semantic shades. Another knowledge discovery is that these dreams are supported by a revolutionary movement with numerous sacrifices.

12. *The Gift of Magi* by O. Henry

i. Summary

This corpus has 1 document with 1,544 total words and 474 unique word forms. Created 25 seconds ago (on 2nd November 2017).

Vocabulary Density: 0.307

Average Words Per Sentence: 10.0

Most frequent words in the corpus: jim (26); hair (18); della (15); said (13); looked (11); watch (11); like (9); look (9); wise (9); gifts (8);

Figure 73 Summary, The Gift of Magi

This corpus utilizes 474 unique words and 1544 total words; therefore, more than three times unique words have been reproduced in the text, and this repetitive vocabulary facilitates intermediate level readers. O. Henry uses 0.307 vocabulary density, and each sentence consists of 10 words which denote small and easy sentences.

ii. Cirrus

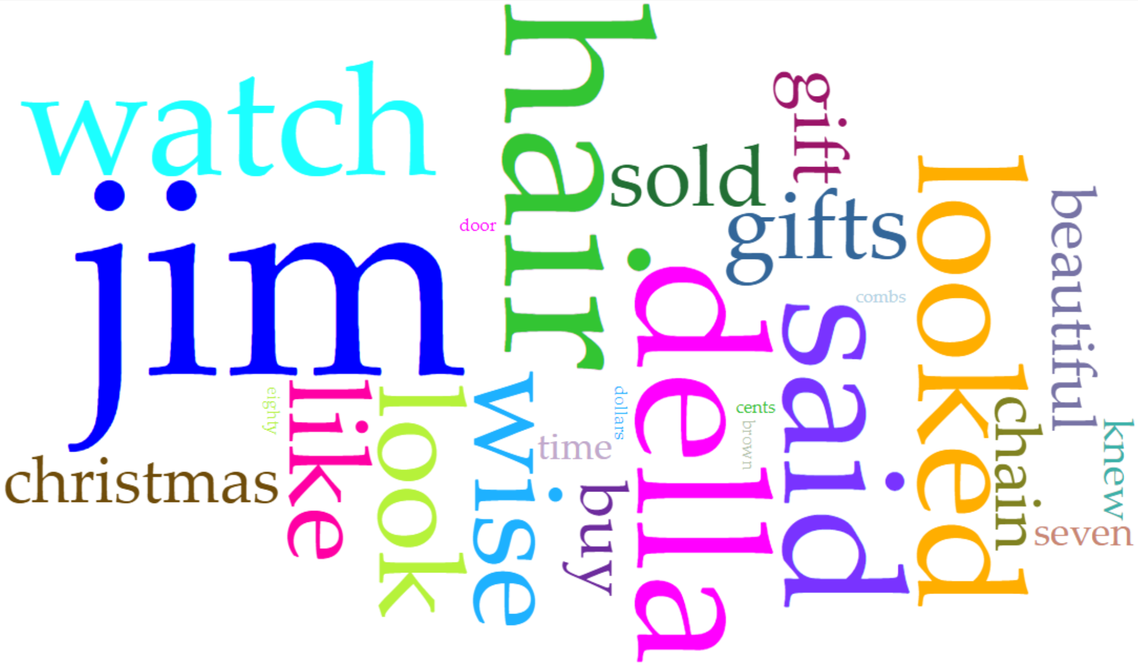


Figure 74 Cirrus, The Gift of Magi

Figure 73 Cirrus shows two major characters, “Jim (26)” and “Della (15)”. Then it highlights the background of “Christmas (6)” and an old tradition of exchanging “gifts (8)” on Christmas which compel both characters to present the best gifts to each other. Both of them are short of money, but they own two valuable things: Della’s long beautiful “hair (18)” and Jim’s gold “watch (11)”. The word “sold (7)” refers to selling their precious things to buy a valuable gift for each other. Therefore, Della buys “chain (6)” for Jim’s watch by selling her long beautiful hair, and Jim buys combs for Della’s long beautiful hair by selling his gold watch. So, two valuable

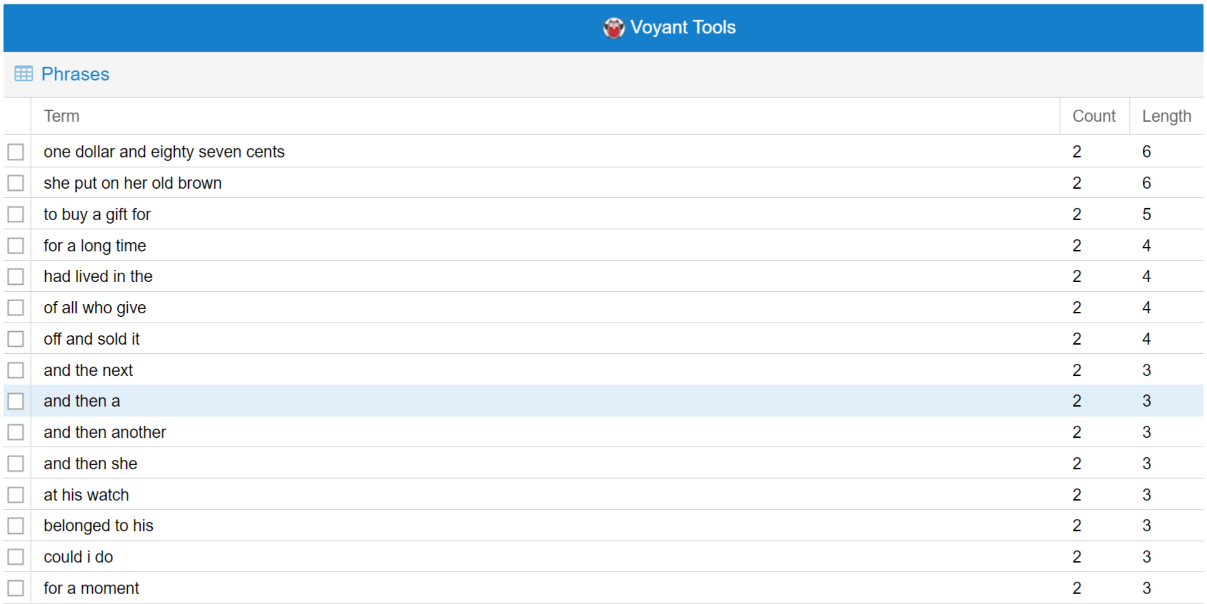
gifts have been bought, but both gifts are unusable for both lovers, because a chain is ineffectual without the watch, and a hair comb is inoperable without long hair.

The theme of “wise (9)” expresses Jim and Della’s selling acts and an extreme feeling of the sacrifice of their dearest things for their beloved ones. Apparently, those gifts are unserviceable, but they are replete with feelings of love for each other as Magi brought unusable gifts of gold, myrrh and frankincense for child Jesus. Therefore, Jim and Della have been equated with Magi regarding their feelings, acts of love and sacrifice.

The word “said (13)” indicates that several dialogues have also been uttered in this short story, and it uses direct conversations between characters. Besides, the narrative technique has also been employed. The theme of “looked (11)” shows two semantic shades: one is their act of looking here and there, and the second is about characters' appearance, especially how Della’s appearance has been elaborated. Similarly, the previous study of Jane Austen’s work (Sinclair, & Rockwell, 2015b) finds key ideas through Cirrus, and the current study does the same.

Human analysis shows two characters, namely husband and wife, and key motifs of the short story are love, sacrifice, husband’s role of breadwinning, pretty, beautiful hair, gold watch (Hussain, 2009, pp. 51-52). Comparing machine and human analysis, the extracted characters are the same in human and machine analysis, while some themes of“watch, chain” are the same, but other themes are different.

iii. Phrases



	Term	Count	Length
□	one dollar and eighty seven cents	2	6
□	she put on her old brown	2	6
□	to buy a gift for	2	5
□	for a long time	2	4
□	had lived in the	2	4
□	of all who give	2	4
□	off and sold it	2	4
□	and the next	2	3
□	and then a	2	3
□	and then another	2	3
□	and then she	2	3
□	at his watch	2	3
□	belonged to his	2	3
□	could i do	2	3
□	for a moment	2	3

Figure 75 Phrases, The Gift of Magi

Standard collocation/ n-gram “she put on” (Prn+V+Prep) expresses the correct use of preposition for wearing something. The phrase “for a long time” (Prep+Art+Adj+N) refers to a long span. These phrases also teach the correct use of prepositions, such as “belonged to” (V+Prep). Another phrase, “for a moment” (Prep+Art+N) denotes a very short time. The phrase “have to look” (Aux+V) implies compulsion. Another phrase, “her hair was” (Prn+N+Aux), teaches the usage of a singular auxiliary verb with the word “hair”. All phrases have been found twice, and they consist of 3 to 6 words.

iv. Links

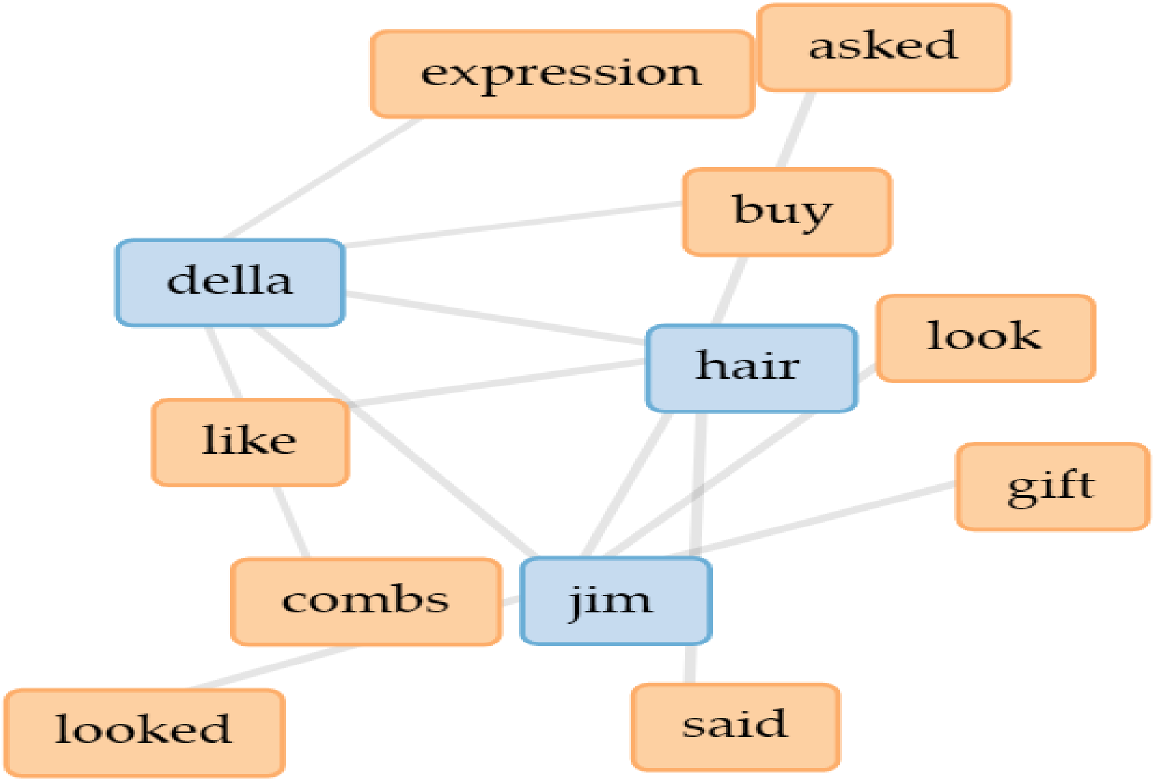


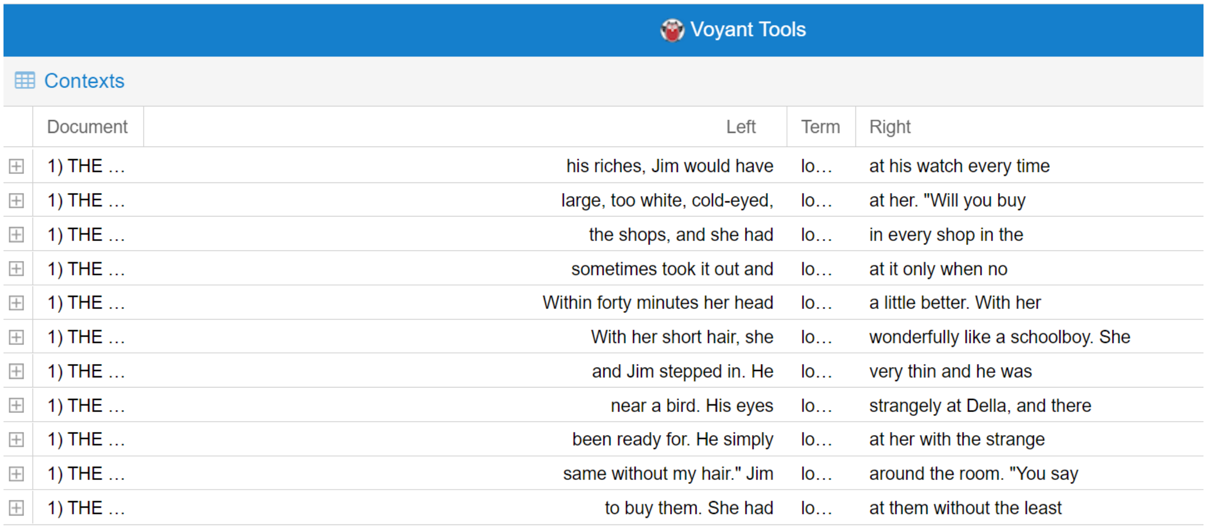
Figure 76 Links, The Gift of Magi

“In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct” (Agrawal, & Psaila, 1995, p. 1). As a human brain understands different knowledge dimensions with the bonding of neurons, similarly, a KG builds multiple links for decoding information from unstructured text. The whole story revolves around the KG, “Della, Jim, gift”. The KG of “Della, like, combs” shows that the gift of comb has been bought to beautify Della’s hair. Another KG of “combs, Jim, hair, Della” refers to Jim’s shopping of combs for Della’s hair as a Christmas gift.

The word “like” has been used nine times in this corpus, and from them, eight times it shows simile, whereas one time it refers to the meaning to “adore” something. Both Jim and Della mostly use the word “like” to show the similarity of things and characters. that is why, the KG of “Jim, like, Della” shows that both Jim and Della passionately love each other, and they sacrifice their most precious things for each other’s pleasures.

Another KG of “looked, Jim, look,” refers to the presence of the word “look” and “looked” in this corpus. So, this frequent occurrence builds the KG from these three words “Jim, look, looked”. One deficiency of Phrases tool is that Jim also buys a gift for Della, but the word “buy” has not been connected with Jim, while it has been connected with only Della. Instead of the word “buy”, the word “gift” has been linked with Jim. Hermeneutica Theory explains that tools “fail in interesting ways” (Rockwell, & Sinclair, 2016, p. 166) and this failure leads to new instrumental developments. Again, the solution to this ambiguity is the extension of the KG, that is why these tools are interactive.

v. Contexts



	Document	Left	Term	Right
⊞	1) THE …	his riches, Jim would have	lo…	at his watch every time
⊞	1) THE …	large, too white, cold-eyed,	lo…	at her. "Will you buy
⊞	1) THE …	the shops, and she had	lo…	in every shop in the
⊞	1) THE …	sometimes took it out and	lo…	at it only when no
⊞	1) THE …	Within forty minutes her head	lo…	a little better. With her
⊞	1) THE …	With her short hair, she	lo…	wonderfully like a schoolboy. She
⊞	1) THE …	and Jim stepped in. He	lo…	very thin and he was
⊞	1) THE …	near a bird. His eyes	lo…	strangely at Della, and there
⊞	1) THE …	been ready for. He simply	lo…	at her with the strange
⊞	1) THE …	same without my hair." Jim	lo…	around the room. "You say
⊞	1) THE …	to buy them. She had	lo…	at them without the least

Figure 77 Contexts, The Gift of Magi

The word “looked” has two meanings: as an action verb, it means to see; and as a linking verb, it means to appear or seem. To differentiate between them, they are analysed with Contexts tool, and figure 77 shows that the word “looked” gives the meaning of “appeared” only two times (in 5th and 7th sentences), for instance, “her head looked a little better” and “he looked very thin”, while it has been used nine times in the sense of seeing.

Hermeneutica Theory directs that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). This study shows that the Contexts tool shows the interactive bidirectional context of any searched word. Then human reflection and cognition play critical roles to comprehend semantic shade, position in the sentence and part of speech of the problematized word. In short, text mining tools just show knowledge patterns, and human beings interpret them according to their reflective capabilities and cognition.

13. *God Be Pleased* by Ahmed Nadeem Qasmi

i. Summary

This corpus has 1 document with 3,181 total words and 1,017 unique word forms. Created 24 seconds ago (on 3rd November 2017).

Vocabulary Density: 0.320

Average Words Per Sentence: 12.1

Most frequent words in the corpus: maulvi (48); abul (38); shamim (21); said (15); ahmed (14); went (12); mehrun(11); villa ge (11); chaudhry (10); house (10);

Figure 78 Summary, God Be Pleased

This story comprises 1017 unique words, and they have been repeated almost three times, so, there are 3181 total words. Thus, the density of vocabulary is 0.320. Ahmed Nadeem Qasmi/ translator writes an average of 12.1 words in a sentence. Thus, sentences of this short story are smaller than the previous story which has 14 words in a sentence. Its most frequent words mention the names of notable characters for example, “Maulvi (48)”, “Shamim (21)” “Ahmed (14)”, “Mehrun (11)”, “Chaudhry (10)”, “sir (9)”, and some other important themes of the shorts tory are “shop (9)”, “cloth (8)”, and “rupees (9)”.

ii. Cirrus

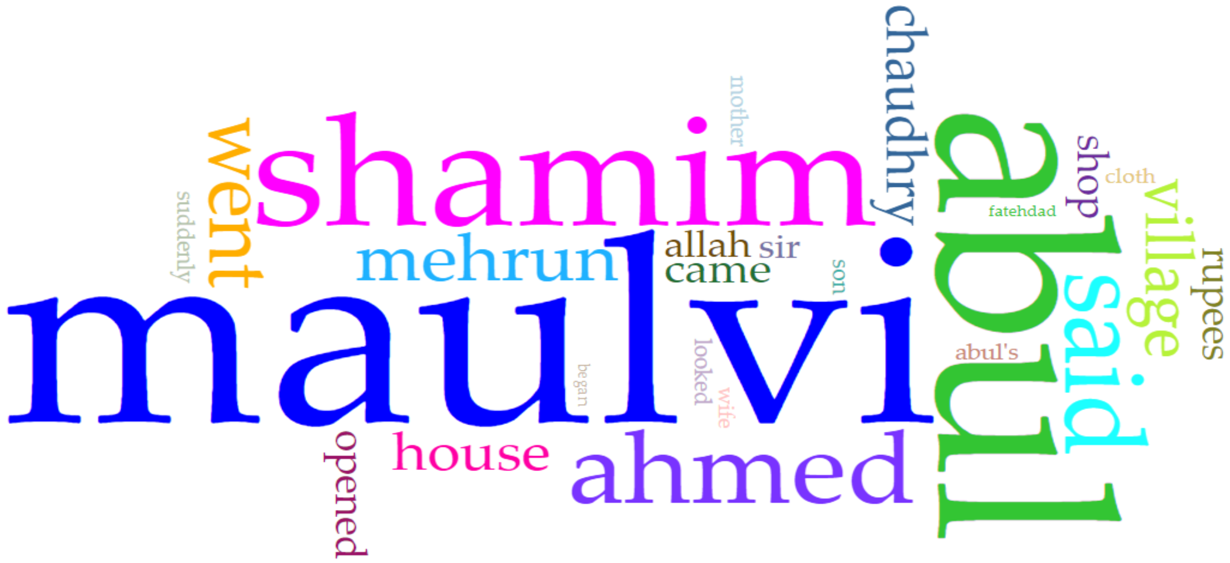


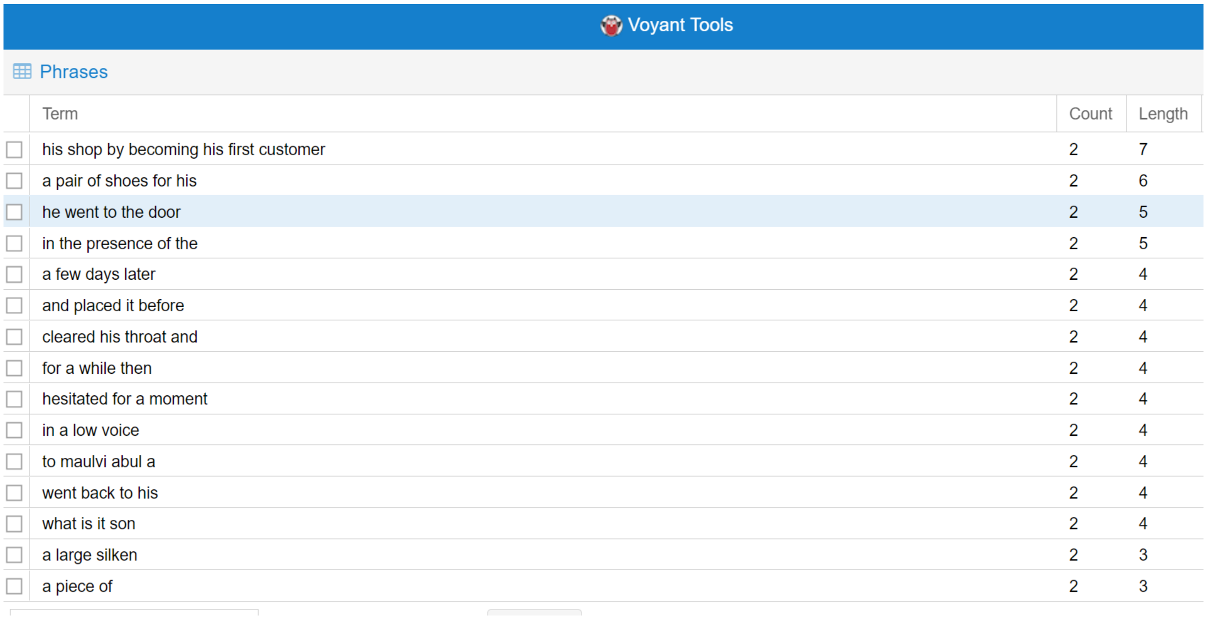
Figure 79 Cirrus, God Be Pleased

This Cirrus has extracted knowledge patterns which highlight major characters, for instance, “Maulvi (48) Abul (38)”, “Shamim (21) Ahmed (14)”, “Mehrun (11)”, “Chaudhry (10)”, “Allah (9)” and “mother (8)”. The theme of “Allah (9)” is prevalent in this short story because Maulvi Abul is a religious person who calls for five times prayer in the mosque and leads prayers to villagers, so he mostly utters this word to express his gratitude for Allah. One address term, “sir (9)” has been used nine times: Shamim Ahmed uses this term 7 times for Maulvi Abul, and it reveals knowledge that he is very respectful to his religious teacher and his would-be father-in-law. Chaudhry Sahib uses the term “sir” 2 times to address Maulvi Abul.

Another theme of “village (11)” shows the rustic setting of the story. Besides, themes of “cloth (8)” and “shop (9)” reveal that Shamim Ahmed opens a cloth shop which is inaugurated by Maulvi Abul. One more theme of “said (15)” indicates that direct dialogues have been included in this short story. Another motif of “rupees (9)” has been used when Maulvi Abul takes his total savings of 43 rupees to buy Mehrun’s cloth piece whose price is 42 rupees. Every time, Maulvi feels the dearth of rupees and weeps bitterly when he sees the naked feet of his youngest daughter. Contrary to these pathetic situations, he was well off before his marriage, and he used to distribute Eid money among destitute people, but after his marriage and continuous birth of children, he is in dire need of money to buy daily commodities for his children.

Human analysis shows that female characters are “helpless”, “female characters are dependent on male characters,” and the prime duty of females is “home keeping”. The main theme of this story is “marriage” (Hussain, 2009, p. 53). Cirrus mentions the names of all characters, while human analysis mentions them with a generalized term of female characters. Cirrus tool highlights different themes, for instance, cloth, shop, said, village, while human analysis ignores all these themes.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	his shop by becoming his first customer	2	7
□	a pair of shoes for his	2	6
□	he went to the door	2	5
□	in the presence of the	2	5
□	a few days later	2	4
□	and placed it before	2	4
□	cleared his throat and	2	4
□	for a while then	2	4
□	hesitated for a moment	2	4
□	in a low voice	2	4
□	to maulvi abul a	2	4
□	went back to his	2	4
□	what is it son	2	4
□	a large silken	2	3
□	a piece of	2	3

Figure 80 Phrases, God Be Pleased

Standard collocation patterns/ n-grams in this short story are: “becoming the first customer” (V+Art+Adj+N), “a pair of shoes” (Art+N+Prep+N), “in the presence” (Prep+Art+N), “a few days later” (Adj+N+Adv), “placed it before” (V+Prn+Adv), “cleared his throat” (V+Prn+N), “for a while” (Prep+Art+N), “hesitated for a moment” (V+Prep+Art+N), “in a low voice” (Prep+Art+Adj+N), “a piece of” (Art+N+Prep) and “what is it, son” (Int Prn+Aux+Prn, +N). These collocation patterns/ n-grams accelerate the learning and comprehension level of the text. The extracted phrases have been repeated twice, and their length ranges from 3 to 7 words.

iv. Links

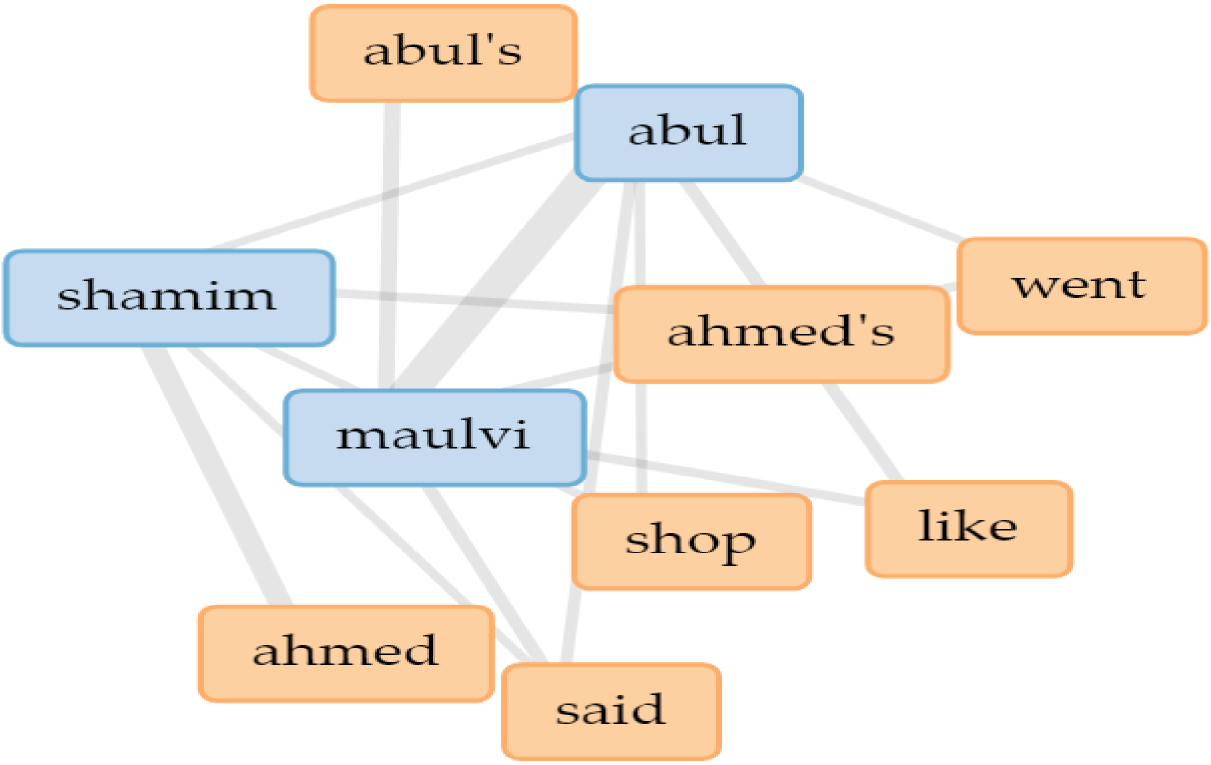
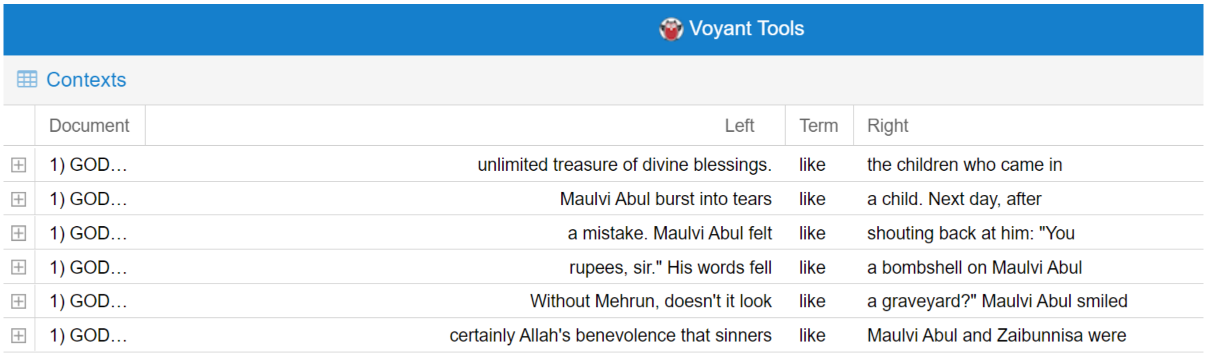


Figure 81 Links, God Be Pleased

Blue coloured KG, “abul, Maulvi, Shamim” , “ Shamim Ahmed” and “Maulvi Abul” show that two characters have discussed repeatedly because both characters interact with each other on several occasions, for instance, at the time of the inaugural session of the shop, and at the time of seeking Mehrun’s match from Maulvi Abul. One more triangular KG of “Shamim, Ahmad’s, shop” refers to his cloth shop business in the village. Another KG of “Maulvi, Abul, like” suggests the use of the simile word “like” 6 times for himself; for instance, he bursts into tears like a child. One more triangular KG of “Maulvi, Abul, went” betokens that the word “went” has been used 12 times; and it has been used nine times for Maulvi Abul, so it is quite appropriate to link this word with Maulvi Abul. Thus, KGs play a pivotal role in text connectivism during the text mining process.

v. Contexts



	Document	Left	Term	Right
⊞	1) GOD…	unlimited treasure of divine blessings.	like	the children who came in
⊞	1) GOD…	Maulvi Abul burst into tears	like	a child. Next day, after
⊞	1) GOD…	a mistake. Maulvi Abul felt	like	shouting back at him: "You
⊞	1) GOD…	rupees, sir." His words fell	like	a bombshell on Maulvi Abul
⊞	1) GOD…	Without Mehrun, doesn't it look	like	a graveyard?" Maulvi Abul smiled
⊞	1) GOD…	certainly Allah's benevolence that sinners	like	Maulvi Abul and Zaibunnisa were

Figure 82 Contexts, God Be Pleased

In figure 82, the word “like” has been used six times. It is ambiguous whether they are similes or main verbs. One postulate of “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). So, to disambiguate word sense, Contexts tool has been used; hence, it reveals a knowledge pattern that the word “like” has been used six times as similes, and it has never been used for the meaning of “adore”.

14. *Overcoat* by Ghulam Abbas

i. Summary

This corpus has 1 document with 1,849 total words and 679 unique word forms. Created 18 seconds ago (on 15th December 2017).

Vocabulary Density: 0.367

Average Words Per Sentence: 14.0

Most frequent words in the corpus: young (19); man (16); overcoat (9); sir (8); evening (7); mall (6); people (6); quite (6); sa id (6); wearing (6);

Figure 83 Summary, Overcoat

“Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). The unique words of this short story are 679, and they have been repeated almost three times; hence, the total words of this short story reach the limit of 1849 words. Accordingly, vocabulary density is calculated at 0.367. Ghulam Abbas/translator usually writes 14 words in a sentence, and it suggests that sentences are moderate in length and comprehensible for intermediate-level learners. The most frequent words are about the main character of the short story, “young (19)” “man (16)”. Moreover, the title of the short story “overcoat (9)” has also been mentioned 9 times in this short story.

ii. Cirrus

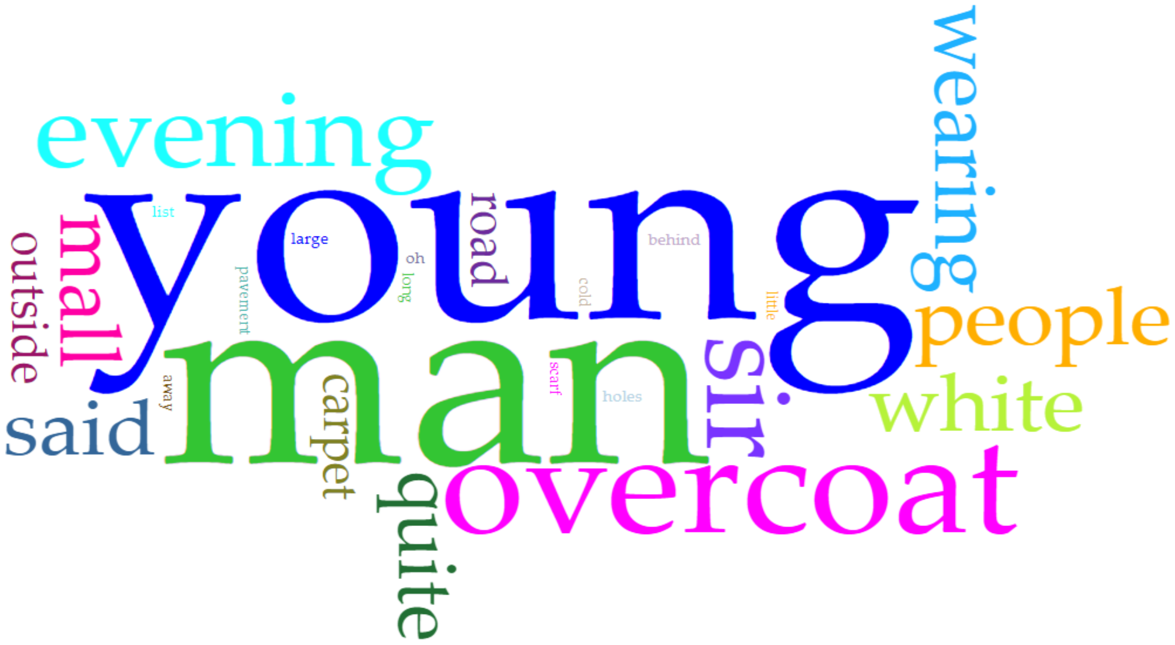
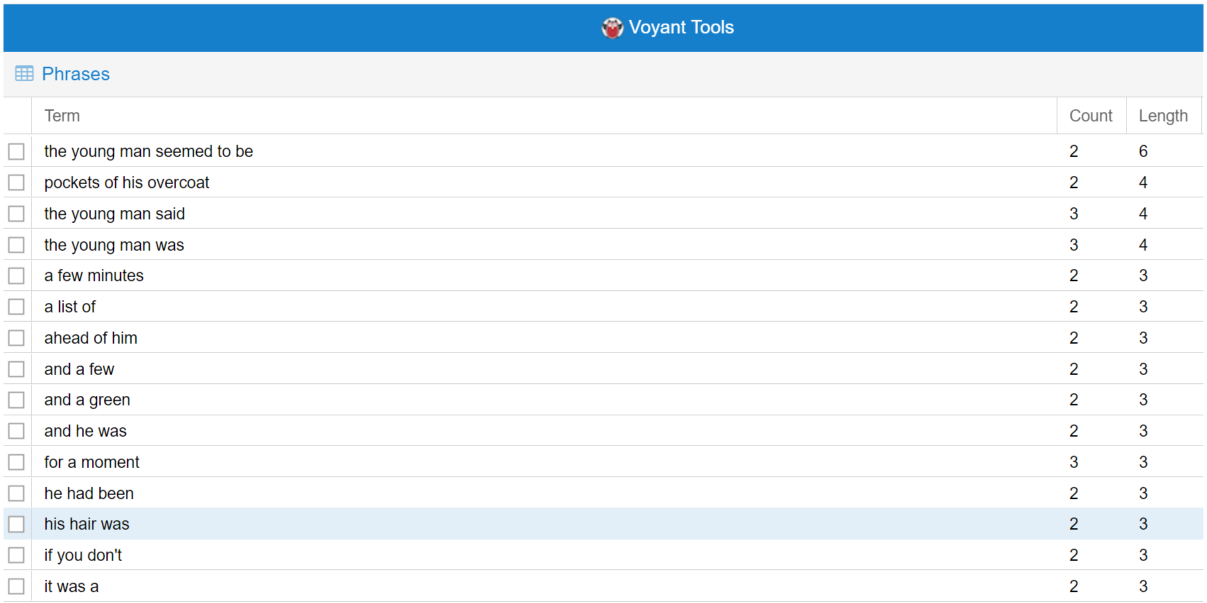


Figure 84 Cirrus, Overcoat

The title of the short story *‘Overcoat’* has been mentioned nine times, and the main character “young (19)” “man (16)” is the most occurring person since the whole story revolves around him. This Cirrus also reveals that the young man is called “sir (8)” by people because his appearance pretends to be a member of the posh class. The young man enjoys the status of the elite class by dint of his “wearing (6)” “white (6)” dress, chic appearance and show off, while several “holes (4)” are present in his vest when his dress is removed after the road accident. In reality, he was in a poverty-stricken condition. This Cirrus also shows the location of the young man’s “evening (7)” stroll on the “mall (6)” road, Lahore. Another theme of “said (6)” indicates the presence of dialogues in this short story. Furthermore, the young man also questions shopkeepers about the “list (4)” of the gramophone record, musical instruments and “carpet (5)” rates, while he is penniless. Apart from these conversations, the nursing staff also converses with each other and comments on the young man's dead body.

Human analysis reveals key characters of a young man, a woman, two nurses, Dr Khan and some unnamed characters. The main themes of this short story are “outing, recreations, tall, short, bulky” (Hussain, 2009, pp. 53-55). Cirrus and human analysis find similar characters, but their identified themes are entirely different.

iii. Phrases



	Term	Count	Length
□	the young man seemed to be	2	6
□	pockets of his overcoat	2	4
□	the young man said	3	4
□	the young man was	3	4
□	a few minutes	2	3
□	a list of	2	3
□	ahead of him	2	3
□	and a few	2	3
□	and a green	2	3
□	and he was	2	3
□	for a moment	3	3
□	he had been	2	3
□	his hair was	2	3
□	if you don't	2	3
□	it was a	2	3

Figure 85 Phrases, Overcoat

Standard phraseology or collocation patterns/ n-grams are: “by now” (Prep+Adv), “for a moment” (Prep+Art+N), “one of the” (N+Prep+Art), “those who” (Prn+Rel Prn), “a few minutes” (Adj+N), “a list of” (Art+N+Prep), “a short of” (Art+N+Prep) and “ahead of him” (Adv+Prep+Prn). The rationale for excluding other phrases from standard phraseology is that they are insignificant and do not come together every time, for instance, “and the”, “at the”. Standard phrases occur 2 to 6 times, and their length covers 2 to 4 words.

iv. Links

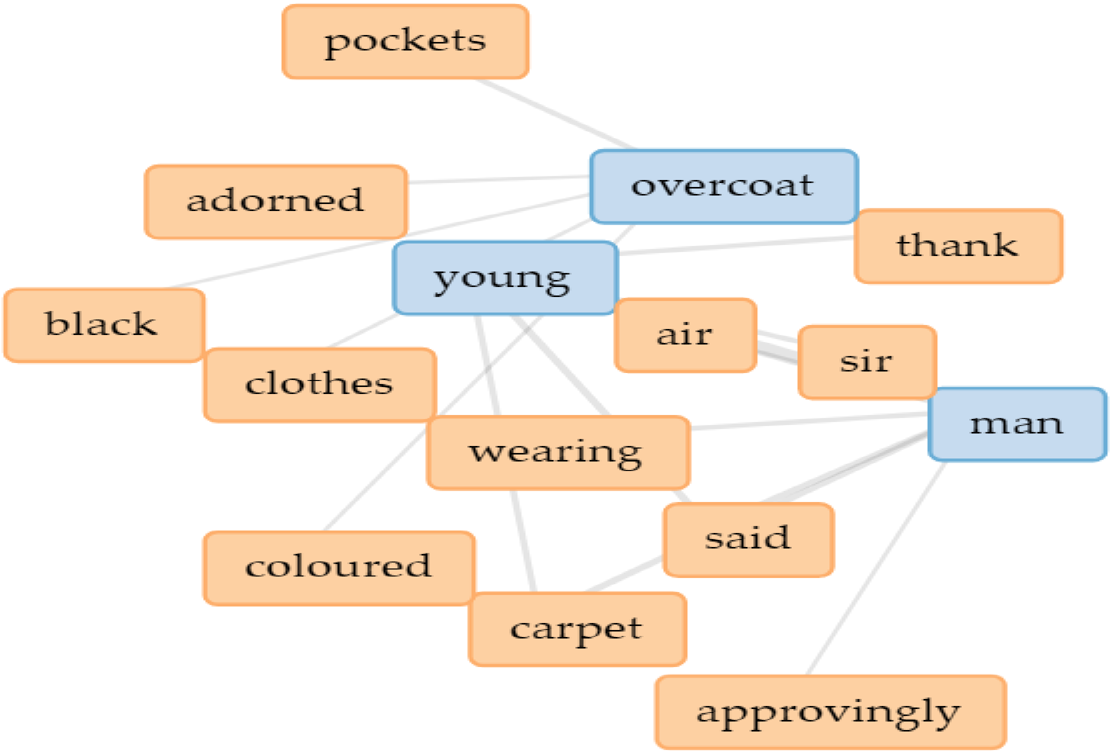
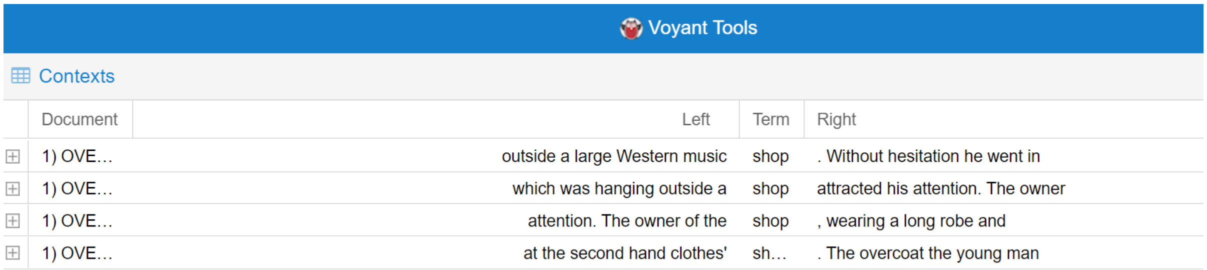


Figure 86 Links, Overcoat

KGs build connections among different entities as a human brain and cognition construct meanings with the wiring of different neural nodes with one another. The title of "overcoat" has a central place in this KG, and the relationship of "overcoat, pockets, young, man" indicates that the young man wears an overcoat during his night stroll, and his hands are in his overcoat pockets. Another KG of "young, man, said, carpet" refers to the conversation between the young man and carpet seller in an approving tone. The KG of "black, overcoat, pockets" refers to two black things of the young man: "black corduroy trousers" and "black comb" in the pocket of his overcoat.

Another KG of "sir, young, man" denotes respectable and fashionable address term "sir" for the ostentatious young man. One more KG of "young, man, thank" guides on habitual uttering of the word "thank" you to all offers, for instance, tanga (horse-driven cart) wala's offer to give him a ride. The KG of "overcoat, young, clothes" reveals knowledge patterns and situations when young man's clothes have been removed after his death. They find holed vest, dirty body, impoverished things under his showy dress and overcoat. Thus, his fake and penurious personality has been completely exposed after this accident. To conclude, "Manipulation is in service of exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166). Here different knowledge graph nodes have been manipulated to explore different hermeneutic patterns which validate textual evidence. Thus, concurrent validity is established.

v. Contexts


Voyant Tools

Contexts

	Document	Left	Term	Right
⊞	1) OVE…	outside a large Western music	shop	. Without hesitation he went in
⊞	1) OVE…	which was hanging outside a	shop	attracted his attention. The owner
⊞	1) OVE…	attention. The owner of the	shop	, wearing a long robe and
⊞	1) OVE…	at the second hand clothes'	sh…	. The overcoat the young man

Figure 87 Contexts, Overcoat

There is semantic and grammatical ambiguity that the word "shop" is a noun or a verb, and whether it refers to a shopping place or an act of buying something. Contexts tool disambiguates word sense, and reveals that in first three sentences, it is used as a singular noun referring to a place of shopping; and fourth time, it is used as a plural noun; and it has not been used as a verb. During word sense disambiguation, tool and human cognition work in collaboration to discover knowledge patterns.

15. *The Angel and the Author – And Others* by Jerome K. Jerome

i. Summary

This corpus has 1 document with 952 total words and 390 unique word forms. Created 9 seconds ago (10th December 2017).

Vocabulary Density: 0.410

Average Words Per Sentence: 14.6

Most frequent words in the corpus: good (9); christmas (6); deeds (5); said (5); angel (4); charity (4); morning(4); time (4); y es (4); agreed (3);

Figure 88 Summary, The Angel and the Author – And Others

Jerome K. Jerome knits the entire story with 390 unique words which have been repeated almost more than two times until total words are counted as 952. Consequently, the density of vocabulary is 0.410. Jerome K. Jerome writes an average of 14.6 words in a sentence, and it is suitable for the intermediate level readers. Most frequent words with their statistical weight show qualitative and quantitative data. To conclude, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). Here all stylistic features have been quantified to present a compact style of the short story writer.

ii. Cirrus

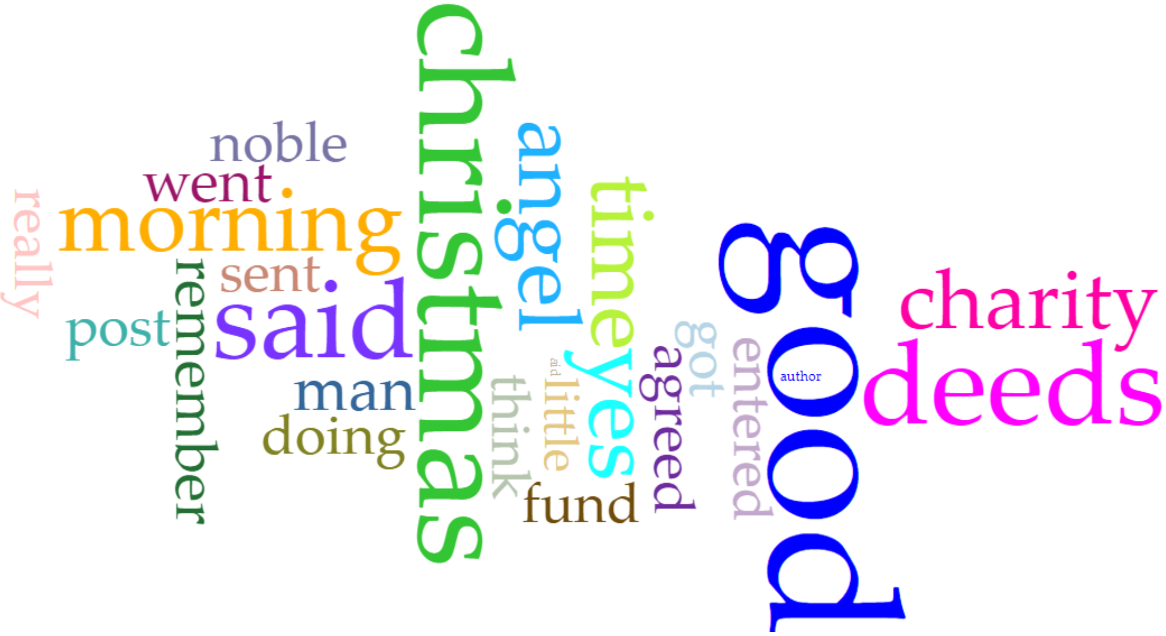
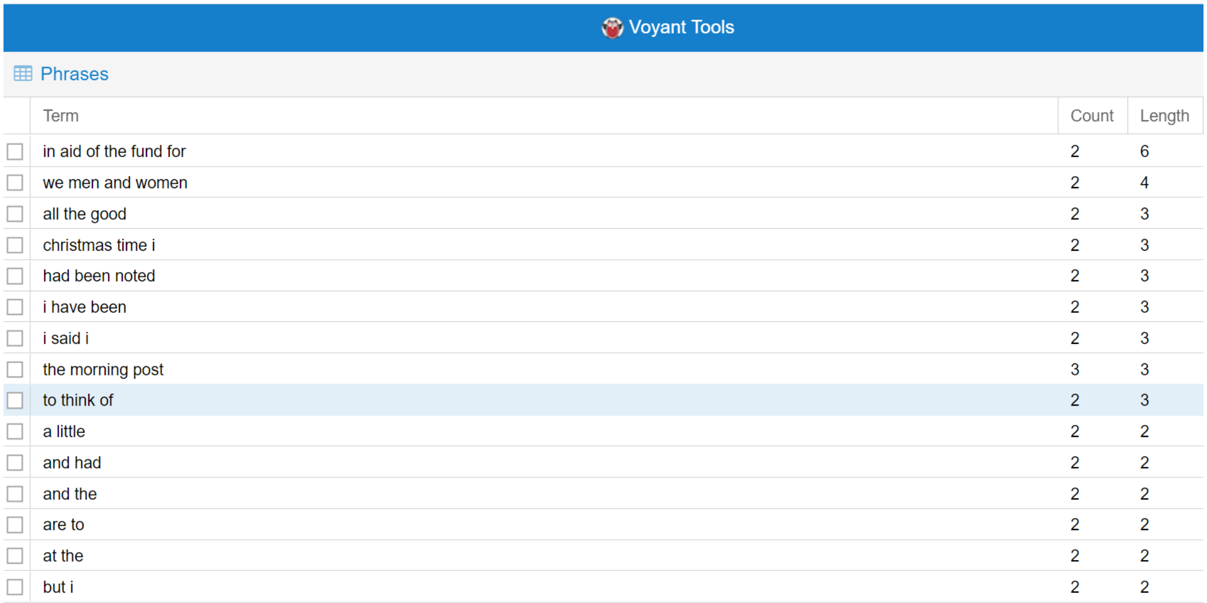


Figure 89 Cirrus, The Angel and the Author – And Others

This Cirrus extracts several thematic patterns from this short story which has been written with “Christmas (6)” background, and all Christians become generous in performing “good (9)” “deeds (5)”. The theme of “remember (3)” is also found in the cluster because the author has thought his list of good deeds repeatedly. The word “said (5)” refers to the plot development through dialogues. The actor verifies his good deeds and “charity (4)” acts whether the Recording angel has “entered (3)” them or not. It exhibits the influence of computers and the internet on language (Crystal, 2001). The angel’s data recording act is just like the data entry of a data entry operator. The theme of “agreed (3)” denotes the agreement of angel with the author’s “noble (3)” acts, for instance, payment of 10 shillings to Daily Graph; four charity dinners; a dozen signed pictures; the performance of Talbot Champneys, four balcony seats for a monster show; sending old unwearable clothes and one wearable coat for the rummage sale; and one raffle for a car. The author seeks confirmation of his good deeds from the written record of the Recording Angel. This is an implied sarcasm over ostentation in the garb of virtues. To conclude, One postulate of the theory states that “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166). This study is delimited to only 15 themes; otherwise, Cirrus tool is capable of exploring maximum of 500 themes along with their statistical weight.

Human analysis shows that the whole story concentrates on the author and his discussion with a recording angel (Hussain, 2009, p. 55). Cirrus tool also finds the same characters, but it unveils more themes than human analysis: Christmas, good deed, remember and noble.

iii. Phrases



	Term	Count	Length
□	in aid of the fund for	2	6
□	we men and women	2	4
□	all the good	2	3
□	christmas time i	2	3
□	had been noted	2	3
□	i have been	2	3
□	i said i	2	3
□	the morning post	3	3
□	to think of	2	3
□	a little	2	2
□	and had	2	2
□	and the	2	2
□	are to	2	2
□	at the	2	2
□	but i	2	2

Figure 90 Phrases, The Angel and the Author – And Others

Among these phrases, the following standard phrases are “noble deeds” (Adj+N), “the morning post” (Art+Adj+N), “a little” (Art+Adj), “Christmas time” (Adj+N). These collocations enhance fluency in speaking and writing, and they are equally useful for constructing new collocation patterns/ n-grams. They occur 2 to 7 times, and their length consists of 2 to 3 words.

iv. Links

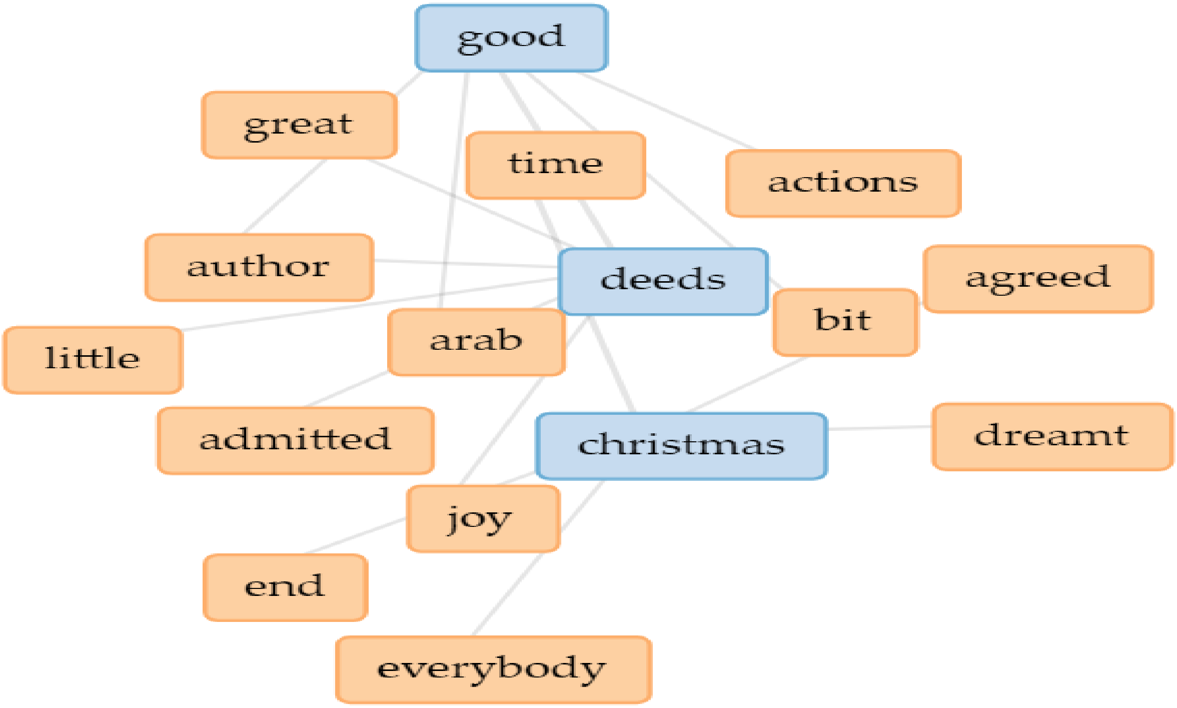
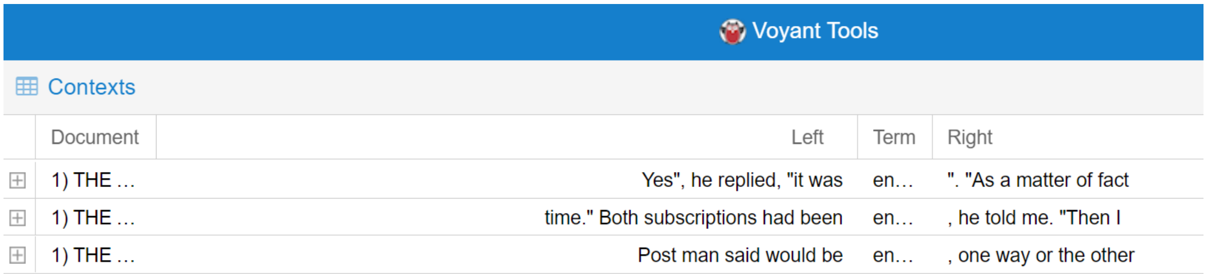


Figure 91 Links, The Angel and the Author – And Others

The KG of "everybody, Christmas, deeds, good, actions" reveals the knowledge pattern that Christmas makes people spiritual, and persuades them to perform holy deeds. Another KG of "good, deeds, Arab" shows that charity works are done for the poor Arab person. Moreover, the KG of "deeds, Christmas, time, dreamt" exposes a knowledge pattern that the author envisions a dream, and talks with Recording Angel. Moreover, the KG of "admitted, Arab, deeds" denotes that the Recording Angel accepts, records, and tallies all charity claims of the author's good deeds. One more KG of "author, good, deeds, joy" indicates that all Christians derive bliss and divine joy by performing spiritual deeds during Christmas time.

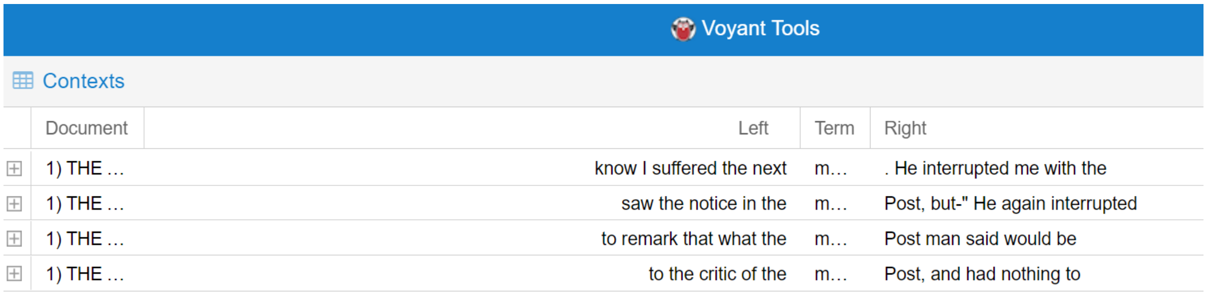
v. Contexts



	Document	Left	Term	Right
⊞	1) THE …	Yes", he replied, "it was	en…	". "As a matter of fact
⊞	1) THE …	time." Both subscriptions had been	en…	, he told me. "Then I
⊞	1) THE …	Post man said would be	en…	, one way or the other

Figure 92 Contexts, The Angel and the Author – And Others

The word “entered” has two main semantic ambiguities, for instance, the phrase, entered the room, means for stepping into the room, and secondly, the word “entered” means for inputting data into a computer repository. The word “entered” has been found three times in the text, so the information retrieval with Contexts tool elucidates the true semantic shade. Three sentences with “entered” refer to the data entry process of good deeds on and after Christmas. These findings are in line with Kwary’s (2018) findings.



	Document	Left	Term	Right
⊞	1) THE …	know I suffered the next	m…	. He interrupted me with the
⊞	1) THE …	saw the notice in the	m…	Post, but-" He again interrupted
⊞	1) THE …	to remark that what the	m…	Post man said would be
⊞	1) THE …	to the critic of the	m…	Post, and had nothing to

Figure 93 Contexts, The Angel and the Author – And Others

There is another semantic ambiguity that the word “morning” refers to the time of dawn, or it is a proper noun. One postulate of “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). Digital tools are also built on the basis of theories. To answer to this question, Contexts tool highlights that in 1st sentence, it refers to the time of dawn, while from 2nd to 4th sentences, it means London origin Morning Times newspaper. There is a deficiency that Contexts tool does not show capitalization of the proper noun “Morning”. Hermeneutic tools “fail in interesting ways” (Rockwell, & Sinclair, 2016, p. 166). This shortcoming can be covered by adding proper nouns and case-sensitive words in the embedded lexicons of Voyant tools.

4.4 Data Analysis of Book III

Book III (Appendix B) has two sections: the first section covers three plays, and the second section includes 20 poems which have been written by British and American poets, whereas, some poems have been translated from Oriental languages to English language.

4.5 Text Mining of One-Act Plays

A playwright writes a play for a theatre, and its plot develops with dialogues.

1.*Heat Lightning* by Robert F. Carroll

i. Summary

This corpus has 1 document with 3,018 total words and 638 unique word forms. Created about 4 minutes ago (on 15th October 2017).

Vocabulary Density: 0.211

Average Words Per Sentence: 8.5

Most frequent words in the corpus: man (125); girl (71); second (44); door (29); light (21); lightning (17); yes (15); right (13); know (12); room (12)

Figure 94 Summary, Heat Lightning

Summarization shows a condensed report of a subset of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45). Large text mining results are shown in a few lines in the shortest possible time. Corpus Summary tool is a means of computational stylistics to reveal the literary style of the play writer. Robert F. Carroll writes 638 unique words, and almost every word is repeated five times; hence, total words are counted as 3018. It proposes that repeated vocabulary has been used in this play, and it creates ease for intermediate-level readers. Average words per sentence also show that small sentences have been written in the play because dialogues are usually written in small sentences. This play uses 8.5 words per sentence which show its suitability for basic level readers. Besides, the most occurring words suggest key characters and themes of the entire play. Furthermore, anyone who highlighted key word works like a hyperlink to retrieve the whole corpus. The foregoing discussion implies a one-word library concept for readers and students. When a teacher wants to send the whole syllabus of intermediate to his/her class, only one highlighted word will be mailed and that word will convey the whole corpus. Findings of this corpus tool for stylometry verify previous findings of stylometric studies in literature, for instance, Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O’Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018.

ii. Cirrus

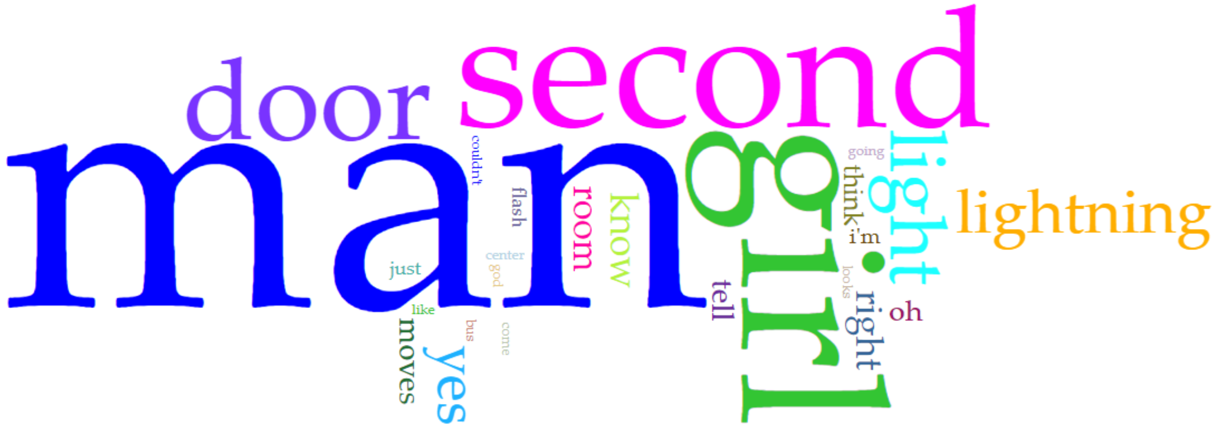


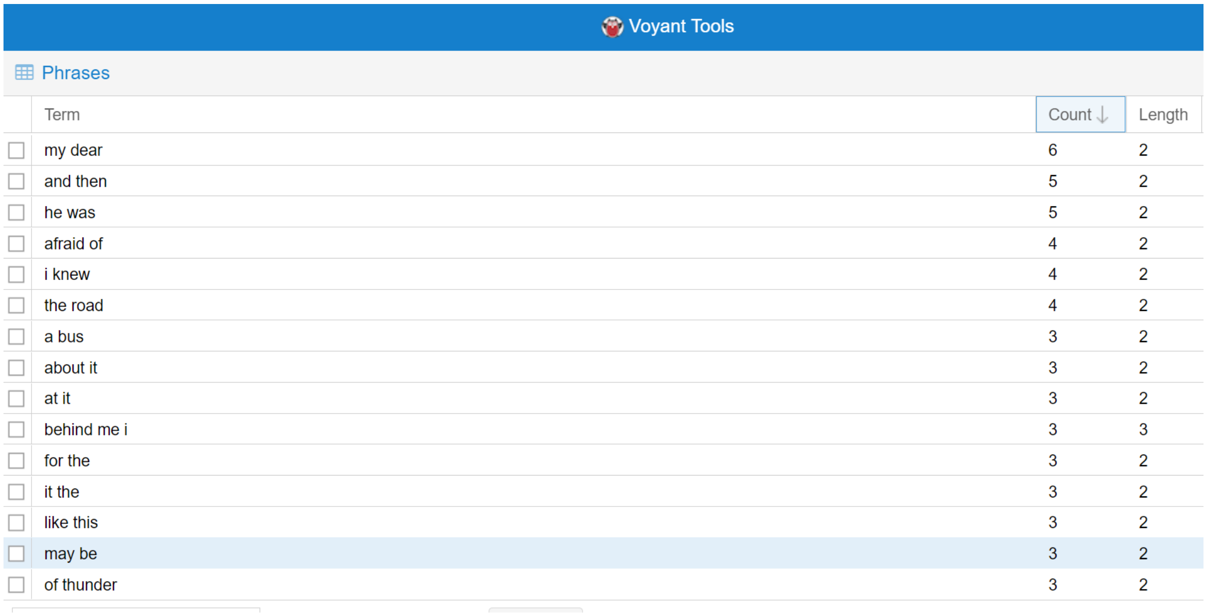
Figure 95 Cirrus, Heat Lightning

Data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). Cirrus quantifies all themes and presents them in the compressed form to highlight the most repeated themes and characters. This Cirrus exposes major characters and themes from the play. Prominent characters have been visualised in the Cirrus as “man (125)”, “girl (71)”. There are two men, first man and second man, that is why the number of word “man” exceeds in totality. The central character in the play is a “girl (71)” whose name occurs 71 times in this word cloud. Another cause of the frequent occurrence of names is that this play is written in the dialogic form, and before their turn, the particular character's name appears. Besides, the entire play revolves around three characters: the girl, the first man and the second man.

An animal character, “dog (3)” plays its concluding role to save the girl's life from the killer. It reveals a knowledge pattern that an animal has turned into a saviour, while the first man has transformed into a beast, so characteristic reversal has been presented here. In non-living characters, “light (21)”, “lightning (17)”, “storm (9)”, “thunder (3)” have been presented to heighten the terrifying effect of the play. Whenever an element of fear (“afraid 6)” is created, lightning and rain are shown time and again to heighten the net effect of horror. Furthermore, the word “door (29)” has been used repeatedly with the car door and waiting room door. The most important knowledgeable clue is “flashlight (6)” which reveals the true identity of the first man as a killer. Likewise, Cirrus also discovers themes from 37 Shakespearean plays with Voyant tools (Sinclair, & Rockwell, 2015b).

Human analysis shows that two male characters are associated with killing, cheating, while one female character is described with the traits of falling, rushing, terrified, out of breath. One female character is seeking help from male characters, hence, key themes reveal male dominance and female subordination. Some other themes are afraid, nervous, upset, muffled, weak, frightened, pressed, stunned, deserted, woman's corpse, night, dim light, thunder, lightning and rain. (Hussain, 2009, pp. 64-67). Comparing human and machine analysis, Cirrus tool extracts the same "man (125)" and "girl (71)" characters, but the character of "dog (3)" has been extracted by only Cirrus. Themes of "light (21)", "lightning (17)", "storm (9)", "thunder (3)" are also the same in both analyses. Some human-generated themes have not been mentioned by Cirrus tool.

iii. Phrases



	Term	Count ↓	Length
□	my dear	6	2
□	and then	5	2
□	he was	5	2
□	afraid of	4	2
□	i knew	4	2
□	the road	4	2
□	a bus	3	2
□	about it	3	2
□	at it	3	2
□	behind me i	3	3
□	for the	3	2
□	it the	3	2
□	like this	3	2
□	may be	3	2
□	of thunder	3	2

Figure 96 Phrases, Heat Lightning

Fifteen most occurring phrases have been chosen to present the standard phraseology, for instance, "my dear" (Prn+Adj) has been used six times. This phrase is addressed to the girl to give her sham solace, and it is a way to ensnare her. Moreover, these phrases also enhance fluency in speaking and writing skills. It has another purpose to teach correct language, for instance, the use of the preposition "afraid of" (Adj+Prep). Standard phraseology also guides the correct use of the auxiliary verb, for instance, "he was" (Prn+Aux), "may be" (Aux). They occur 3 to 6 times, and their length ranges from 2 to 3 words.

iv. Links

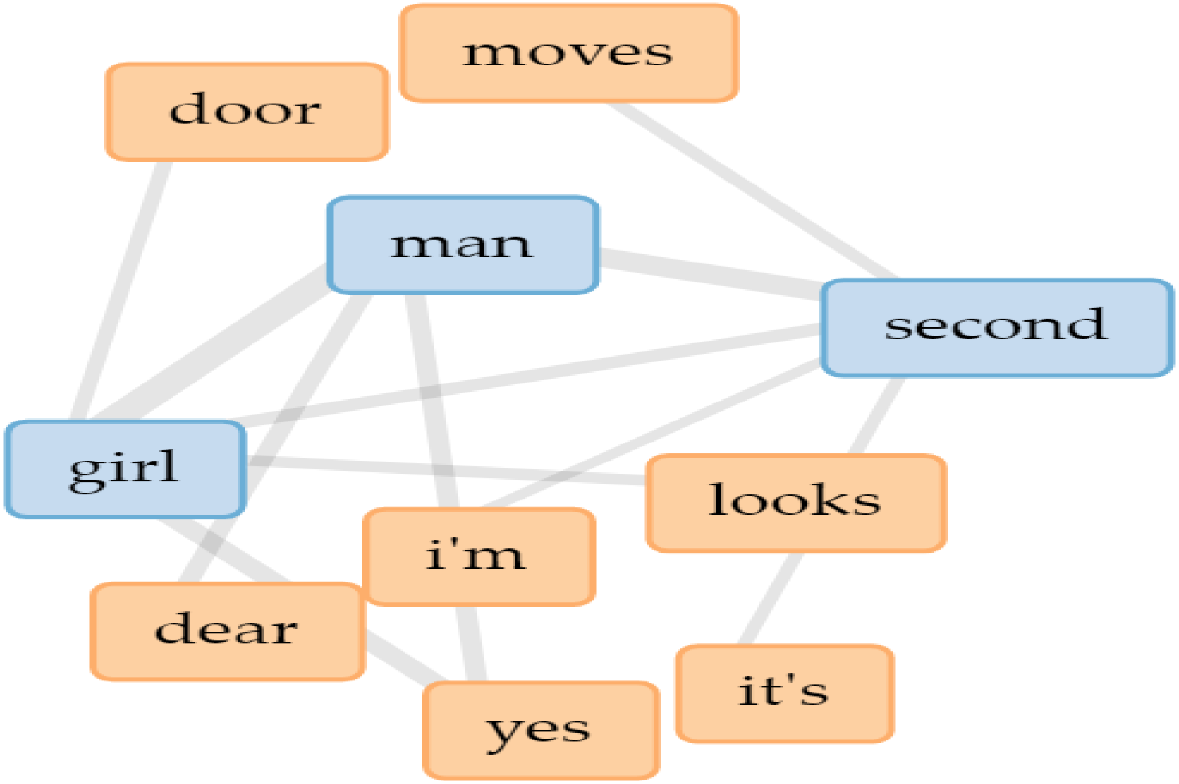


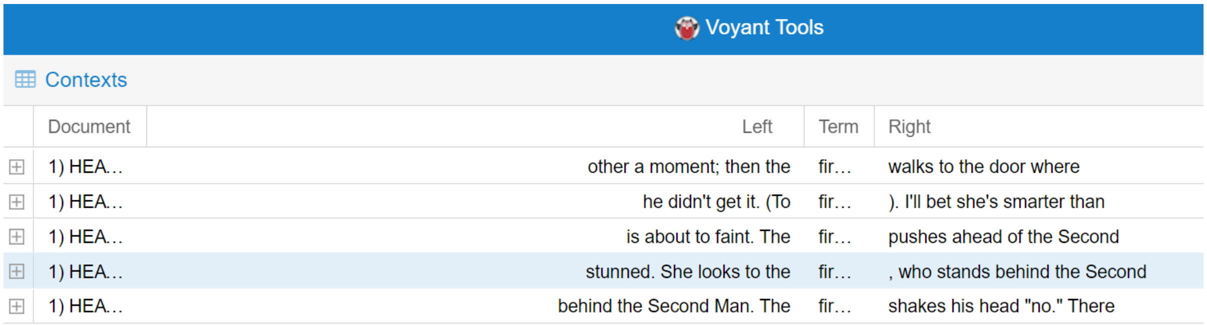
Figure 97 Links, Heat Lightning

The KG "second, man, girl" exhibit main characters (girl, first man, second man) who develop the whole plot of the play, and their KG has been aptly sketched here. Subsequently, the KG discovers the knowledge pattern of this plot. The KG of "moves, second, man, girl" indicates the textual reality that the second man moves towards the girl frequently to start close affinity with her. Another KG of "girl, man, dear" suggests that the man addresses her with the word "dear" frequently, but it is a pseudo feeling just to win her confidence and victimize her since she is the only witness of his act of throwing a girl's dead body on the road.

The KG of "girl, looks" shows that the girl's look is a paragon of fear and horror. Her terrified condition is apparent from her appearance, facial expressions and terrified conversation. Another KG of "girl, door" shows that she bolts the door while looking at the door. Her dread overwhelms her till the end of the play, that is why she slams the door to save her life. Most of the time, she looks at the door whether the murderer is coming or not.

Hermeneutic tools are “not like black boxes” (Rockwell, & Sinclair, 2016, p. 166) because they do not check the accuracy of background programming coding, NLP toolkits and python libraries. Hermeneutica Theory focuses on linguistic and other knowledge patterns to conduct a deeper level of interpretation for knowledge discovery. Text mining tools save the labour of counting and data visualization with these embedded computer programmes.

v. Contexts

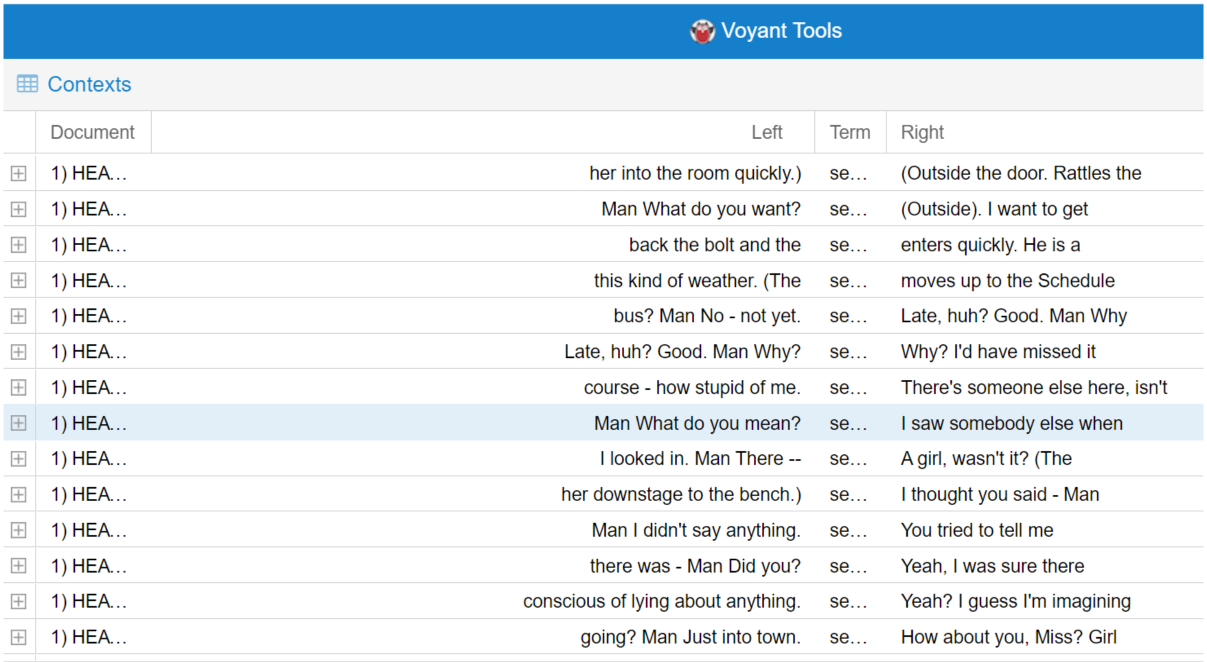


Voyant Tools

Contexts

	Document	Left	Term	Right
⊞	1) HEA…	other a moment; then the	fir…	walks to the door where
⊞	1) HEA…	he didn't get it. (To	fir…	). I'll bet she's smarter than
⊞	1) HEA…	is about to faint. The	fir…	pushes ahead of the Second
⊞	1) HEA…	stunned. She looks to the	fir…	, who stands behind the Second
⊞	1) HEA…	behind the Second Man. The	fir…	shakes his head "no." There

Figure 98 Contexts, Heat Lightning



Voyant Tools

Contexts

	Document	Left	Term	Right
⊞	1) HEA…	her into the room quickly.)	se…	(Outside the door. Rattles the
⊞	1) HEA…	Man What do you want?	se…	(Outside). I want to get
⊞	1) HEA…	back the bolt and the	se…	enters quickly. He is a
⊞	1) HEA…	this kind of weather. (The	se…	moves up to the Schedule
⊞	1) HEA…	bus? Man No - not yet.	se…	Late, huh? Good. Man Why
⊞	1) HEA…	Late, huh? Good. Man Why?	se…	Why? I'd have missed it
⊞	1) HEA…	course - how stupid of me.	se…	There's someone else here, isn't
⊞	1) HEA…	Man What do you mean?	se…	I saw somebody else when
⊞	1) HEA…	I looked in. Man There --	se…	A girl, wasn't it? (The
⊞	1) HEA…	her downstage to the bench.)	se…	I thought you said - Man
⊞	1) HEA…	Man I didn't say anything.	se…	You tried to tell me
⊞	1) HEA…	there was - Man Did you?	se…	Yeah, I was sure there
⊞	1) HEA…	conscious of lying about anything.	se…	Yeah? I guess I'm imagining
⊞	1) HEA…	going? Man Just into town.	se…	How about you, Miss? Girl

Figure 99 Contexts, Heat Lightning

In this play, two unnamed characters, “first man” and “second man” have been mentioned. To clarify their roles and utterances, Contexts tool is used, and it extracts exact bidirectional context without a thorough reading of the play. Figure 98 refers to the context of the word “first”, and figure 99 describes the second man’s dialogues.

A learner can easily comprehend and differentiate between the roles and dialogues of both men. The first man is the murderer and the second man just tries to be frank with the girl. Total 125 times word “man” is used, and the word “second” is used 44 times. It reveals that it refers to the first man 81 times, it refers to the second man 81 times. To save space, only figure 98 and figure 99 have been shown. Apart from them, once the word “second” is used for flashlight which serves as a clue to detect the true murderer. The significance of context is evident because Hermeneutica Theory is “embedded in a context” (Rockwell & Sinclair, 2016, p. 166).

2. *Visit to a Small Planet* by Gore Vidal

i. Summary

This corpus has 1 document with 4,315 total words and 1,004 unique word forms. Created about 4 minutes ago (on 15th October 2017).

Vocabulary Density: 0.233

Average Words Per Sentence: 7.9

Most frequent words in the corpus: spelding (95); kreton (94); powers (78); ellen (45); it's (35); john (34); know (31); i'm (3
0); oh (22); yes (22)

Figure 100 Summary, Visit to a Small Planet

Gore Vidal’s total vocabulary items in this corpus are 1004, and they have been recycled three times, accordingly, the total words of this corpus are 4315. Vocabulary density is derived with Inverse Absolute (Simpson, 2000) (a division of total words by unique words). As a result, its vocabulary density is 0.233 which indicates intermediate-level vocabulary. Gore Vidal writes 7.9 words per sentence, and it denotes small sentences to enhance the readability of basic level readers. Most frequent words exhibit major characters of the play, for example, “Spelding (95)”, “Kreton (94)”, “Powers (78)”, “Ellen (45)” and “John (34)”.

Hermeneutica Theory directs that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). The extraction of characters and themes is knowledge discovery and then, human reflection interprets them. Furthermore, Summary tool exhibits knowledge about quantified stylometric features, and human beings use them to differentiate between writers. Furthermore, stylometric features are used for author identification in case multiple writers claim authorship of any text. Summary tool can also lead to the analysis of wills, and it can contribute significantly to forensic linguistics.

ii. Cirrus

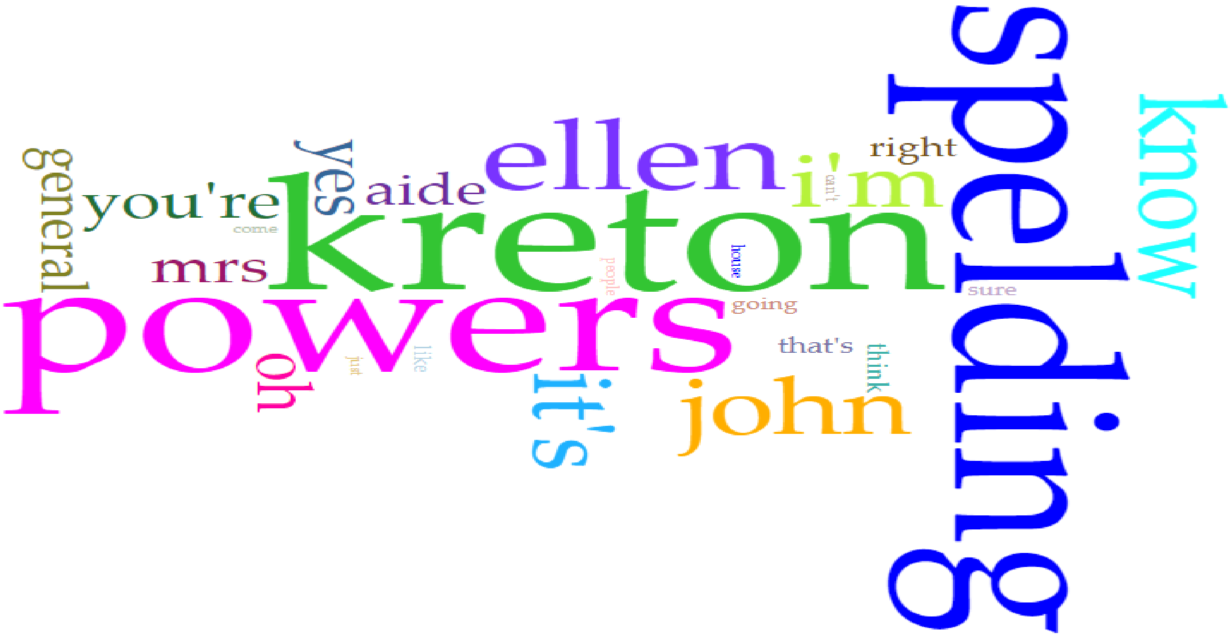


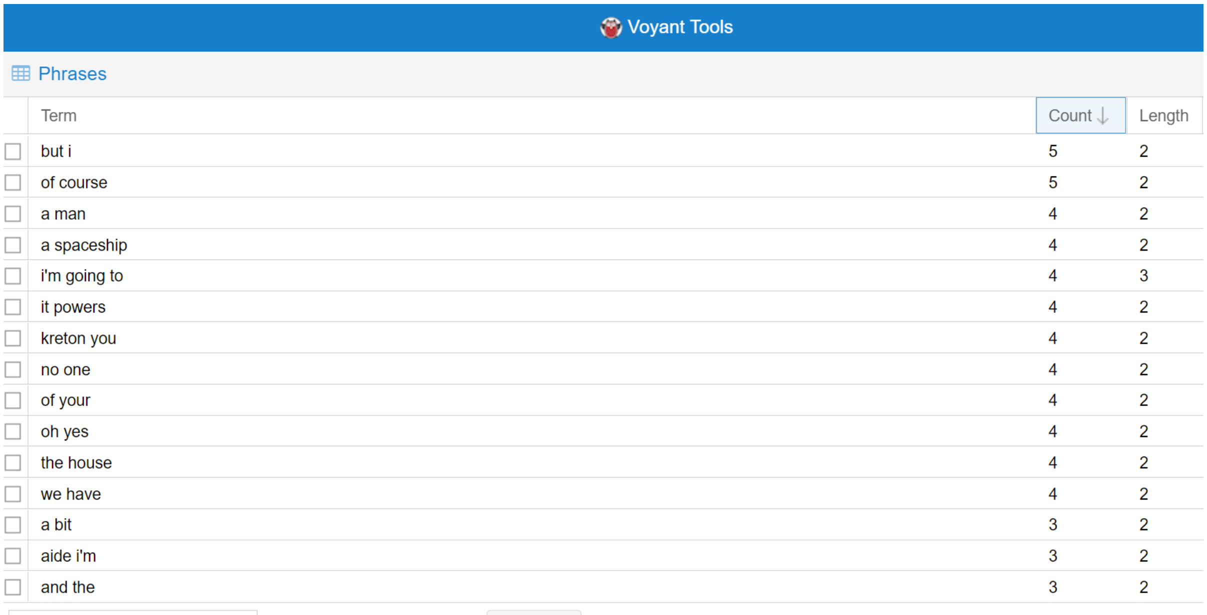
Figure 101 Cirrus, Visit to a Small Planet

This Cirrus shows clustering of major characters of the play, for instance, Mr and Mrs "Spelding (95), "Kreton (94)", General "Powers (78)", Ellen (45)" and "John (34)". The whole play revolves around the themes of Kreton's arrival, his magical abilities and his interactions with other characters. Being a central figure, Kreton's name has been used 94 times because of his dominant role. "Roger" is the first name of Mr Spelding; however, it has been used ten times in this corpus.

The interjection "oh" has been used 22 times, and it has been uttered on occasions of pain, grief and extreme surprise. This interjection is associated with Kreton's incredible acts of mind-reading, immortality, the building of an invisible protective wall and an unbelievable intergalactic vehicle which moves without fuel and steering wheel. Besides, the word "planet" has been used ten times, and it refers to the home of Kreton. The word "know" is used 31 times, and it shows the inquisitive nature of journalist, Mr Spelding, who always yearns to know different pieces of news to become an initiator of breaking news. On the other side, members of his family remain ignorant about his fame and achievements.

Human analysis reveals that there are two female characters, namely Mrs Spelding and Ellen; and seven male characters, including Mr Spelding who is the breadwinner of his family. The main themes of female characters are bored, vague and lively, while masculine-associated themes are unctuous, resonant and vigorous (Hussain, 2009, pp. 68-69). Comparing these characters and themes with Cirrus-generated characters, Cirrus shows more exact characters, for instance, Mr and Mrs "Spelding (95), "Kreton (94)", General "Powers (78)", Ellen (45)" and "John (34)". Cirrus key themes are different from humanly-extracted themes because the machines use counting and subtraction of common words, whereas human beings use their cognition and inclination in search of themes. Consequently, human themes would differ after every reading, while machine-generated themes would remain the same after every analysis. Human analysis is not quantified, while machine analysis is quantified and precise.

iii. Phrases



	Term	Count ↓	Length
□	but i	5	2
□	of course	5	2
□	a man	4	2
□	a spaceship	4	2
□	i'm going to	4	3
□	it powers	4	2
□	kreton you	4	2
□	no one	4	2
□	of your	4	2
□	oh yes	4	2
□	the house	4	2
□	we have	4	2
□	a bit	3	2
□	aide i'm	3	2
□	and the	3	2

Figure 102 Phrases, Visit to a Small Planet

In standard frequent phraseology, these collocation patterns/ n-grams are used for example, "of course" (Adv), "a man" (Art+N), "a spaceship" (Art+N), "I'm going to" (Prn+Aux+V+Prep), "no one" (Prn), "oh yes" (Int+Adv), "the house" (Art+N), "a bit" (Art+N). These four bigrams consist of articles and nouns. Their occurrence count is 3 to 5, and most of them are bigrams except for one trigram.

iv. Links

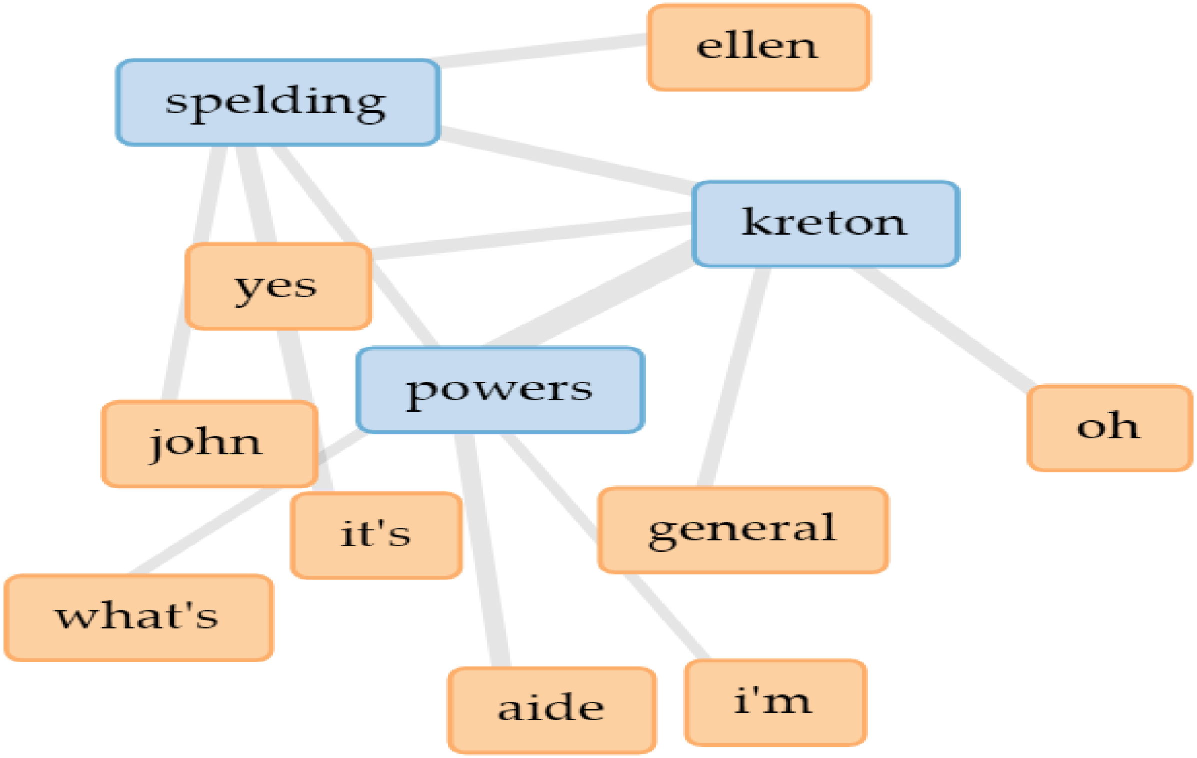


Figure 103 Links, Visit to a Small Planet

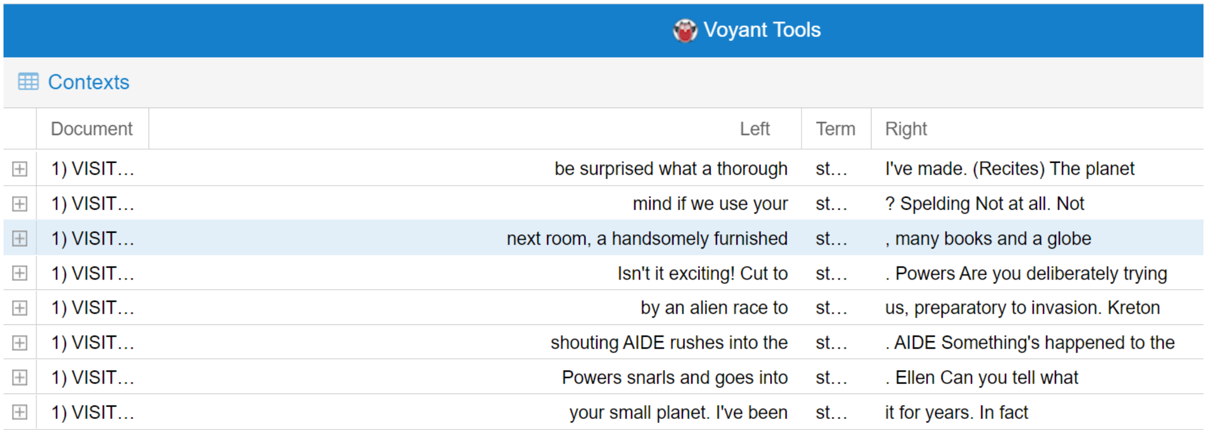
Three main characters have been visualized in the KG, “Spelding, Kreton, Powers”. The KG of “what’s, Powers, Kreton” expresses General Powers’ inquiries from Kreton about his plans and potentials to conquer the world, supernatural abilities of mind reading and immortality, technological advancements, driving the strange intergalactic vehicle without fuel and instrument panel.

Challenging and raising a question means challenging the authority of others. The interrogative pronoun “What’s” has been shown in the KG of “what’s, Powers, Kreton, Spelding, John” because Powers, Mr Spelding, Mrs Spelding and John ask questions from Kreton. Contrary to them, very interestingly and rightly, “What’s” is not linked with “Aide”, a subordinate of General Powers, because neither Aide raises any question, nor he talks to any other character except General Powers. Aide’s sole duty is to implement the orders of General Powers, and to report him back. Thus, relationship mining among relevant variables (Barahate, 2012, p. 13) has been proved. Hermeneutic tools are “not like black boxes” (Rockwell, & Sinclair, 2016, p. 166).

It is not a system that can be viewed as its inputs and outputs rather Hermeneutica Theory leads to a deeper level of interpretation for knowledge discovery.

The pronoun "I" is used to express authority and supremacy on others; therefore, the KG "I'm, General" shows a superior and aggressive attitude of General towards Kreton, and General Powers always considers himself right in his feelings and thoughts; hence, this aptitude makes him authoritative and overwhelming.

v. Contexts



	Document	Left	Term	Right
⊞	1) VISIT…	be surprised what a thorough	st…	I've made. (Recites) The planet
⊞	1) VISIT…	mind if we use your	st…	? Spelding Not at all. Not
⊞	1) VISIT…	next room, a handsomely furnished	st…	, many books and a globe
⊞	1) VISIT…	Isn't it exciting! Cut to	st…	. Powers Are you deliberately trying
⊞	1) VISIT…	by an alien race to	st…	us, preparatory to invasion. Kreton
⊞	1) VISIT…	shouting AIDE rushes into the	st…	. AIDE Something's happened to the
⊞	1) VISIT…	Powers snarls and goes into	st…	. Ellen Can you tell what
⊞	1) VISIT…	your small planet. I've been	st…	it for years. In fact

Figure 104 Contexts, Visit to a Small Planet

There are lexical and semantic ambiguities whether the word "study" is a noun or a verb or library or reading process. To disambiguate word sense, Contexts tool shows that word "study" is used 8 times in figure 104 and 1st sentence, noun "study" means a complete research work. 2nd, 3rd, 4th, 5th, 6th and 7th sentences refer to a library meaning, and they are used as nouns. In 8th sentence, the word "studying" has been used as a verb, and the context of "have been" validated the finding of a verb. To conclude, Contexts tool clarifies parts of speech and semantic shade of any ambiguous word.

3.*The Oyster and the Pearl* by William Saroyan

i. Summary

This corpus has 1 document with 5,738 total words and 964 unique word forms. Created 8 seconds ago (on 15th October 2017).

Vocabulary Density: 0.168

Average Words Per Sentence: 8.4

Most frequent words in the corpus: harry (185); clay (74); miss (61); mccutcheon (53); clark (41); oyster (36); writer (32); girl (31); haircut (31); know (28)

Figure 105 Summary, The Oyster and the Pearl

"Summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). William Saroyan writes 964 unique words, and almost the same words have been repeated six times more until their total number reaches 5738 words. Therefore, its vocabulary density is 0.168 which is suitable for advanced-level readers. In one line, almost 8.4 words have been written by William Saroyan who writes short sentences for beginner level readers. In a nutshell, William Saroyan uses less repeated vocabulary in small sentences. Most of the frequent words consist of the names of the characters of the play. Furthermore, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009), and these elements facilitate the presentation of quantified criticism and specific characteristics of stylistics. These features also show peculiar linguistic qualities of any writer or any piece of writing.

ii. Cirrus

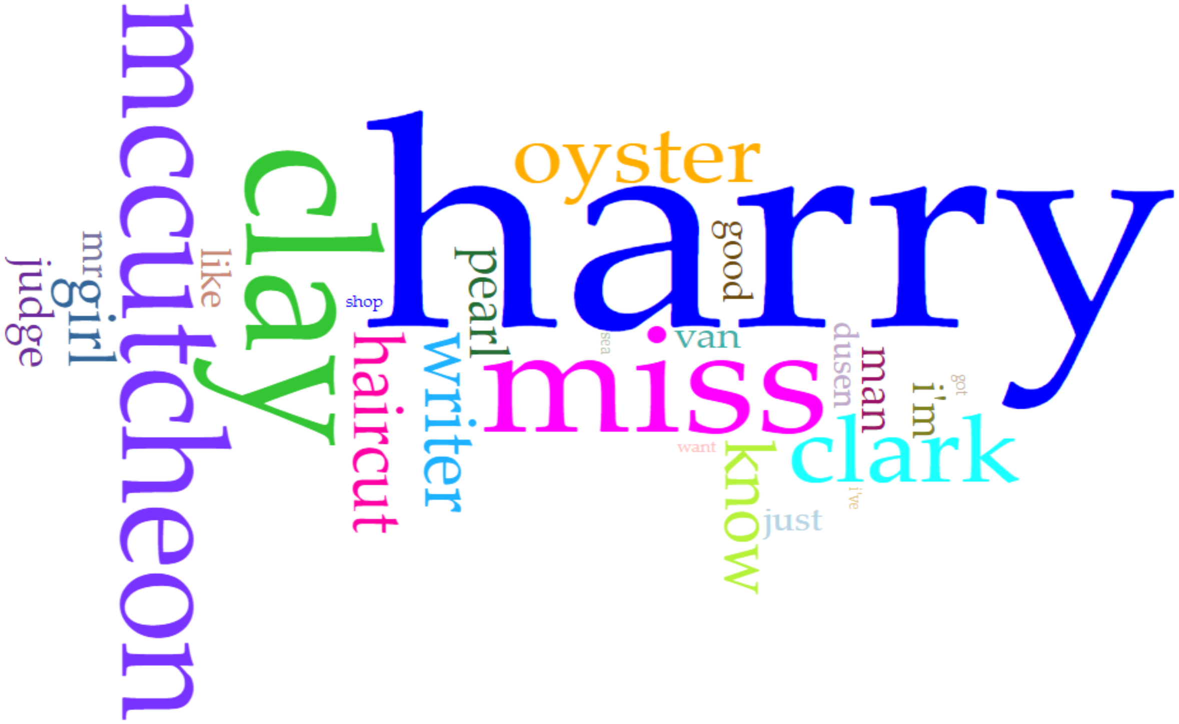


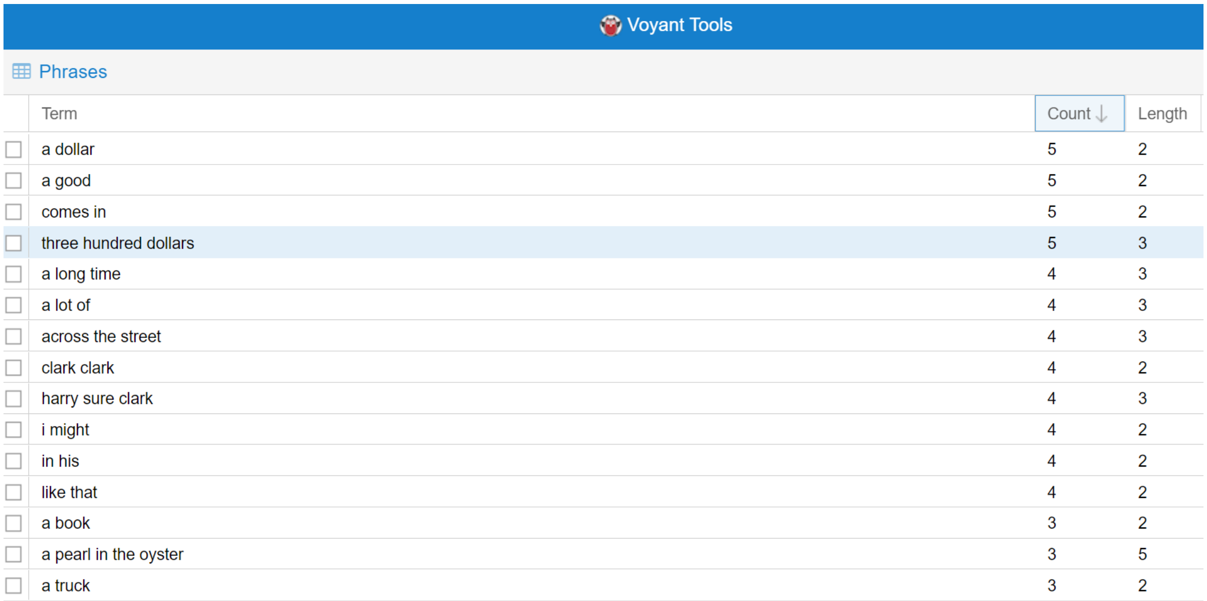
Figure 106 Cirrus, The Oyster and the Pearl

This Cirrus reveals major characters of the play, for instance, “Harry (185)”, “Clay (74)”, “Miss (61)” “McCutcheon (53)”, “Clark (41)”, “Wozzeck (14)”, “Van Dusen (23)”, “girl (31)” and “writer (32)”. Harry is a barber by profession, so the word “haircut (31)” has been used frequently in this play. Almost the entire setting of the play is in Harry’s barber shop; hence, it occurs most of the time, and all characters of the play have deep-seated ties with Harry and his shop.

The theme of “money (16)” has been discussed in the play because money determines the behaviours of people. Clay’s father Clark leaves his home due to absolute poverty and domestic violence. Owing to his miserable circumstances, Clay yearns to get money to resolve parental conflicts and to unite his parents. The themes of “oyster (36)” and “pearl (27)” are chief points of discussion. It has been discussed in Harry’s shop whether the “oyster (36)” is fake or real, and eventually, the “writer (32)” buys it for three hundred dollars to fulfil financial desires of Clay’s family, and to keep illusions of Clay intact, since illusions support weak creatures who cannot face bitter realities of life. Ibsen’s play *‘Wild Duck’* proves that illusions are necessary for weak human beings to sustain in life; otherwise, bitter realities devastate feeble creatures (Ibsen, 1890).

Human analysis shows that there are ten male and three female characters. The most dominating female character is Miss McCutcheon, a teacher, and the second famous female character is Roxana Larrabee, who brings seashells from the sea. Miss McCutcheon is a nervous and disturbed character. Male barber Harry and his shop play a pivotal role in the play. The main themes are blonde hair, too pretty, too young, bathing beauty, well dressed, neatly, swift and bewildered (Hussain, 2009, pp. 69-71). Cirrus analysis is more detailed, and it shows all characters, including human analysis characters. Cirrus themes are different from humanly analysed themes.

iii. Phrases



Voyant Tools

Phrases

	Term	Count ↓	Length
☐	a dollar	5	2
☐	a good	5	2
☐	comes in	5	2
☐	three hundred dollars	5	3
☐	a long time	4	3
☐	a lot of	4	3
☐	across the street	4	3
☐	clark clark	4	2
☐	harry sure clark	4	3
☐	i might	4	2
☐	in his	4	2
☐	like that	4	2
☐	a book	3	2
☐	a pearl in the oyster	3	5
☐	a truck	3	2

Figure 107 Phrases, The Oyster and the Pearl

This corpus exhibits standard phraseology, for instance, “a dollar” (Art+N), “comes in” (V+Prep), “three hundred dollars” (Nu+Adj+N), “a long time” (Art+Adj+N), “a lot of” (Adv) “across the street” (Prep+Art+N), “I might” (Prn+Mod), “like that” (Prep+Prn) and “a book and a pearl in the oyster” (Art+N+Conj+Art+N+Prep+Art+N). The most occurring collocation pattern/ n-grams, “a dollar” discovers the knowledge pattern of financial problems faced by characters. Statistical knowledge is displayed with a 3 to 5 occurrence range and word length ranges from 2 to 5 words.

iv. Links

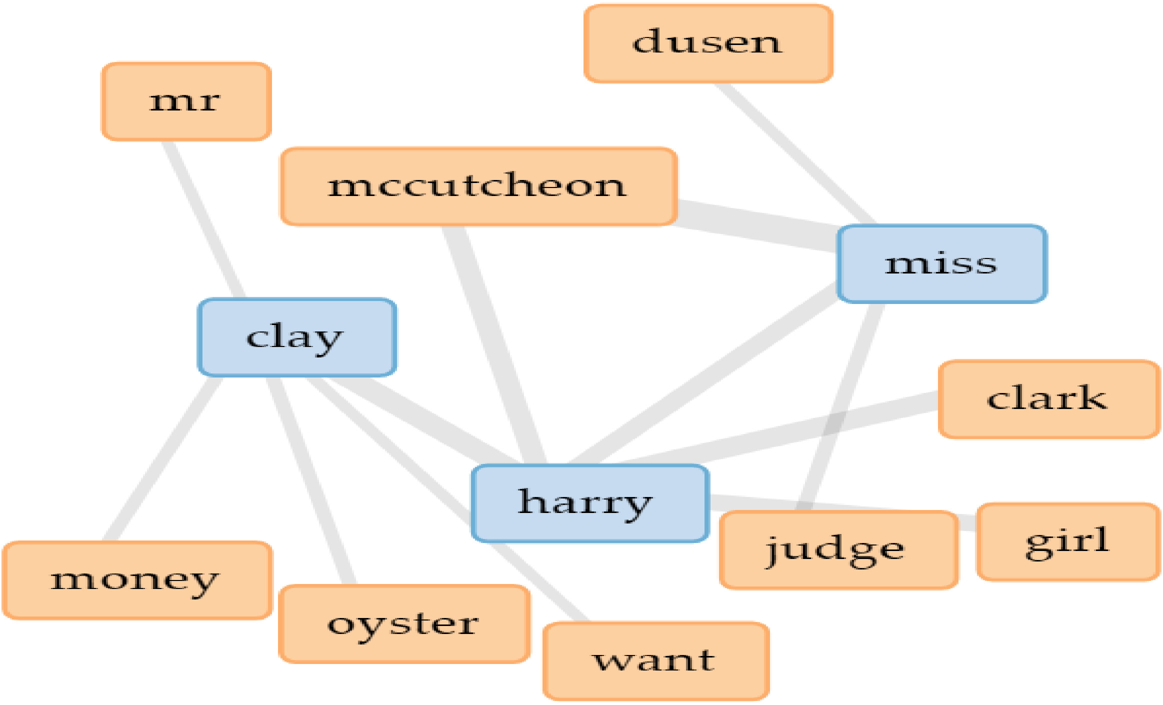
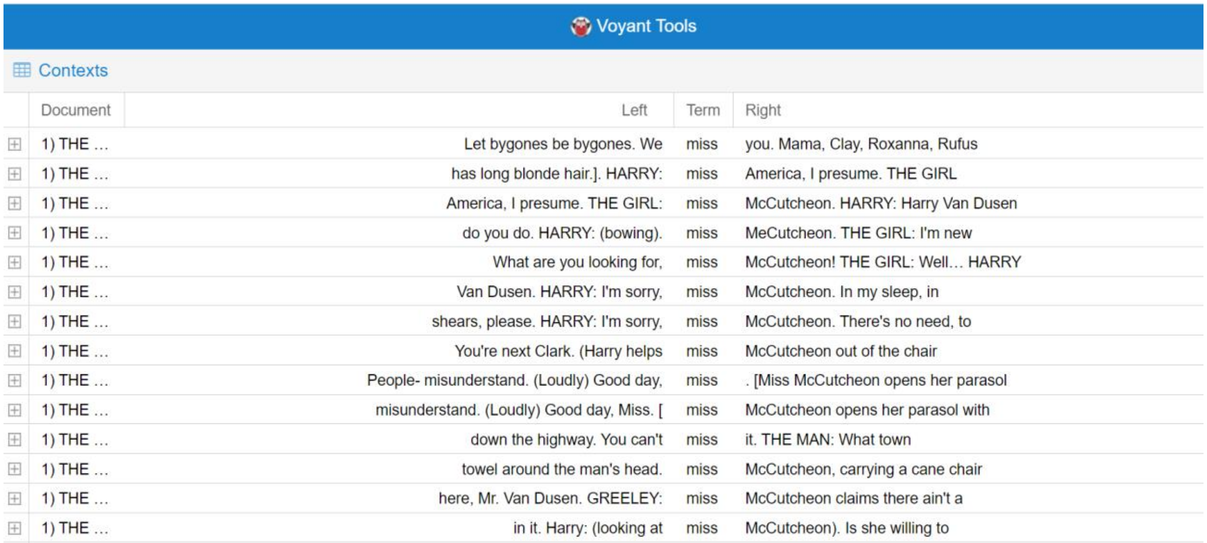


Figure 108 Links, The Oyster and the Pearl

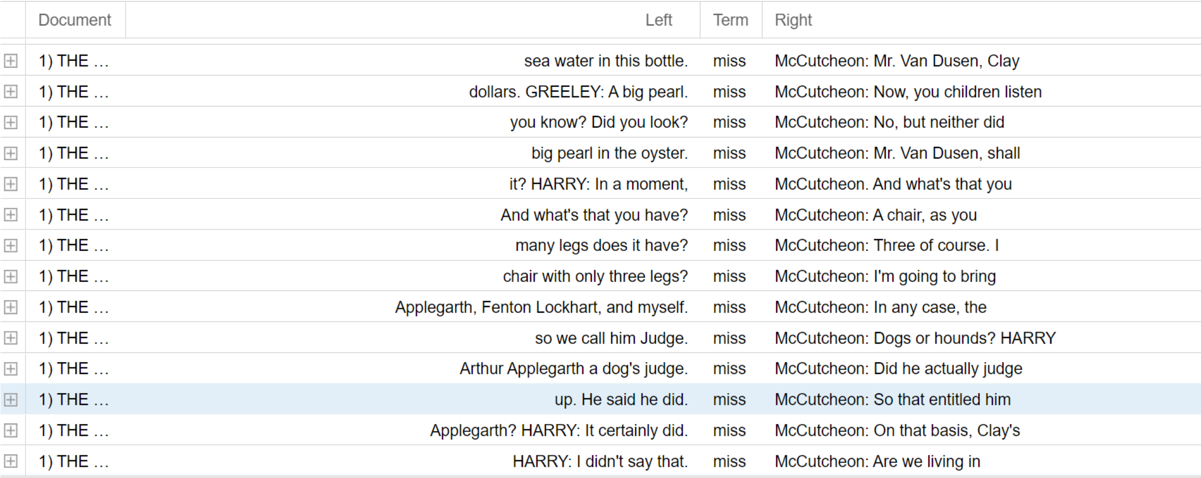
The KG of "oyster, Clay, money" builds an interesting and truthful pattern that Clay requires money to settle his father and mother's quarrelsome financial issues after selling the oyster. Moreover, he wants to bring his father home by dint of this money. Moreover, it also exposes the poverty-stricken conditions of this disturbed family. Continuing the same flow, the KG of "Clay, money, oyster,want" has deep-rooted ties with Clay who yearns to sell the oyster for getting money.

"Harry" or "Dusen" names refer to the same person, whereas "miss", "McCutcheon" and "girl" refer to one person. The KG of other characters, namely "judge, Clark, Miss, McCutcheon, Harry, Dusen" visit the barber's shop, a geographical setting of the play and the centre of all local activities. Consequently, all characters visit him for a haircut and leisurely activities. The same barbershop functions as a courtroom to decide all matters, including the debate of pearl in the oyster. The same barber's shop is an origin of taking life easy philosophy.

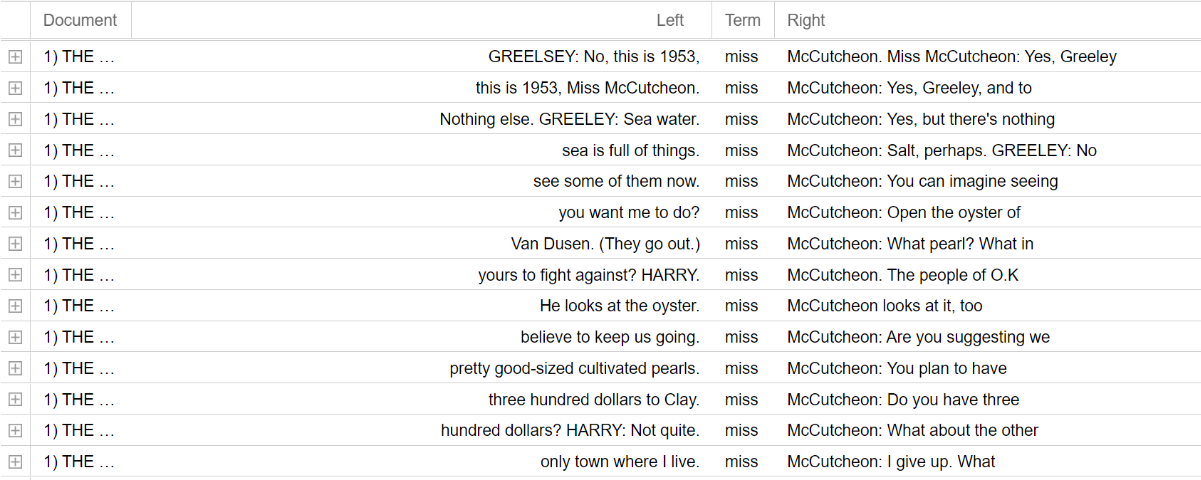
v. Contexts



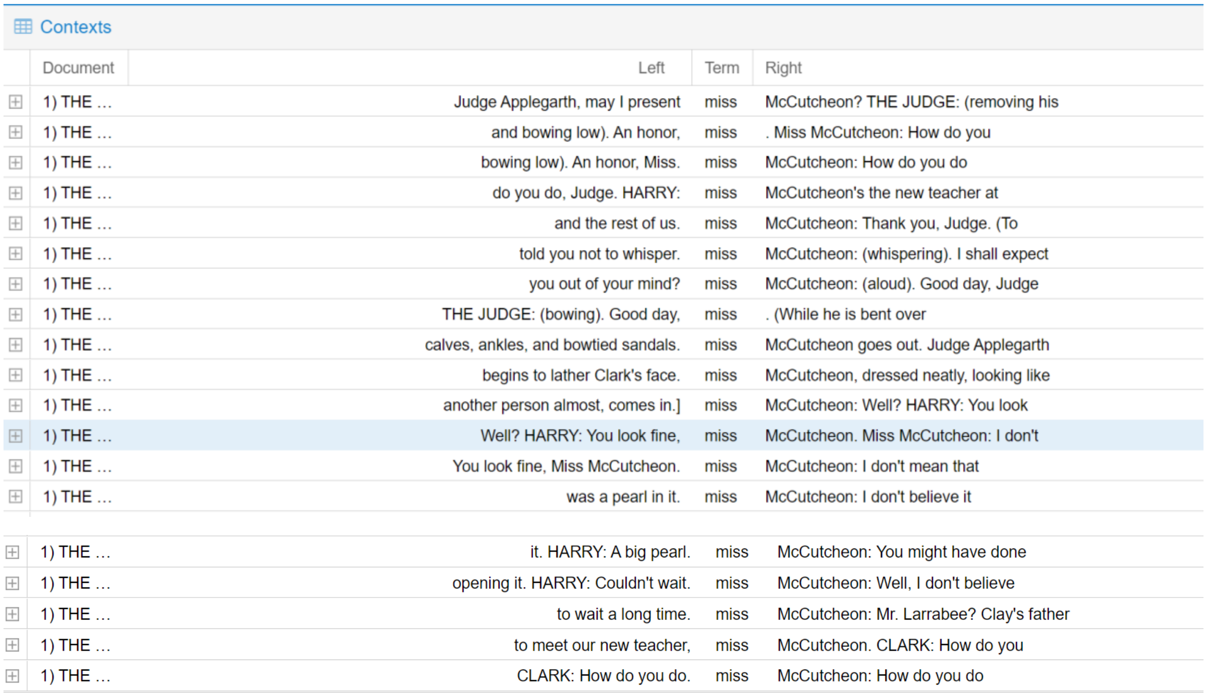
	Document	Left	Term	Right
⊞	1) THE …	Let bygones be bygones. We	miss	you. Mama, Clay, Roxanna, Rufus
⊞	1) THE …	has long blonde hair.]. HARRY:	miss	America, I presume. THE GIRL
⊞	1) THE …	America, I presume. THE GIRL:	miss	McCutcheon. HARRY: Harry Van Dusen
⊞	1) THE …	do you do. HARRY: (bowing).	miss	MeCutcheon. THE GIRL: I'm new
⊞	1) THE …	What are you looking for,	miss	McCutcheon! THE GIRL: Well… HARRY
⊞	1) THE …	Van Dusen. HARRY: I'm sorry,	miss	McCutcheon. In my sleep, in
⊞	1) THE …	shears, please. HARRY: I'm sorry,	miss	McCutcheon. There's no need, to
⊞	1) THE …	You're next Clark. (Harry helps	miss	McCutcheon out of the chair
⊞	1) THE …	People- misunderstand. (Loudly) Good day,	miss	. [Miss McCutcheon opens her parasol
⊞	1) THE …	misunderstand. (Loudly) Good day, Miss. [	miss	McCutcheon opens her parasol with
⊞	1) THE …	down the highway. You can't	miss	it. THE MAN: What town
⊞	1) THE …	towel around the man's head.	miss	McCutcheon, carrying a cane chair
⊞	1) THE …	here, Mr. Van Dusen. GREELEY:	miss	McCutcheon claims there ain't a
⊞	1) THE …	in it. Harry: (looking at	miss	McCutcheon). Is she willing to



	Document	Left	Term	Right
⊞	1) THE …	sea water in this bottle.	miss	McCutcheon: Mr. Van Dusen, Clay
⊞	1) THE …	dollars. GREELEY: A big pearl.	miss	McCutcheon: Now, you children listen
⊞	1) THE …	you know? Did you look?	miss	McCutcheon: No, but neither did
⊞	1) THE …	big pearl in the oyster.	miss	McCutcheon: Mr. Van Dusen, shall
⊞	1) THE …	it? HARRY: In a moment,	miss	McCutcheon. And what's that you
⊞	1) THE …	And what's that you have?	miss	McCutcheon: A chair, as you
⊞	1) THE …	many legs does it have?	miss	McCutcheon: Three of course. I
⊞	1) THE …	chair with only three legs?	miss	McCutcheon: I'm going to bring
⊞	1) THE …	Applegarth, Fenton Lockhart, and myself.	miss	McCutcheon: In any case, the
⊞	1) THE …	so we call him Judge.	miss	McCutcheon: Dogs or hounds? HARRY
⊞	1) THE …	Arthur Applegarth a dog's judge.	miss	McCutcheon: Did he actually judge
⊞	1) THE …	up. He said he did.	miss	McCutcheon: So that entitled him
⊞	1) THE …	Applegarth? HARRY: It certainly did.	miss	McCutcheon: On that basis, Clay's
⊞	1) THE …	HARRY: I didn't say that.	miss	McCutcheon: Are we living in



	Document	Left	Term	Right
⊞	1) THE …	GREELSEY: No, this is 1953,	miss	McCutcheon. Miss McCutcheon: Yes, Greeley
⊞	1) THE …	this is 1953, Miss McCutcheon.	miss	McCutcheon: Yes, Greeley, and to
⊞	1) THE …	Nothing else. GREELEY: Sea water.	miss	McCutcheon: Yes, but there's nothing
⊞	1) THE …	sea is full of things.	miss	McCutcheon: Salt, perhaps. GREELEY: No
⊞	1) THE …	see some of them now.	miss	McCutcheon: You can imagine seeing
⊞	1) THE …	you want me to do?	miss	McCutcheon: Open the oyster of
⊞	1) THE …	Van Dusen. (They go out.)	miss	McCutcheon: What pearl? What in
⊞	1) THE …	yours to fight against? HARRY.	miss	McCutcheon. The people of O.K
⊞	1) THE …	He looks at the oyster.	miss	McCutcheon looks at it, too
⊞	1) THE …	believe to keep us going.	miss	McCutcheon: Are you suggesting we
⊞	1) THE …	pretty good-sized cultivated pearls.	miss	McCutcheon: You plan to have
⊞	1) THE …	three hundred dollars to Clay.	miss	McCutcheon: Do you have three
⊞	1) THE …	hundred dollars? HARRY: Not quite.	miss	McCutcheon: What about the other
⊞	1) THE …	only town where I live.	miss	McCutcheon: I give up. What



Contexts

	Document	Left	Term	Right
⊞	1) THE ...	Judge Applegarth, may I present	miss	McCutcheon? THE JUDGE: (removing his
⊞	1) THE ...	and bowing low). An honor,	miss	. Miss McCutcheon: How do you
⊞	1) THE ...	bowing low). An honor, Miss.	miss	McCutcheon: How do you do
⊞	1) THE ...	do you do, Judge. HARRY:	miss	McCutcheon's the new teacher at
⊞	1) THE ...	and the rest of us.	miss	McCutcheon: Thank you, Judge. (To
⊞	1) THE ...	told you not to whisper.	miss	McCutcheon: (whispering). I shall expect
⊞	1) THE ...	you out of your mind?	miss	McCutcheon: (aloud). Good day, Judge
⊞	1) THE ...	THE JUDGE: (bowing). Good day,	miss	. (While he is bent over
⊞	1) THE ...	calves, ankles, and bowtied sandals.	miss	McCutcheon goes out. Judge Applegarth
⊞	1) THE ...	begins to lather Clark's face.	miss	McCutcheon, dressed neatly, looking like
⊞	1) THE ...	another person almost, comes in.]	miss	McCutcheon: Well? HARRY: You look
⊞	1) THE ...	Well? HARRY: You look fine,	miss	McCutcheon. Miss McCutcheon: I don't
⊞	1) THE ...	You look fine, Miss McCutcheon.	miss	McCutcheon: I don't mean that
⊞	1) THE ...	was a pearl in it.	miss	McCutcheon: I don't believe it
⊞	1) THE ...	it. HARRY: A big pearl.	miss	McCutcheon: You might have done
⊞	1) THE ...	opening it. HARRY: Couldn't wait.	miss	McCutcheon: Well, I don't believe
⊞	1) THE ...	to wait a long time.	miss	McCutcheon: Mr. Larrabee? Clay's father
⊞	1) THE ...	to meet our new teacher,	miss	McCutcheon. CLARK: How do you
⊞	1) THE ...	CLARK: How do you do.	miss	McCutcheon: How do you do

Figure 109 Contexts, The Oyster and the Pearl

The word “miss” is ambiguous because as a noun, it refers to a girl; and as a verb, it means to think about some absent person in a nostalgic mood. The theory postulate guides that “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). To resolve this ambiguity, the context of the word has been searched, and it shows that the word “miss” occurs 61 times in this play. Word “miss” as a verb is used two times in the 1st and 11th sentence, and it gives meaning to think somebody passionately. The word “miss” as a common noun has been used 59 times; and among them, the word “miss” is used 57 times before the proper noun “McCutcheon”. Once the word “miss” has been used before the proper noun “America” in the 2nd sentence. Once the word “miss” is used without a proper noun in the 30th sentence. In this way, the word “miss” has been disambiguated semantically, grammatically and lexically.

4.6 Text Mining of Poems

Poems express intense, lofty and powerful emotions with rhythm, literary devices, diction and imagery.

1. *The Rain* by W.H. Davies

i. Summary

This corpus has 1 document with 63 total words and 45 unique word forms. Created now (on 27th August 2017)

Vocabulary Density: 0.714

Average Words Per Sentence: 31.5

Most frequent words in the corpus: drop (3); hear (3); leaves (3); rain (3); drinking (2)

Figure 110 Summary, The Rain

Summarization shows a condensed report of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45). So, the unique words of the poem are 45, whereas the total words are 63. Vocabulary density is very low which suggests less repeated vocabulary. The vocabulary density is derived with Inverse Absolute (Simpson, 2000) (a division of total words by unique words). It seems that this poem comprises very long sentences because it uses 31.5 words per sentence, but it is a wrong perception. The poem has very small poetic lines and after six lines, a full stop is marked. This extension of thoughts without a period is termed enjambment (the continuation of a sentence without a pause after a verse or stanza). The NLP system counts a sentence between two full stops or from beginning to a full stop. To conclude, four frequent words of the poem (drop, hear, leaves, rain) occur three times in the poem, and among them, one word "rain" is the title of the poem. To conclude, these findings verify some previous literary studies on stylometry, for instance, Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O'Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018.

ii. Cirrus

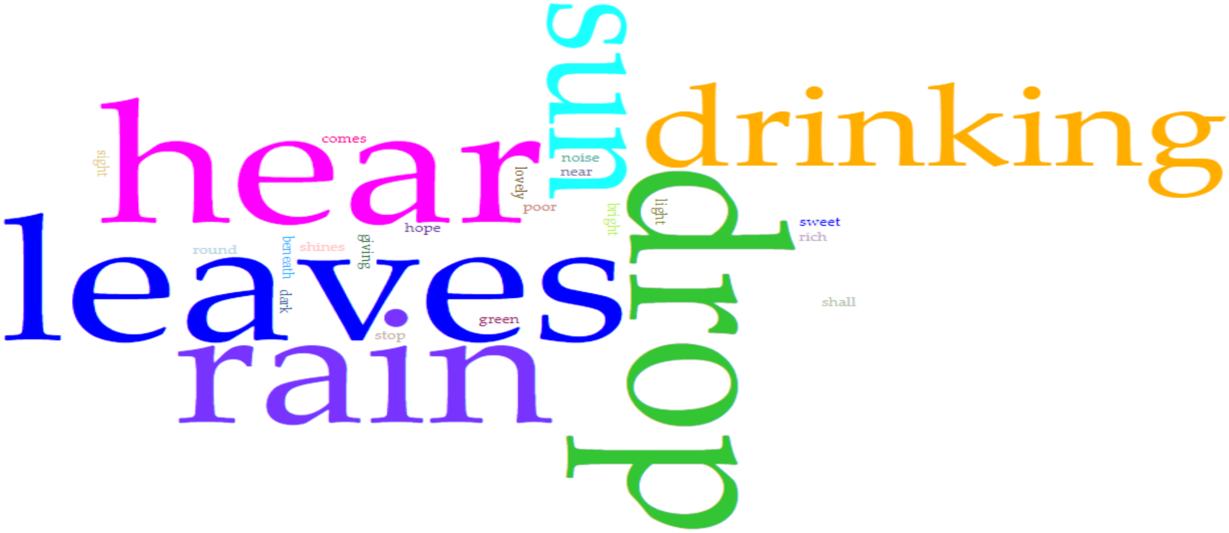
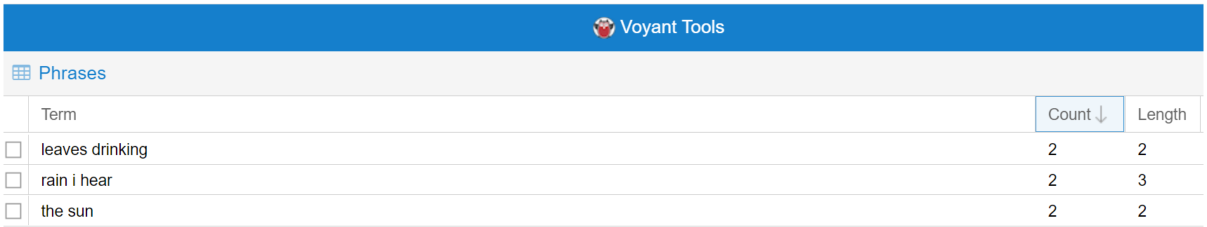


Figure 111 Cirrus, The Rain

Data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009), and they are evident in Cirrus because all themes and characters have been presented with precision and quantification. Major themes of the poem are “rain (3)”, “drop (3)”, “leaves (3)” and “hear (3)” since raindrops fall on leaves to produce musical sound. The word “drinking (2)” occurs twice, and it shows that leaves are drinking drops of rain. The second stanza revolves around the sun and its lustrous effects on the environment after rain. The theme of “sun (2)” is found twice in the poem that how sunshine produces a colourful rainbow after the rain.

Human analysis shows that rain is the key theme, but no human character has been presented (Hussain, 2009, p. 71). Cirrus tool shows themes of “rain (3)”, “drop (3)”, “leaves (3)” and “hear (3)”, but the human analysis did not concentrate on them.

iii. Phrases



	Term	Count ↓	Length
□	leaves drinking	2	2
□	rain i hear	2	3
□	the sun	2	2

Figure 112 Phrases, The Rain

These phrases**,** “leaves drinking” (N+V), “rain I hear” (N+Prn+V) and “the sun” (Art+N), are found twice in the poem. It suggests that the definite article “the” should be used before heavenly bodies. They occur twice, and their length consists of 2 to 3 words.

iv. Links

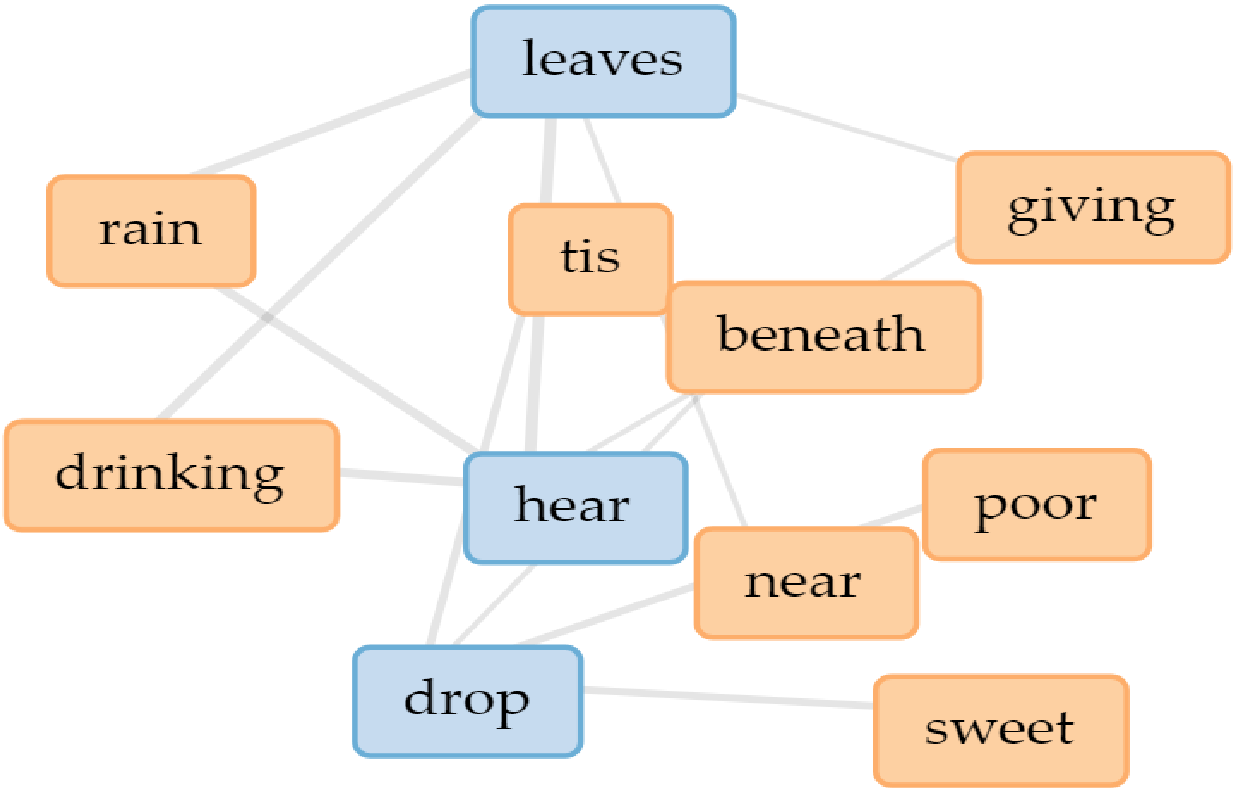
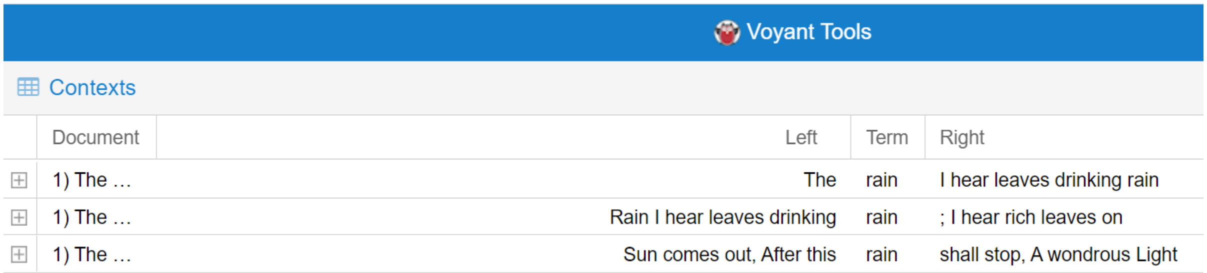


Figure 113 Links, The Rain

The KG of "rain, hear, drop" has been linked showing the association of themes. Moreover, drops go beneath poor leaves, therefore "rain" and "hear" are linked because rain creates audible natural rhythmic musicality, and this situation has been displayed in the KG of "beneath, hear, drop". The KG of "sweet, drop" refers to sweet noise which can be named as pleasing music. "Drop" of rain and water represents the "giving" process. That is why they are linked with each other in the KG of "giving, beneath, drop". "Leaves" have their association in the KG of "drinking, leaves, near, poor". Three rhyming words "near, hear, poor" show their acoustic relationship. Besides, Hermeneutic tools are "not like black boxes" (Rockwell, & Sinclair, 2016, p. 166). It is not a system which can be viewed as only its inputs and outputs. Hermeneutica Theory leads to a deeper level of interpretation for knowledge discovery. It does not check the efficacy of embedded coding in text mining tools.

v. Contexts



	Document	Left	Term	Right
⊞	1) The …	The	rain	I hear leaves drinking rain
⊞	1) The …	Rain I hear leaves drinking	rain	; I hear rich leaves on
⊞	1) The …	Sun comes out, After this	rain	shall stop, A wondrous Light

Figure 114 Contexts, The Rain

There is a grammatical ambiguity whether the word “rain” is used as a noun or a verb in this poem. To disambiguate the word rain, the context of the word “rain” has been extracted, and the word “rain (3)” has been found three times in the poem, for instance, first time in the title, the second time in the first sestet, and the third time in the second sestet. Three times the word “rain” is used as a noun. The significance of context is evident because Hermeneutica Theory is “embedded in a context” (Rockwell, & Sinclair, 2016, p. 166).

2. *Night Mail* by W.H. Auden

i. Summary

This corpus has 1 document with 114 total words and 85 unique word forms. Created now (on 27th August 2017).

Vocabulary Density: 0.746

Average Words Per Sentence: 19.0

Most frequent words in the corpus: letters (2); mail (2); night (2); passes (2); turn (2)

Figure 115 Summary, Night Mail

W.H. Auden almost repeats one word twice because its unique words are 85, and the total words are 114. Vocabulary is so easy that its vocabulary density is 0.746. This density enables readers to deduce knowledge that it is most appropriate for kids or beginner level readers due to its easy diction. Poetic lines are small, but the sense of an idea finishes after some verses, that is why, it has 19 words per sentence, and such poetic structure is termed enjambment.

ii. Cirrus

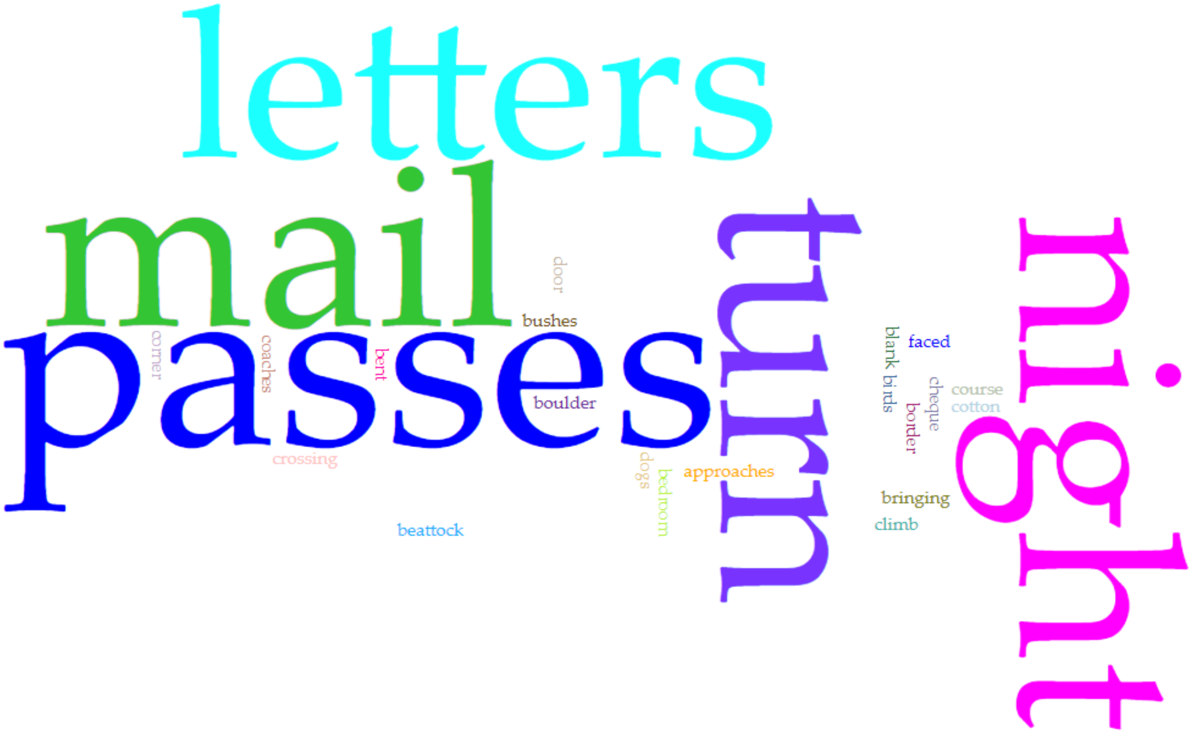
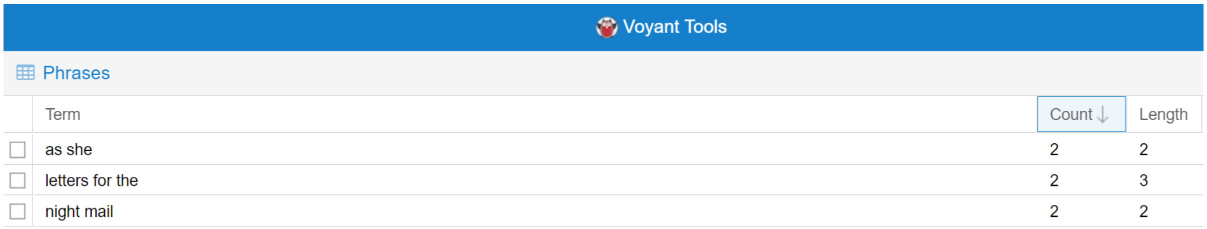


Figure 116 Cirrus, Night Mail

"In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). Its major themes are "night (2)", "mail (2)", "letters (2), "turn (2)" and "passes (2)". Night mail carries different sorts of letters, including business mail and love letters for different types of recipients. The movement of Night Mail is elaborated with its "turns" and "passes" from different memorable scenographic places of fields and moorland boulders.

Human analysis reveals that no human character is present (Hussain, 2009, pp. 71-72). Cirrus shows these themes "night (2)", "mail (2)", "letters (2), "turn (2)" and "passes (2)".

iii. Phrases



	Term	Count ↓	Length
□	as she	2	2
□	letters for the	2	3
□	night mail	2	2

Figure 117 Phrases, Night Mail

Three bigrams and trigrams have been extracted from this poem: “letters for the” (N+Prep+Art), “night mail” (Adj+N) and “as she” (Conj+Prn). Moreover, the key ideology of the poem is evident that night mail brings letters for people belonging to different genders, professions and age groups to meet their social and business necessities. These collocations/ n-grams occur twice, and their length ranges from 2 to 3 words.

iv. Links

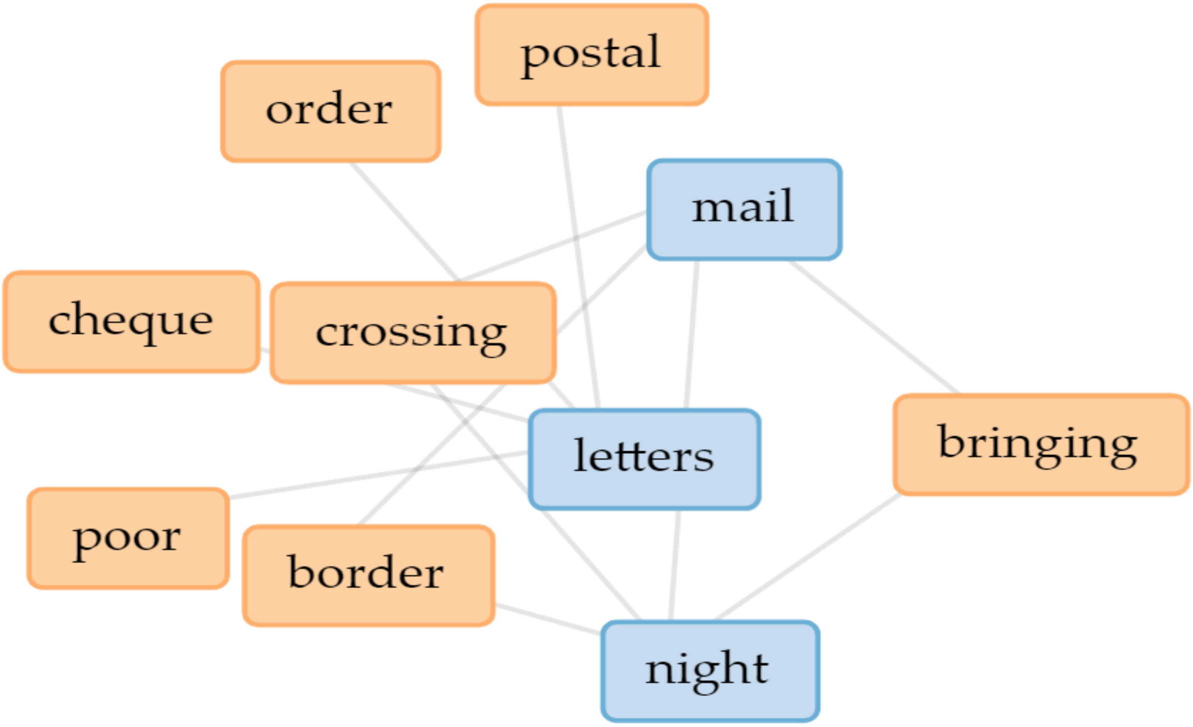
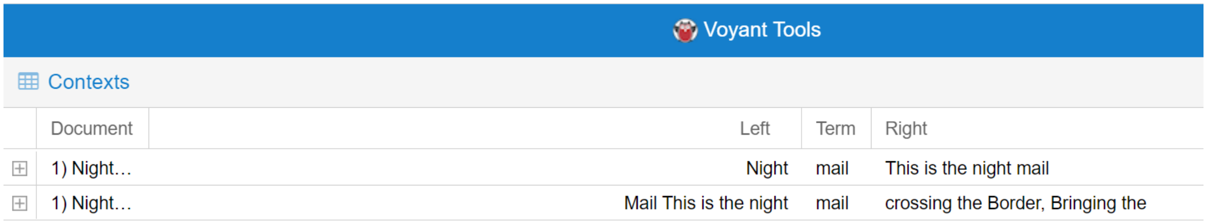


Figure 118 Links, Night Mail

Links tool shows the KG of words “night, letters, mail” which covers the title of the poem too, and this title occurs as a group throughout the poem. Night Mail performs some advantages in the KG of “bringing, mail, letters” since all of them have been interconnected like brain neurons. The act of “crossing” the “border” by Night Mail has been linked in the KG, “night, mail, crossing, border”. This technique interconnects all persons, events and movements of Night Mail correctly. Apart from its authenticity, one deficiency is also visible that Links tool separates one meaningful word “postal order” and considers it two separate words. Machines find “postal” and “order” as two separate words. The same possibility of shortcomings has been mentioned in Hermeneutica that tools “fail in interesting ways” (Rockwell, & Sinclair, 2016, p. 166); therefore, these failures suggest new tool developments through improved coding.

v. Contexts



	Document	Left	Term	Right
⊞	1) Night…	Night	mail	This is the night mail
⊞	1) Night…	Mail This is the night	mail	crossing the Border, Bringing the

Figure 119 Contexts, Night Mail

In this poem, lexical ambiguity prevails whether the word “mail” is a noun or a verb. To disambiguate the word sense, the context of the key word “mail” is retrieved in figure 119. The title of the poem, *‘Night Mail’* has been used twice in the poem, and every time the word “mail” is used as a common noun.

3. *Loveliest of Trees, the Cherry Now* by A. E. Housman

i. Summary

This corpus has 1 document with 80 total words and 56 unique word forms. Created now (on 28th August 2017).

Vocabulary Density: 0.700

Average Words Per Sentence: 26.7

Most frequent words in the corpus: cherry (3); bloom (2); fifty (2); hung (2); loveliest (2)

Figure 120 Summary, Loveliest of Trees, the Cherry Now

The unique words of the poem are 56, and the total words are 80. When unique words are divided by total words, 0.700 vocabulary density is derived. Its low vocabulary density expresses its beginner level ease and the use of simple diction in the poem. So, 26.7 average words per sentence suggest that a full stop is placed at the end of a quatrain, not after each line. That is why, sentences are long, and it is termed enjambment. The most frequent words encompass the title of the poem and the loveliest look of cherry and its bloom.

ii. Cirrus

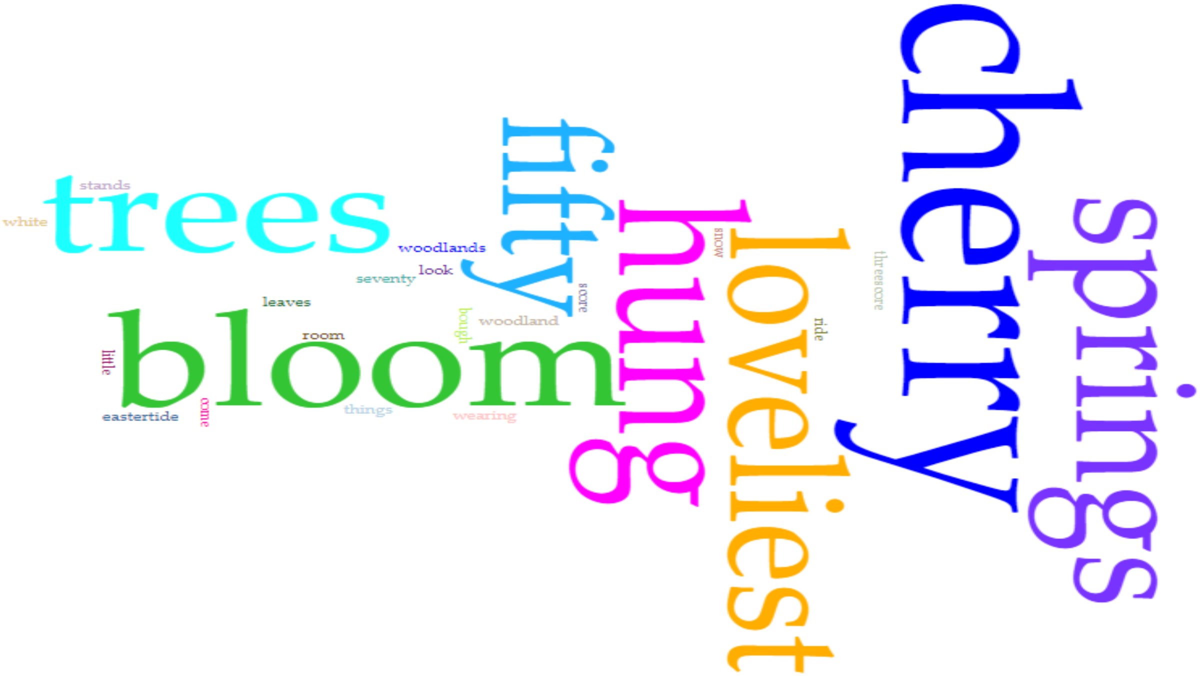
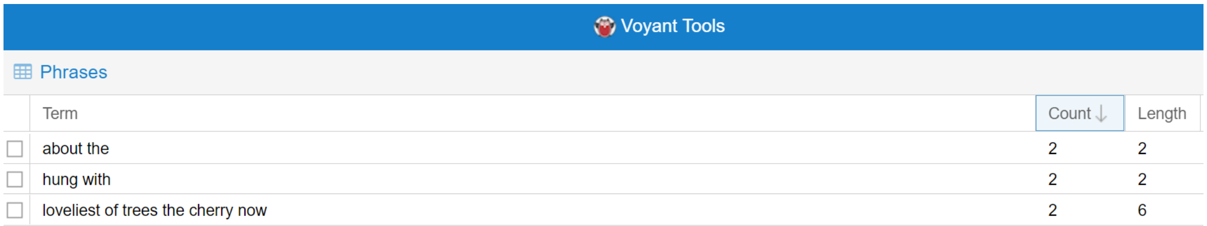


Figure 121 Cirrus, Loveliest of Trees, the Cherry Now

This poem discusses character and themes of “cherry (3)”, its “bloom (3)”, “springs (2)”, “trees (2)”, “woodland (1)”, “loveliest (2)”, “snow (1)” and just in one visual, the whole thematic image has been presented. The beauty of vegetation in “spring (2)” is at its peak, and it has been conveyed with the word “loveliest (2)”. Likewise, all words denote natural beauty, plantation and its bewitching magical effect for beholders of aesthetic sense. Besides, the Christian religious theme of “Eastertide (1)” and “trees (2)” have been manifested in the poem.

Human analysis shows that no human character is present in the poem (Hussain, 2009, pp. 71-72). Cirrus tool shows inhuman characters of “cherry (3)”, and “trees (2)”. Key themes of the poem are “bloom (3)”, “springs (2)”, “trees (2)”, “woodland (1)”, “loveliest (2)”, “snow (1)”. To conclude, machine analysis is more comprehensive, precise and accurate, and nothing can be overlooked, while human cognition ignores some thematic aspects.

iii. Phrases



Voyant Tools

Phrases

	Term	Count ↓	Length
☐	about the	2	2
☐	hung with	2	2
☐	loveliest of trees the cherry now	2	6

Figure 122 Phrases, Loveliest of Trees, the Cherry Now

The poem title phrase "loveliest of trees the cherry now" (Adj+Prep+N+Art+N+Adv) has been repeated twice. The collocation/ n-gram "hung with" (V+Prep) expresses the correct use of the preposition. They occur twice, and their length varies from 2 to 6 words.

iv. Links

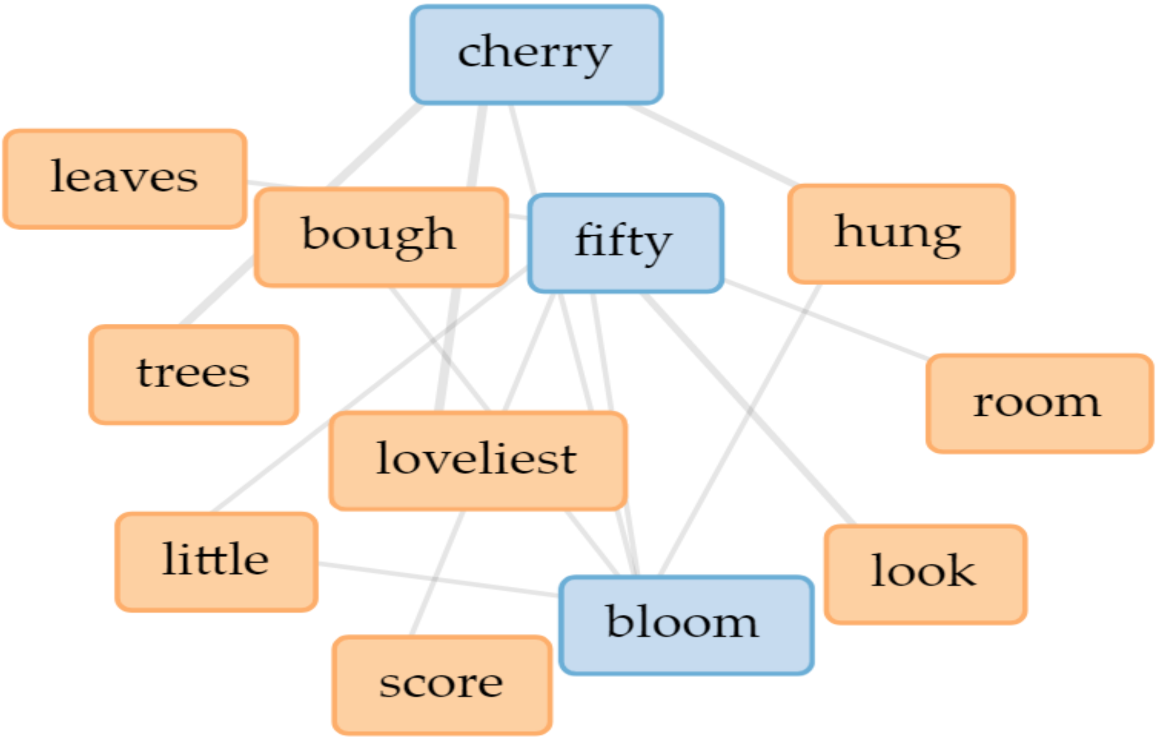
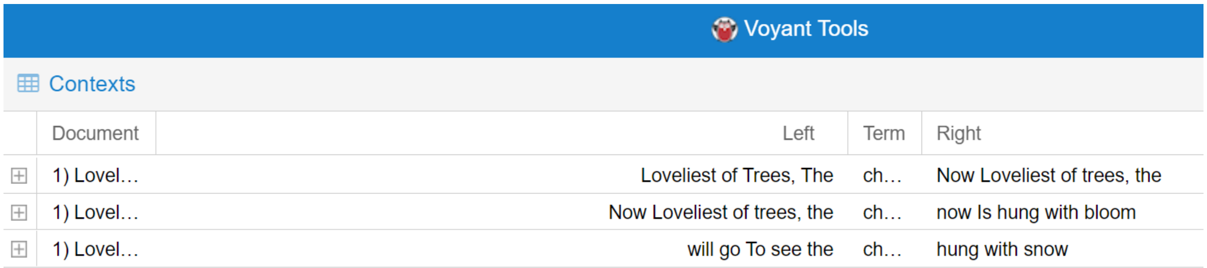


Figure 123 Links, Loveliest of Trees, the Cherry Now

This KG uses colour coding and sky-blue words are primary words and more frequent, while orange words are less occurring in the corpus. The KG of "bloom, hung, cherry, bough, trees" builds such a relationship which is quite evident with the physical appearance of the cherry tree. In another KG, "trees, Cherry, loveliest" is linked with "loveliest" to validate the knowledge pattern regarding the title of the poem. One postulate of the theory informs that "Manipulation is in service of exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166). When different nodes are linked with other themes, they create multifaceted hermeneutic patterns. These knowledge graphs can also be extended for the exploration of new semantic and hermeneutic linkages. Thus, data visualization also changes with the change of inquiry.

Straight-line KG of "little, fifty, room" suggests a short span of fifty years to relish the beauty of cherry trees. Another KG of "fifty, score" is also connected to the span from which a score (20 years) has been subtracted from the average human age of seventy years, and the remaining age is "fifty" springs which are very brief to adore beauteous features of cherry. Natural beauty is abundant and numerous as compared to a limited human life span.

v. Contexts



Document	Left	Term	Right
1) Lovel…	Loveliest of Trees, The	ch…	Now Loveliest of trees, the
1) Lovel…	Now Loveliest of trees, the	ch…	now Is hung with bloom
1) Lovel…	will go To see the	ch…	hung with snow

Figure 124 Contexts, Loveliest of Trees, the Cherry Now

Grammatically, cherry can be used as an adjective or noun. The point of inquiry is that how it has been used in this poem. Theoretical underpinning guides thus: "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). So, this lexical ambiguity is resolved by Contexts tool which follows information retrieval for the word "cherry". The word "cherry" is found three times in this poem, once in the title of the poem, and twice within the poem. Context of the title suggests that cherry is the loveliest of trees, and it is hung with a bloom laden with snow. Moreover, Contexts tool clarifies the searched theme and finds how the poet expands its thematic title, *'Loveliest of Trees, the Cherry Now'.* The current study about WSD is aligned with Kwary (2018) in chapter 2.

4. *O Where Are You Going?* By W.H. Auden

i. Summary

This corpus has 1 document with 135 total words and 80 unique word forms. Created now (on 28rh August 2017).

Vocabulary Density: 0.593

Average Words Per Sentence: 19.3

Most frequent words in the corpus: said (6); farer (2); fearer (2); going (2); hearer (2)

Figure 125 Summary, O Where Are You Going?

One postulate of the theory, “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). So, human cognition and understanding derive stylometric features from this corpus summary. The stylometric features of one poet are different from other poets; thus, it becomes a thumb sign like the quality of a writer. Furthermore, these features facilitate the author's identity of any text. Moreover, they can also be used in forensic linguistics. The unique words of this poem are 80, and the total words are 135. By dividing unique words with total words, vocabulary density is generated. Its vocabulary density suggests that a new word occurs after 0.593 words. In this poem, a thought ends after a quatrain; hence, a full stop is placed after the quatrain. That is why its sentences are long, and on average, each sentence consists of 19.3 words. Each poetic line consists of a few words, but in one stanza, two participants utter their dialogues in a single stanza.

ii. Cirrus

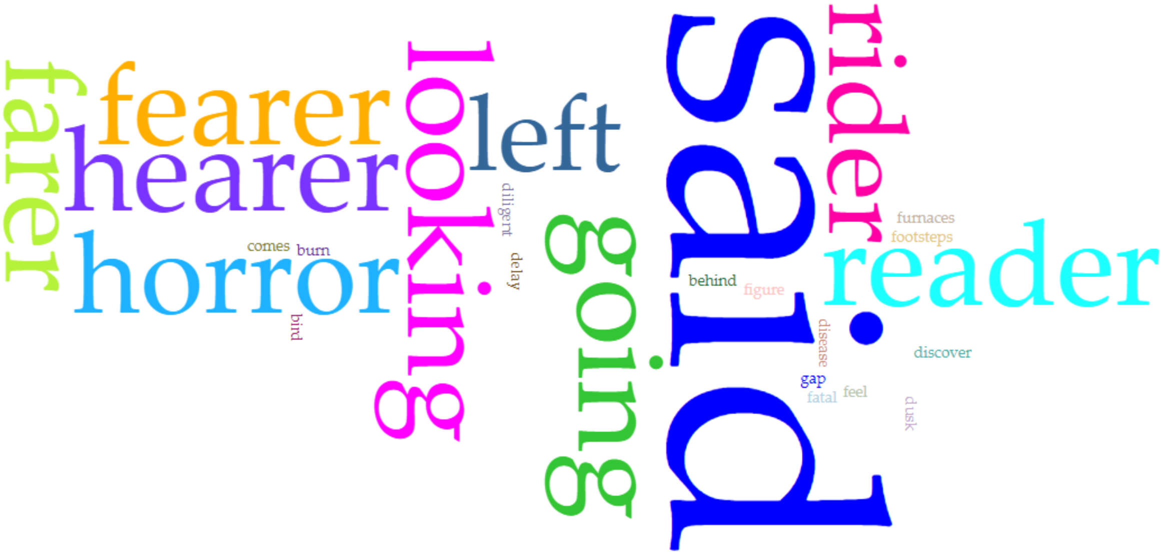
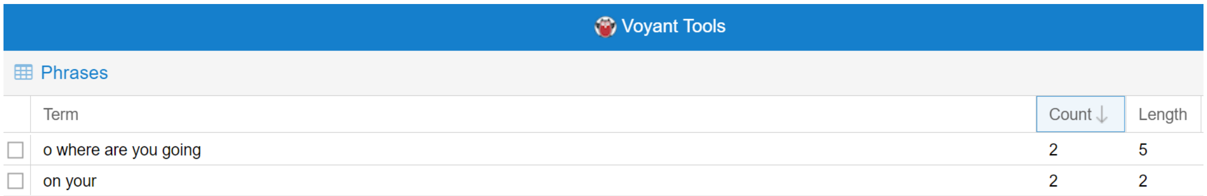


Figure 126 Cirrus, O Where Are You Going?

Topic modelling is the basic task of word clouds; therefore, this Cirrus shows six major conversation partners who have participated in this conversational poem. These characters are “reader (2)”, “rider (2)”, “horror (2)”, “hearer (2)”, “fearer (2)” and “farer (2)”. Knowledge Discovery Theory is defined as “the extraction of implicit, previously unknown and potentially useful information from data” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). The word “said” occurs six times, and it clarifies that characters are saying and talking to one another in this poem. This poem is in a dialogic form like Allama Iqbal’s Urdu poem “PAHAR OR GULEHRE” (Iqbal, 1924). Besides, gloomy and pessimistic themes of the poem are apparent through these words: “fearer (2)”, “furnace (1)”, “burn (1)”, “shocking (1)”, “left (2)”, “grave (1)”, “horror (2)”, “disease (1)” and “odors (1)”.

Human analysis shows the nonexistence of male and female characters. (Hussain, 2009, pp. 71-72). Cirrus tool shows key characters with their statistical weight, for instance, “reader (2)”, “rider (2)”, “horror (2)”, “hearer (2)”, “fearer (2)” and “farer (2)”. Cirrus tool conducts a deeper analysis of themes than human analysis, for instance, themes of “furnace (1)”, “burn (1)”, “shocking (1)”, “left (2)”, “grave (1)”, “horror (2)”, “disease (1)” and “odors (1)” have been mentioned.

iii. Phrases



	Term	Count ↓	Length
□	o where are you going	2	5
□	on your	2	2

Figure 127 Phrases, O Where Are You Going?

The title phrase of the poem, “O where are you going” (Int+Adv+Aux+Prn+V) and “on your” (Prep+Prn) have been repeated twice, and their length is five words and two words respectively.

iv. Links

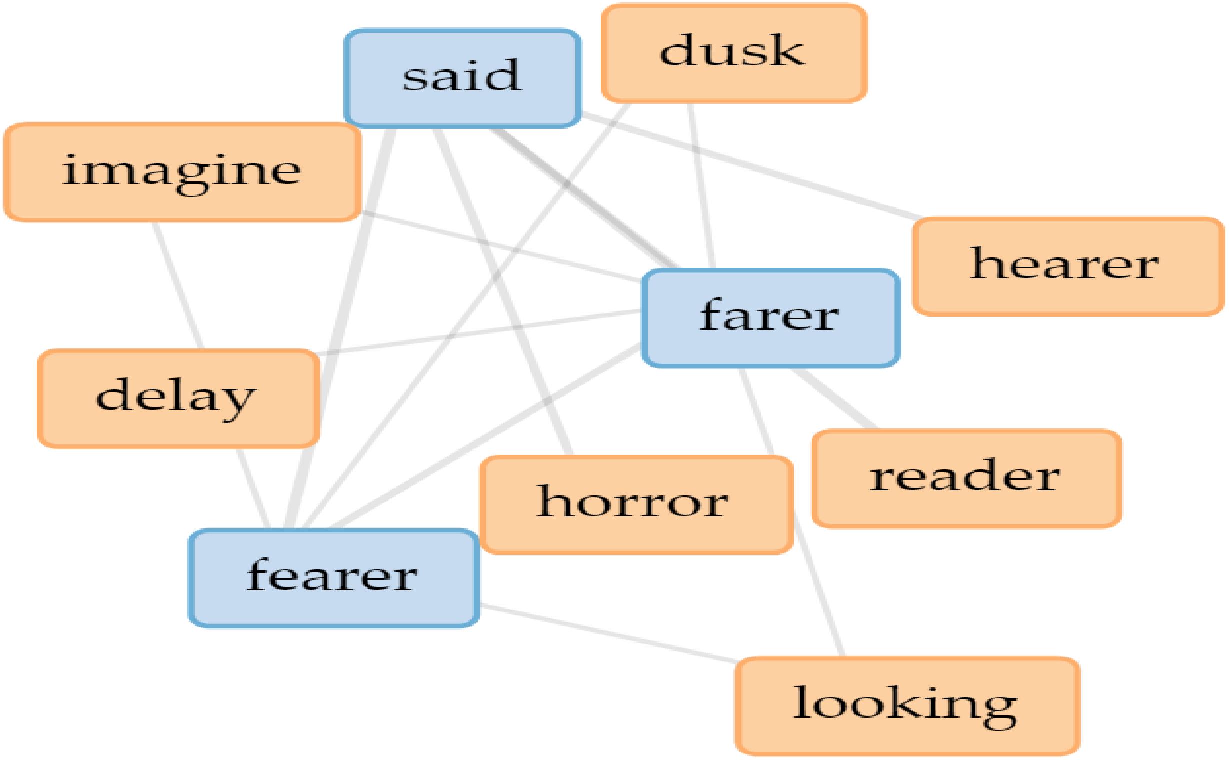
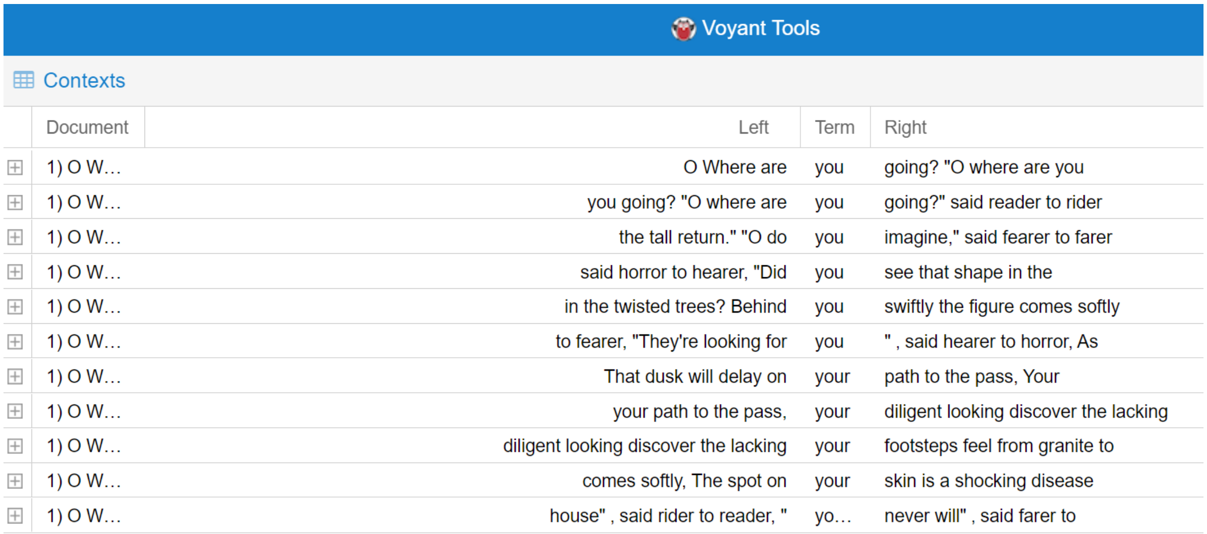


Figure 128 Links, O Where Are You Going?

The word "said" is in the centre which informs about the dialogic nature of the poem, that is why KG, "fearer, farer, said" was drawn. The KG of "hearer, said, horror" has been linked through their poetic conversation. One KG, "reader, said, fearer" reveals that reading makes a person law abiding and fearful to take any wrong step. The KG of "dusk, farer, delay" is also causing misery and trouble for the travellers of untrodden destinations. The word "looking" is also referring to find some faults. The fearer raises the question about "imagine"; consequently, the KG of "imagine, fearer" has been connected.

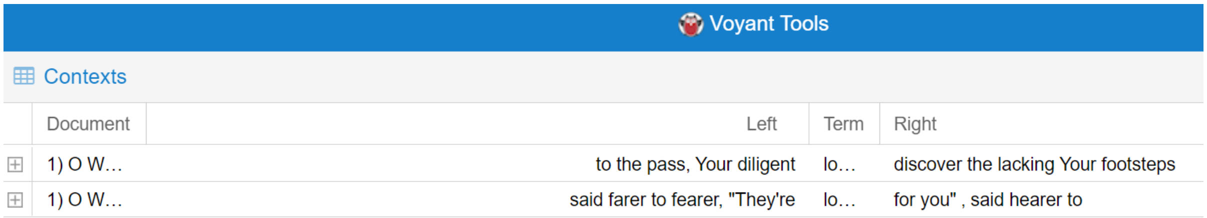
v. Contexts



	Document	Left	Term	Right
⊞	1) O W…	O Where are	you	going? "O where are you
⊞	1) O W…	you going? "O where are	you	going?" said reader to rider
⊞	1) O W…	the tall return." "O do	you	imagine," said fearer to farer
⊞	1) O W…	said horror to hearer, "Did	you	see that shape in the
⊞	1) O W…	in the twisted trees? Behind	you	swiftly the figure comes softly
⊞	1) O W…	to fearer, "They're looking for	you	" , said hearer to horror, As
⊞	1) O W…	That dusk will delay on	your	path to the pass, Your
⊞	1) O W…	your path to the pass,	your	diligent looking discover the lacking
⊞	1) O W…	diligent looking discover the lacking	your	footsteps feel from granite to
⊞	1) O W…	comes softly, The spot on	your	skin is a shocking disease
⊞	1) O W…	house" , said rider to reader, "	yo…	never will" , said farer to

Figure 129 Contexts, O Where Are You Going?

Deictic pronouns usually create ambiguity because of their reference in the poem. The human mind cannot retain all referents just after one reading, especially if a poem is dialogic. So, Contexts tool disambiguates true semantic shades and accurate interpretation. The deictic pronoun “you” has been used 11 times in the poem (figure 129), and it can be easily understood through the Contexts tool that who is the addresser and who is the addressee.



	Document	Left	Term	Right
⊞	1) O W…	to the pass, Your diligent	lo…	discover the lacking Your footsteps
⊞	1) O W…	said farer to fearer, "They're	lo…	for you" , said hearer to

Figure 130 Contexts, O Where Are You Going?

Another lexical ambiguity is present with the word "looking" which has been used twice: the first time it is used as a noun, and the second time it is used as a verb. So, the difference between a noun and a verb can also be taught and learnt with Contexts tool (figure 130). When a part of speech has been determined, its semantic shade is easy to decide. In the 2nd chapter, Bhala, & Abirami (2014) have already discussed it. Finding parts of speech and true semantic shade is a process of knowledge discovery and WSD.

5. *In the Street of Fruit Stalls* by Jan Stallworthy

i. Summary

This corpus has 1 document with 92 total words and 64 unique word forms. Created about a minute ago (on 29th August 2017).

Vocabulary Density: 0.696

Average Words Per Sentence: 30.7

Most frequent words in the corpus: dark (3); street (3); fruit (2); gold (2); guava (2)

Figure 131 Summary, In the Street of Fruit Stalls

“Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). A small corpus of a total of 92 words has only 64 unique words. Basically, the bigger corpus one builds, the better, uniformed and comprehensive results one generates. This poem is very small, so its corpus results are also very limited. The unique words of this poem are 64, and the total number of words of the poem are 92. Vocabulary density 0.696 shows very less appearance of new words in the text. It means that beginner level easy vocabulary is given in the poetic text. In reality, there are two quintets/ quintains (five lines in a stanza) and one quatrain in the poem; subsequently, a full stop has been placed at the end of each quintet or quatrain, that is why sentences have 30.7 words per line, and technically it is called enjambment.

ii. Cirrus

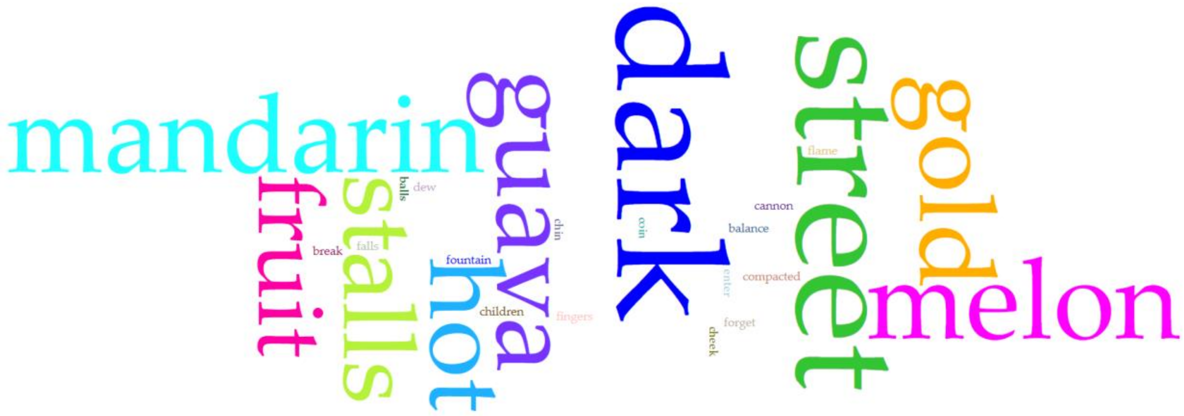


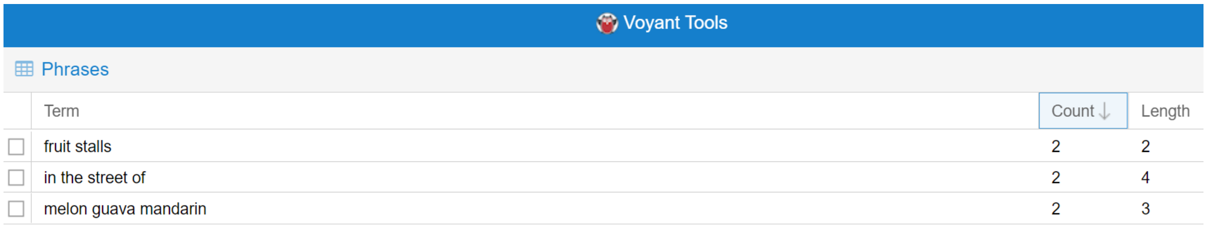
Figure 132 Cirrus, In the Street of Fruit Stalls

Key themes of this poem have been clustered as “street (3)”, “dark (3)”, “fruit (2)”, “melon (2)”, “guava (2)”, “mandarin (2)”, “gold (2)”, and they give an overview that this poem discusses a dark street with a fruit stall vendor who sells melon, guava and mandarin. The theme of “gold” refers to the golden colour of the fruit. Thus, Cirrus elaborates the central idea in the form of colour-coded words.

The themes of “cannon balls (1)” draw viewers' attention that the theme of war is leading within the poem. Furthermore, “pyramid (1)” reference takes us to war-loving Egyptian empire which fought wars during ancient, Ottoman and modern eras against its own people, Ottoman Empire, England, Israel, Ethiopia and Libya (Williams, 2008).

Human analysis reveals that the characters of some children have been portrayed in the poem (Hussain, 2009, p. 71). Cirrus does not show children as its human characters because of their least occurrence. Cirrus extracts certain themes, for instance, “street (3)”, “dark (3)”, “fruit (2)”, “melon (2)”, “guava (2)”, “mandarin (2)”, “gold (2)” and they can be validated from the poetic text.

iii. Phrases



	Term	Count ↓	Length
☐	fruit stalls	2	2
☐	in the street of	2	4
☐	melon guava mandarin	2	3

Figure 133 Phrases, In the Street of Fruit Stalls

Three common phrases “in the street of” (Prep+Art+N+Prep), “fruit stalls” (Adj+N) and three fruit names, “melon, guava, mandarin” (N+N+N) are found. Another dimension is that many more new collocations/ n-grams can be constructed on the pattern of standard phrase (“in the street of”), for instance, in the home of, in the shop of. Similarly, bigram “fruit stalls” can be replicated with other newly coined bigrams, for instance, vegetable stalls, crockery stalls and dry fruit stalls. To conclude, Phrases tool not only extracts standard collocations/ n-grams, but also helps in coining new bigrams, trigrams/ quadgrams. Their occurrence is twice, but their length ranges from 2 to 4 words.

iv. Links

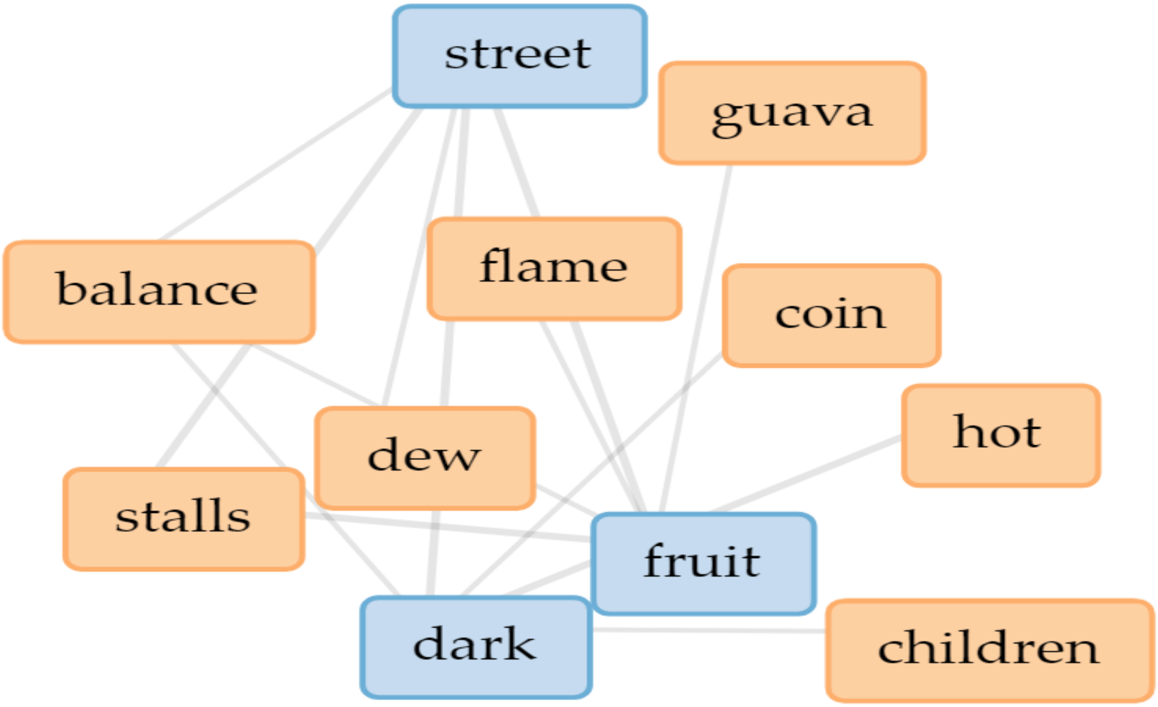
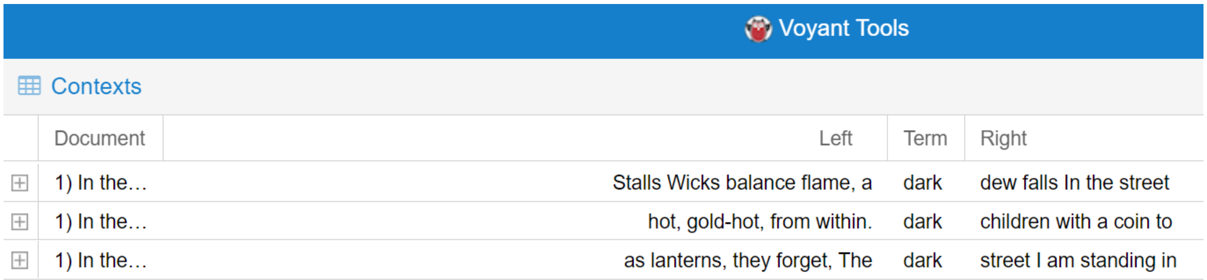


Figure 134 Links, In the Street of Fruit Stalls

The KG of “Street, dark, flame” exhibits that the street has been enveloped with pitch-black darkness; therefore, a flame is lit to lessen the intensity of the murky street. One more KG of “flame, dark, balance” informs about balancing of flame with wicks, and it is also happening with the light of fruit stall. Another KG links “stalls, fruit, guava” because fruit and guava have a whole part relationship, and these fruit are placed on fruit stalls.

The KG nodes “children, dark, coin” reveal a knowledge pattern about children who go to fruit stalls in the darkness, and they pay coins to buy different types of fruit. From the straight line of KG, a sub-node is “dark, children” which portrays the dirty blackish poverty-stricken appearance of the children.

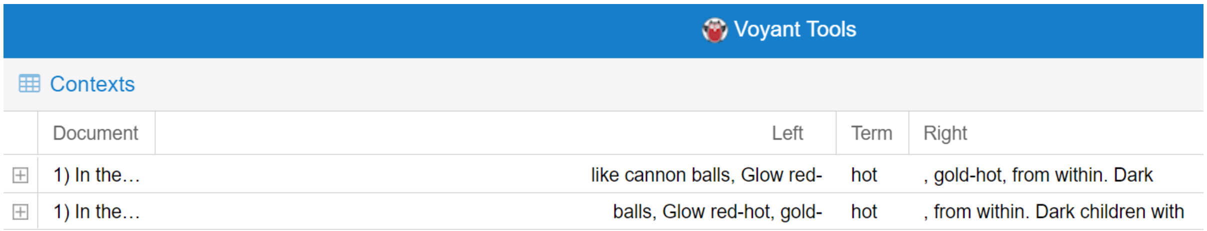
v. Contexts



	Document	Left	Term	Right
⊞	1) In the…	Stalls Wicks balance flame, a	dark	dew falls In the street
⊞	1) In the…	hot, gold-hot, from within.	dark	children with a coin to
⊞	1) In the…	as lanterns, they forget, The	dark	street I am standing in

Figure 135 Contexts, In the Street of Fruit Stalls

The word "dark" in context shows three uses as "dark dews", "dark children" and "dark street", and they testify that dark is an adjective; hence, Contexts tool facilitates in determining different parts of speech of the same grapheme to remove lexical ambiguity. Theoretical support strengthens the use of this technique thus, "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166).



	Document	Left	Term	Right
⊞	1) In the…	like cannon balls, Glow red-	hot	, gold-hot, from within. Dark
⊞	1) In the…	balls, Glow red-hot, gold-	hot	, from within. Dark children with

Figure 136 Contexts, In the Street of Fruit Stalls

There is also a semantic ambiguity about the word "hot", since it has literal and figurative semantic shades that which thing is hot and what sort of hot is meant by the poet. To disambiguate, Contexts tool retrieves context in figure 136. It reveals that the word "hot" does not mean heated, but it exemplifies the red and gold colours of fruit as a figurative meaning.

6. *A Sindhi Woman* by Jan Stallworthy

i. Summary

This corpus has 1 document with 69 total words and 56 unique word forms. Created now (30th August 2017).

Vocabulary Density: 0.812

Average Words Per Sentence: 34.5

Most frequent words in the corpus: bare (1); bazaar (1); beneath (1); blown (1); cloth (1)

Figure 137 Summary, A Sindhi Woman

Among the first six poems, this poem has 0.812 vocabulary density because it does not repeat its words except stop words. It is evident in the most frequent word list that all words occur only once. It also suggests quantifiable stylometric features of Stallworthy who uses diversified diction in this poem. His stylistic quality is evident by the use of 56 unique words from 69 total words. It unveils a knowledge pattern that only stop words have been repeated, but no content word has been repeated. This fact can be verified by another Phrases tool which shows that no repeated phrase is present in this poem. The average sentence length is 34.5 words because the poet marks a full stop after each sestet, and this is termed enjambment

ii. Cirrus

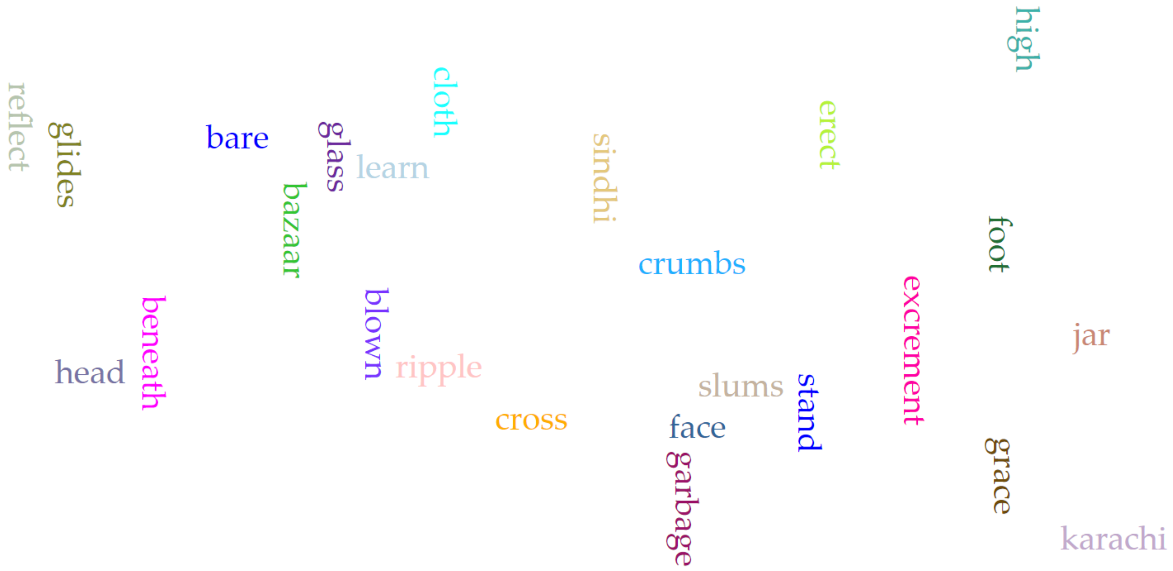


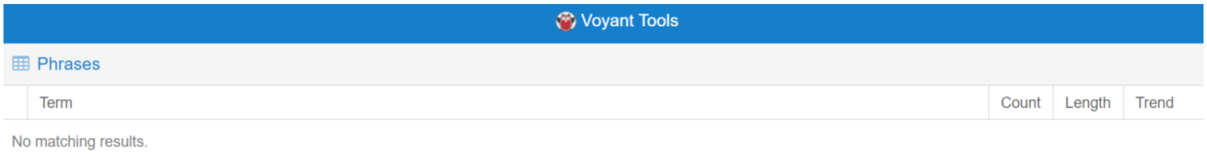
Figure 138 Cirrus, A Sindhi Woman

All words of the poem are used once; that is why they are in the same size and at the same distance. The rule is that size of a word increases by its statistical weight in any corpus. Quantification or statistical value is equal; that is why they are in equal size. The central character of the poem is a Sindhi vendor woman. Main themes occur once in this corpus. At first glance, the Cirrus portrays destitute conditions with “garbage (1)”, “stone (1)”, “glass (1)”, “slums (1)”, “bare (1)”, “excrement (1)” and “crumbs (1)”.

Human analysis shows that a Sindhi working woman is carrying a stone jar on her head. Moreover, she is so poor that she walks bare footed. Key themes are stones, garbage, excrement, crumbs of glass, undulant grace, glide with the stone jar and ripple in her tread (Hussain, 2009, pp.

71-73). Comparing both analyses, common themes are glass, bare, crumbs, whereas themes of excrement and slums have been extracted by only Cirrus.

iii. Phrases



Voyant Tools

Phrases

Term	Count	Length	Trend

No matching results.

Figure 139 Phrases, A Sindhi Woman

"Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166), even if no result has been shown, for example, this poem does not show any phrase or collocation pattern/ n-gram because, except stop words or function words, no two words have been repeated in this corpus. Another knowledge discovery is that the poet uses diversified diction that he even does not repeat the title in the poem.

iv. Links

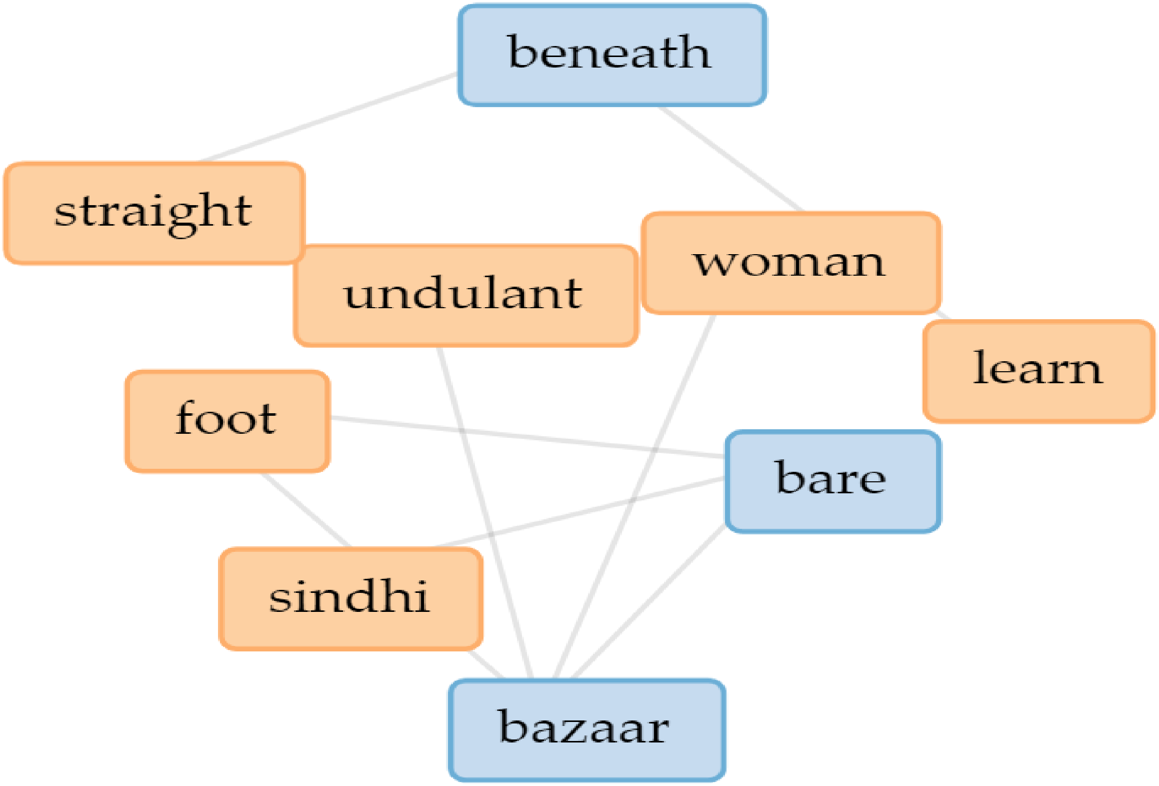
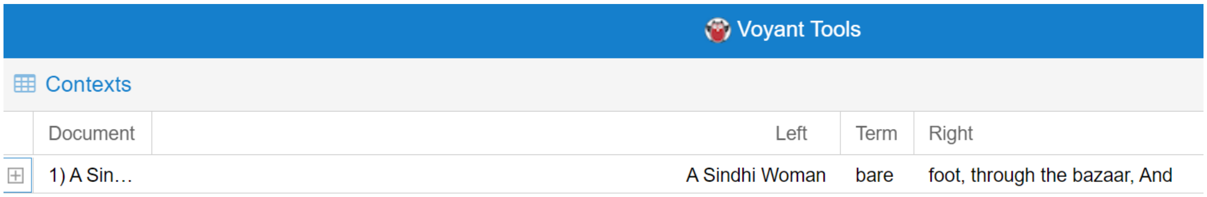


Figure 140 Links, A Sindhi Woman

The triangular KG of "Sindhi, bare, foot" depicts a working woman, and this KG suggests her miserable plight and impoverished predicament. "Manipulation is in service of exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166). Data visualization provides different nodes for further exploration, and these knowledge graphs can be extended for further elaboration. Another KG of "Woman, bazaar, Sindhi" shows that she is a working woman, and she walks through the bazaar to sell commodities placed on her head in a big jar. One more KG of "Sindhi, bare, foot" refers to her graceful walk despite her bare feet. This KG discovers a knowledge pattern that the poet wants to express that grace has no connection with riches; rather a barefoot walk is also graceful. Bashfulness, chastity, good moral bearing, self-reliance and hard work give elegance and grace to her gait. Besides, the word "learn" indicates that there is a moral lesson in the poem. Therefore, the KG of "straight, beneath, woman, learn" is the underpinning of the poem that those who learn to walk under the weight of sufferings, social and financial challenges, they stand straight and succeed in the neck-breaking competitive world.

v. Contexts



Document	Left	Term	Right
1) A Sin…	A Sindhi Woman	bare	foot, through the bazaar, And

Figure 141 Contexts, A Sindhi Woman

The word "bare" has different semantic shades, for instance, without clothes, most basic, minimum and empty. To explore true meaning, information is retrieved for key word "bare". Figure 141 shows that the phrase "bare foot" occurs once, and it clarifies that Sindhi working woman is not wearing shoes due to her absolute poverty.

7. *Times* (From Ecclesiastes, 3, 1-12)

i. Summary

This corpus has 1 document with 122 total words and 44 unique word forms. Created now (on 30th August 2017).

Vocabulary Density: 0.361

Average Words Per Sentence: 122.0

Most frequent words in the corpus: time (21); away (2); cast (2); stones (2); born (1)

Figure 142 Summary, Times

Its unique words are 44, and almost every word has been repeated three times in this corpus. Its vocabulary density is 0.361, and it refers to the use of repeated and easy words in this poem. Average words per sentence are 122; therefore, it means that a full stop is placed at the end of the poem, and the whole poem comprises only one sentence; hence, it is called enjambment.

ii. Cirrus

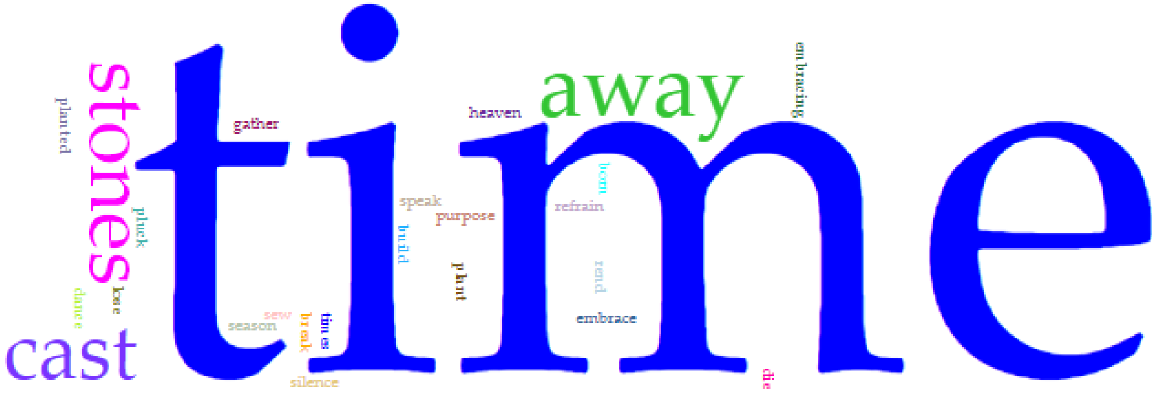
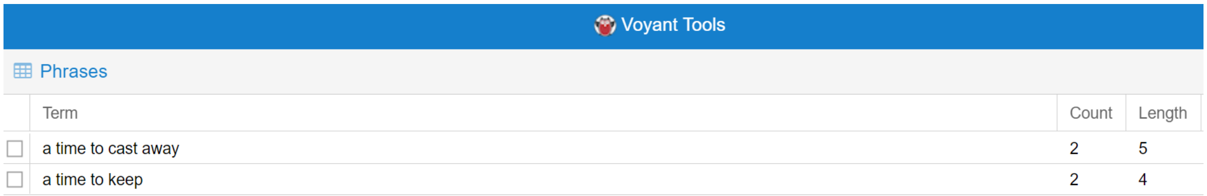


Figure 143 Cirrus, Times

The title and central theme of "time (21)" is the most dominating feature. Time must be obeyed in various phases and situations of life to gain acme. Actions after due time bear no fruits for doers of action. The phrase "cast away stones (2)" occurs twice in the corpus. Then several other topics, for instance, "pluck (1)", "born (1)", "break (1)" and "lose (1)" occur once to exhibit minor topics of the discussion.

Thesis and antithesis, for example, plant and pluck; born and die; have been delineated in the poem. It discovers the contrasting style of the poem in which opposing words have been juxtaposed to show minute differences.

iii. Phrases



	Term	Count	Length
□	a time to cast away	2	5
□	a time to keep	2	4

Figure 144 Phrases, Times

Only two standard phrases are found, for instance, “a time to cast away” (Art+N+Inf V+Prep) and “a time to keep” (Art+N+Inf V). By rule, the article is not used before time, but the Bible translator uses the supremacy of poetic license to write an article before time. This knowledge pattern is revealed by collocation patterns/ n-grams. Moreover, it suggests that the positive flow of actions or advantages of punctuality and anti-flow of actions/ harms of unpunctuality have been discussed in this poem.

iv. Links

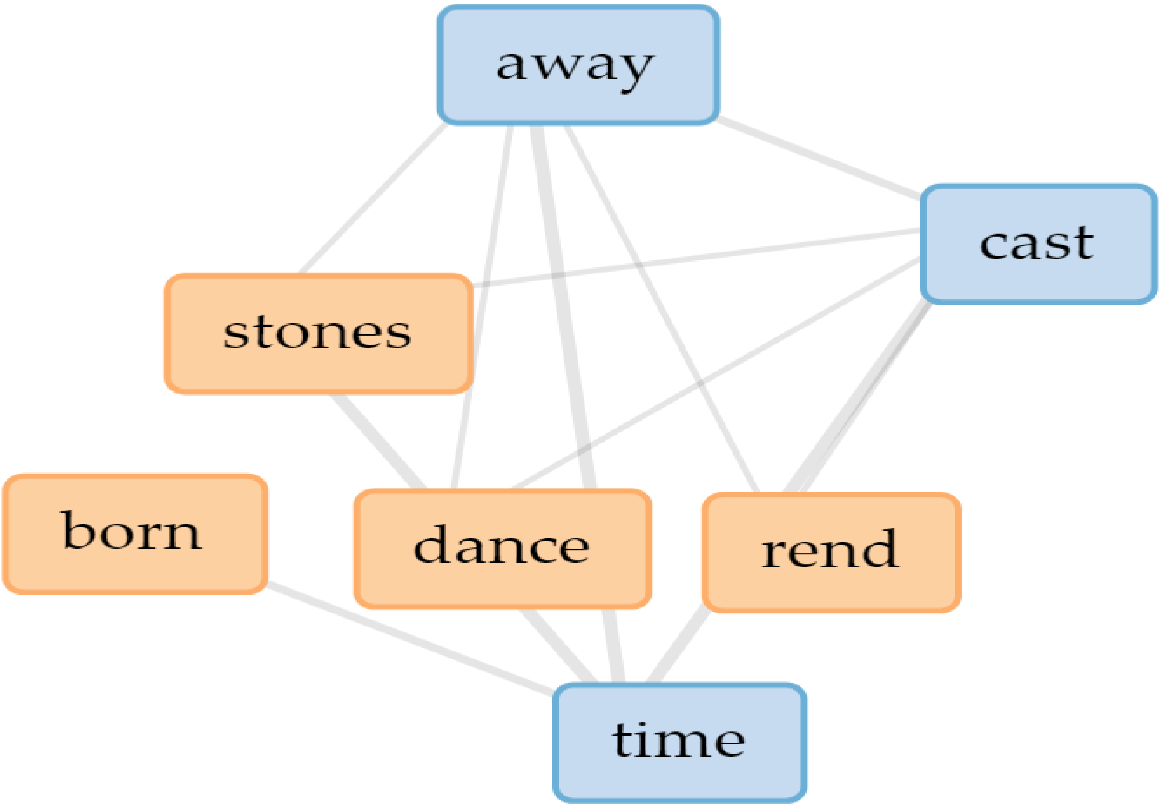
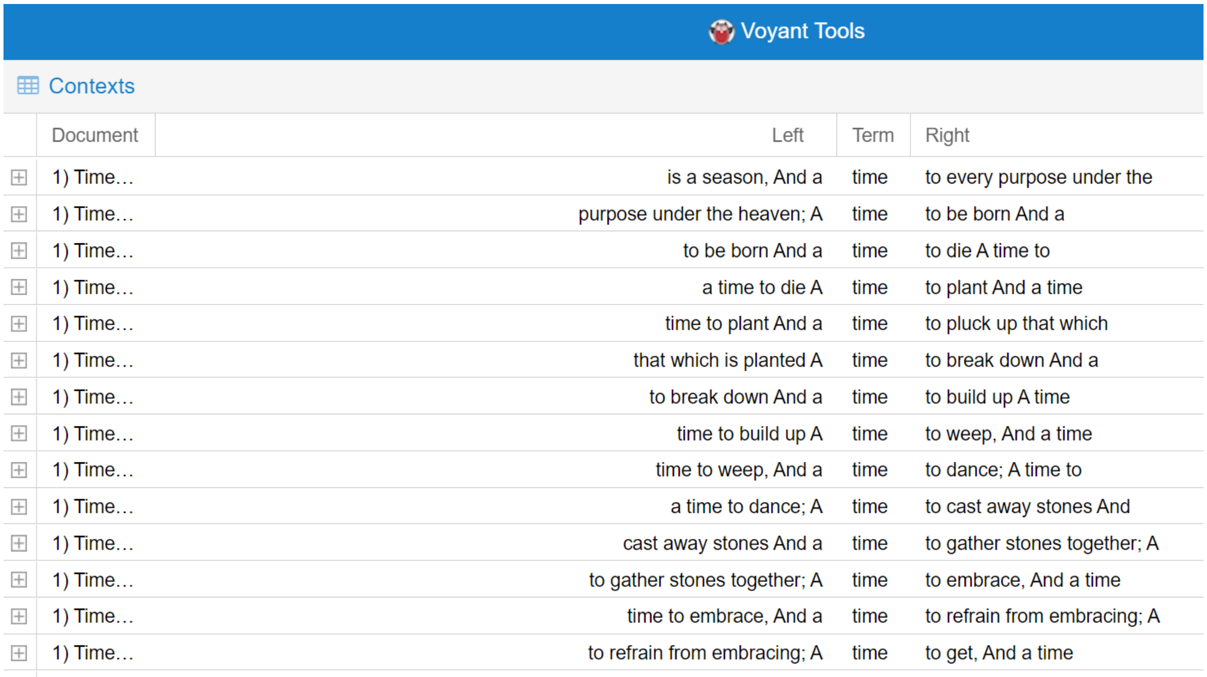


Figure 145 Links, Times

KGs are found in straight lines and in multi-pronged form too. One multi-pronged KG “time, stones, cast, away” refers to the poetic line, “A Time to cast away stones” which is shown in the sky-blue colour to exhibit the most occurring words. Consequently, blue coloured words are the most occurring words and one-time occurring words are in orange colour. The poetic line, “A time to be born” and a KG proves it by linking “time, born” with a straight-line node. Another poetic line is, “A time to rend”, consequently, the KG of “time, rend” is joined by filtering the word “cast” which is in the middle. It is done in the light of the filtering process and Hermeneutica

Theory directs that "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166). Human reflection and cognition guide about filtering and zooming process. The KG "time, dance" interconnects and proves the text quote, "A time to dance".

v. Contexts



	Document	Left	Term	Right
⊞	1) Time…	is a season, And a	time	to every purpose under the
⊞	1) Time…	purpose under the heaven; A	time	to be born And a
⊞	1) Time…	to be born And a	time	to die A time to
⊞	1) Time…	a time to die A	time	to plant And a time
⊞	1) Time…	time to plant And a	time	to pluck up that which
⊞	1) Time…	that which is planted A	time	to break down And a
⊞	1) Time…	to break down And a	time	to build up A time
⊞	1) Time…	time to build up A	time	to weep, And a time
⊞	1) Time…	time to weep, And a	time	to dance; A time to
⊞	1) Time…	a time to dance; A	time	to cast away stones And
⊞	1) Time…	cast away stones And a	time	to gather stones together; A
⊞	1) Time…	to gather stones together; A	time	to embrace, And a time
⊞	1) Time…	time to embrace, And a	time	to refrain from embracing; A
⊞	1) Time…	to refrain from embracing; A	time	to get, And a time

Figure 146 Contexts, Times

"Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). This postulate elucidates that context disambiguates word sense and provides hermeneutic support by compiling the most relevant information on one topic. Likewise, these apparently similar terms have different usages. The most repeated word "time" has been used in different contexts in the poem. Contexts tool also facilitates compiling a sequential progression of any theme, for instance, a note can be written on any topic or title. An interesting knowledge pattern is that the article "a" has been used 14 times before "time", on the other hand, English grammar forbids the use of articles before an abstract noun. To disambiguate the issue, the poetic license permits poets to use non-grammatical language in poems.

8. *Ozymandias* by P. B. Shelley

i. Summary

This corpus has 1 document with 112 total words and 85 unique word forms. Created now (on 30th August 2017).

Vocabulary Density: 0.759

Average Words Per Sentence: 28.0

Most frequent words in the corpus: ozymandias (2); antique (1); appear (1); away (1); bare (1)

Figure 147 Summary, Ozymandias

The summary tool is a means to discover the stylometry of P. B. Shelley, who is considered an artist in his poems due to the use of diversified vocabulary items. Except function words (stop words) and title Ozymandias, no word has been repeated; that is why unique words are 85, and total words are 112. It means that stop words have been mentioned 25 times, and the poem title has been repeated twice respectively. The same rich vocabulary leads to a higher vocabulary density of 0.759.

Technically, this poem is a 14 lined sonnet. Text mining summary shows that 28 words are written per line on average. It means that one full stop has been placed at the end of this poem, but one dash and one dotted line also show slight pauses. Likewise, the 7th poem, ‘Times’ marks one full stop at the end of the poem and is named enjambment. Practically, several sentences have been constructed in the poem *‘Ozymandias’*. Simultaneously, words “said” and “tell” reveal a knowledge pattern that there are different speakers (king, ancient traveller) who utter different notions in their various sentences.

ii. Cirrus

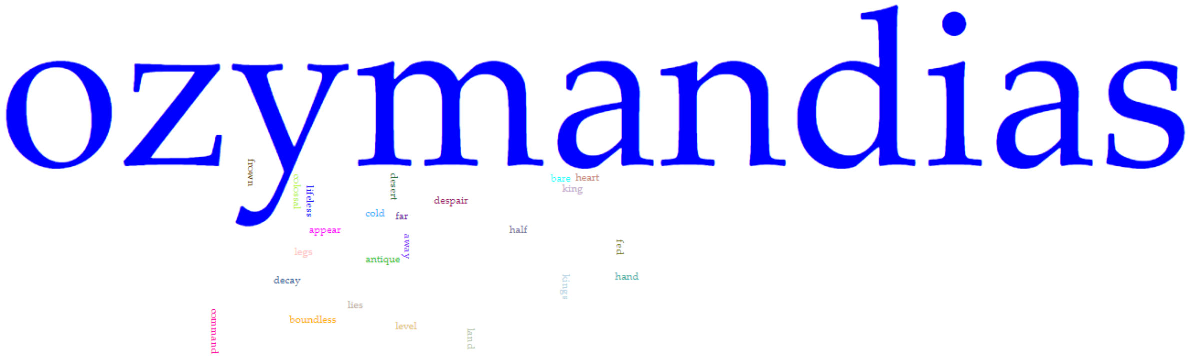
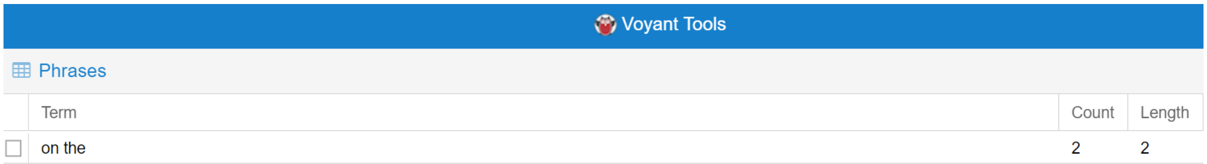


Figure 148 Cirrus, Ozymandias

“In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct” (Agrawal, & Psaila, 1995, p. 1). Cirrus tool mines text and performs the task of topic modelling which is applied to extract key themes and characters from the corpus. Only the king's name "Ozymandias" occurs twice; that is why only this word appears in bold letters, whereas other words are in a very small font, and they scatter lonely in the rectangle of the word cloud. It is Shelley's dexterous style that no word has been repeated in the poem except the title word. Its other topics of discussion are "statue (1)", "lip (1)", "face (1)", "sculptor (1)", "shatter'd (1)" ,"decay (1)" and "lone (1)" which reveal a story of demise and wreck of Ozymandias, the king of kings. Apart from key themes, knowledge patterns of "tell (1)" and "said (1)" express that somebody or traveller from ancient land has reported about the king Ozymandias, his acme and his wretched fall.

iii. Phrases



	Term	Count	Length
□	on the	2	2

Figure 149 Phrases, Ozymandias

One key finding of poems is that very less, or sometimes no repeated phrase is found. In this poem, only bigram "on the" (Prep+Art) collocation pattern/ n-gram is found twice, and it is not a standard collocation because it is devoid of meanings. It also informs that the poet uses a variety of dictions in this poem.

iv. Links

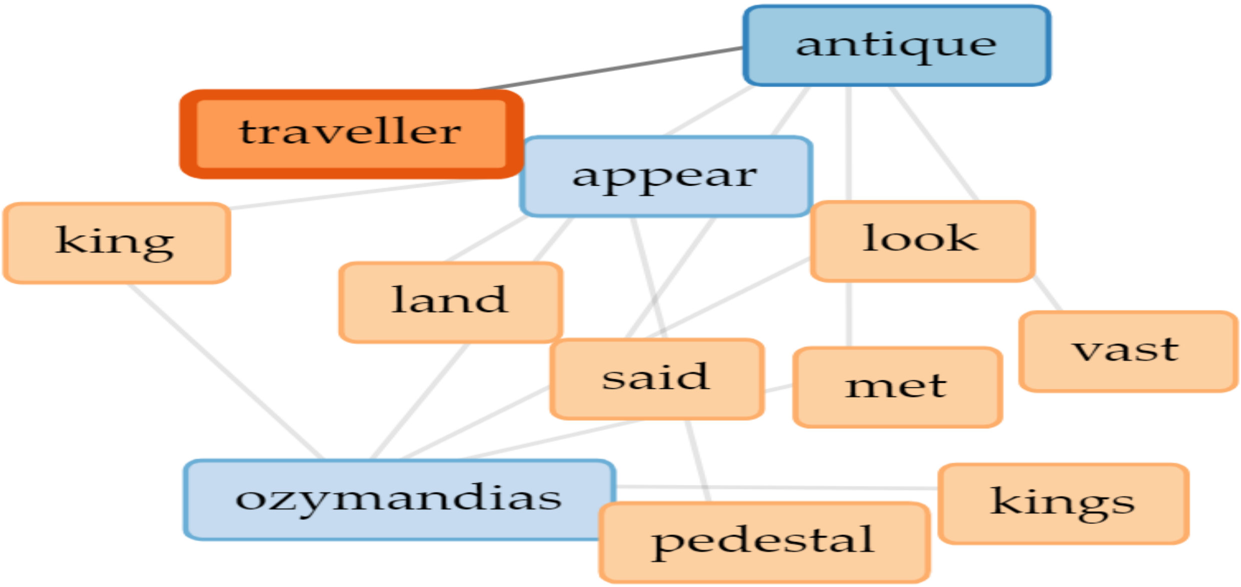
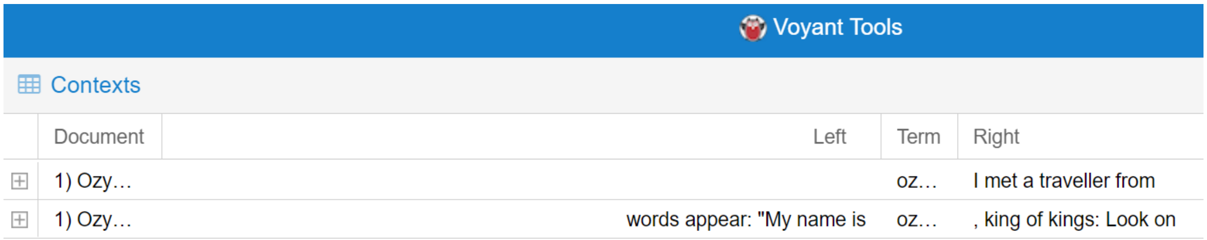


Figure 150 Links, Ozymandias

Ozymandias king is called by the title of “king of Kings”; thereupon, KG consists of “king, Ozymandias, kings” in Links toy/panel/skin. Another KG “king, Ozymandias, pedestal” informs us about the erection of his statue on the high pedestal. One more KG shows the relation of “look, said, Ozymandias” because he said, look at my works.

One very interesting and accurate knowledge pattern has been discovered in harmony with the text of the poem: “And on the pedestal, these words appear; My name is Ozymandias, king of kings:” The poetic text verifies the KG “appear, pedestal, Ozymandias, king”. P.B. Shelley mentions “a traveller from an antique land”, and its evidence is found in the KG of “traveller, antique, land”. Another KG “traveller, antique, said” proves and follows the text of “a traveller from an antique land Who said”. To conclude, KGs and poetic text verify each other; consequently, they establish the concurrent validity of this research. To conclude, knowledge discovery and interpreting some text from different angles is a very extensive process. That is why one postulate of the theory emphasizes that “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166). Extension, elaboration and extraction of multifaceted layers of meanings are crucial elements of hermeneutics.

v. Contexts



	Document	Left	Term	Right
⊞	1) Ozy…		oz…	I met a traveller from
⊞	1) Ozy…	words appear: "My name is	oz…	, king of kings: Look on

Figure 151 Contexts, Ozymandias

The proper noun “Ozymandias” is an unfamiliar name for readers. To comprehend its relevant words and meanings, its context is explored. This word occurs twice: first time in the title and second time during his introduction as a “king of kings”.

9. *The Feed* by Ahmed Nadeem Qasmi

i. Summary

This corpus has 1 document with 101 total words and 55 unique word forms. Created now (on 11th September 2017).

Vocabulary Density: 0.545

Average Words Per Sentence: 16.8

Most frequent words in the corpus: grain (5); beak (3); feed (3); ones (3); young (3)

Figure 152 Summary, The Feed

Text analytics imparts key information about words, and this textual knowledge pattern reveals the stylometry of any poet. This corpus has 55 unique words, and they are repeated almost twice; hence, the total words of the poem are 101. Its vocabulary density is 0.545 which also refers to the use of simple diction for beginner level readers. A full stop is placed after three or four lines; that is why the average words per sentence are 16.8. The most frequent words “grain”, “beak”, “feed” and “young ones” revolve around the topic of the poem *‘Feed’*. Overall, the Summary tool presents the stylistic qualities of Ahmed Nadeem Qasmi and his poem *‘Feed’.*

ii. Cirrus

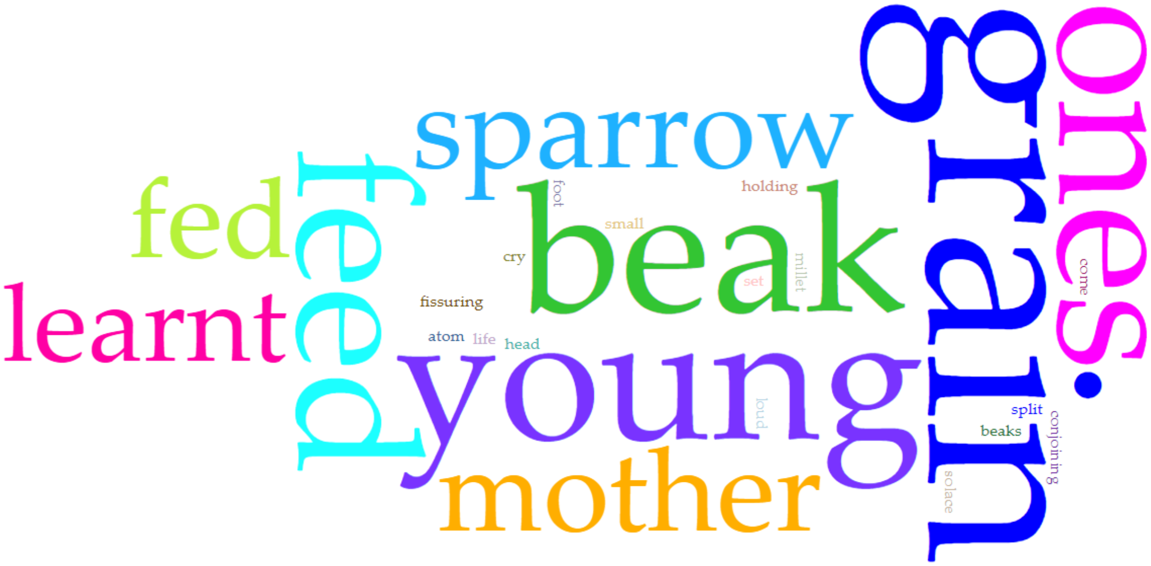
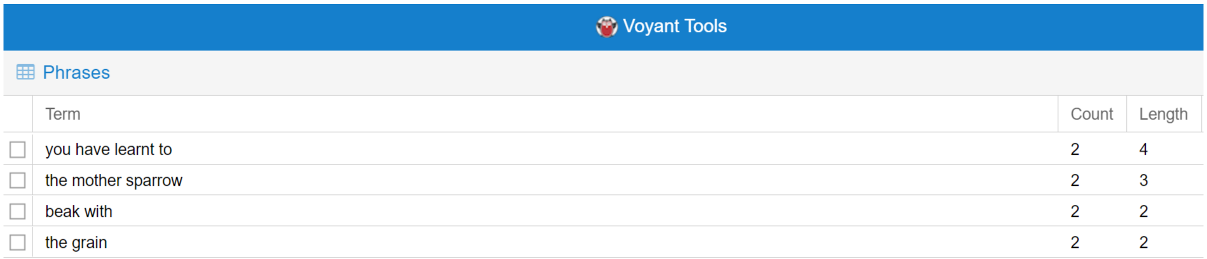


Figure 153 Cirrus, The Feed

Primarily, Cirrus is generated for topic modelling purpose; hence, the most occurring topic of the poem is “grain (5)”, and it has been placed in the centre due to its highest statistical weight. Then themes of “beak (3)”, “feed (3)” and “young (3)” “ones (3)” are visible around the theme of “grain (5)”. The poem expresses how a mother sparrow feeds her ten young sparrows by joining her beak with their beaks. With a cursory glance, it shows an alarming situation of scarcity of food for human beings and animals in this age. Every forthcoming day brings less water, lowering water surface, toxic water, insufficient water reservoirs, less agricultural lands, deforestation and growing trouble of feeding young ones with a single grain. The word “grain” is singular; however, the word "ones" indicates plural. In terms of economics, food production and supply are less, while food demand is more, so this situation leads to price hike. Moreover, such a predicament is disastrous for the survival of the next generations.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	you have learnt to	2	4
□	the mother sparrow	2	3
□	beak with	2	2
□	the grain	2	2

Figure 154 Phrases, The Feed

Figure 153 mentions collocation patterns/ n-grams of "you have learnt to" (Prn+Aux+Inf V), "the mother sparrow" (Art+Adj+N), "beak with" (N+Prep), and "the grain" (Art+N). The first phrase suggests that it is a didactic poem that food is becoming limited with an ongoing flow of time; hence, it is better to produce more food for a massively growing population. All collocations/ n-grams occur twice, and their length varies from 2 to 4 words.

iv. Links

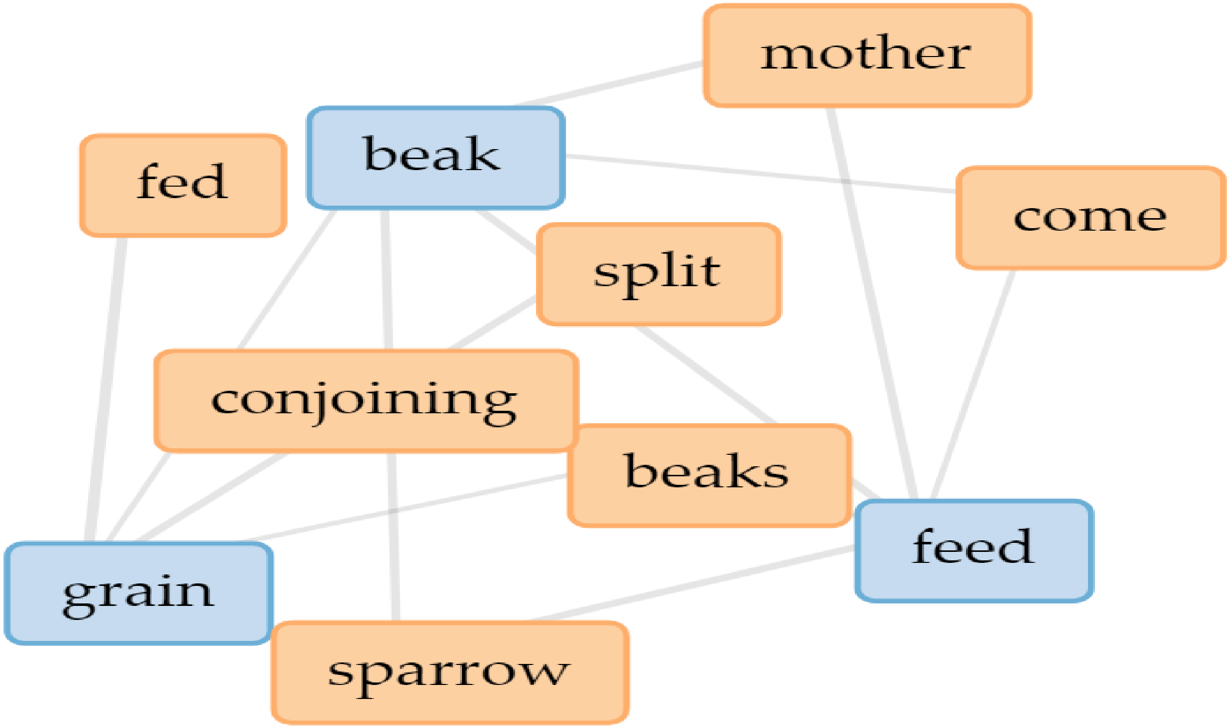
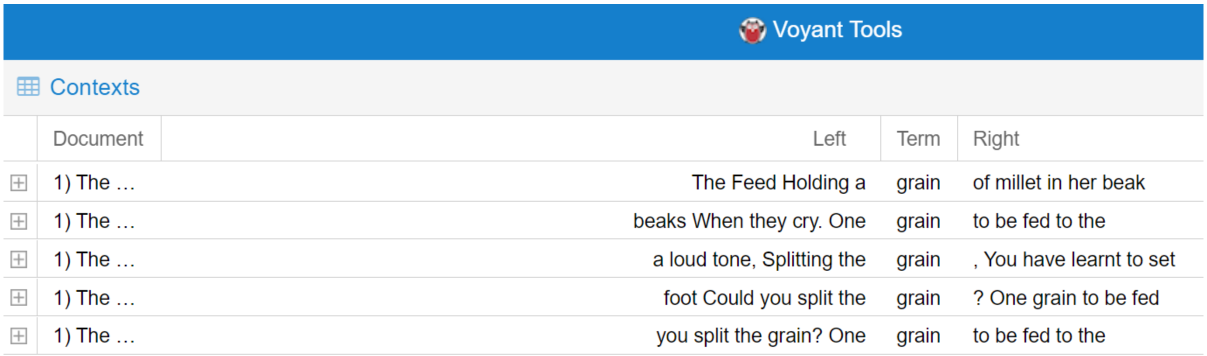


Figure 155 Links, The Feed

Here KG consists of “grain, beak, split” because the grain for young sparrows has been split with mother sparrow’s beak, as it is evident in the KG of “mother, beak,sparrow”. An interesting knowledge pattern is visible with co-occurring nodes in the KG of “grain, beak, beaks, feed” which expresses the purpose of joining mother sparrow’s one beak with ten beaks of young sparrows. One more relevant KG, “sparrow, feed” shows that the mother sparrow feeds the children sparrows with her beak. Knowledge Discovery Theory is defined as “the extraction of implicit, previously unknown and potentially useful information from data” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). Present food scarcity indicates great challenges of food production.

The KG of “grain, split, beak” refers to splitting, fissure and fusion of atom grain to produce nuclear energy for peaceful purposes. Furthermore, preparation of nitrogen fertilizers, biodiversity to change the genetic makeup of grains to produce more grain types, soil evaporation for agriculture and more food production to alleviate the dearth of food from the global village. These technologies can enhance food production to feed the growing populations of the entire world.

v. Contexts



	Document	Left	Term	Right
⊞	1) The …	The Feed Holding a	grain	of millet in her beak
⊞	1) The …	beaks When they cry. One	grain	to be fed to the
⊞	1) The …	a loud tone, Splitting the	grain	, You have learnt to set
⊞	1) The …	foot Could you split the	grain	? One grain to be fed
⊞	1) The …	you split the grain? One	grain	to be fed to the

Figure 156 Contexts, The Feed

Grain means a seed from a plant and a unit of mass. During word sense disambiguation, theoretical underpinning guides that “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). To determine the exact word sense, the context of the word “grain” has been displayed from five different lines of the poem. Five times the word “grain” refers to a seed of a plant or “a grain of millet”. The line “One grain to be fed to the ten young ones” has been repeated twice. The intensity of the food crisis for the growing population has been discussed that how one small millet can be distributed to ten children of the sparrow. Information retrieval (IR) of one key word facilitates in comprehending the true semantic shade, and Kwary (2018) also applies IR for WSD.

10. *The Hollow Men* by T.S. Eliot

i. Summary

This corpus has 1 document with 87 total words and 63 unique word forms. Created about 10 minutes ago (on 19th September 2017).

Vocabulary Density: 0.724

Average Words Per Sentence: 43.5

Most frequent words in the corpus: men (5); hollow (3); dry (2); stuffed (2); together (2);

Figure 157 Summary, The Hollow Men

"Summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). Vocabulary density of the poem is 0.724, and it means that a new word occurs after several words, and the poet does not repeat content words. It shows that T.S. Eliot's diction is full of unique linguistic variety which informs his vast linguistic knowledge and extensive reading. The average words per sentence are 43.5, and it informs that full stop has been marked after some verses. In fact, there are only two full stops, the first full stop is marked in the 4th line, and the second full stop is placed in the 18th line. They show that the first thought has been completed in the first four lines, whereas the 2nd thought has been completed from the 5th line to 18th line of the poem. T.S. Eliot's stylistic analysis shows that only "men, hollow, dry and stuffed" and some stop words have been repeated; that is why in 87 words corpus, 63 words are unique. It reveals that Eliot uses unique words in this poem.

ii. Cirrus

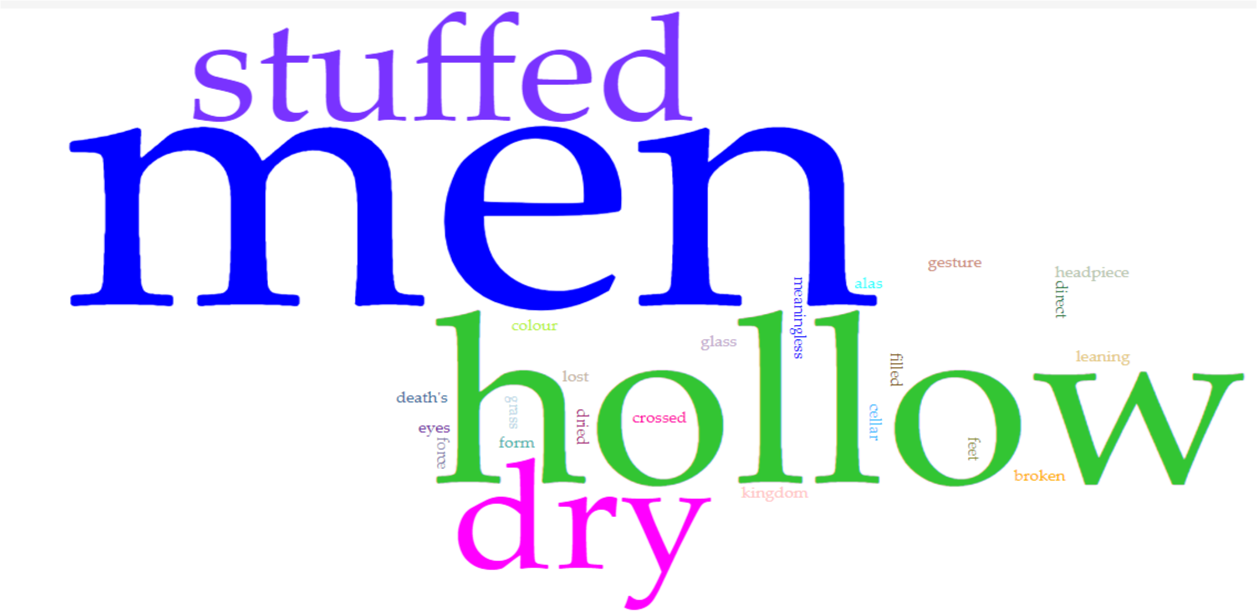
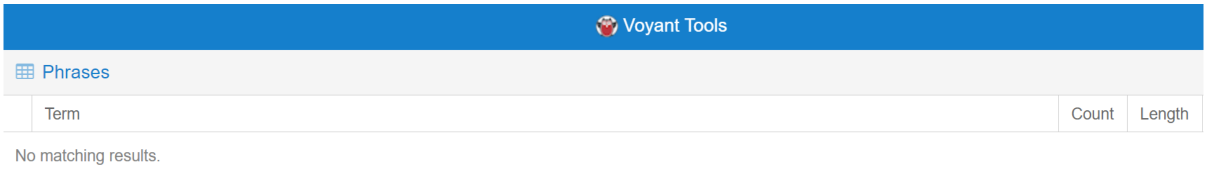


Figure 158 Cirrus, The Hollow Men

This Cirrus shows major themes of “hollow (3)” “men (5)” who are “stuffed (2)” men with “dry (2)” and lifeless voices. Other words are in a very small font because they occur just once in the corpus. To summarize, the poem revolves around the central characters, theme, and title “hollow men”. So, title, content and statistics verify the text and the generated Cirrus.

iii. Phrases


Voyant Tools

Phrases

Term	Count	Length

No matching results.

Figure 159 Phrases, The Hollow Men

Not even a single phrase is matched any other phrase in this poem. It refers to T.S. Eliot’s poetic style of unique diction; even no word has been repeated except function words.

iv. Links

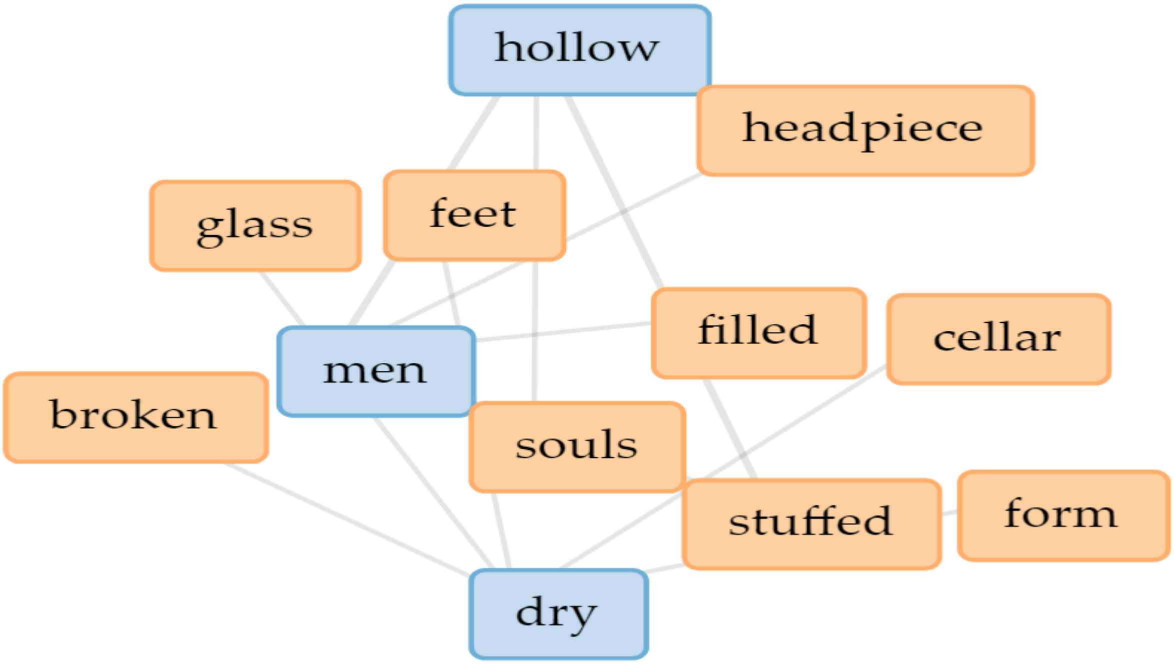
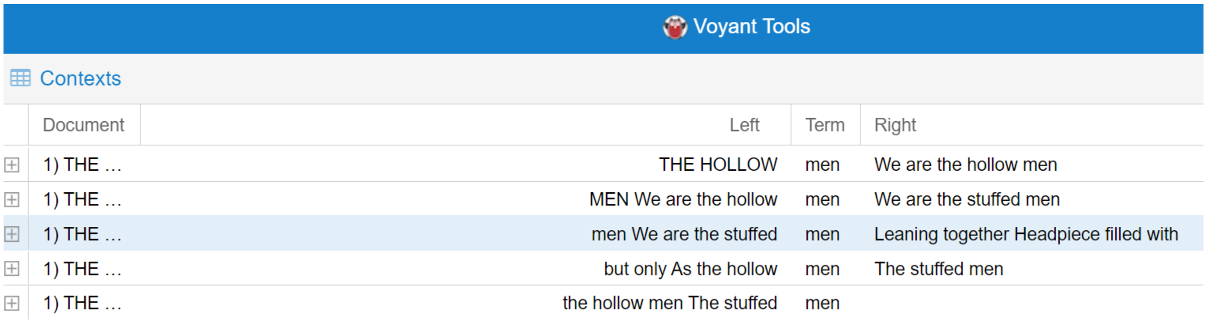


Figure 160 Links, The Hollow Men

This poetic KG links “men, hollow, stuffed” qualities that human beings are void of characteristics of meaningfulness, rationality, deep reflection and essence of humanity. Instead of bona fide humane characteristics, human beings are replete with artificiality. Moreover, the KG of “form, stuffed” expresses the poet’s self-acceptance about the stark and bitter reality that we are the very stuffed and hollow men. Stuffed things are not real living beings. Theoretical underpinning states that “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166), and knowledge graphs can be further extended for further understanding or to discover new hermeneutic patterns.

v. Contexts



Document	Left	Term	Right
1) THE …	THE HOLLOW	men	We are the hollow men
1) THE …	MEN We are the hollow	men	We are the stuffed men
1) THE …	men We are the stuffed	men	Leaning together Headpiece filled with
1) THE …	but only As the hollow	men	The stuffed men
1) THE …	the hollow men The stuffed	men	

Figure 161 Contexts, The Hollow Men

The common noun “men” is an ambiguous referent, so it is retrieved with Contexts tool which reveals adjective “hollow” has been associated with “men” three times (in 1st, 2nd and 4th sentences), and the adjective “stuffed” is connected with “men” twice (in 3rd and 5th sentences). These adjectives and their further examples humiliate the modern man because Eliot highlights certain traits of hollow men.

11. *Leisure* by W.H. Davies

i. Summary

This corpus has 1 document with 93 total words and 56 unique word forms. Created now (on 21st September 2017).

Vocabulary Density: 0.602

Average Words Per Sentence: 31.0

Most frequent words in the corpus: time (6); stand (3); stare (3); care (2); life (2)

Figure 162 Summary, Leisure

Computational stylistics counts total and unique words to generate vocabulary density. In this poem, W.H. Davies uses 56 unique words which have been repeated almost twice and the total words are 93. Vocabulary density is 0.602 which shows the simplicity of the poem. While depicting the beauty of Nature, Davies uses natural diction in a very simple language, and this quality resembles Wordsworth's poems. Almost two lines or one verse makes a sentence, and on average 31words per sentence have been written by Davies.

ii. Cirrus

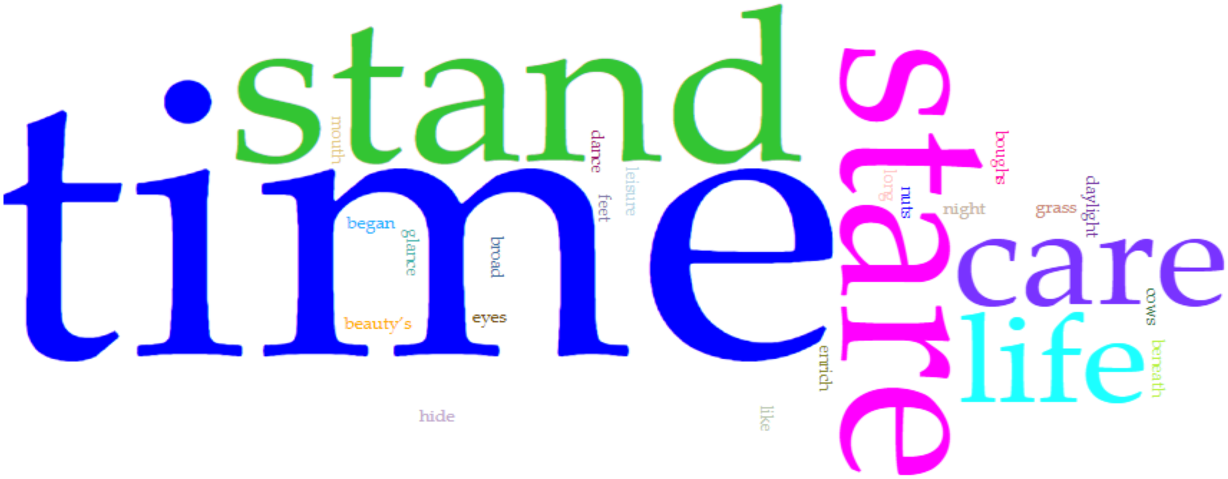
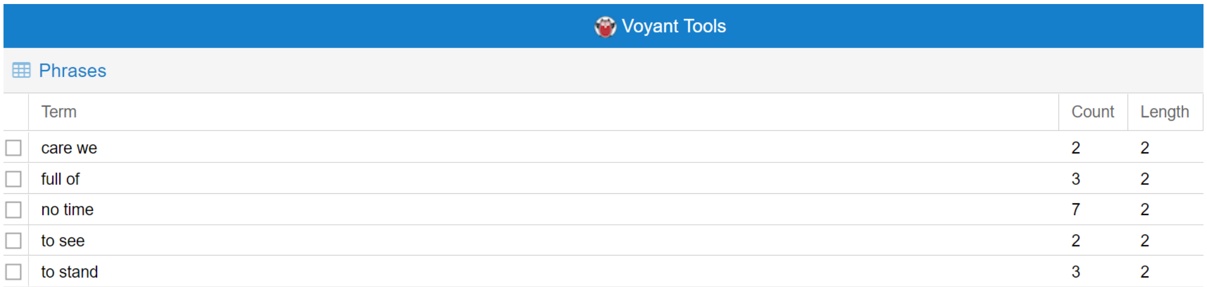


Figure 163 Cirrus, Leisure

A major theme concentrates on the dearth of "time (6)", and statistics also buttresses the same point. Man does not find any time to "stand (3)" and "stare (3)" to enjoy pleasures of leisure. Deprivation of leisure leads to man's "life (2)" to "care (2)" and worries in pursuit of materialism and the worship of Mammon, the Syrian deity of riches and material things. To summarize, only six prominent words in Cirrus successfully elucidate the key themes of the poem since topic modelling is the basic function of word clouds. One knowledge discovery is that no human character has been mentioned in this poem.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	care we	2	2
□	full of	3	2
□	no time	7	2
□	to see	2	2
□	to stand	3	2

Figure 164 Phrases, Leisure

This poem finds the occurrence of bigrams, for instance, "no time" (Det+N), "full of" (Adj+Prep) and "to stand" (Inf V) 6, 3, 3 times respectively. Standard phraseology also guides the correct use of a preposition, for instance, "full of" (Adj+Prep). Moreover, collocations/ n-grams teach grammatical rules that the first form should be used after "to".

iv. Links

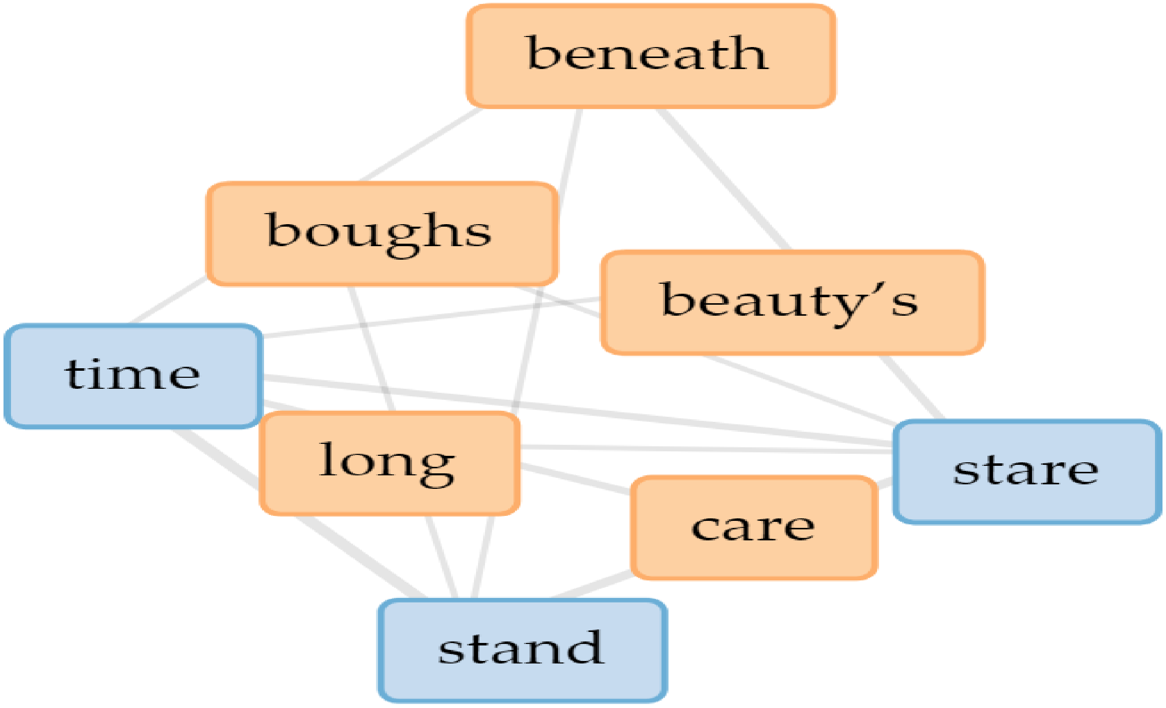
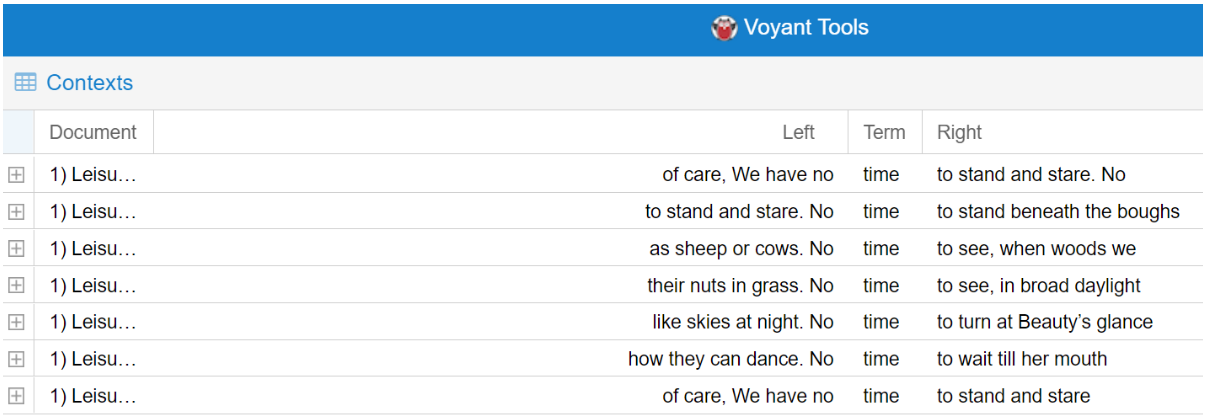


Figure 165 Links, Leisure

KGs interlink relevant ideas and themes to reveal knowledge patterns and new associations. The title of the poem and the most occurring theme of "Time" is connected with actions of "stand" and "stare" in the KG of "time, stand, stare". It means that busy people of the modern industrial age do not have time to stand leisurely and stare at phenomena of nature. Moreover, the KG of "time, beauty" reveals that man is deprived of time to watch abundant natural beauty in his/her surroundings. Another KG of "beneath, stand, long, boughs" exposes that modern man has no time to stand beneath boughs of trees like animals. So, animals enjoy more natural beauty than busy human beings.

Hermeneutica Theory expresses that "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166). Links tool is a knowledge-bearing tool which shows knowledge graphs, and then human reflection and critique extracts knowledge patterns; for instance, modern man is so preoccupied with the worldly chores that human beings do not have time to stand and stare to enjoy different phenomena of nature. Another aspect of hermeneutica is to explore new possible options of deriving semantic shades and to interpret various themes.

v. Contexts



	Document	Left	Term	Right
⊞	1) Leisu…	of care, We have no	time	to stand and stare. No
⊞	1) Leisu…	to stand and stare. No	time	to stand beneath the boughs
⊞	1) Leisu…	as sheep or cows. No	time	to see, when woods we
⊞	1) Leisu…	their nuts in grass. No	time	to see, in broad daylight
⊞	1) Leisu…	like skies at night. No	time	to turn at Beauty's glance
⊞	1) Leisu…	how they can dance. No	time	to wait till her mouth
⊞	1) Leisu…	of care, We have no	time	to stand and stare

Figure 166 Contexts, Leisure

"Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). Different interpretations require different contexts to explicate problematized linguistic entities. As a result of theory, new tools have been designed to disambiguate word sense by human reflection. Time is used as a noun and verb, and to disambiguate this word sense, the context of 'time' has been shown in figure 166. The phrase "no time" is used six times as a noun, and it shows the scarcity of time for modern man in this technological era. Furthermore, the phrase "No time to stand" has been used three times that people are rushing so mechanically that they have been deprived of leisure and its jubilations. "No time to turn", "No time to wait" and "no time to see" have been found once in the corpus. Some microseconds should be consumed in visualising the beauties of Nature, but this iota of time is also extinct in the hectic schedule of modern man.

12. *Ruba'iyat* by Allama Muhammad Iqbal

i. Summary

This corpus has 1 document with 87 total words and 70 unique word forms. Created about 2 minutes ago (on 21st September 2017).

Vocabulary Density: 0.805

Average Words Per Sentence: 17.4

Most frequent words in the corpus: faith (3); abraham (1); age's (1); bends (1); blends (1)

Figure 167 Summary, Ruba'iyat

This corpus includes 70 unique words and 87 total words. By dividing unique words with total vocabulary, 0.805 vocabulary density is derived. Though it is a translated poem by the national poet, Allama Muhammad Iqbal, yet it manifests his stylistic features. He uses diversified vocabulary; hence, mostly stop words have been repeated. He writes 17.4 words in a sentence, and it is appropriate for advanced level readers. To summarise, long sentences, complex vocabulary, and sublime philosophical thoughts have been used in this poem to awaken the nation from a deep slumber.

ii. Cirrus

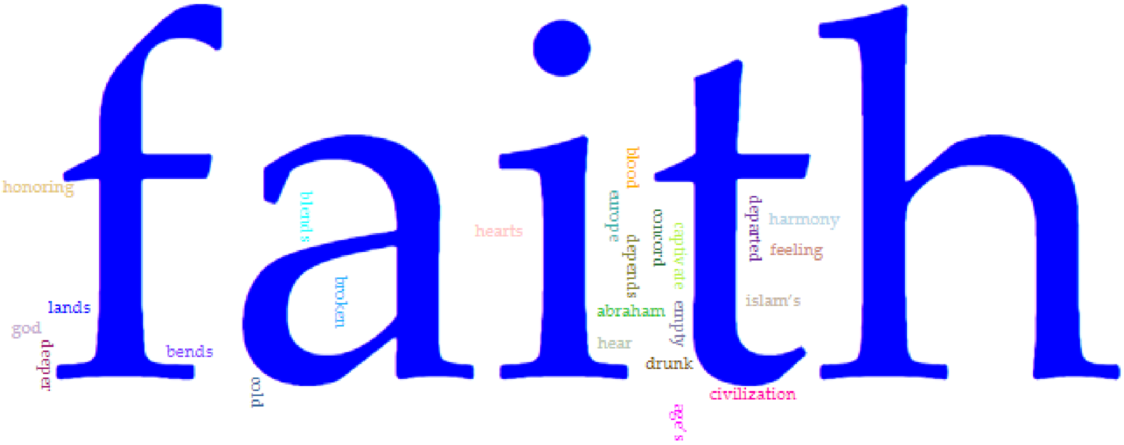
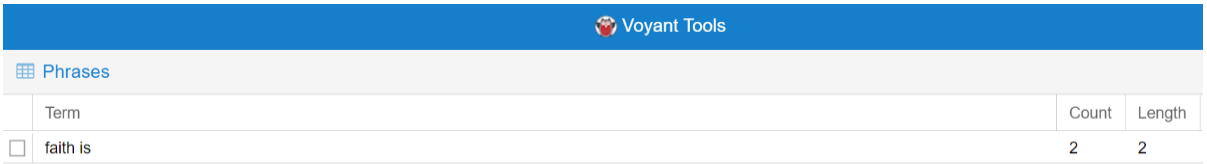


Figure 168 Cirrus, Ruba'iyat

Figure 167 Cirrus performs the task of topic modelling from any dataset. The first Rubai revolves around the single key theme of "faith (3)" which validates the foundation of Sufism in three quatrains. Several other minor themes have also been mentioned only once for example, "God (1)", "Abraham (1)" (A.S), "Islam (1)", "Muslims (1)", "prayers (1)", "Makkah (1)", "love (1)", "madness (1)", "Europe (1)" and "civilization (1)". The first six minor themes are concerned with Islam and its mystic values, whereas themes of Europe and its civilization are juxtaposed as a foil. The major theme of "faith (3)" reveals its poet, Allama Muhammad Iqbal, whose poetry is embedded in mysticism and philosophy. It also relates to the nature of Ruba'iyat that each quatrain is thematically complete; hence, three Ruba'iyat do not show any uniformed thematic progression; consequently, minor themes scatter in this Cirrus.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	faith is	2	2

Figure 169 Phrases, Ruba'iyat

Only one bigram, "faith is" (N+Aux) occurs twice in this corpus, and it describes the narratology of the true manifestation of faith. These poetic lines define faith, its effects on entire life. "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166). Human reflection and Phrases tool lead to the construction of narratology by repetition, which is also an element of rhetoric.

iv. Links

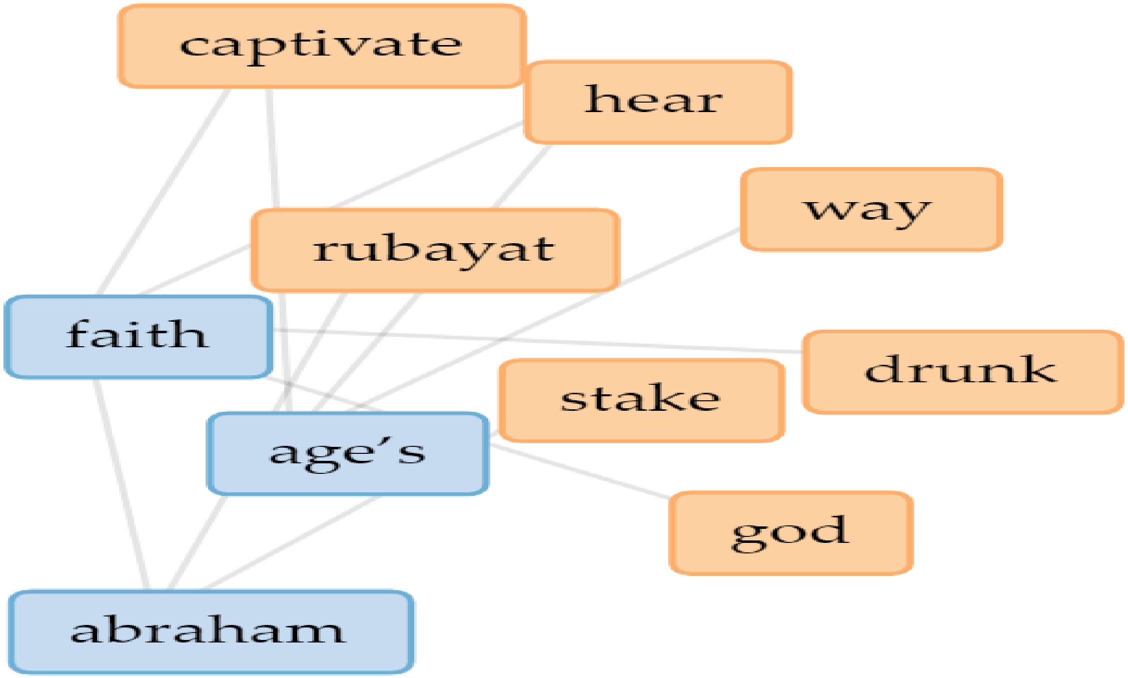
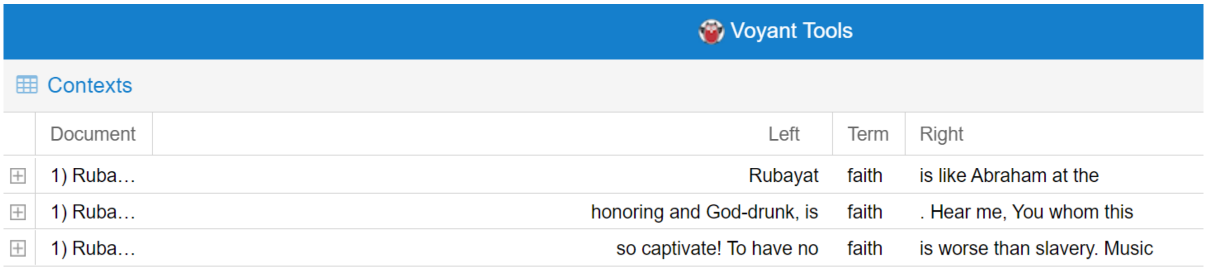


Figure 170 Links, Ruba'iyat

The KG ties several themes in an associated string to make them meaningful and to testify the text. The first line of *'Ruba'iyat'* is "Faith is like Abraham at the stake:" so, the KG of "faith, Abraham, stake" has been associated. In another KG, "drunk, faith, God" entities have been connected with the word "faith", and faith is defined as "God-drunk". Another set of KG proves this poetic line "… this age's way so captivate!" with the KG of "captivate, age's, way". One shortcoming of Links tool is that it takes apostrophe as a separated word; therefore, the word Ruba'iyat is shown as two words here. The theory postulate, "Hermeneutic tools fail in interesting ways" (Rockwell, & Sinclair, 2016, p. 166), informs that this shortcoming can be overcome by improving the algorithm and embedded lexicon. Another reason is that the word 'Ruba'iyat' is Arabic, and the tool has not been trained for such poetic terms. Its solution is to train the model and algorithm of Voyant tools in terms of literary terms.

v. Contexts



	Document	Left	Term	Right
⊞	1) Ruba…	Rubayat	faith	is like Abraham at the
⊞	1) Ruba…	honoring and God-drunk, is	faith	. Hear me, You whom this
⊞	1) Ruba…	so captivate! To have no	faith	is worse than slavery. Music

Figure 171 Contexts, Ruba'iyat

Faith means a set of ideologies about God or trust in Him. To clarify semantic ambiguity, Contexts tool extracts complete contextual information regarding "faith", used three times in this poem. It shows that true faith has been exemplified with the faith of Hazrat Abraham (A.S). Secondly, the situation of God-drunk is faith, and if a person is deprived of faith, such a state of mind is inferior to the enhancement of body and soul. The faith does not mean "trust in"; rather it stands for staunch Islamic faith.

13. *A Tale of Two Cities* by John Peter

i. Summary

This corpus has 1 document with 135 total words and 90 unique word forms. Created 8 seconds ago (on 22nd September 2017).

Vocabulary Density: 0.667

Average Words Per Sentence: 27.0

Most frequent words in the corpus: burnt (2); great (2); afraid (1); arms (1); ashes (1)

Figure 172 Summary, A Tale of Two Cities

"Summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). This compact description is generated by Summary tool which has analysed the stylistic qualities of John Peter's poem, *'A Tale of Two*

Cities' with the distant reading approach. Unique words of this poem are 90, and half of those words (45) are reused in the poem. Its high vocabulary density is 0.667 which implies that the text is very simple and quite appropriate for beginner-level readers. Usually, a full stop is marked at the end of a quatrain; that is why the average number of words per sentence is 27. Most frequent words are exclaiming the drastic conditions of the atomic bomb explosion for example, "burnt, afraid, arms, ashes". The diction of the poem is simple, but it depicts the drastic effects of the second world war.

ii. Cirrus

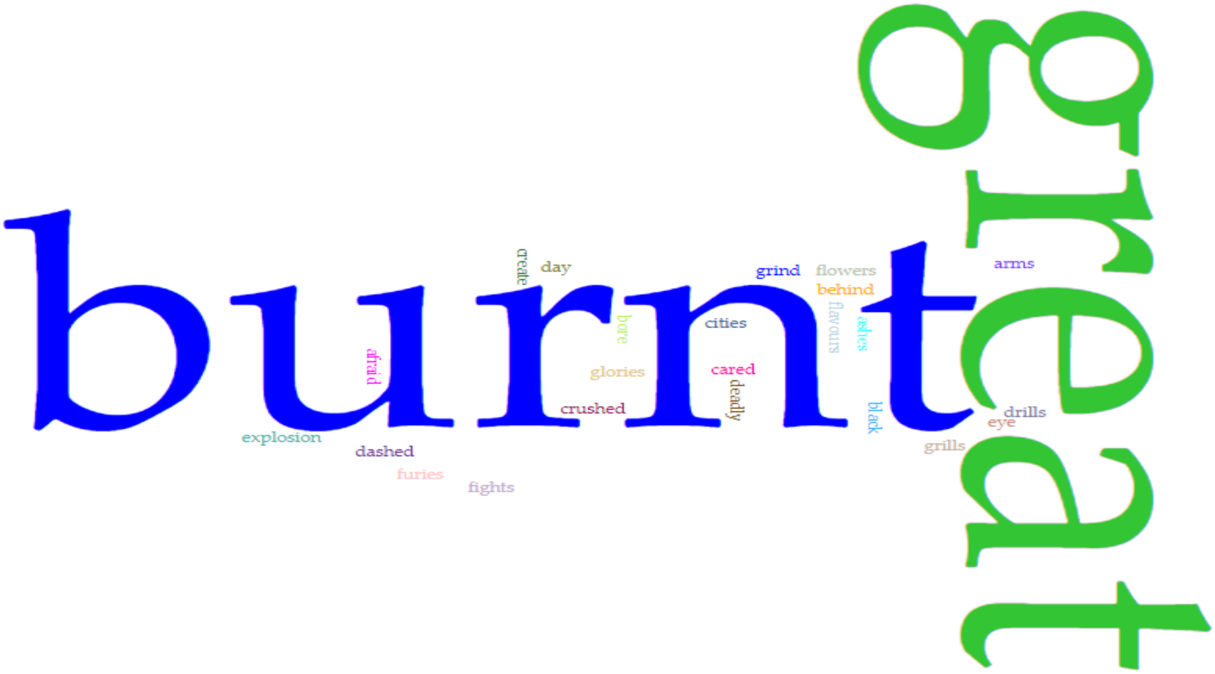
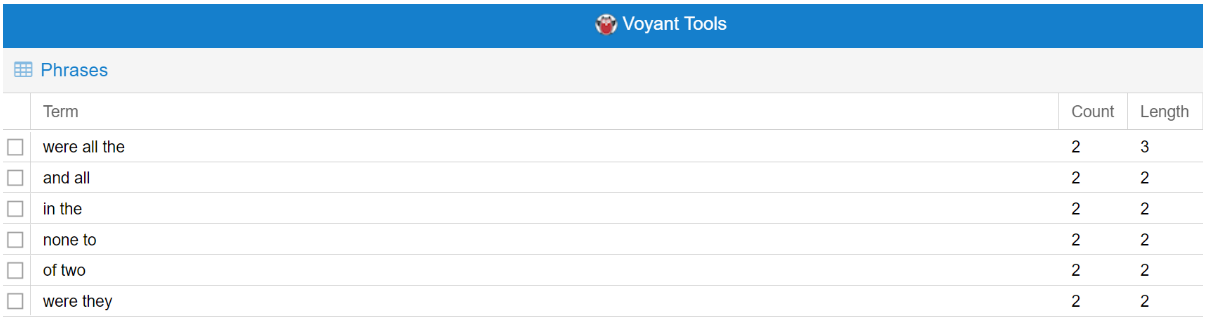


Figure 173 Cirrus, A Tale of Two Cities

Input textual data have been transformed into a data visualization in the form of a word cloud which reveals key topics with the help of topic modelling. This Cirrus shows two words, "burnt (2)" and "great (2)" which suggest burning of the cities on a large scale. On one side, there is destruction on a large scale, and on the other side, the poet exalts those "great (2)" people who suffered a lot and faced scars on their bodies and souls during Hiroshima and Nagasaki nuclear attacks in 1945. Some other themes, "afraid (1)", "arms (1)" and "ashes (1)" contribute to enhance the tragic effect of the second world war.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	were all the	2	3
□	and all	2	2
□	in the	2	2
□	none to	2	2
□	of two	2	2
□	were they	2	2

Figure 174 Phrases, A Tale of Two Cities

From figure 173, some standard bigrams “none to” (Prn+Prep), “of two” (Prep+Adj), “in the” (Prep+Art) are very useful to learn standard collocation patterns/ n-grams, and their length ranges from 2 to 3 words.

iv. Links

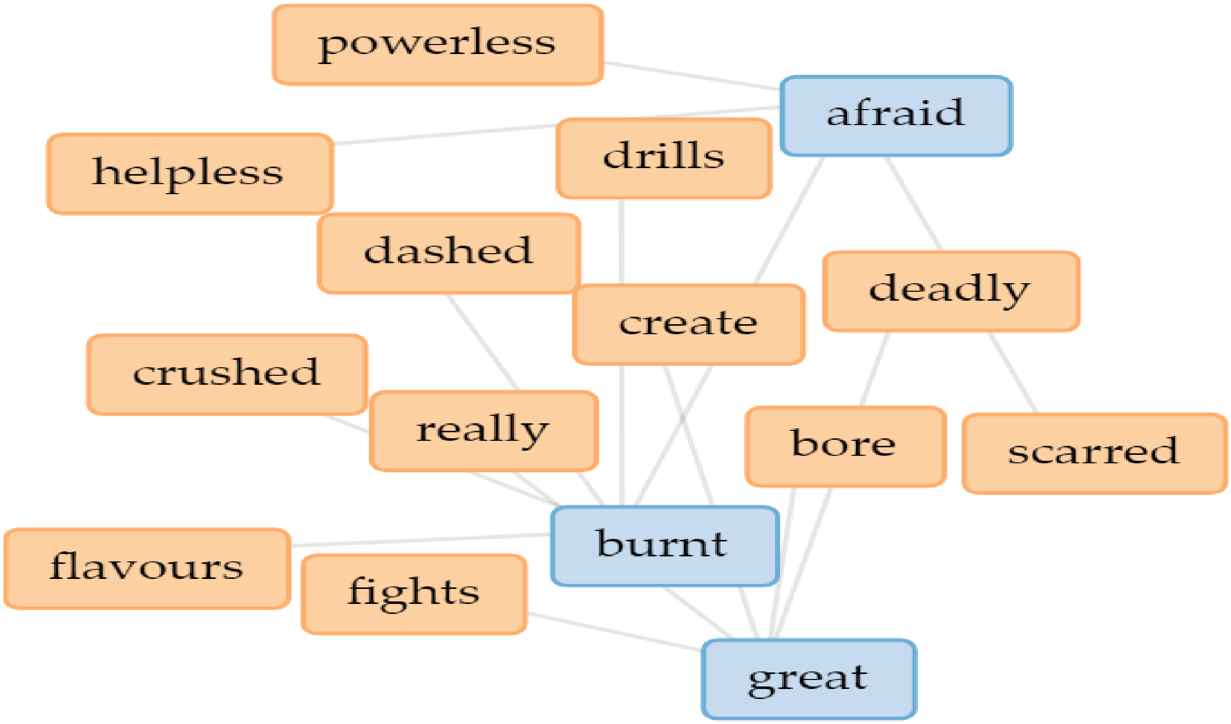
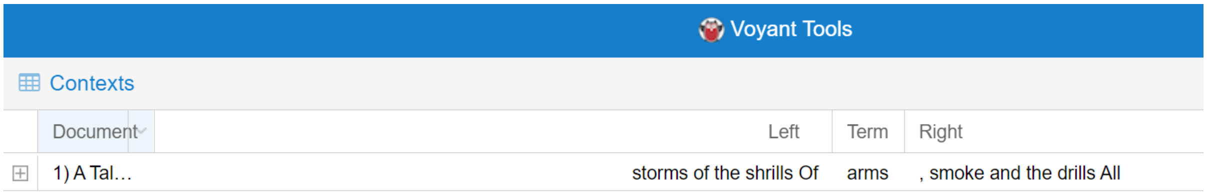


Figure 175 Links, A Tale of Two Cities

The KG of “scarred, afraid, burnt,” depicts a terrified condition of war-stricken people whose bodies and souls are scarred and intimidated. One more KG of “create, great” concludes that great nations produce great people because nations become strong by dint of noble people of the society. Another KG of "afraid, helpless, powerless" expresses that the weak nations become friendless, penniless and helpless in conditions of extreme fear, but powerful nations change their circumstances. Another KG "crushed, burnt, drills" reveals a knowledge pattern about soldiers who learn their best assault techniques to devastate their foes. Consequently, ruthless training leads to ruthless war tactics to destroy enemy on a large scale.

v. Contexts



Document	Left	Term	Right
1) A Tal…	storms of the shrills Of	arms	, smoke and the drills All

Figure 176 Contexts, A Tale of Two Cities

The semantic shade of the word "arms" is ambiguous whether it refers to the plural of a body organ, or it means weapons. To disambiguate it, Contexts tool retrieves relevant context. The tool has been designed on the premise of a theoretical framework. "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). It reveals that the word "arms" conveys the narratology of weapons and destruction on a large scale because shrills, smoke and drills portray war catastrophes.

14. *My Neighbour Friend Breathing His Last* by Bullah Shah

i. Summary

This corpus has 1 document with 97 total words and 57 unique word forms. Created now (on 23rd September 2017).

Vocabulary Density: 0.588

Average Words Per Sentence: 10.8

Most frequent words in the corpus: aghast (5); god (5); breathing (2); friend (2); bullah (1)

Figure 177 Summary, My Neighbour Friend Breathing His Last

Summary tool reveals stylometric qualities of Bullah's poem. This corpus has 57 unique words and 97 total words. Its vocabulary density is 0.588 which shows new words are repeated less; hence, it is appropriate for beginner level readers. Simple diction is a prominent characteristic of mystic poets because common people are their target audience. Furthermore, their poetry always conveys sublime purposes to link humanity with God, elucidating the materialistic world's transient nature. The average words per sentence are 10.8 which show small and beginner level poetic lines.

ii. Cirrus

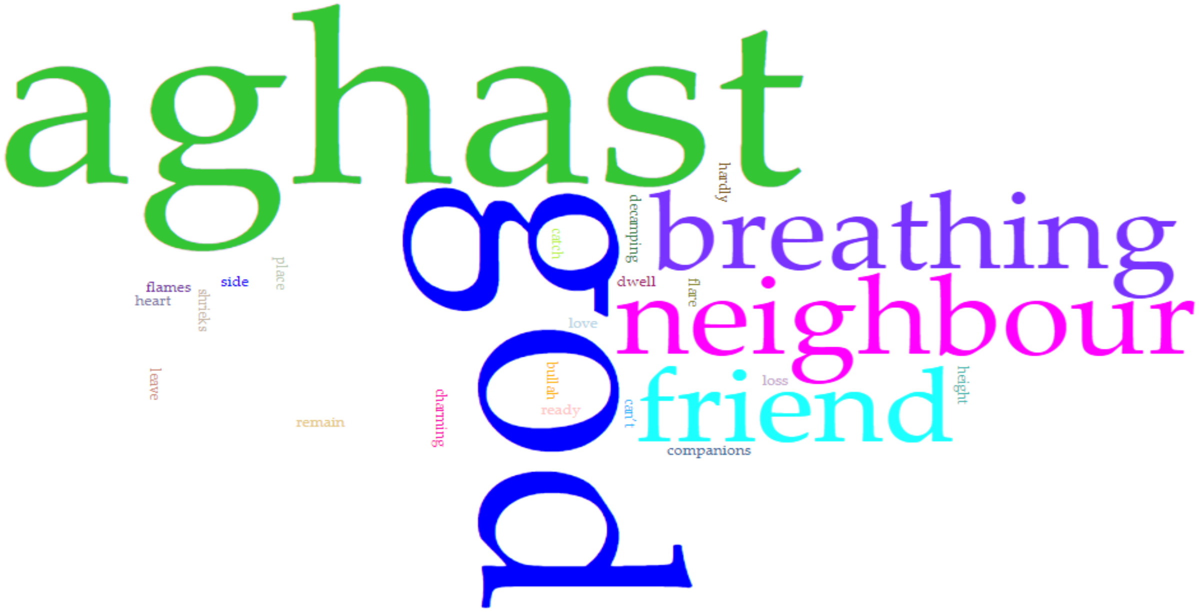
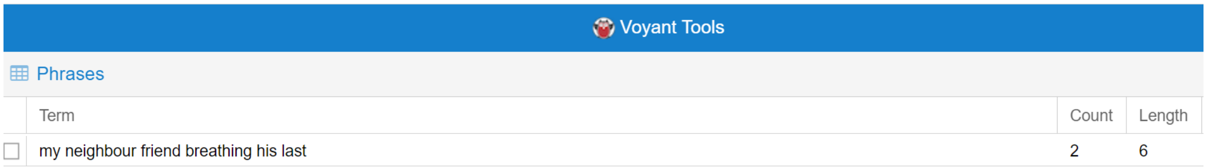


Figure 178 Cirrus, My Neighbour Friend Breathing His Last

Topic modelling in this Cirrus shows the most frequent word “aghast (5)” which refers to the saddest and gloomiest theme of the poem in which the last “breathing (2)” of dear “friend (2)” has been mentioned. Last “breathing (2)” and “decamping (1)” indicate the advent of imminent death which is always painful and perturbing for kith and kin. Cultural values and linguistic norms also elaborate text for hermeneutic purposes. One theory postulate states that “It is supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166). The word “God” exhibits a mystic message, and the poet utters it five times “O God” in the poem. This is an address term to directly implore God for His help to save the life of his spiritual guide. Simultaneously, “O God” phrase expresses a fearful state of mind when the dearest one is on the verge of imminent death. If we refer back to its context, Bullah writes these verses on the eternal departure of his Peer and Murshid (spiritual guide) Baba Shah Inayat Qadiri Shatari from whom Bulleh Shah and Waris Shah learnt mysticism.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	my neighbour friend breathing his last	2	6

Figure 179 Phrases, My Neighbour Friend Breathing His Last

Only one quadgram, “friend breathing his last” (N+V+Prn+N) is repeated twice: Once in the title, and once in the poem. It expresses the ideology of death around which the whole elegiac poem revolves.

iv. Links

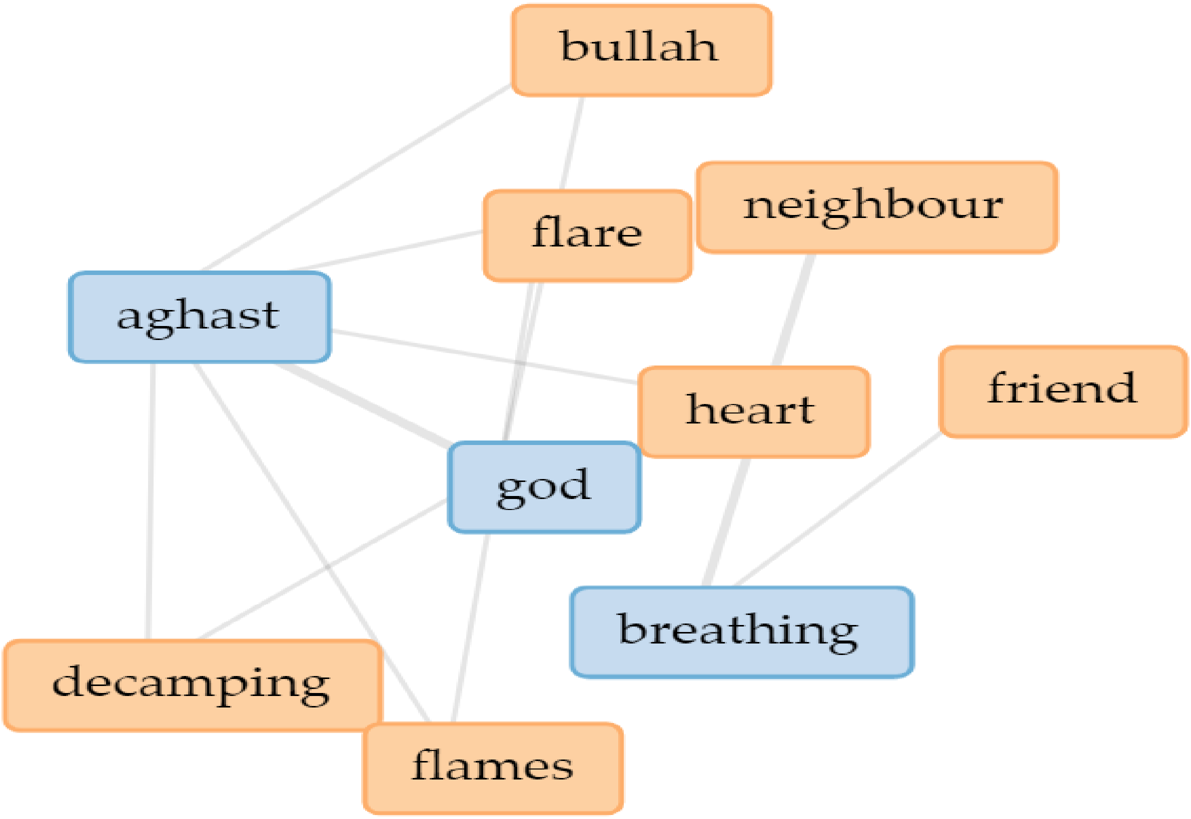
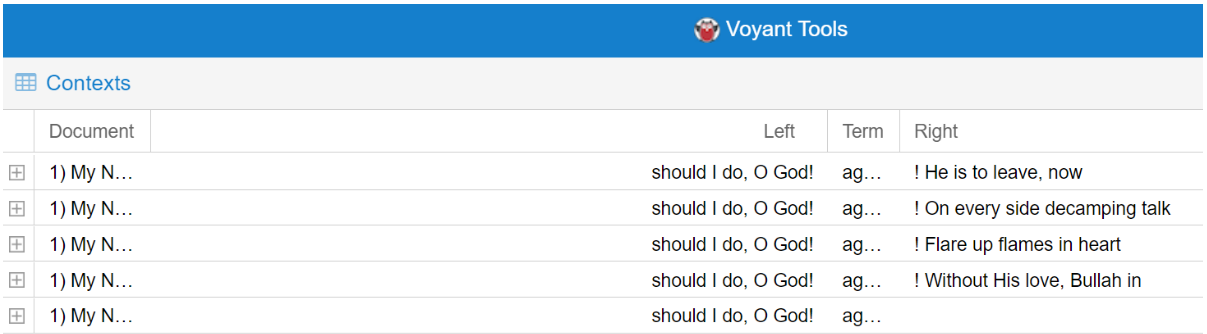


Figure 180 Links, My Neighbour Friend Breathing His Last

As a human brain connects different things to reach a comprehensive idea, the KG draws links to present a comprehensive network of ideas and thoughts from the poem. One theory postulates that “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166), and it emphasizes understanding of deep hermeneutic patterns. As an inquiry expands, knowledge graphs also expand. Hermeneutics becomes deeper by looking into different perspectives of the text. These examples elaborate these points further. Title words "neighbor" and "friend" have been mutually linked; then both of them have been connected with the word "breathing". These links connect the message of the eternal departure of Inayat Ali Shah from this world. Furthermore, both themes, "decamping" and "breathing" are interlinked because there are talks about the eternal departure or death of Bullah's Murshid, Baba Shah Inayat Qadiri Shatari.

The phrase "O God! Aghast!" is used five times in the poem. That is why three words are interconnected with one another. The KG of "Aghast, Bullah" is linked because Bullah is in a shocked and grieved state of mind, and no other mourner has been mentioned in this poem. The KG, "flare, flames" validates the poetic line, "Flare up flames in heart …".

v. Contexts



	Document	Left	Term	Right
⊞	1) My N…	should I do, O God!	ag…	! He is to leave, now
⊞	1) My N…	should I do, O God!	ag…	! On every side decamping talk
⊞	1) My N…	should I do, O God!	ag…	! Flare up flames in heart
⊞	1) My N…	should I do, O God!	ag…	! Without His love, Bullah in
⊞	1) My N…	should I do, O God!	ag…	

Figure 181 Contexts, My Neighbour Friend Breathing His Last

The uncommon interjection "Aghast" is ambiguous for readers; consequently, its contextual information is searched to comprehend its true semantic shade. The most occurring word "aghast" is used five times in the corpus, and each time a sign of exclamation is used after them. It shows an extreme heartfelt feeling of terror and horror which indicates the imminent death of Murshid Inayat Ali Shah. Furthermore, every time the phrase "what should I do, O God!" precedes before the interjection "aghast". Bullah is in a meaningless and miserable condition with the demise of his spiritual teacher.

15. *He Came to Know Himself* by Sachal Sarmast

i. Summary

This corpus has 1 document with 76 total words and 50 unique word forms. Created now (on 23rd September 2017)

Vocabulary Density: 0.658

Average Words Per Sentence: 38.0

Most frequent words in the corpus: came (2); just (2); know (2); love (2); able (1)

Figure 182 Summary, He Came to Know Himself

This poem comprises 50 unique words, and almost its half words (+26) are repeated. Vocabulary is as highly dense as 0.658, and less words have been repeated. Simple diction is used in this translated poem; hence, it is easy for beginner level readers. On average, each sentence comprises 38.0 words because in 12 lines, only two full stops are placed, and this literary phenomenon of extending quatrain beyond lines is termed enjambment. To conclude, the Summary tool analyses stylometry of Sachal Sarmast and his poem in a very compact and quantified form.

ii. Cirrus

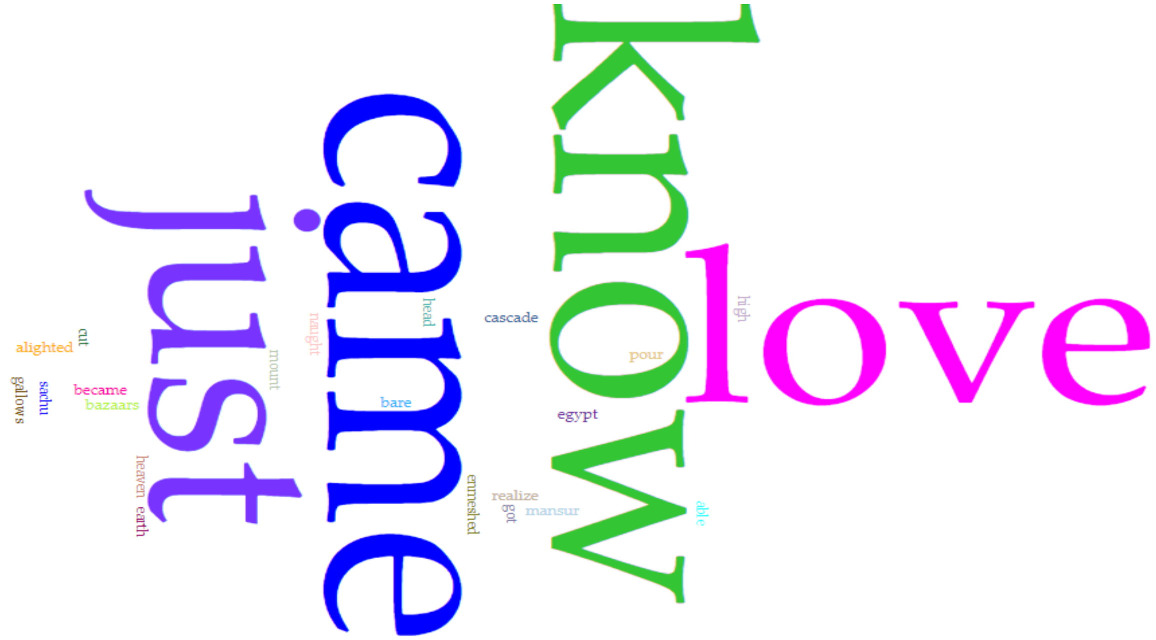
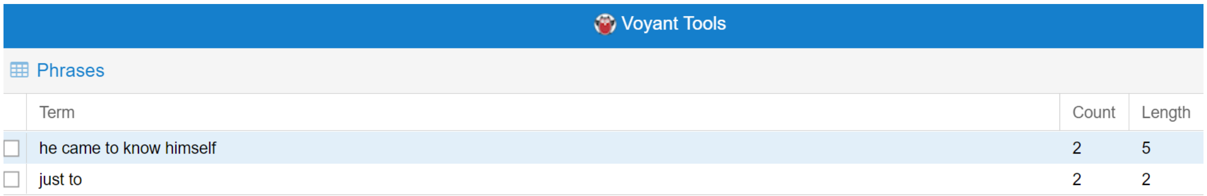


Figure 183 Cirrus, He Came to Know Himself

Topic modelling is deeply associated with Cirrus which shows the most prominent theme of “know (2)” thyself from this poem. Self-awareness and self-accountability lead to accomplish all hierarchies of self-actualisation. Nowadays, the best-selling books focus on self-discovery and self-organization for optimum productivity. Secondly, “love (2)” to God theme is also the result of self-accountability. Divine love is the commonest feature of mysticism, that is why all saints, Sufis and prophets preach divine love through all religions and scriptures. Subsequently, Sachal Sarmast's mystic messages carry universal appeal.

In fact, the phrase "came to know" is a trigram, and it refers to one word "know" semantically, but Cirrus tool shows three different words: came, to, know because they exist as separate words in digital English lexicons. Actually, there is no theme of "came" in the poem. This is the limitation of the Voyant tool that it cannot differentiate that sometimes bigram, trigram and quadgram present only one theme. "Hermeneutic tools fail in interesting ways" (Rockwell, & Sinclair, 2016, p. 166). To overcome this deficiency, the dataset should be re-trained to address these deficiencies, and standard collocations should be included during training of machine learning models.

iii. Phrases



	Term	Count	Length
□	he came to know himself	2	5
□	just to	2	2

Figure 184 Phrases, He Came to Know Himself

The title phrase "He came to know himself" (Prn+V+Inf V+Ref Prn) and "just to" (Adv+Prep) phrases have been repeated twice, and their length consists of 5 and 2 words respectively. The first collocation pattern/ n-gram reveals a knowledge pattern to use "to" before the second verb in a sentence, and the first form of the verb is used after "to". As a result, first collocation/ n-gram teaches learners the correct use of grammatical rules in the inductive method.

iv. Links

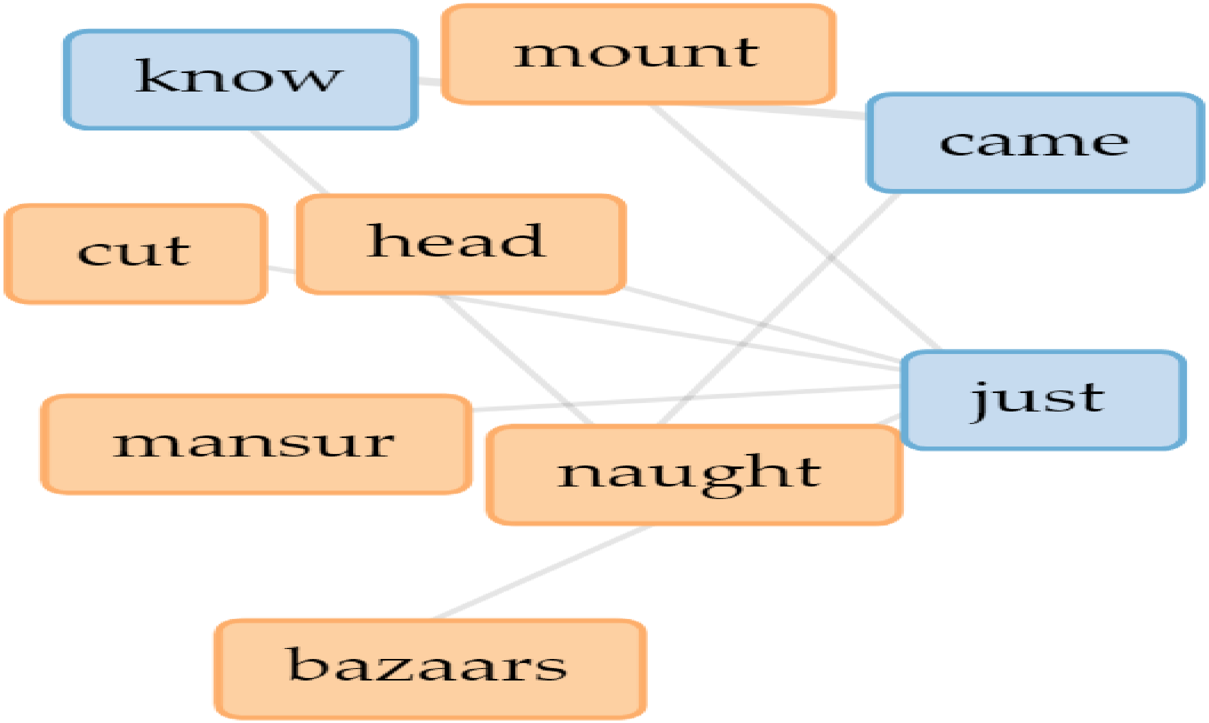
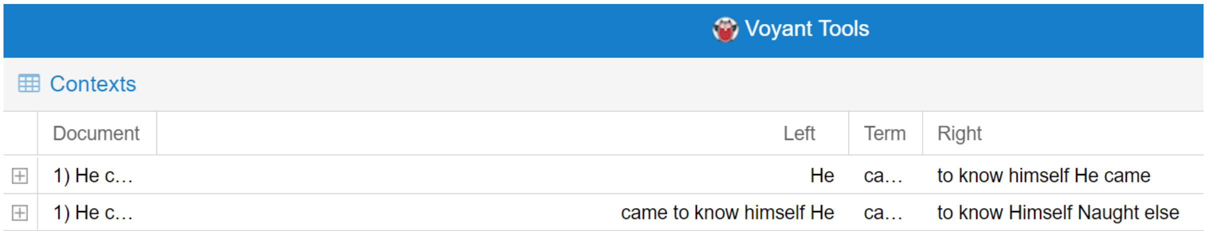


Figure 185 Links, He Came to Know Himself

The KG interweaves issues and themes to exhibit a more comprehensive understanding of the poem. The KG of "Mansur, just, head, cut" builds a relationship of laying life of Mansur Hallaj (858A.D. to 922 A.D), a Persian poet and Sufi, who chanted the slogan of "I'm truth" and shallow-minded people hanged him on the pretext of the claim of deity, while it was the expression of self-annihilation and submergence for God. Moreover, it reveals that the saint's own body and soul has diminished, and it has transformed into God's being, until all distances and alienation has come to an end. In reality, it is the lofty stage of mysticism that is beyond the comprehension of common worldly people.

Text mining cannot be replaced with human hermeneutic analytics in its entirety. These digital tools impart "an aid" in the process of hermeneutics. The link of "bazaars" is incomplete because it should have been linked with both words: "sold" and "slave". One postulate of the theory informs us that "Hermeneutic tools fail in interesting ways" (Rockwell, & Sinclair, 2016, p. 166). Such deficiencies are found only in poetic works because themes of poems are not recurrent in their text. Therefore, longer texts do not show such erroneous results.

v. Contexts



	Document	Left	Term	Right
⊞	1) He c…	He	ca…	to know himself He came
⊞	1) He c…	came to know himself He	ca…	to know Himself Naught else

Figure 186 Contexts, He Came to Know Himself

When the word "came" appears; apparently a person deduces the meaning of movement or arrival of somebody at some place. When an ambiguity occurs, the theoretical framework supports resolving the issue. "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166). To know its true semantic shade, the context of the word "came" is searched, and the word "came" occurs twice: once in the title and once in the first line of the poem. Consequently, the phrase "came to know" means to reveal reality and not for coming or reaching somewhere physically. To conclude, KWIC disambiguates syntactic and semantic ambiguity; and the same has been proved in an earlier study by Fischer (1971).

16. *God's Attributes* by Jalaluddin Rumi

This poem is written in Middle Persian or Pahlavi, and Dr Nicholson translates it.

i. Summary

This corpus has 1 document with 83 total words and 50 unique word forms. Created now (23rd September 2017).

Vocabulary Density: 0.602

Average Words Per Sentence: 20.8

Most frequent words in the corpus: god (4); calls (3); end (3); attributes (2); god's (2)

Figure 187 Summary, God's Attributes

Summary tool digitally quantifies stylistic attributes, including total words, unique words, vocabulary density, the average length of sentences and the most occurring words of Molana Rumi's poem, *'God's Attributes'*. Unique words are 50, and almost half words (+23) are repeated in the poem. Vocabulary density is 0.602 which is high; therefore, it suggests that this poem employs the most repeated vocabulary for beginner level readers. Usually, a full stop is marked after each second line; subsequently, 20.8 words in a sentence or a verse have been shown in the corpus summary.

ii. Cirrus

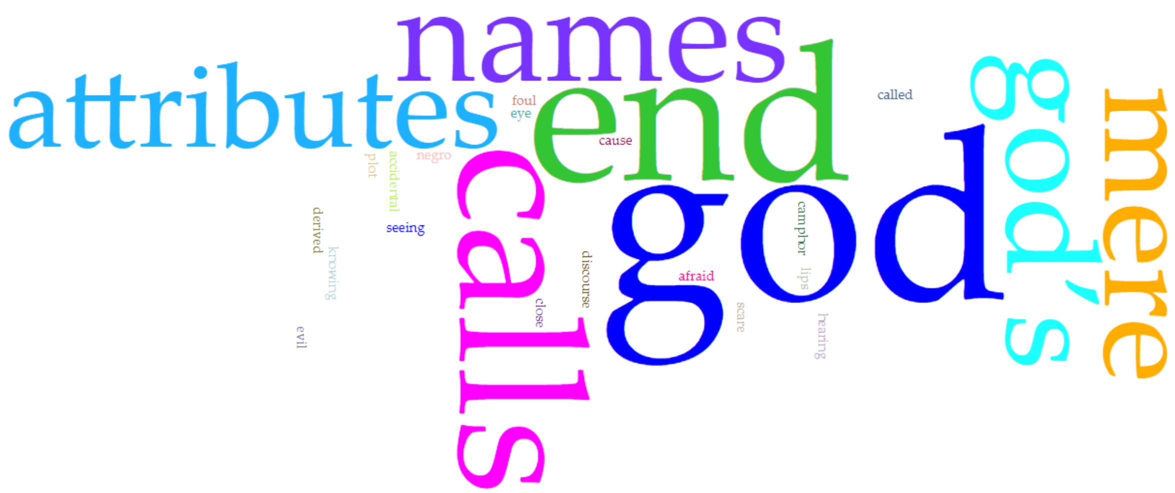
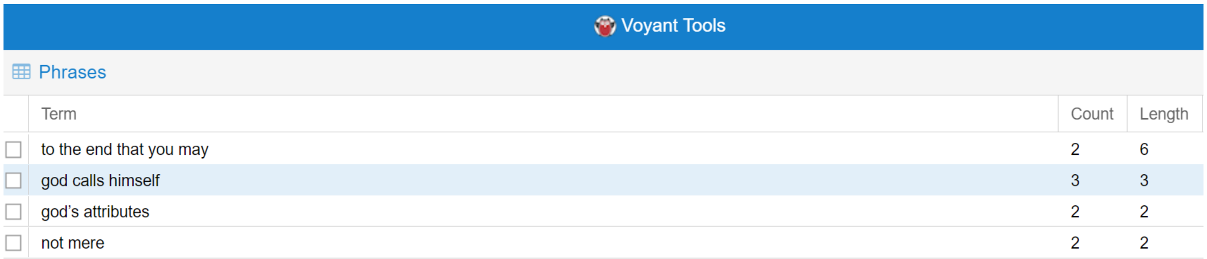


Figure 188 Cirrus, God's Attributes

Just a cursory glance at the word cloud/Cirrus discovers the knowledge pattern of the poem that it covers themes of "God (4)", "God's (2)" "attributes (2)" and His "names (2)". Therefore, the entire poem is about the attributes and names of Allah. The word "end (2)" discovers a knowledge pattern that each attribute has its extreme perfection, acme and matchlessness.

Other themes of the poem are "Knowing (1)", "Seeing (1)" and "Hearing (1)". Owing to their less occurrence, they are not written boldly in the Cirrus. Primarily these three divine characteristics or themes have been discussed in the poem. In conclusion, the statistical topic modelling formula functions in the word cloud to extract key themes. Sometimes Statistical information is misleading because attributes of "Knowing (1)", "Seeing (1)" and "Hearing (1)" occur once in the corpus, but these three divine attributes dominate in the poem.

iii. Phrases



	Term	Count	Length
□	to the end that you may	2	6
□	god calls himself	3	3
□	god's attributes	2	2
□	not mere	2	2

Figure 189 Phrases, God's Attributes

Four standard phrases are found in the poem: "God calls himself" (N+V+Ref Prn), "God's attributes" (N+Apo+N), "not mere" (Adv+Adj) and "to the end that you may" (Prep+Art+N+Prn+Prn+Mod). They occur 2 to 3 times, and their length varies from 2 to 6 words. Collocations/ n-grams also represent an ideology and contribute to linguistic fluency.

iv. Links

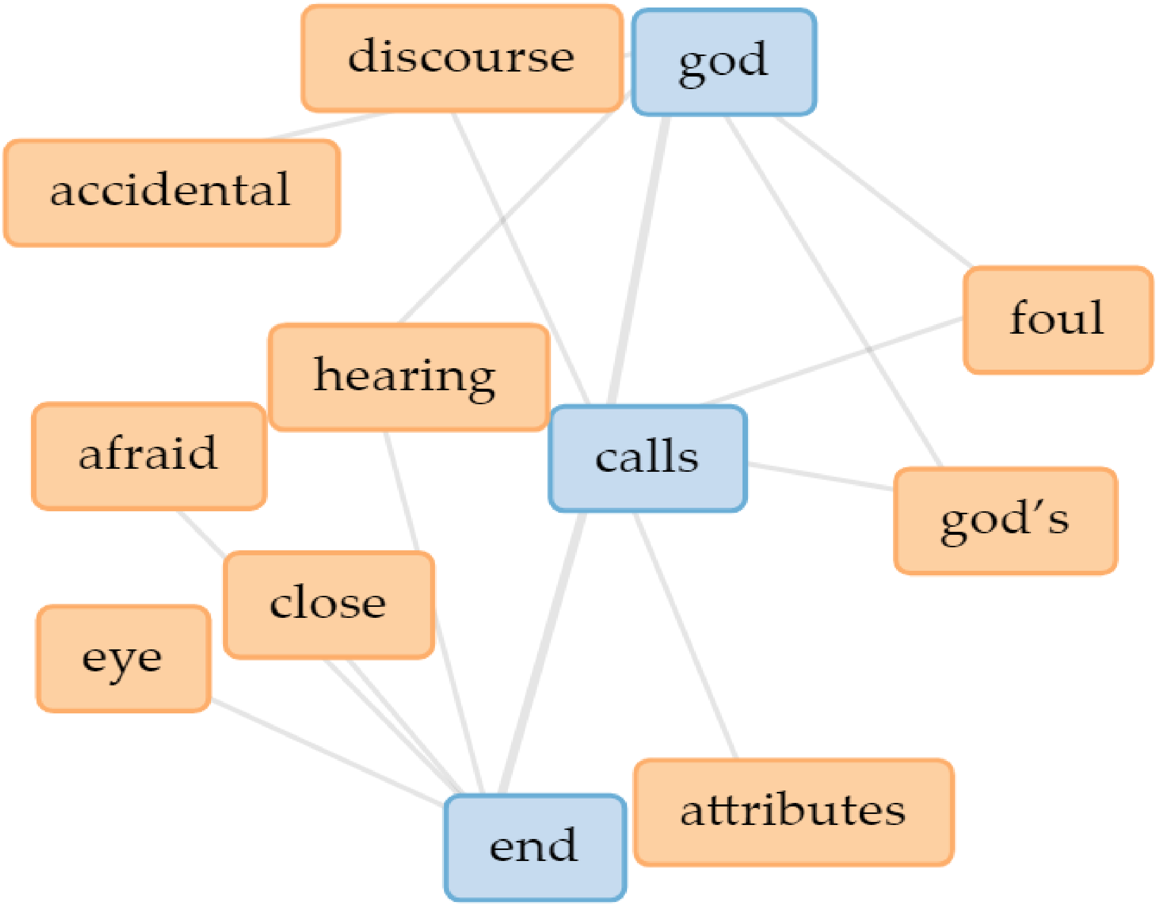
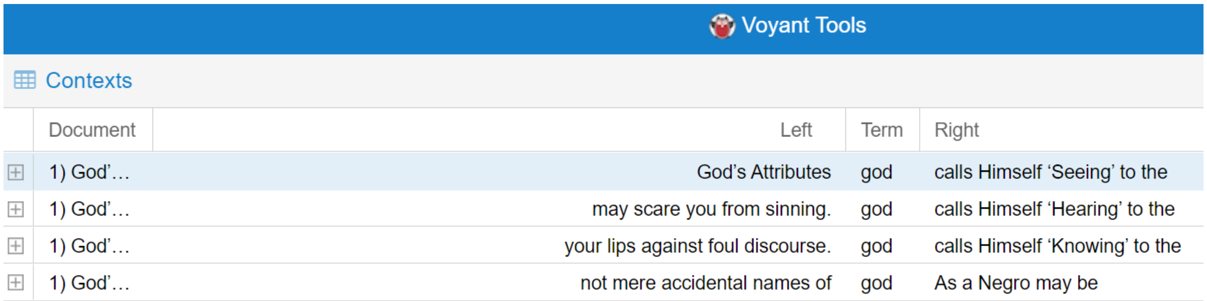


Figure 190 Links, God's Attributes

"In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). The KG of "God, hearing, end" suggests that God owns the matchless quality of hearing beyond all limits. One more KG "God, calls, end, eye" informs that God possesses the dexterous and incomparable characteristic of visualising; even nothing can hide itself from God's most extensive and unfathomable sight. Another query is who has named these divine qualities? It answers that God calls himself Master of all creatures because He owns the best and sublime characteristics, that is why its evidence is present in the KG of “God, calls, attributes”. Another dimension is an inquiry about the effects of these qualities on human beings. Human beings, by nature, intend to hide their wrongdoings for a face-saving strategy. Since God sees, hears and knows everything completely, human beings shun and remain fearful to commit any evil or foul act, and the same has been mentioned in the KG of “foul, God, hearing, afraid”.

v. Contexts



	Document	Left	Term	Right
⊞	1) God’…	God’s Attributes	god	calls Himself ‘Seeing’ to the
⊞	1) God’…	may scare you from sinning.	god	calls Himself ‘Hearing’ to the
⊞	1) God’…	your lips against foul discourse.	god	calls Himself ‘Knowing’ to the
⊞	1) God’…	not mere accidental names of	god	As a Negro may be

Figure 191 Contexts, God’s Attributes

“Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). Different religions believe in different attributes of God, and to clarify the concept of God, the context of the word “God” is searched, and it reveals three basic attributes of Allah: “Seeing”, “Hearing” and “Knowing”. God’s attributes are practical characteristics, and these names have been selected rationally but not arbitrarily.

17. *The Delight Song* by N. Scott Momaday

i. Summary

This corpus has 1 document with 169 total words and 68 unique word forms. Created now (24th September 2017).

Vocabulary Density: 0.402

Average Words Per Sentence: 84.5

Most frequent words in the corpus: alive (4); good (4); relation (4); stand (4); bright (2)

Figure 192 Summary, The Delight Song

Computational stylistics mines all textual data without close reading, and it precisely informs us about major stylistic characteristics. As N. Scott Momaday’s literary style is concerned, this poem has 68 unique words which have been repeated almost more than twice to count 169 total words. Based on total words and unique words, its vocabulary density is 0.402 which is suitable for beginner level readers.

Average words per sentence are 84.5, and figure 195 indicates the presence of extra-long sentences, but in reality, the first full stop is marked after 14 lines, and the second full stop is marked after six lines; therefore, it is named as enjambment in literary terms. In poems, completion of one thought requires a full stop; otherwise, usually, poetic lines are not so long. Apart from the word “bright (2)”, the most frequent words, “alive (4)”, “stand (4)”, “good (4)” and “relation (4)” are found in the last sestet of the poem.

ii. Cirrus

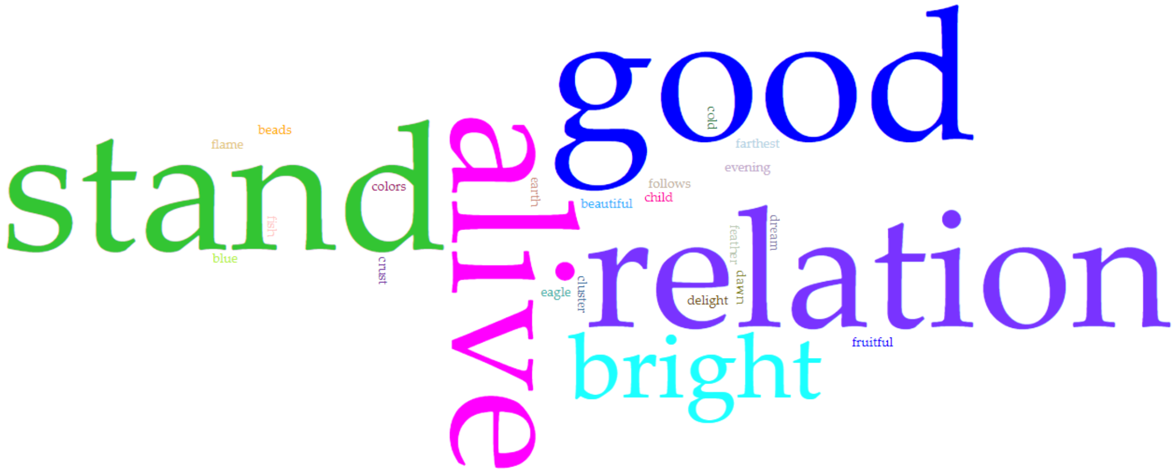
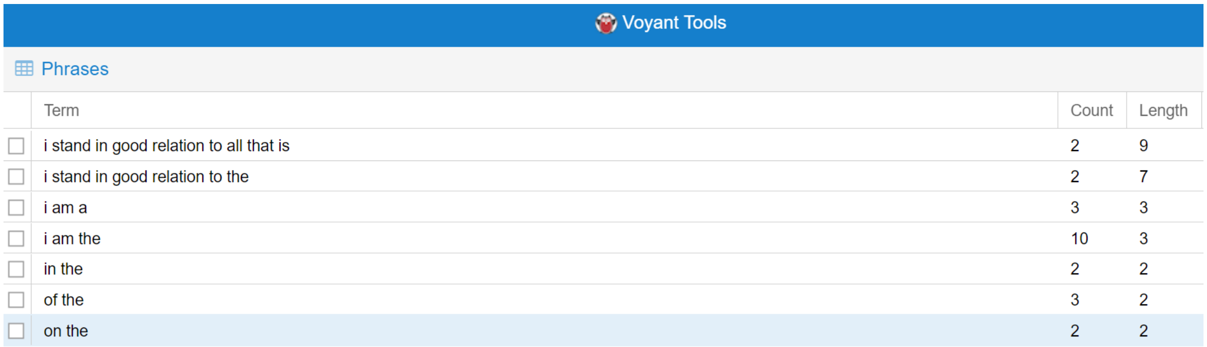


Figure 193 Cirrus, The Delight Song

Data visualization of a word cloud is based on statistical weight and topic modelling. These prominent words are “alive (4)”, “good (4)”, “relation (4)” and “stand (4)”. Usually, the most occurring word is a central theme, but this poem is an exception. In fact, deriving delight from natural beauty is the key theme of the poem, but it is shown as a minor theme due to its less statistical weight. Its minor themes show “delight morning (1)”, “snow (1)”, “meadows (1)”, “cluster (1)”, “fish (1)”, “moon (1)”, “song (1)”, and all of them are associated with pleasures of the phenomenon of nature.

iii. Phrases



	Term	Count	Length
□	i stand in good relation to all that is	2	9
□	i stand in good relation to the	2	7
□	i am a	3	3
□	i am the	10	3
□	in the	2	2
□	of the	3	2
□	on the	2	2

Figure 194 Phrases, The Delight Song

This poem generates seven repeated phrases, and among them, two phrases are long. The standard phraseology is “I am the” (Prn+Aux+Art), “I am a” (Prn+Aux+Art), “I stand in good relation to all that is” (Prn+V+Prep+Adj+N+Prep+Adv+Prn+Aux), and “I stand in good relation to the” (Prn+V+Prep+Adj+N+Prep+Art). They occur 2 to 10 times; however, their length varies from 2 to 9 words. Collocations/ n-grams enhance linguistic fluency in four language skills. They communicate the poet's ideology that the use of the pronoun “I” reveals the knowledge pattern that the poet talks most of the time about himself with a self-centred approach.

iv. Links

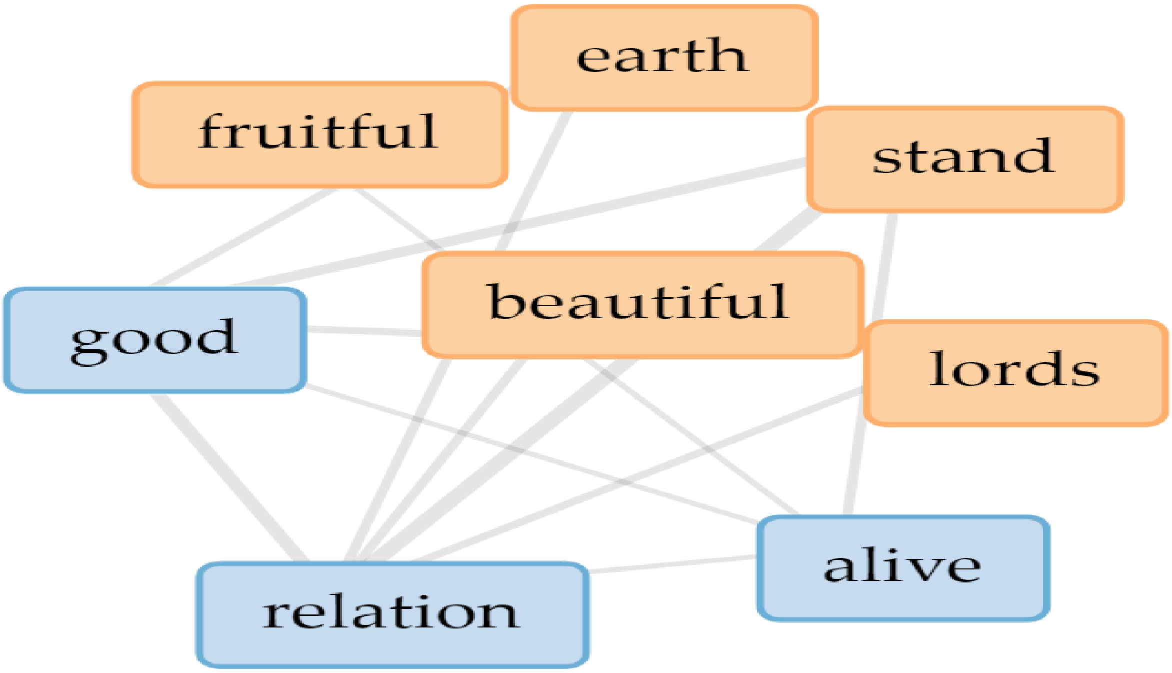
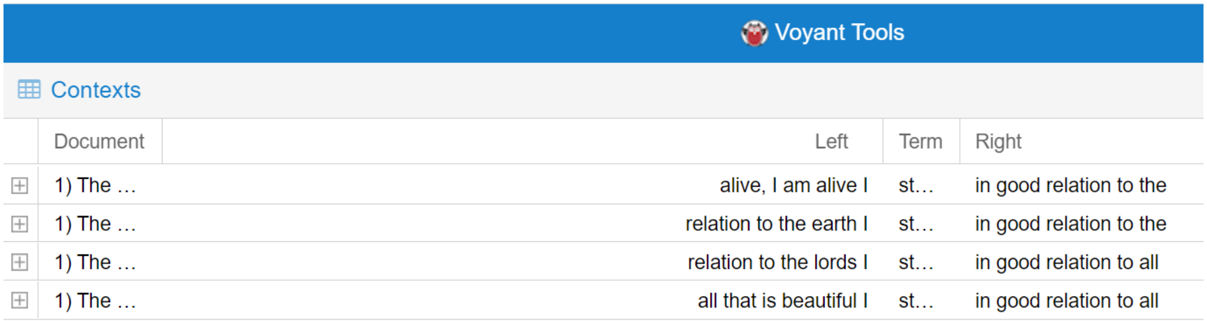


Figure 195 Links, The Delight Song

The KG of "stand, good, relation, earth" is based on the poetic line "I stand in good relation to the earth". Second KG consists of "stand, lords, relation, good" which verifies the poetic line, "I stand in good relation to the lords". Apart from it, the poetic line, "I stand in good relation to all that is beautiful" is visualised with the KG of "stand, beautiful, relation, good". Then another KG about "stand, relation, good, fruitful" is connected to harmonise this poetic line "I stand in good relation to all that is fruitful".

Here one shortcoming is that the Links tool develops links only for the last sestet, ignoring the first fourteen lines. "Hermeneutic tools fail in interesting ways (Rockwell, & Sinclair, 2016, p. 166). The solution to this shortcoming is the extension of KGs. Likewise, "Hermeneutica Theory emphasizes that "They can be extended to expose new things" (Rockwell, & Sinclair, 2016, p. 166).

v. Contexts



	Document	Left	Term	Right
⊞	1) The …	alive, I am alive I	st…	in good relation to the
⊞	1) The …	relation to the earth I	st…	in good relation to the
⊞	1) The …	relation to the lords I	st…	in good relation to all
⊞	1) The …	all that is beautiful I	st…	in good relation to all

Figure 196 Contexts, The Delight Song

The word "stand" has dual meanings: to be in a vertical position or to have an opinion. Theoretical underpinning "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166); therefore, context should be searched for word sense disambiguation. The word "stand" is retrieved four times in this poem as "I stand in good relation", and every time, its semantic shade is to have an opinion. The poet has developed good relationships with various phenomena of Nature.

18. *Love – An Essence of All Religions* by Jalaluddin Rumi

Originally Jalaluddin Rumi wrote this poem in Pahalvi (Middle Persian). Later, it was translated into English by Dr Nicholson.

i. Summary

This corpus has 1 document with 77 total words and 41 unique word forms. Created now (on 24th September 2017).

Vocabulary Density: 0.532

Average Words Per Sentence: 77.0

Most frequent words in the corpus: love (12); becomes (9); become (1); burning (1); butter (1)

Figure 197 Summary, Love – An Essence of All Religions

The stylometric analysis quantifies the literary style of Molana Rumi's poem, *'Love – An Essence of All Religions'*. These tools are designed with the training of algorithms (Argamon et al., 2003; Zhang et al., 2002), and the application of SVM to digital tools (Stamatatos, 2009). In this poem, 41 unique words and 77 total words have been used. Each unique word is used almost twice in this poem. The average words per sentence are 77.0 which show the use of enjambment, because the full stop is marked at the end of the entire poem. The most frequent word is "love" which changes negative situations into positive situations.

ii. Cirrus

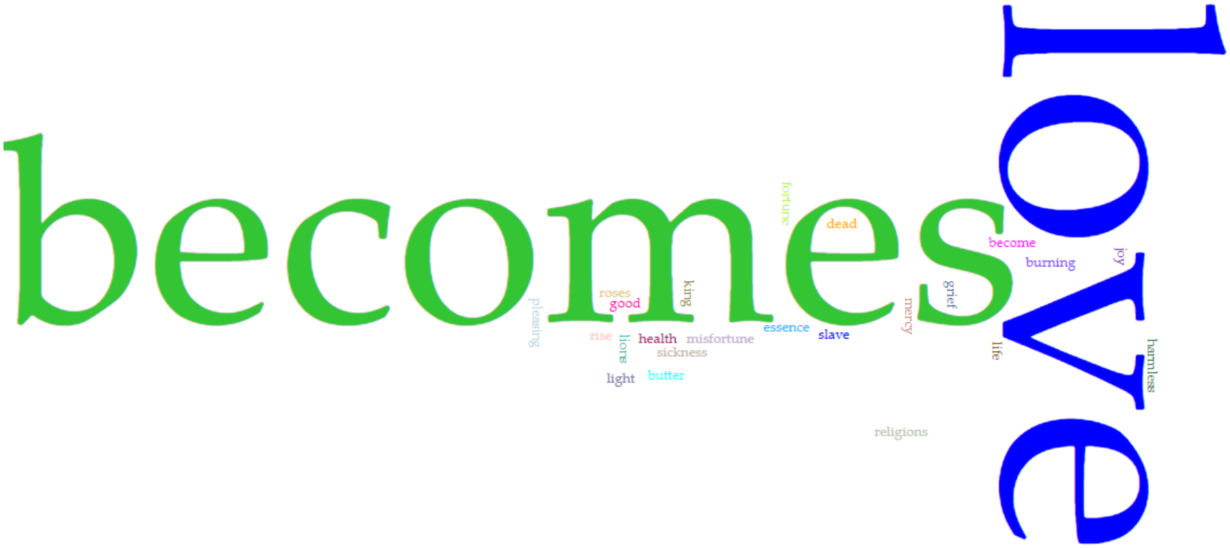
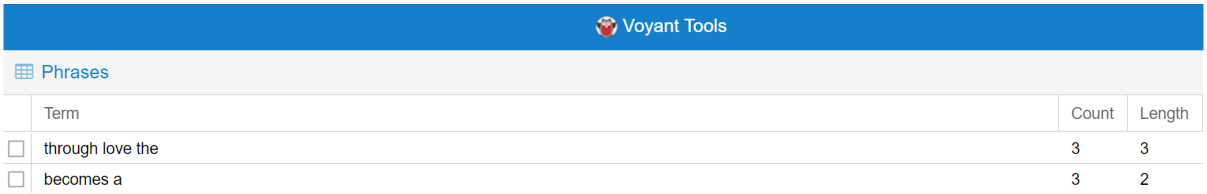


Figure 198 Cirrus, Love – An Essence of All Religions

Primarily, the topic modelling process in this Cirrus highlights only two words ("love", "becomes"), and minor themes have been mentioned in a very small font. This mystic poem centres upon the theme of "love (12)", and its magical transforming power to change negative situations into positive conditions. The second theme is narrated by the word "becomes (9)" which suggests radical changes caused by love. Minor themes are in pairs, and they are a foil to each other, for instance, “thorns (1)” and “roses (1)”; “vinegar (1)” and “wine (1)”; “misfortune (1)” and “good fortune (1)”; “grief (1)” and “joy (1)”; “dead (1)” and “rise (1)”; and “wrath (1)” and “mercy (1)”. To conclude, thesis and antithesis exhibit a spectrum of contrasting themes in this Cirrus.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	through love the	3	3
□	becomes a	3	2

Figure 199 Phrases, Love – An Essence of All Religions

Meaningless phraseology “becomes a” has been excluded from the list of standard phraseology because it is useless in academic settings. The standard trigram “through love the” (Prep+N+Art) occurs three times in the poem. Phraseology teaches grammar, correct language and fluency in four language skills.

iv. Links

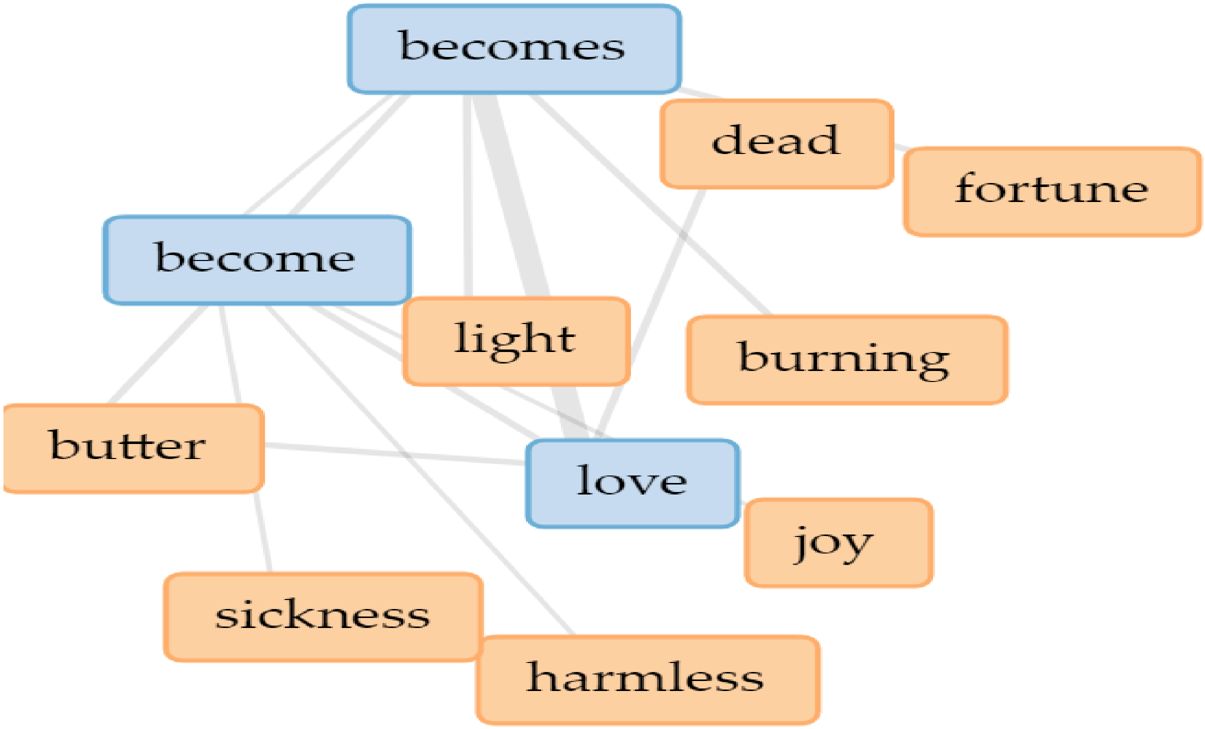
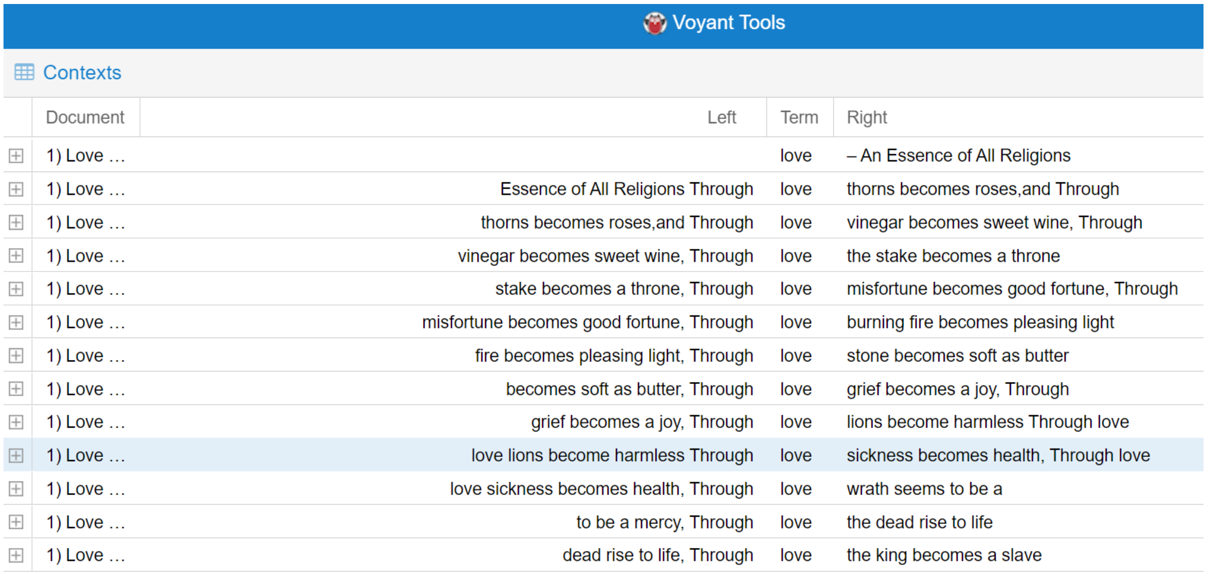


Figure 200 Links, Love – An Essence of All Religions

The KG of “burning, becomes, light” certifies the poetic line “Through love burning fire becomes pleasing light”. Firstly, the noun is removed, and its adjective is kept, then the adjective is removed, and its noun is kept in figure 199. Two possibilities can be functional behind it: firstly, in machine learning, an algorithm is trained to do the aforementioned procedure; secondly, there are a few flaws in Links tool. One postulate of the theory admits thus: “Hermeneutic tools fail in interesting ways” (Rockwell, & Sinclair, 2016, p. 166). This failure leads to the development of digital tools with the training of more data. The current study suggests a solution to this issue by deleting both adjectives (burning, pleasant) and linking two nouns (fire, light).

The theme of “become” shows the transformation from one situation to another, how love miraculously changes one negative characteristic into one positive attribute. That is why love is a change-producing agent in human beings and animals. One KG, “dead, love, become” reveals that the passion of love influences three negative things to change them into positive ones. Positive effects and fruits of love are shown in another KG with the nodes of “fortune, becomes, light”. To conclude, relationship mining among relevant variables (Barahate, 2012, p. 13) has been done through KGs.

v. Contexts



	Document	Left	Term	Right
⊞	1) Love …		love	– An Essence of All Religions
⊞	1) Love …	Essence of All Religions Through	love	thorns becomes roses,and Through
⊞	1) Love …	thorns becomes roses,and Through	love	vinegar becomes sweet wine, Through
⊞	1) Love …	vinegar becomes sweet wine, Through	love	the stake becomes a throne
⊞	1) Love …	stake becomes a throne, Through	love	misfortune becomes good fortune, Through
⊞	1) Love …	misfortune becomes good fortune, Through	love	burning fire becomes pleasing light
⊞	1) Love …	fire becomes pleasing light, Through	love	stone becomes soft as butter
⊞	1) Love …	becomes soft as butter, Through	love	grief becomes a joy, Through
⊞	1) Love …	grief becomes a joy, Through	love	lions become harmless Through love
⊞	1) Love …	love lions become harmless Through	love	sickness becomes health, Through love
⊞	1) Love …	love sickness becomes health, Through	love	wrath seems to be a
⊞	1) Love …	to be a mercy, Through	love	the dead rise to life
⊞	1) Love …	dead rise to life, Through	love	the king becomes a slave

Figure 201 Contexts, Love – An Essence of All Religions

Firstly, there is a lexical ambiguity whether the word “love” is a verb or a noun. Secondly, writing a comprehensive note on “love” and finding textual quotes from big data is a big challenge. To answer to both queries, Contexts tool extracts contextual information from KWIC that the word "love" is used 12 times in this poem. Furthermore, figure 201 reveals that the passion for love brings drastic changes from negative predicaments to noble things. Secondly, the word "love" is used twelve times as a noun. These findings are in harmony with the work of Fischer (1971).

19. *A Man of Words and Not of Deeds* by Charles Perrault

Originally this poem was written in French, and later it was translated into English by Robert Samber.

i. Summary

This corpus has 1 document with 121 total words and 47 unique word forms. Created now (on 24th September 2017).

Vocabulary Density: 0.388

Average Words Per Sentence: 17.3

Most frequent words in the corpus: like (7); it's (6); begins (5); dead (3); bird (2)

Figure 202 Summary, A Man of Words and Not of Deeds

"Summarization involves methods for finding a compact description for a subset of data" (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). Computational stylistics extracts precise and accurate knowledge patterns regarding the literary style of any poet or translator. Robert Samber uses 47 unique words, and they are repeated almost three times in the entire poem, so the total number of words of the poem are 121. Its vocabulary density is 0.388 which is suitable for intermediate level readers. The average number of words per sentence is 17.3 because the full stop is placed after each quatrain. The most frequent word in the poem is "like (7)", and it refers to the excessive use of similes to exemplify the message.

ii. Cirrus

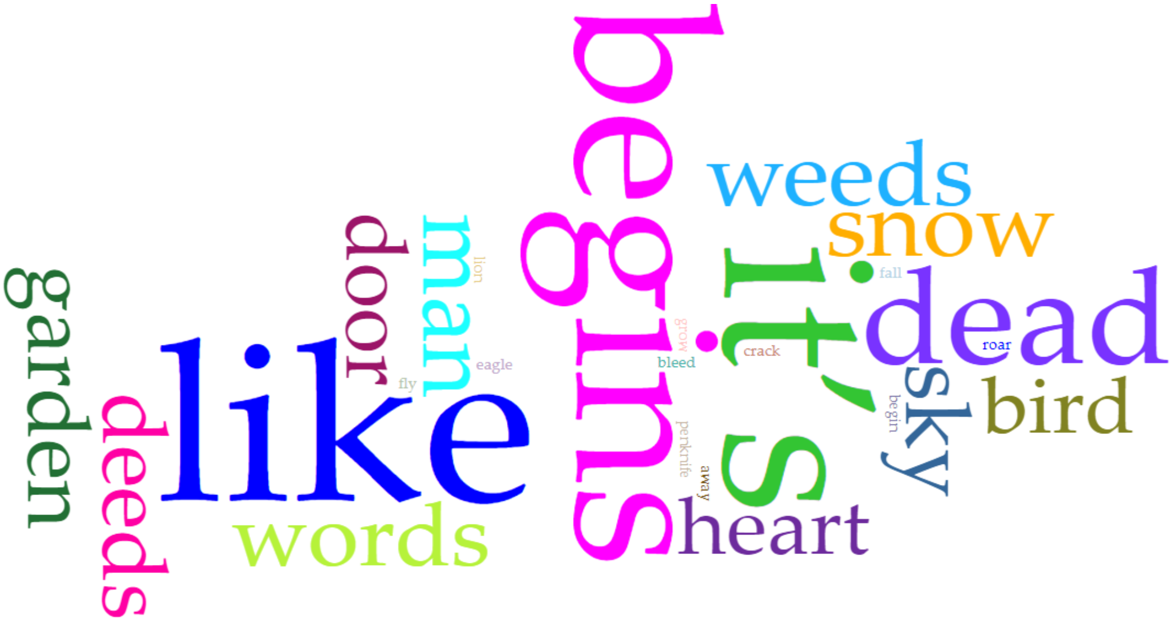
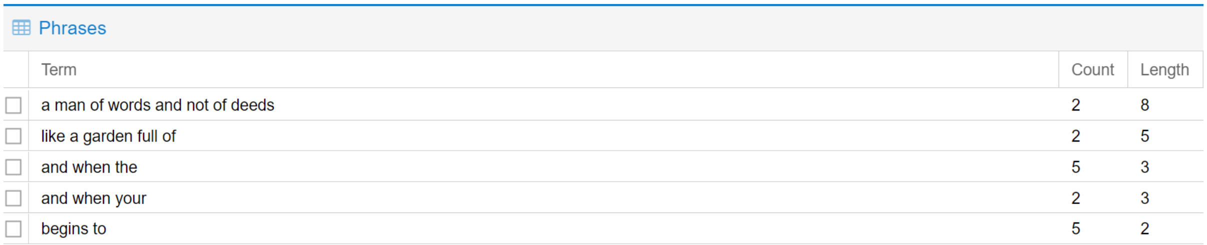


Figure 203 Cirrus, A Man of Words and Not of Deeds

This Cirrus reveals key themes of the poem with the topic modelling process in NLP. This word cloud exposes several bold words, so its statistical reason is that each main theme occurs twice in the poem. The translator and the poet use one word in the first line, then they extend the idea with the same word in the next line; hence, one word is used twice in the poem, that is why most of the Cirrus words are in bold font. On the other hand, Cirrus of the 6th poem *'A Sindhi Woman'* is in thin font because it has not been repeated twice.

The word "like (7)" reveals the knowledge pattern that seven similes have been used in the poem. The phrase "it is (6)" shows neutrality, and the word "dead (3)" is used three times in the last line; hence, it indicates that a man of words may bleed and die morally and spiritually. Some other themes are "weeds (2)", "deeds (2)", "bird (2)", "heart (2)", "sky (2)", "garden (2)" and "snow (2)" which discuss other themes of the poem.

iii. Phrases



Phrases

	Term	Count	Length
□	a man of words and not of deeds	2	8
□	like a garden full of	2	5
□	and when the	5	3
□	and when your	2	3
□	begins to	5	2

Figure 204 Phrases, A Man of Words and Not of Deeds

This poem uses the phraseology of "begin to" (V+Prep), "a man of words and not of deeds" (Art+N+Prep+N+Conj+Adv+Prep+N), and "like a garden full of weeds" (Prep+Art+N+Adj+Prep+N). They occur 2 to 5 times, and their length ranges between 2 to 8 words. The ideology is conveyed that just lip service and shrinking practical work devastate a person's life completely. Such bluff words bring forth only weeds and useless shrubs, while real hard work bears fruit.

iv. Links

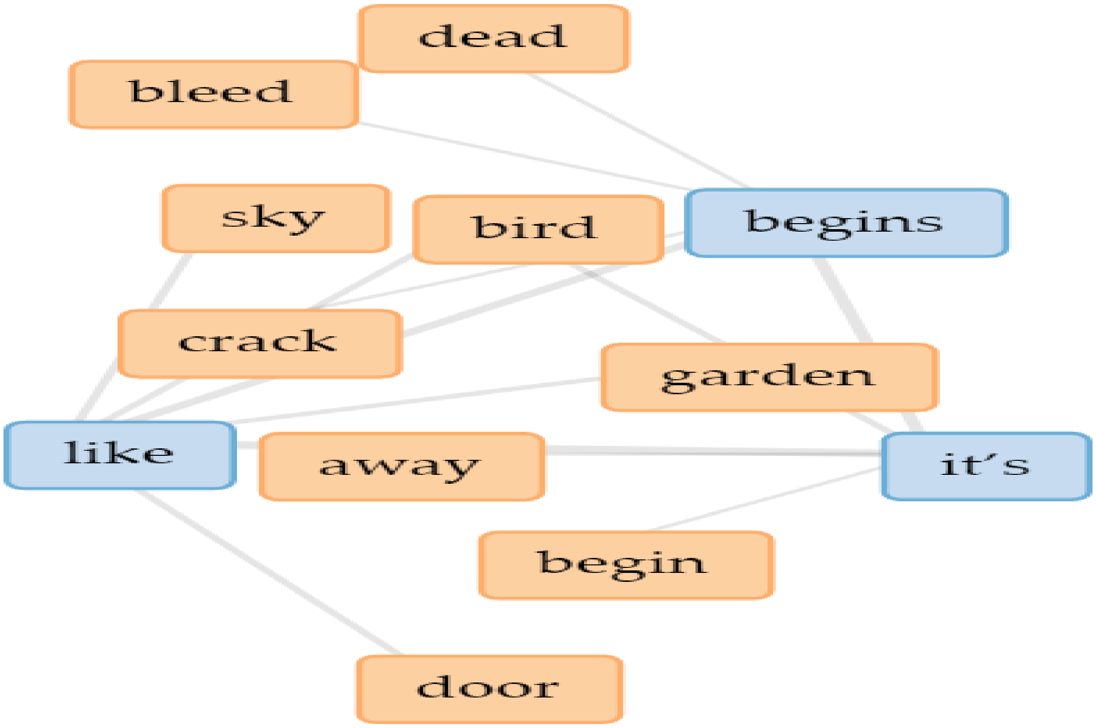
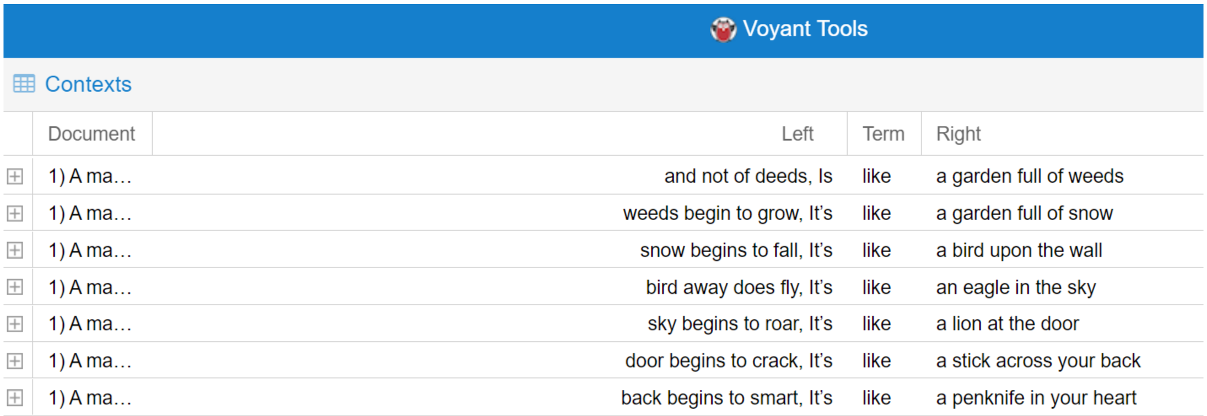


Figure 205 Links, A Man of Words and Not of Deeds

The KG of "like, garden" is an evidence of poetic text "Is like a garden full of weed", but six other similes are not shown in these poetic lines, for instance, "like a garden full of snow", "like a bird upon the wall", "like an eagle in the sky", "like a lion at the door", "like a stick across your back" and "like a pen in your heart". One poetic line, "And when the bird away does fly" finds the KG of "bird, like, away". The poetic line "And when the door begins to crack" has been mentioned in the KG of "crack, like, door". Another textual phrase "begins to bleed" is shown in the KG of "bleed, begins". "It is supplemented by other materials" (Rockwell, & Sinclair, 2016,

p. 166). To conclude, hermeneutic emphasizes the use of textual evidence for supplementing the data visualization.

v. Contexts



	Document	Left	Term	Right
⊞	1) A ma…	and not of deeds, Is	like	a garden full of weeds
⊞	1) A ma…	weeds begin to grow, It's	like	a garden full of snow
⊞	1) A ma…	snow begins to fall, It's	like	a bird upon the wall
⊞	1) A ma…	bird away does fly, It's	like	an eagle in the sky
⊞	1) A ma…	sky begins to roar, It's	like	a lion at the door
⊞	1) A ma…	door begins to crack, It's	like	a stick across your back
⊞	1) A ma…	back begins to smart, It's	like	a penknife in your heart

Figure 206 Contexts, A Man of Words and Not of Deeds

The word “like” creates semantic and grammatical ambiguity for distant readers. Theoretical underpinning “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). Disambiguation is essential for the true explanation of any text. To answer to this lexical confusion, Contexts tool retrieves the context of the word “like” in figure 206. It is found that seven similes are used with the word “like”, for instance, “like a garden full of weeds”, “like a garden full of snow”, “like a bird upon the wall”, “like an eagle in the sky”, “like a lion at the door”, “like a stick across your back” and “like a pen in your heart”. Furthermore, the word “like” is not used as an action verb or with a semantic shade of “adore”. To conclude, Contexts tool aptly differentiates between similes and main verbs. Moreover, it also extracts all similes from the big data of a literary database.

20. *In Broken Images* by Robert Graves

i. Summary

This corpus has 1 document with 112 total words and 43 unique word forms. Created now (22nd September 2017).

Vocabulary Density: 0.384

Average Words Per Sentence: 16.0

Most frequent words in the corpus: images (9); broken (4); fact (4); relevance (4); clear (3)

Figure 207 Summary, In Broken Images

Summary tool elucidates stylistic qualities of Robert Graves' poem, *'Broken Images'*. He uses 43 unique words, and he repeats them almost three times; hence, the total words are 112. Vocabulary density is 0.384 which is appropriate for beginner level readers. The average number of words per sentence are 16.0 because the full stop is placed at the end of each verse. The most recurring themes are "images" which are abstract and subjective in nature. There is a further discussion about "broken", "clear" "images", "facts". In this world, seeking meanings and clarity is the biggest challenge for contemplative human beings.

ii. Cirrus

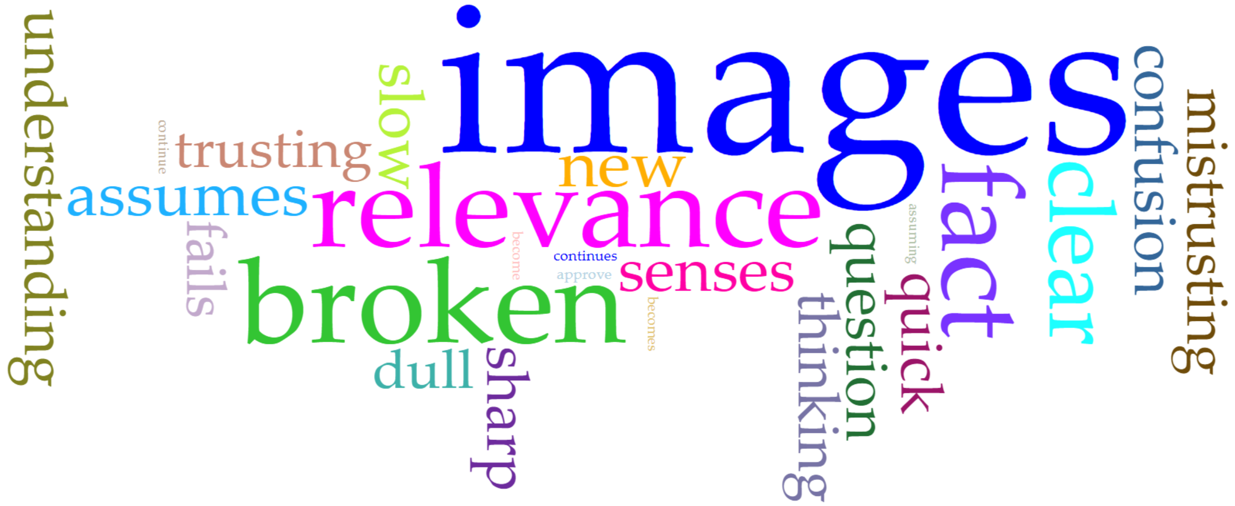
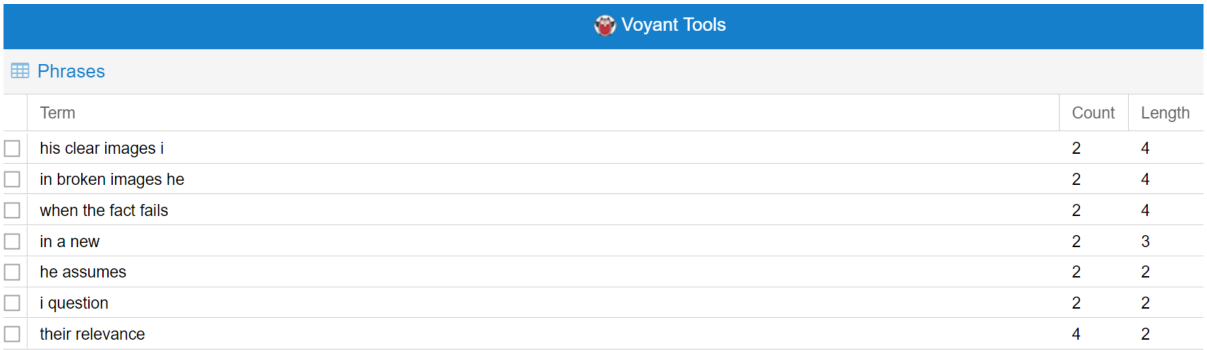


Figure 208 Cirrus, In Broken Images

This poem mentions two unnamed characters, namely "he (7)" and "I (7)", and they converse mutually and logically in this dialogic poem. Cirrus has been generated with the use of topic modelling, and a very interesting knowledge pattern is found that several opposing themes have been delineated in the poem, for instance, "trusting (2)" and "mistrusting (2)"; "assumes (2)" and "question (2)"; "sharp (2)" and "dull (2)"; and "confusion (2)" and "understanding (2)". These opposing themes inform about two opposing dialogues by two different characters who are a foil to each other. Themes of "clear (3)" and "broken (4)" "images (9)" and their "relevance (4)" have been discussed in the poem.

iii. Phrases



	Term	Count	Length
□	his clear images i	2	4
□	in broken images he	2	4
□	when the fact fails	2	4
□	in a new	2	3
□	he assumes	2	2
□	i question	2	2
□	their relevance	4	2

Figure 209 Phrases, In Broken Images

Standard phrases of this poem are “their relevance” (Prn+N), “he assumes” (Prn+V), “his clear images” (Prn+Adj+N), “I question” (Prn+V), “in a row” (Prep+Art+N), “in broken images” (Prep+Adj+N) and “when the fact fails” (Adv+Art+N+V). There are seven collocation patterns/ n-grams and among them, three bigrams, three trigrams and one quadgram have been used, and they occur 2 to 4 times, whereas their length ranges from 2 to 4 words.

Students can also enhance their grammatical competence and performance through phraseology; for instance, after a possessive pronoun, the use of noun is compulsory as it is used in the phrase “their relevance”. If a subject is a third-person pronoun or a singular noun, inflection “s” is added with a verb as it happens in the phrase “he assumes” and “when the fact fails”, while inflection “s” is not added with a first and second person for example “I question”. To conclude, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). In the collocation extraction process, all repeated collocations have been counted and presented in the compact tabular form. Thus, the linguistic data have been transformed into quantified knowledge patterns.

iv. Links

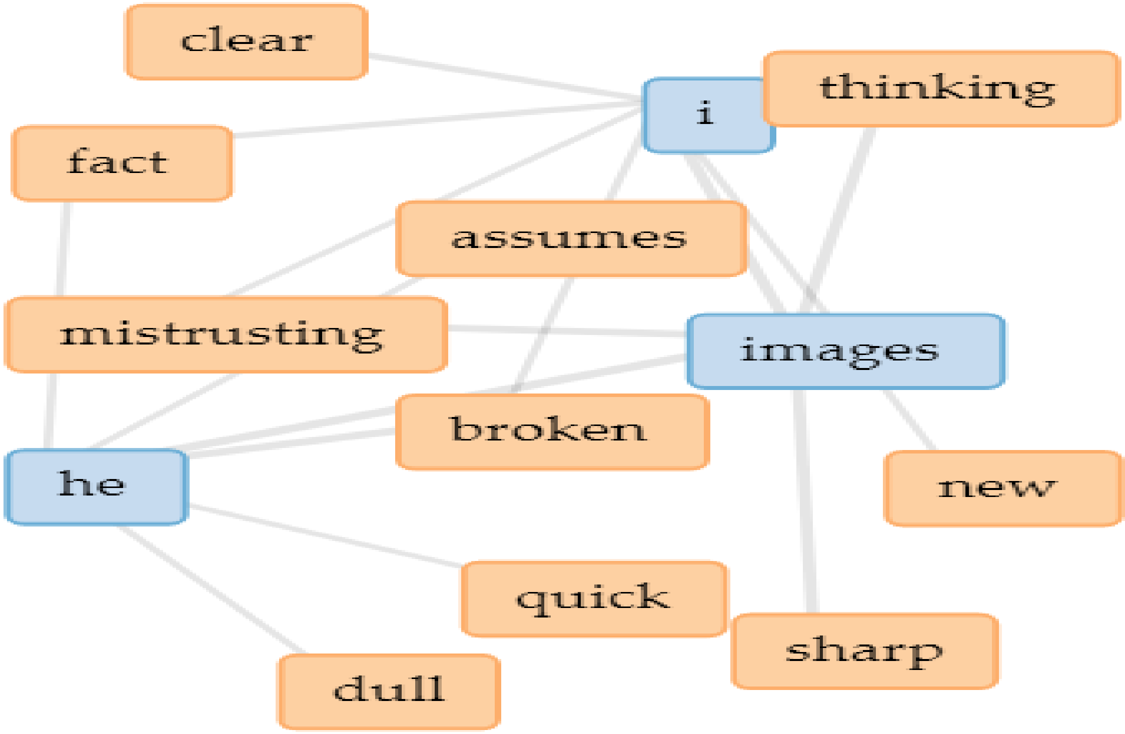
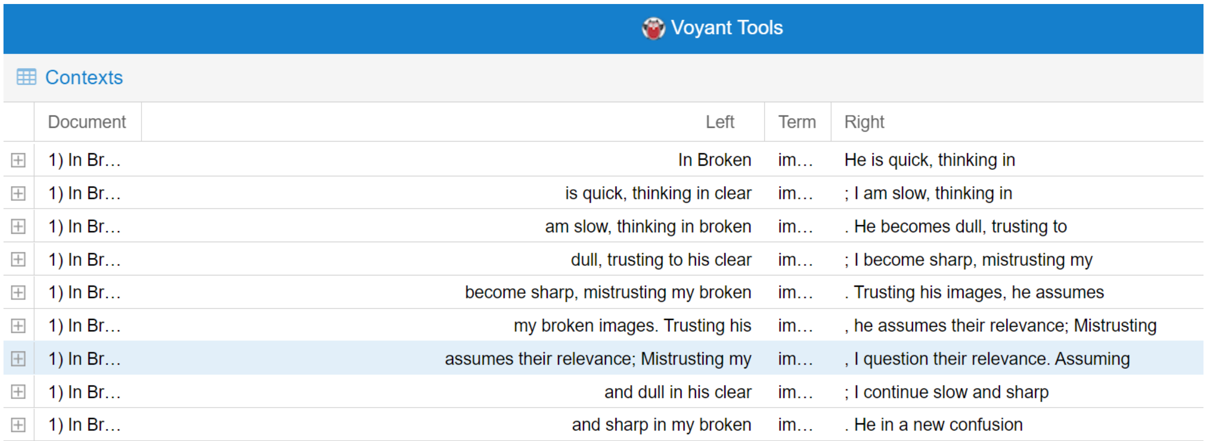


Figure 210 Links, In Broken Images

The KG interlinks different characters and themes to discover new knowledge patterns. One KG links "I, he" since both speakers progress the poem with a poetic turn-taking process. Another poetic line, "He is quick, thinking in clear images;" finds the relevant KG of "he, quick, clear". One KG of "I, images, broken" verifies the poetic line "I am slow, thinking in broken images." Hermeneutica Theory is "supplemented by other materials" (Rockwell, & Sinclair, 2016, p. 166). Here KGs have been supplemented and verified by the poetic lines. Moreover, it establishes concurrent validity for knowledge graphs.

Three-pronged KG of "images, mistrusting" is generated under the influence of the 6th line, "Mistrusting my images, I question their relevance." Apart from it, the KG of "images, he, dull" verifies the poetic line, "He continues quick and dull in his clear images;"

v. Contexts



Document	Left	Term	Right
1) In Br…	In Broken	im…	He is quick, thinking in
1) In Br…	is quick, thinking in clear	im…	; I am slow, thinking in
1) In Br…	am slow, thinking in broken	im…	. He becomes dull, trusting to
1) In Br…	dull, trusting to his clear	im…	; I become sharp, mistrusting my
1) In Br…	become sharp, mistrusting my broken	im…	. Trusting his images, he assumes
1) In Br…	my broken images. Trusting his	im…	, he assumes their relevance; Mistrusting
1) In Br…	assumes their relevance; Mistrusting my	im…	, I question their relevance. Assuming
1) In Br…	and dull in his clear	im…	; I continue slow and sharp
1) In Br…	and sharp in my broken	im…	. He in a new confusion

Figure 211 Contexts, In Broken Images

The word "image" carries the meaning of a picture and an abstract impression. To disambiguate word sense, Contexts tool is employed. Figure 211 reveals that the word "images" has been used 9 times in this dialogic poem, and the dialogue continues between two unnamed characters "he" and "I". Both characters argue the pros and cons of clear and broken images. In fact, both characters argue three times each; that is why both words "clear" and "broken" are used three times in the poem. Once the word "image" is used in the title, and it is used twice with personal possessive pronouns, namely "my" and "his". Besides, this tool also finds adjectives before the searched noun.

4.7 Data Analysis of Book II

Book II (Appendix C) covers 10 literary essays and 5 biographical essays. It is a very old and literary compilation written by British prose writers.

4.8 Text Mining of Literary Essays

It is the composition of a literary essay focusing on one topic, and the essayist gives his/her own opinions and understanding about one topic.

1. *The Dying Sun* by Sir James Jeans

i. Summary

This corpus has 1 document with 1,046 total words and 376 unique word forms. Created about 3 minutes ago (on 31st December 2017).

Vocabulary Density: 0.359

Average Words Per Sentence: 23.8

Most frequent words in the corpus: life (13); space (12); sun (10); star (8); universe (7); earth (6); stars (6); hot (5); immense (5); like (5)

Figure 212 Summary, The Dying Sun

“Summarization shows a condensed report of mined data” (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45), and it reflects stylistic patterns with statistical evaluation. Sir James Jeans writes 376 unique words which are repeated about three times, and 1046 total words are found in this essay. Its vocabulary density is 0.359, and it is calculated by dividing unique words with a denominator of total words. Higher vocabulary density refers to less repetitive and challenging vocabulary, whereas lower vocabulary density suggests frequently repetitive and easy vocabulary usage. It reveals that one word is repeated almost three times in this essay, and this repeated vocabulary causes ease for readers. Sir James Jeans constructs 23.8 words long sentences on average, and they are suitable for highly advanced level readers. The most frequent words do concept mining and express “life (13)”, “space (12)”, “sun (10)”, “stars (6)” and “earth (6)”, that how the sun causes the existence of life on the earth, and how it maintains life by sustaining normal temperature which is integral for human survival. To conclude, findings of the current stylometric study harmonise with some previous studies: Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O’Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018 as mentioned in Ch.2.

ii. Cirrus

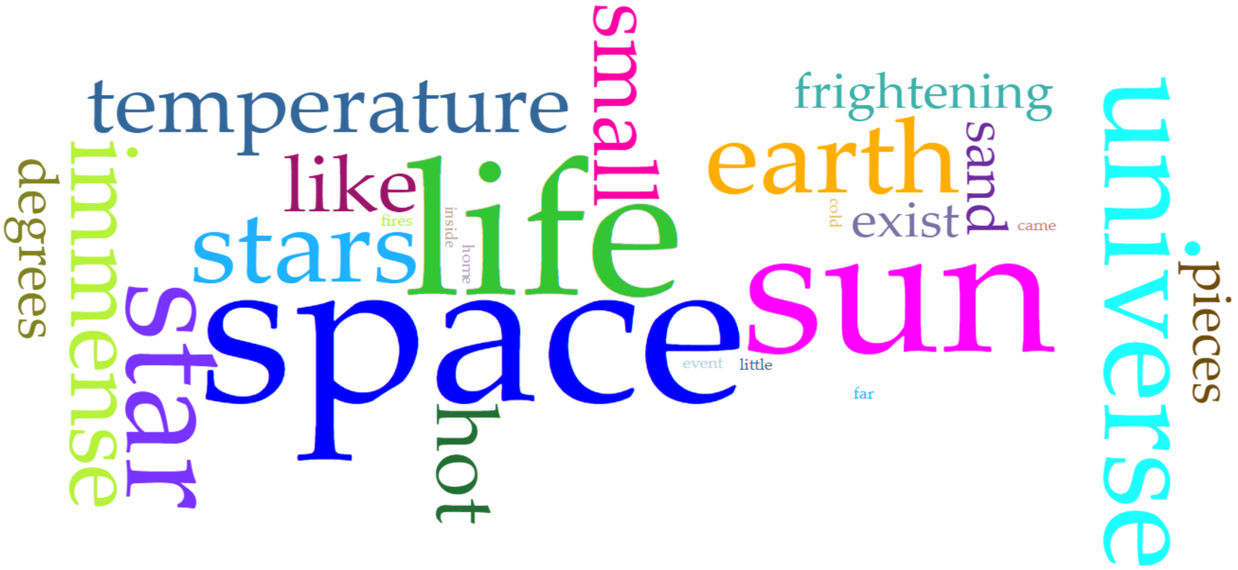
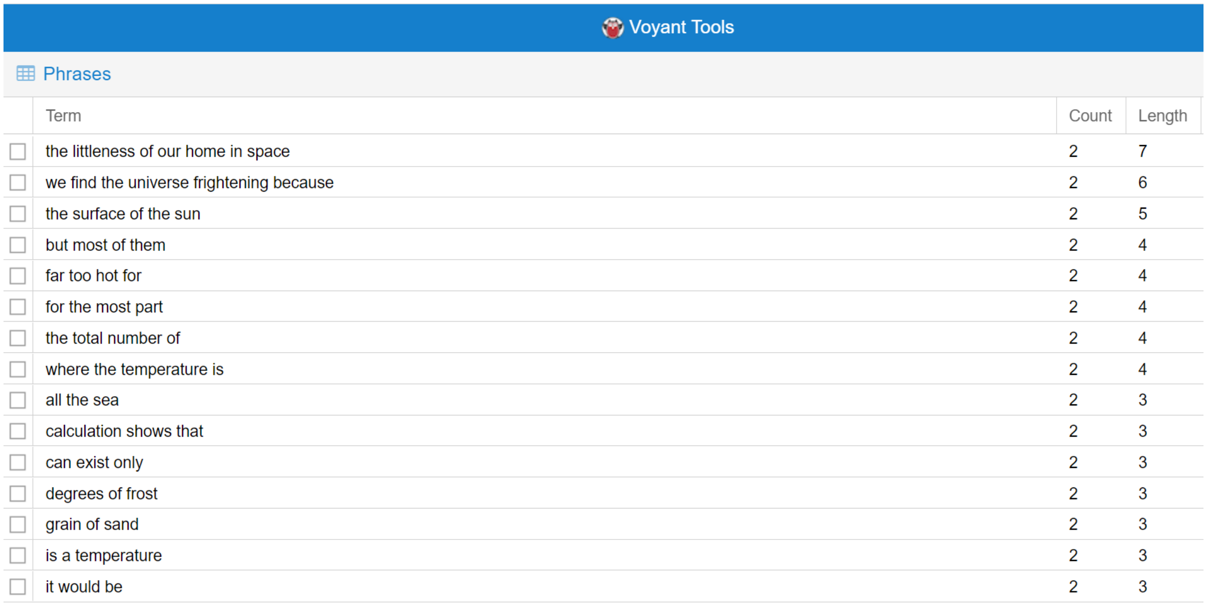


Figure 213 Cirrus, The Dying Sun

Data compression, quantification and linguistics are vital features of Information Theory (Shannon, 2009). Here all textual data have been presented in the compact as well as in the quantified form. Cirrus extracts vital topics from the text with statistical weight and topic modelling. The most occurring theme of this corpus is “life (13)” because the distance and pivotal function of the sun sustains life on the earth. In fact, the “sun (10)” or the “star (8)” is a unifying force that keeps all planets revolving around it; consequently, the entire solar system is functioning smoothly to maintain life on the earth. It seems to scientists that life is not aimed at the creation of the universe because it is void of life and its requirements. Numerous stars have been exemplified with grains of “sand (4)”. Characteristically, the “sun (10)” is too “hot (5)” to survive life on it or in its close neighbourhood. Life can exist at some distance where water can remain liquid, and such life belts are present only on “earth (6)”. Beyond these life belts, “universe (7)” is either so “frozen (2)” or “extremely (2)” “hot (5)”. The theme of “immense (5)” “temperature (5)” also explains the same situation.

There are different pieces of information about the sun and the earth (Hussain, 2009, p. 56). Cirrus tool shows broad themes of life, sun, star and earth, while human analysis is devoid of these themes.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	the littleness of our home in space	2	7
□	we find the universe frightening because	2	6
□	the surface of the sun	2	5
□	but most of them	2	4
□	far too hot for	2	4
□	for the most part	2	4
□	the total number of	2	4
□	where the temperature is	2	4
□	all the sea	2	3
□	calculation shows that	2	3
□	can exist only	2	3
□	degrees of frost	2	3
□	grain of sand	2	3
□	is a temperature	2	3
□	it would be	2	3

Figure 214 Phrases, The Dying Sun

One theoretical postulate is “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). To refine phrases, meaningless phrases have been omitted. So, the exploration with tools and understanding with human cognition continues simultaneously. These are standard phrases which constitute collocation patterns/ n-grams, for instance, “surface of the sun” (N+Prep+Art+N), “Most of them” (Adj+Prep+Prn, “too hot” (Adj+N), “the most part” (Art+Adj+N), “total number of” (Adj+N+Prep), “calculation shows” (N+V), “degrees of frost” (N+Prep+N), “grain of sand” (N+Prep+N) and “it would be” (Prn+Mod). As their count and length are concerned, they occur twice, while their length ranges from 3 to 7 words.

iv. Links

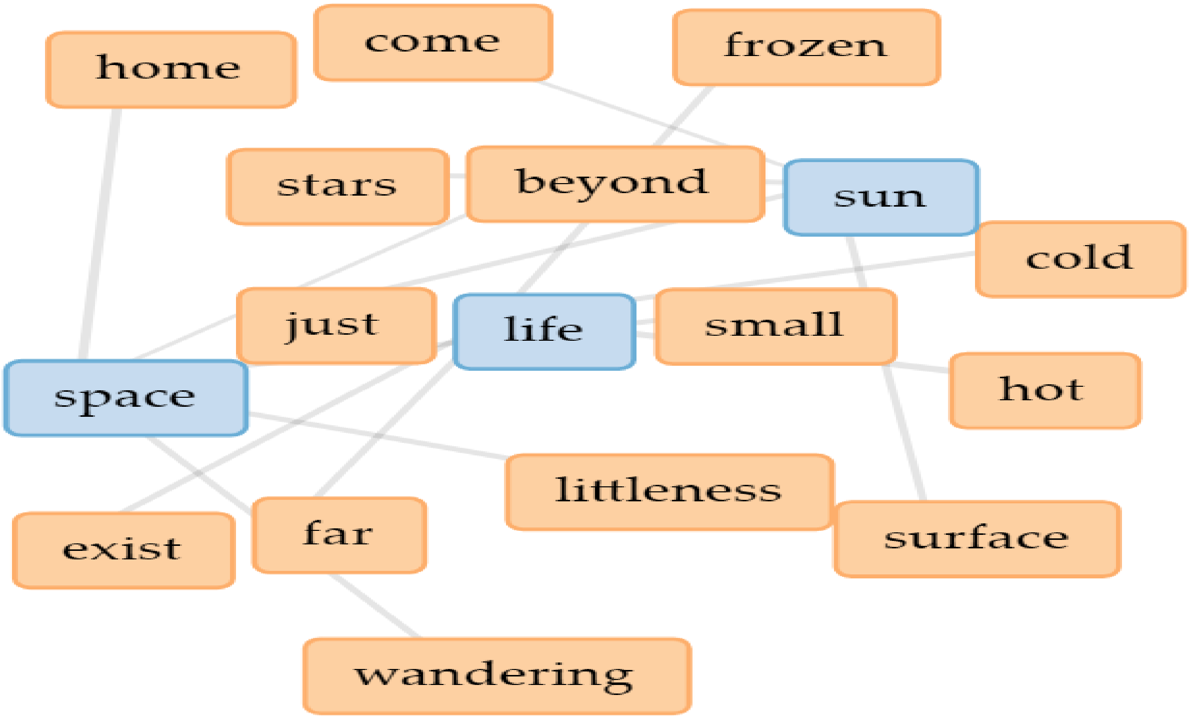
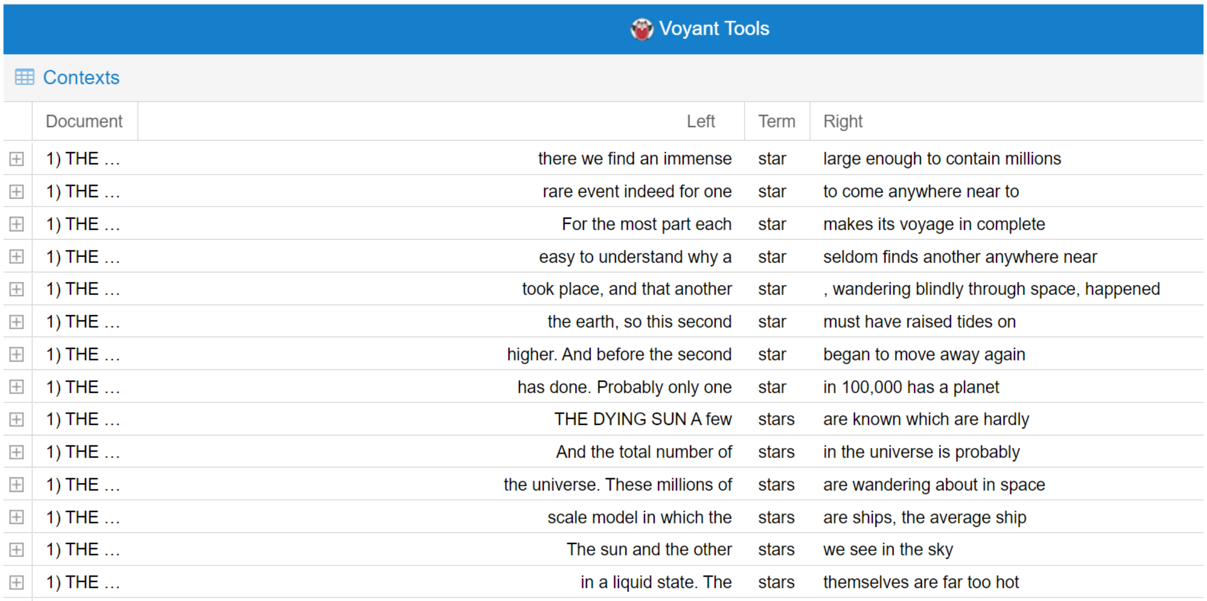


Figure 215 Links, The Dying Sun

A KG produces an interconnected web of themes and characters with its nodes. The blue coloured nodes are the most frequent words which have been connected with other nodes. KG of “littleness, space, home” exposes the knowledge pattern that our earth is so tiny as compared to the other heavenly bodies in the cosmos. Another five-pronged KG is “hot, life, frozen”, and it discovers knowledge that long distance from the sun leads to freezing temperature, while the shorter distance from the sun produces unbearable hot temperature. So, between these two temperature extremes, the moderate temperature is essential for human life and is found in some regions of the earth. One more KG of “wandering, space, stars” indicates that the advancement of a wandering star near the sun causes the separation of some mass from the sun, and they change into planets to revolve around it. Another KG of “sun, beyond, stars” expresses the fact that the sun is also a star because all stars have their own light. To conclude, hermeneutic tools are “not like black boxes” (Rockwell, & Sinclair, 2016, p. 166). It means that it does not evaluate embedded computer programmes; rather it concentrates on linguistic output and knowledge patterns in qualitative and quantitative form. These knowledge bearing tools lead to a deeper level of interpretation for knowledge discovery.

v. Contexts



Document	Left	Term	Right
1) THE …	there we find an immense	star	large enough to contain millions
1) THE …	rare event indeed for one	star	to come anywhere near to
1) THE …	For the most part each	star	makes its voyage in complete
1) THE …	easy to understand why a	star	seldom finds another anywhere near
1) THE …	took place, and that another	star	, wandering blindly through space, happened
1) THE …	the earth, so this second	star	must have raised tides on
1) THE …	higher. And before the second	star	began to move away again
1) THE …	has done. Probably only one	star	in 100,000 has a planet
1) THE …	THE DYING SUN A few	stars	are known which are hardly
1) THE …	And the total number of	stars	in the universe is probably
1) THE …	the universe. These millions of	stars	are wandering about in space
1) THE …	scale model in which the	stars	are ships, the average ship
1) THE …	The sun and the other	stars	we see in the sky
1) THE …	in a liquid state. The	stars	themselves are far too hot

Figure 216 Contexts, The Dying Sun

The word “star” has various semantic shades, for instance: i. Heavenly object ii. Famous showbiz or sports celebrity iii. High-rank officer such as a four-star general iv. Quality of hotel, for example, five-star hotel. There is a semantic ambiguity as to what type of star has been used in this corpus. To disambiguate, Contexts tool is employed to search the key word “star” as shown in figure 216. It shows that eight times “star” and six times “stars” are used. Once the word “started” is used in the 15th sentence, and it has common letters s, t, a, r, but the word “started” has a separate lexical and semantic entity. In a nutshell, the word “star” is used 14 times in the sense of the heavenly body. The significance of context is evident because Hermeneutica Theory is “embedded in a context” (Rockwell, & Sinclair, 2016, p. 166).

2. *Using the Scientific Method* by Darrel Barnard & Lon Edwards

i. Summary

This corpus has 1 document with 1,276 total words and 498 unique word forms. Created 15 seconds ago (on 1st January 2018).

Vocabulary Density: 0.390

Average Words Per Sentence: 19.3

Most frequent words in the corpus: people (15); method (11); use (11); water (10); scientific (9); today (9); foods(8); better (6); years (6); ago (5); cities (5);

Figure 217 Summary, Using the Scientific Method

Darrel Barnard & Lon Edwards use 498 unique words, and they utilize the same words almost three times in this essay until they write a total of 1276 words. This repetition generates a 0.390 density of vocabulary that refers to ease for beginner level readers. Darrel Barnard & Lon Edwards write sentences having an average length of 19.3 words. The most occurring words deal with key themes of this essay, for instance, “scientific (9)”, “methods (11)”, “foods (8)”, “water (10)” and “cities (5)”.

ii. Cirrus

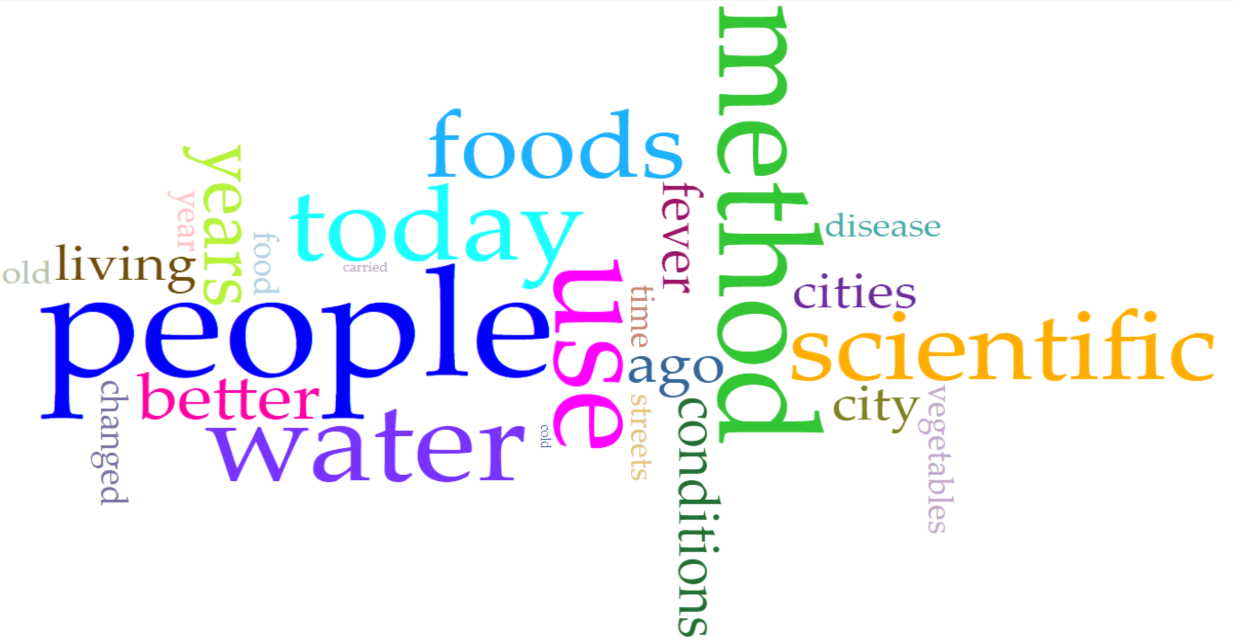


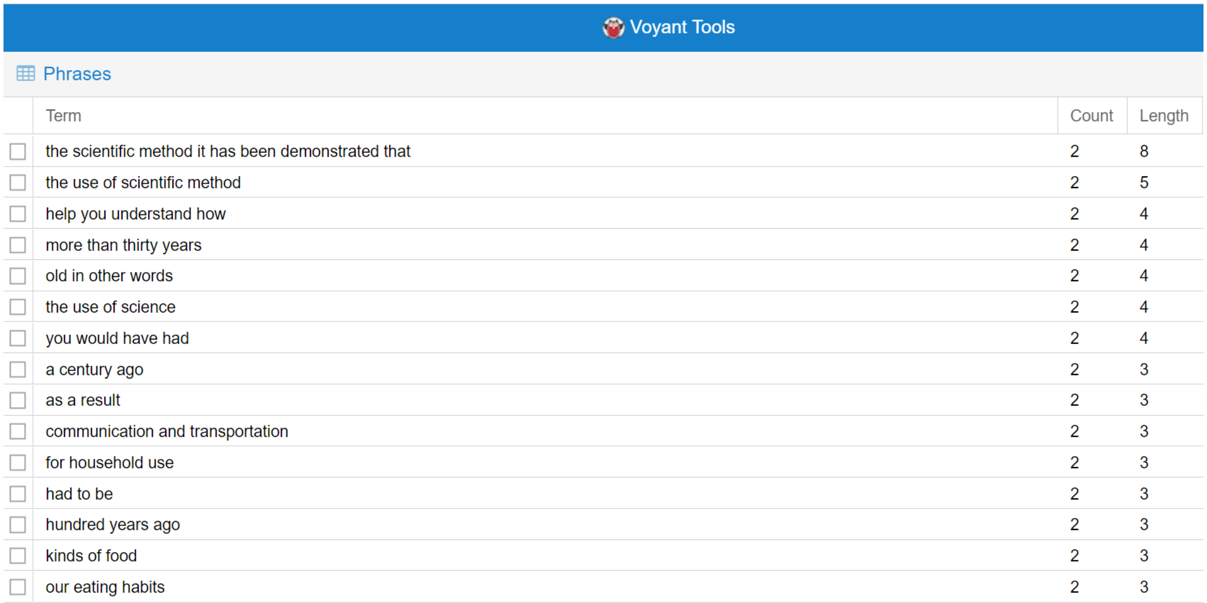
Figure 218 Cirrus, Using the Scientific Method

Cirrus does concept mining, and it can also be used as previewing and brainstorming techniques for any large text. Analysing this essay, the most occurring theme is “people (15)” because most of the scientific inventions and discoveries improve lifestyle; that is why the entire universe is brought to existence for the service of human beings. On the other hand, this essay also refers to the miserable human predicament before the advent of science, leading to a prosperous, healthy life after implementing valuable scientific discoveries and inventions. In the past, they had to bring “water (10)” from far-flung areas, so they had to use it sparingly, but now abundant water is used for cleaning and industrial purposes.

Before the advent of science, many people died because of hunger and famine, but now “food (8)” is being produced on a large scale so that scarcity of food is not a problem throughout the world. Means of transportation have also become so easy that in the case of famine, things can be transported to any country speedily. Science produces abundant and high-quality food, a luxurious lifestyle and effective medicines for human beings. The same is the case with the “diseases (4)” which have ruined cities, but now magical medicines relieve human beings from suffering. Now the average human life has also been exceeded because of science. Furthermore, in the olden days, our “streets (4)” were full of garbage and animals were feeding on it. Due to scientific methods, our drains are cemented and clean. Several antibacterial medicines are being sprayed to protect human beings from toxic elements.

One female and four male characters are present in this essay. Thrifty housewives preserved their home-grown vegetables and fruit. Main themes are home keeping and women's non-productive and reproductive activities (Hussain, 2009, pp. 56-57). Cirrus tool covers all characters with the word “people”, and it does not mention them separately. Main themes of Cirrus are water, food, life and diseases, whereas human analysis does not mention these themes.

iii. Phrases



	Term	Count	Length
□	the scientific method it has been demonstrated that	2	8
□	the use of scientific method	2	5
□	help you understand how	2	4
□	more than thirty years	2	4
□	old in other words	2	4
□	the use of science	2	4
□	you would have had	2	4
□	a century ago	2	3
□	as a result	2	3
□	communication and transportation	2	3
□	for household use	2	3
□	had to be	2	3
□	hundred years ago	2	3
□	kinds of food	2	3
□	our eating habits	2	3

Figure 219 Phrases, Using the Scientific Method

Standard collocations/ n-grams of this essay are “in other words” (Prep+Adj+N), “use of science” (N+Prep+N), “as a result” (Adv+Art+N), “communication and transportation” (N+Conj+N), “household use” (Adj+N), “kinds of food” (N+Prep+N) and “eating habits” (Adj+N). They occur twice, and their length varies from 3 to 8 words. One common knowledge pattern is that the above-mentioned collocations/ n-grams have one or more nouns in them.

iv. Links

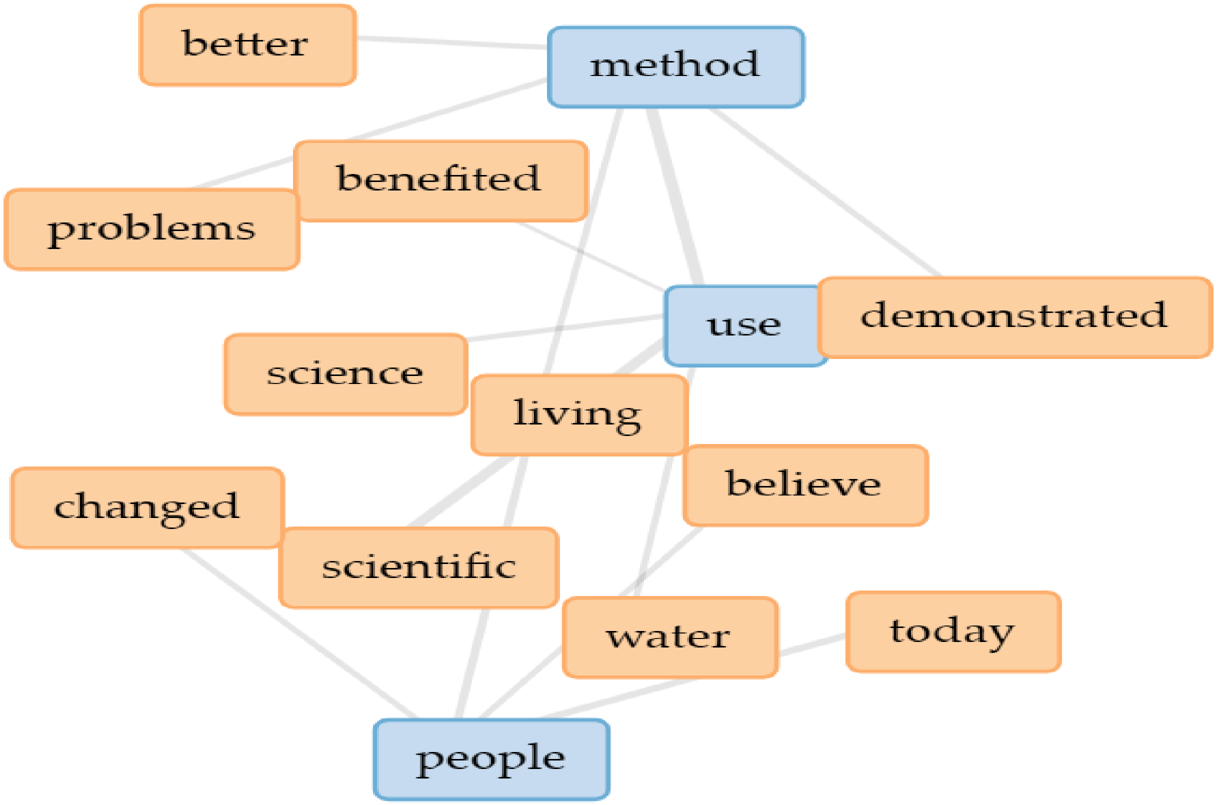
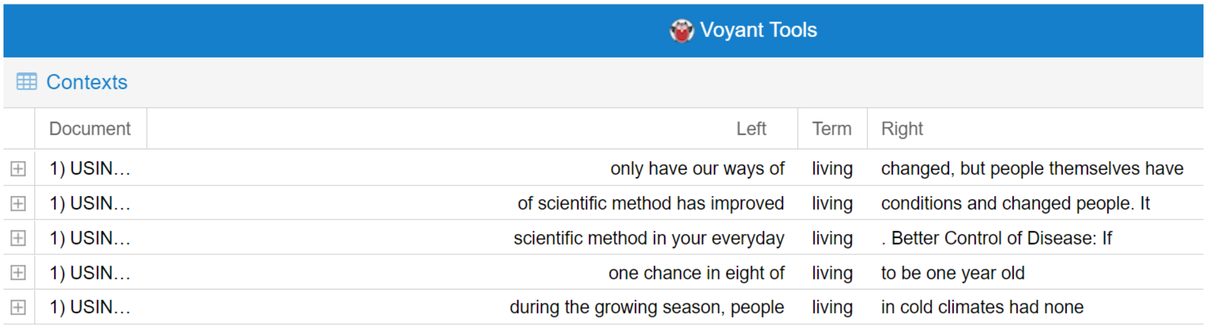


Figure 220 Links, Using the Scientific Method

The KG of “scientific, use, method, better” suggests that scientific methods have utterly improved human ways of life, and they get abundant water for drinking and cleaning purposes. Moreover, they get more and better food to satiate their daily nourishment. One KG of “changed, people, water, use” refers to the extensive use of water for bathing and cleaning purposes, while in the past, it was considered as wastage of water, or they could not use water for cleaning their baths. The KG of “living, people, changed” indicates that science changes people’s lifestyle, appearance and thinking patterns. As a result, scientific methods and critical thinking have become a hallmark of modern people’s attitudes. In addition to it, people abandon superstitions due to scientific methods. One more KG of “science, use, benefitted” reveals the knowledge pattern that science brings benefits by changing their living style from mud houses to cemented, durable and luxurious homes. Hermeneutic knowledge discovery is pervasive, and the extension of a knowledge graph finds and interprets new knowledge patterns. The same has been mentioned in one postulate of the theory thus: “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166).

v. Contexts



Voyant Tools

Contexts

	Document	Left	Term	Right
⊞	1) USIN…	only have our ways of	living	changed, but people themselves have
⊞	1) USIN…	of scientific method has improved	living	conditions and changed people. It
⊞	1) USIN…	scientific method in your everyday	living	. Better Control of Disease: If
⊞	1) USIN…	one chance in eight of	living	to be one year old
⊞	1) USIN…	during the growing season, people	living	in cold climates had none

Figure 221 Contexts, Using the Scientific Method

There is semantic and grammatical ambiguity whether the word “living” is a noun, adjective, present participle or gerund. To differentiate among them and disambiguate them, Contexts tool is employed to find the context of the key word “living” used five times. Three times the word “living” has been used as a noun in the phrases “ways of living”, “everyday living”, “one chance in eight of living”. Once the word “living” is used as an adjective in “living conditions”. Once the word “living” is used as a gerund in the phrase “people living in cold climate”. To conclude, “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166), so, context of the problematized word and human reflection facilitate in word sense disambiguation. After clarity of textual meaning, deeper hermeneutic patterns can be explored.

3. *Why Boys Fail in College* by Herbet E. Hawkes

i. Summary

This corpus has 1 document with 1,667 total words and 561 unique word forms. Created 3 seconds ago (on 3rd January 2018).

Vocabulary Density: 0.337

Average Words Per Sentence: 24.9

Most frequent words in the corpus: boy (23); college (21); boys (11); work (10); parents (8); reason (7); ability (6); failure (6); good (6); ought (5);

Figure 222 Summary, Why Boys Fail in College

Herbert E. Hawkes writes 561 unique words in his essay, and he utilises this basic vocabulary three times, which becomes a total of 1667 words. Due to three times of repetition, its vocabulary density is 0.337 words. Sentences are constructed longer than the previous essays; that is why, on average, 24.9 words are written in one sentence, and they are suitable for highly advanced level learners. Most occurring words of this essay guide about causes of failure and their possible solutions.

Hermeneutica Theory directs that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166), and here Voyant tool expresses knowledge by counting all stylometric features. Roberto Busa spent more than two decades quantifying the text, and now such works can be done within a few minutes; hence, these tools are replete with statistical and linguistic knowledge. Afterwards, it is human reflection’s role to extract innovative insights about stylometry, the total vocabulary of any writer, repeated themes and vocabulary density. This reflection leads to actual author identification and forensic linguistics.

ii. Cirrus

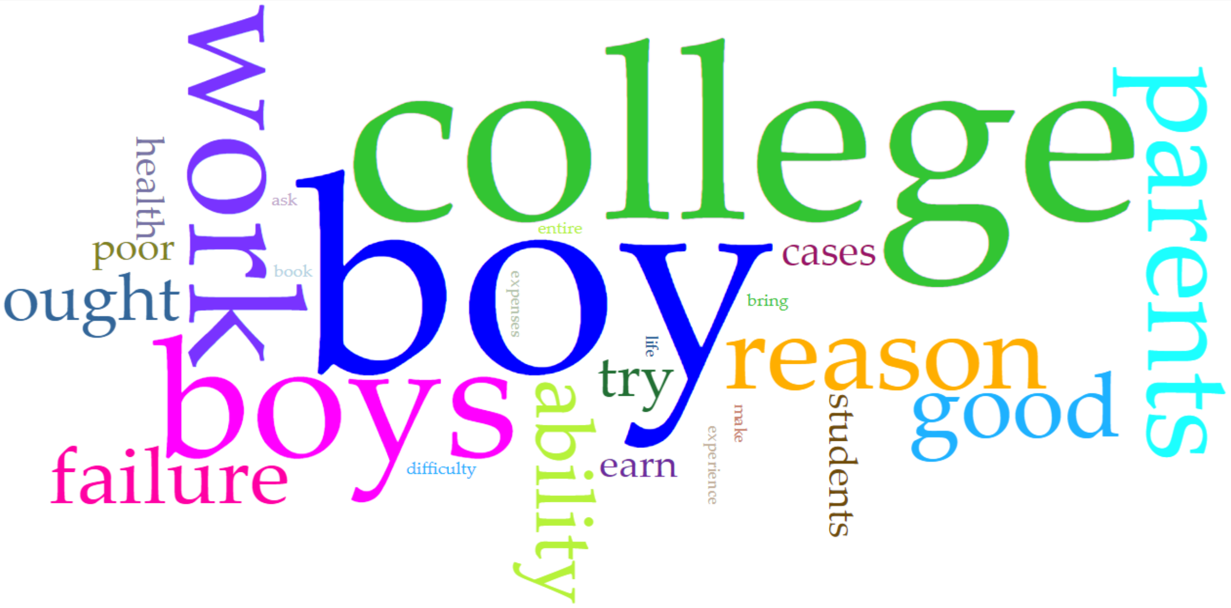
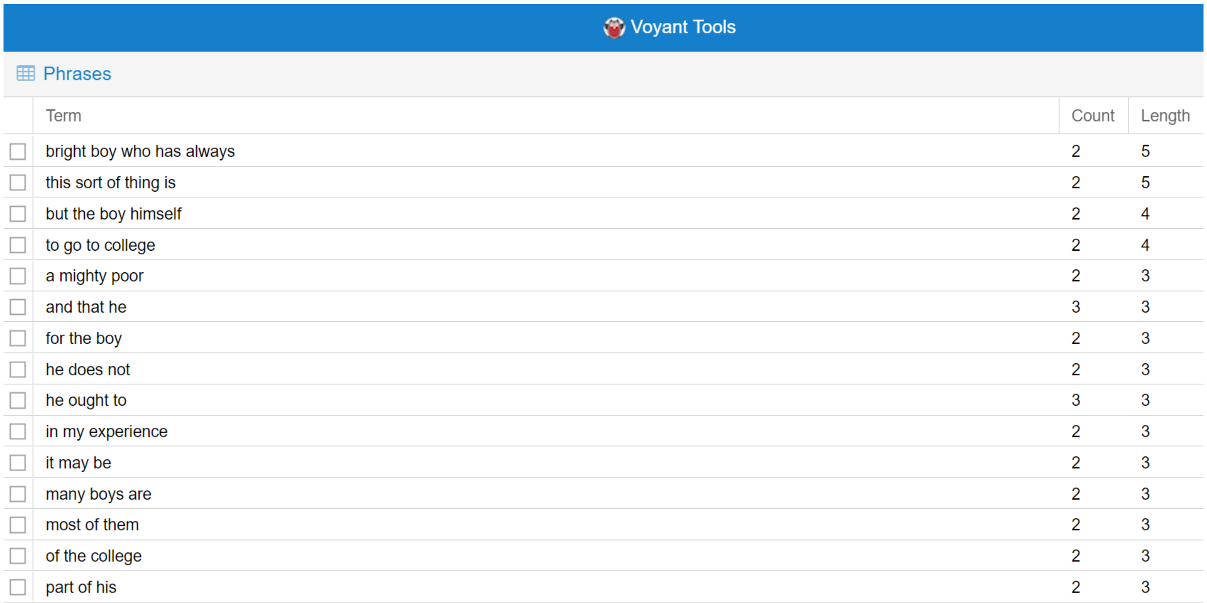


Figure 223 Cirrus, Why Boys Fail in College

The central characters of this essay are “boys (11)”/ “students (6)” who have failed during their “college (21)” life. Several reasons for their “failure (6)” have been clustered through topic modelling; for instance, in some cases, boys have to “earn (4)” their food and other academic requirements. Therefore, they have to work from dawn to dusk; consequently, they waste their precious time in earning efforts and fail in their college studies. Some “students (4)” have poor health; therefore, the college doctor should address their health issues to save them from their failure. Some students have the ability, but they do not work due to mistaken ambition determined by their “parents (8)”. Another theme of “ought to (5) suggests that the writer not only points out reasons for the failure but also gives the solution to their problems so that failures can be diminished from students' academic careers.

Human analysis reveals that no female character has been mentioned except college boys (Hussain, 2009, p. 56). Cirrus tool also finds a detailed analysis of characters, such as boys, students, and parents. Cirrus shows key themes, for instance, failure, college, earn and ought to. Consequently, Cirrus analyses text in a more detailed manner as compared to the manual analysis.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	bright boy who has always	2	5
□	this sort of thing is	2	5
□	but the boy himself	2	4
□	to go to college	2	4
□	a mighty poor	2	3
□	and that he	3	3
□	for the boy	2	3
□	he does not	2	3
□	he ought to	3	3
□	in my experience	2	3
□	it may be	2	3
□	many boys are	2	3
□	most of them	2	3
□	of the college	2	3
□	part of his	2	3

Figure 224 Phrases, Why Boys Fail in College

Phrases tool provides collocation patterns/ n-grams and provides insight for learning repeated linguistic aspects. As standard phraseology is concerned, they are: "bright boy" (Adj+N), " this sort of thing" (Adj+N+Prep+N), " go to college" (V+Prep+N)," a mighty poor" (Art+Adj+N),

" ought to " (Aux), " in my experience" (Prep+Prn+N), " it may be" (Prn+ Mod) and " many boys"(Adj+N). They occur twice, and their length ranges from 3 to 6 words.

In addition to it, as knowledge patterns, these phrases teach repeated language patterns; for instance, “who” as a relative pronoun is used only with human beings. Another grammatical point is that reflexive pronouns always match their subjects. After infinitive to, the first form of a verb is used. The auxiliary verb “does” is used with the subject "he". Being a source of language learning, collocation/ n-gram “many boys” refer to the matching of plural adjective and plural noun.

iv. Links

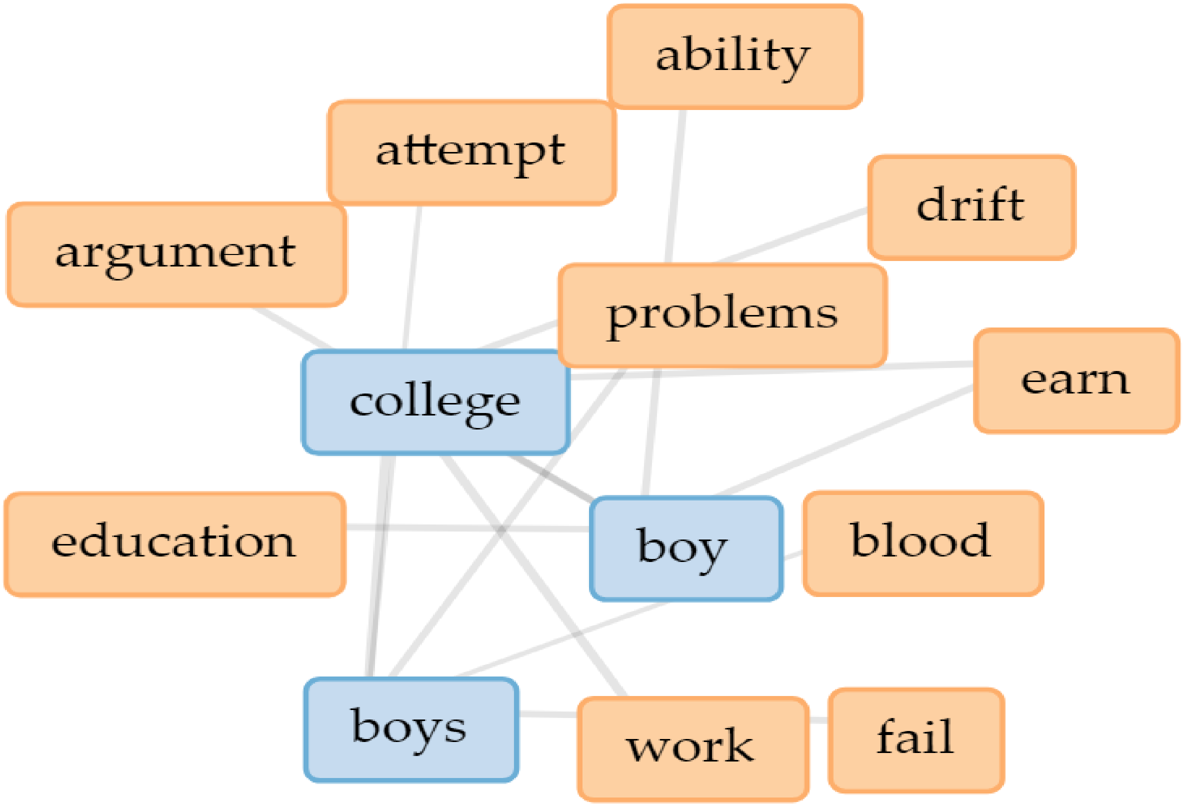
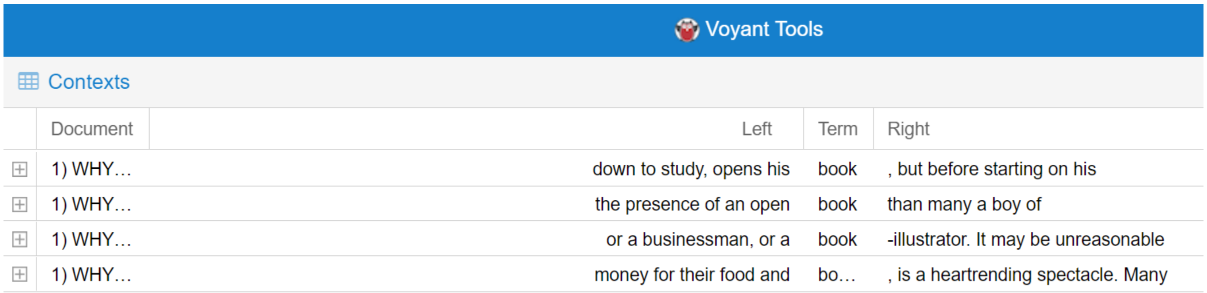


Figure 225 Links, Why Boys Fail in College

The blue coloured words are the most occurring themes, while orange-coloured words are less occurring themes in the corpus. The KG of "college, boys, work, fail" reveals the knowledge pattern about the reasons for boys’ failure in the college. Now some problems have been enumerated, for example, KG of "earn, boy, blood" suggests that boys work in factories for long hours or someone sold blood. They cannot find time to study, so economic pressure becomes the main reason for their failures in college. Another reason for failure is that some boys can pass the examination, but they fail because their parents force them to study some distasteful subjects; consequently, this information is embedded in the KG of "fail, boy, ability".

v. Contexts



	Document	Left	Term	Right
⊞	1) WHY…	down to study, opens his	book	, but before starting on his
⊞	1) WHY…	the presence of an open	book	than many a boy of
⊞	1) WHY…	or a businessman, or a	book	-illustrator. It may be unreasonable
⊞	1) WHY…	money for their food and	bo…	, is a heartrending spectacle. Many

Figure 226 Contexts, Why Boys Fail in College

The word "book" as a noun means readable book, while it means the reservation of a seat or place as a verb. To disambiguate between dual word senses, Contexts tool shows that word "book" has been used four times as a noun and every time it refers to reading material. Significance of context can be understood that "Hermeneutica Theory is embedded in a context" (Rockwell, & Sinclair, 2016, p. 166) because context reveals true sense which is required for hermeneutic purposes. Initial clarity will lead to dig deeper and multi-layered hermeneutic patterns.

4. *End of Term* by David Daiches

i. Summary

This corpus has 1 document with 897 total words and 416 unique word forms. Created 19 seconds ago (on 5th January 2018).

Vocabulary Density: 0.464

Average Words Per Sentence: 35.9

Most frequent words in the corpus: school (13); week (7); end (6); summer (6); days (5); friday (5); holidays (5); monday (5) ; morning (5); time (5);

Figure 227 Summary, End of Term

The summary tool presents the stylometry of David Daiches that he writes 416 unique words which have been repeated twice; hence, this essay comprises 897. By dividing unique words with total words, vocabulary density 0.464 is derived; hence, it shows a higher difficulty level for the readers. Moreover, David Daiches writes very long and highly advanced level sentences, so, 35.9 words are found in one sentence on average. The most frequent words are concerning with “school (13)”, “week (7)”, “end (6)”, “summer (6)” and “holidays (5)” which indicate that this essay is dealing with memories of childhood, school and its holidays.

ii. Cirrus

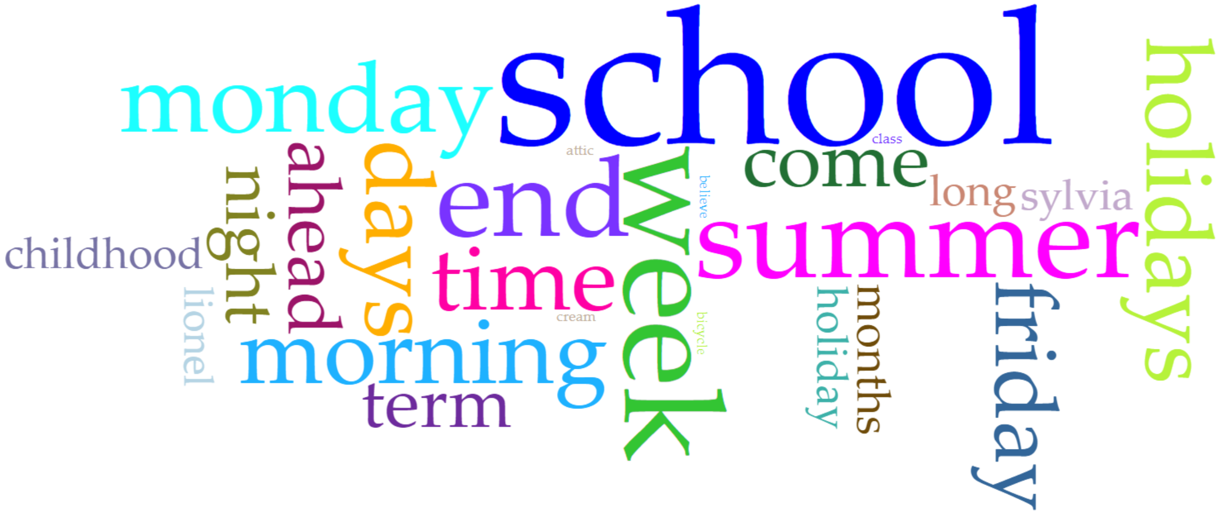
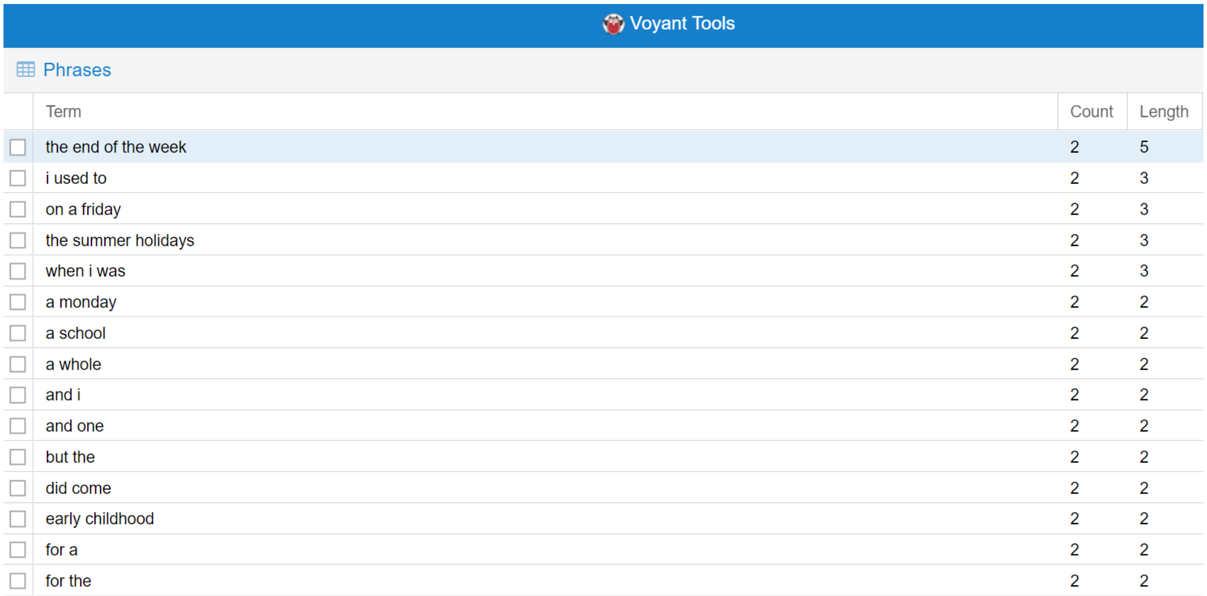


Figure 228 Cirrus, End of Term

This Cirrus informs about two significant characters, "Lionel (3)" and "Sylvia (3)". As significant themes are concerned, "school (13)" is the central place that gives students jubilation and anxiety. Usually, getting up early in the “morning (5)” causes uneasiness and frustration for Daiches, while holidays and summer vacations give him immense exultation. Furthermore, the condition of Daiches and students have been portrayed on "Friday (5)", Saturday and Sunday, that on Friday students anxiously wait for the full weekend, and on Sunday “night (4)”, they are fearful about blue "Monday (5)" and its workload. Habitually students derive pleasures on the advent of "holidays (3)". Other holidays, for instance, Christmas holidays and unexpected respites give them more happiness, but real long-awaited holidays are of "summer (6)" "holidays (5)" which are very “long (3)”, enjoyable and a source of excitement for the students. Long holidays become a source of pleasures, while children visit their relatives in far-flung areas. To conclude, David Daiches undergoes all the aforementioned joys and sorrows of childhood. This Cirrus successfully derives key themes through topic modelling as Scrivener and Davis (2017) have extracted key themes from the text.

Human analysis shows that the main characters are maid servant, and the main themes are waking up in the morning, laughing stock, deep funeral tones and ominous tread (Hussain, 2009, pp. 57-58). On the other hand, Cirrus extracts the names of the main characters Lionel and Sylvia. Moreover, Cirrus extracts entirely different main motifs: school, Friday, Monday, holidays, night and morning. Only one theme, “morning” , is common between human and machine analysis.

iii. Phrases



	Term	Count	Length
□	the end of the week	2	5
□	i used to	2	3
□	on a friday	2	3
□	the summer holidays	2	3
□	when i was	2	3
□	a monday	2	2
□	a school	2	2
□	a whole	2	2
□	and i	2	2
□	and one	2	2
□	but the	2	2
□	did come	2	2
□	early childhood	2	2
□	for a	2	2
□	for the	2	2

Figure 229 Phrases, End of Term

Collocation patterns/ n-grams “end of the week” (N+Prep+Art+N), “used to” (Mod), “on a Friday” (Prep+Art+N), “summer holidays” (Adj+N), “a whole” (Art+Adj) and ‘early childhood” (Adj+N) have been extracted from this essay. They occur twice, and their length ranges from 2 to 5 words. Besides, these phrases teach fluency and correct grammatical patterns; for example, collocation/ n-gram “on Friday” teaches preposition of time “on” with days of the week. Another collocation pattern/ n-gram of “did come” (Aux+V) teaches that the first form of the verb is used after the auxiliary verb “did”. One theoretical postulate is “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). To conclude, it is human understanding which separates meaningful and meaningless collocation. Consequently, meaningful collocations are presented as knowledge patterns.

iv. Links

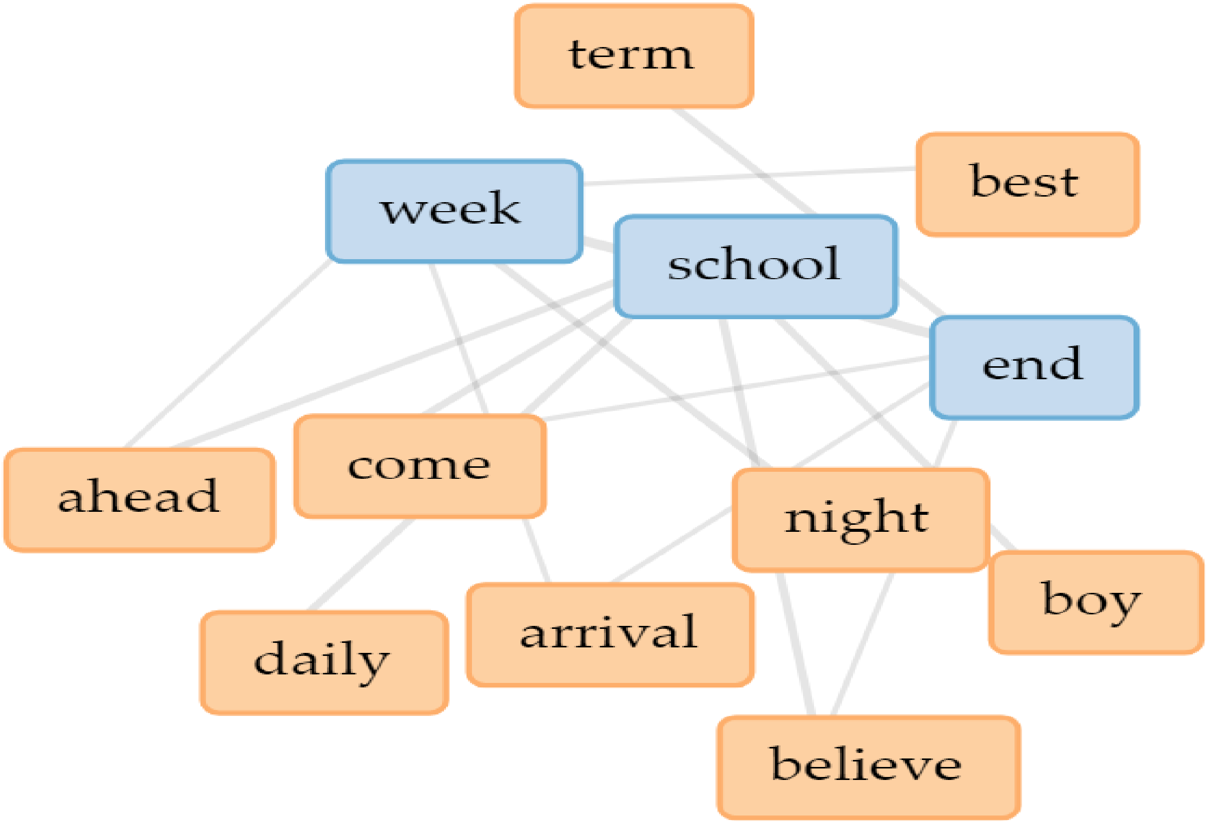
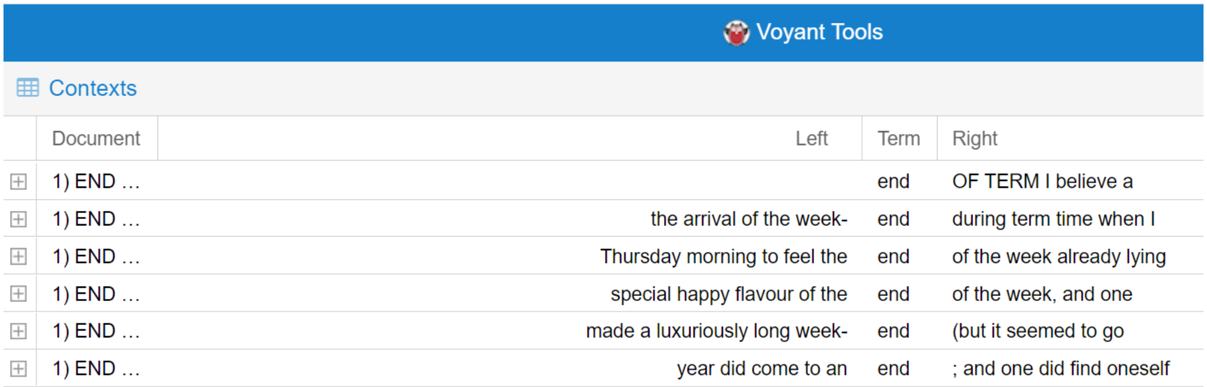


Figure 230 Links, End of Term

“In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct” (Agrawal, & Psaila, 1995, p. 1). Here data visualization and its interpretation in the light of the research question are significant. One triangular KG of “arrival, week, end” shows different mental and psychological situations of students pertaining to the arrival of the weekend; for instance, Friday night is promising because of two intact holidays, whereas Sunday night is full of fears due to loss of jocund weekend time. One KG of “boy, school, week” shows that several situations of the essay reveal the mental conditions of a school boy. Another triangular KG of “boy, school, believe” expresses that this essay is about the beliefs and thoughts of school-going boys. Another KG of “term, school, end, boy” shows that they get a luxurious long weekend with Monday off during term time.

v. Contexts



	Document	Left	Term	Right
⊞	1) END ...		end	OF TERM I believe a
⊞	1) END ...	the arrival of the week-	end	during term time when I
⊞	1) END ...	Thursday morning to feel the	end	of the week already lying
⊞	1) END ...	special happy flavour of the	end	of the week, and one
⊞	1) END ...	made a luxuriously long week-	end	(but it seemed to go
⊞	1) END ...	year did come to an	end	; and one did find oneself

Figure 231 Contexts, End of Term

The word "end" has meanings of last/ending point, some key objective or purpose. To disambiguate word sense, it is necessary to understand the context of a key word shown in figure 231. In this corpus, the word "end" is used six times as a noun, and every time it gives a semantic shade of finishing point.

5. *On Destroying Books* by J. C. Squaire

i. Summary

This corpus has 1 document with 1,207 total words and 536 unique word forms. Created 50 seconds ago (on 5th January 2018).

Vocabulary Density: 0.444

Average Words Per Sentence: 19.5

Most frequent words in the corpus: books (12); people (5); thought (5); sack (4); bridge (3); came (3); cold (3); embankment (3); merely (3); near (3);

Figure 232 Summary, On Destroying Books

“Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). This compact description quantifies all text. Stylometry also reveals the author’s average length of sentences, total and unique words and the most prominent themes. In this essay, J. C. Squaire writes 536 unique words, and they are repeated more than twice until the total word count reaches 1207. By dividing unique words with total words, 0.444 vocabulary density is derived, and it is appropriate for basic level readers. J. C.

Squaire’s sentences comprise an average of 19.5 words which are also suitable for advanced level readers. The most frequent word in this essay is "books (12)", which refers to destroying books by applying different methods, and lastly, throwing books into the river method is applied. Consequently, this method works well to get rid of unwanted books though the writer repents his act later.

ii. Cirrus

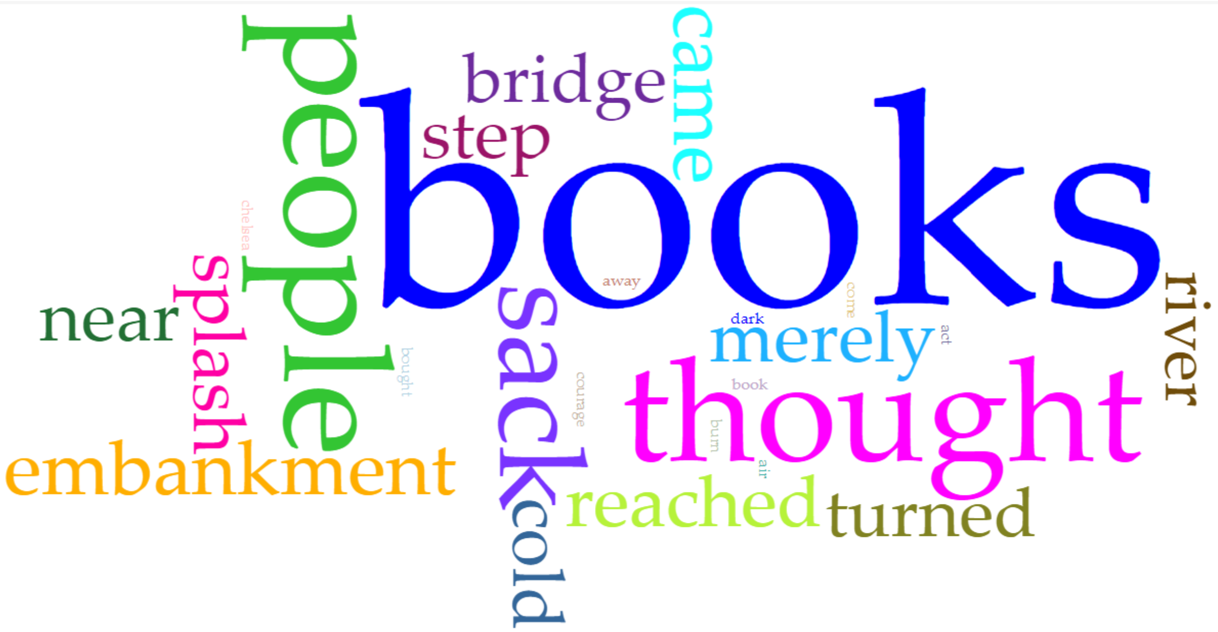
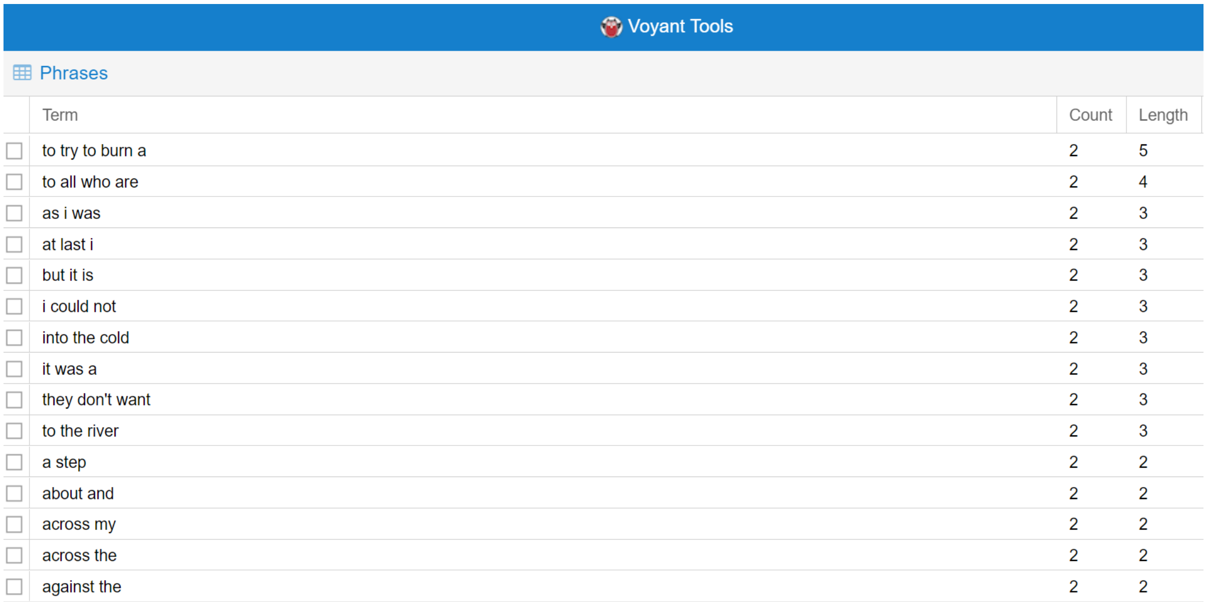


Figure 233 Cirrus, On Destroying Books

In this Cirrus, the most prominent theme is "books (12)" because the writer "thought (5)" about different methods of destroying books, and he has reached this point whether books or he can stay in the small flat. The topic modelling process in Cirrus reveals that “books” are the central theme; hence, the title of the essay supports the same stance. The theme of "came (3)" and "reach" are also found that the writer packs books in a "sack (4)", reaches the “embankment (3)” of the “river (3)” at night, and throws his books into the river with a big “splash (3)”. The theme of "thought (5)" indicates that the writer is thinking more and acting less, so he is not the man of action rather he is a man of words. Knowledge Discovery Theory is defined as “the extraction of implicit, previously unknown and potentially useful information from data” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). He faces a “to be or not to be” situation while nurturing several apprehensions and fears about the sack, night time and people. The findings of the current study are the same as Muhammad (2012) extracts Cirrus from Google Books Corpus N-grams.

Human analysis shows that there is no comparison of male and female characters, rather it focuses on destroying books (Hussain, 2009, p. 56). Cirrus unveils “people” to cover all direct and indirect characters, while human analysis mentions males and females. Cirrus tool presents more themes, while human analysis mentions a theme of the destruction of books.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	to try to burn a	2	5
□	to all who are	2	4
□	as i was	2	3
□	at last i	2	3
□	but it is	2	3
□	i could not	2	3
□	into the cold	2	3
□	it was a	2	3
□	they don't want	2	3
□	to the river	2	3
□	a step	2	2
□	about and	2	2
□	across my	2	2
□	across the	2	2
□	against the	2	2

Figure 234 Phrases, On Destroying Books

The standard phrases are “at last” (Prep+N) and “into the cold” (Prep+Art+N). These multi-word expressions are generated through voyanting (the use of Voyant tools to mine text, study, teach, and conduct research hermeneutically). Some other grammatical knowledge patterns are derived from these phrases; for instance, "who" is a relative pronoun, and it can be used after "all" (human beings). The phrase "to try to burn a" (Inf V+Inf V+Art) guides that if two verbs are written in a sentence, add "to" before the second verb and after "to", the first form of the verb must be used. They occur twice, and their length ranges from 2 to 5 words. Some collocations/ n-grams, for instance, “across the" are excluded because they do not show any meaningful sense. The same has been directed in KDD, “In active data mining paradigm,… rules are discovered” (Agrawal, & Psaila, 1995, p. 1).

iv. Links

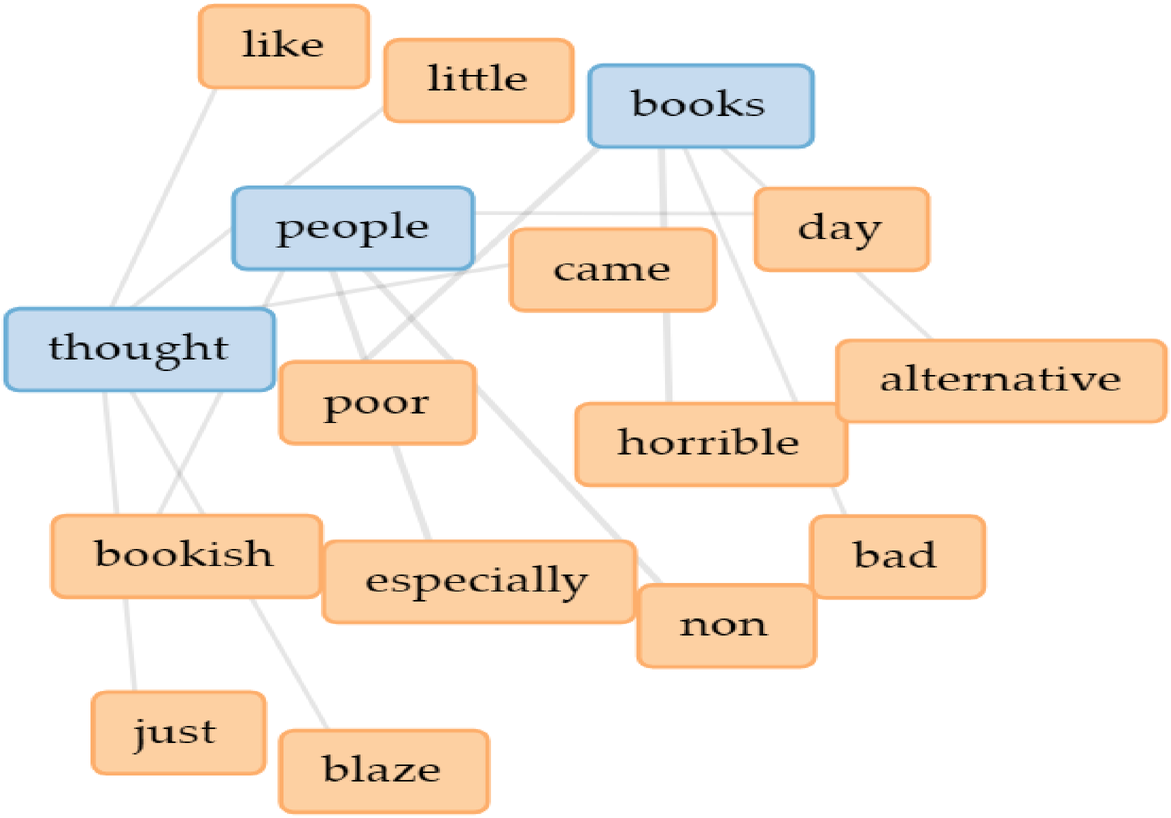
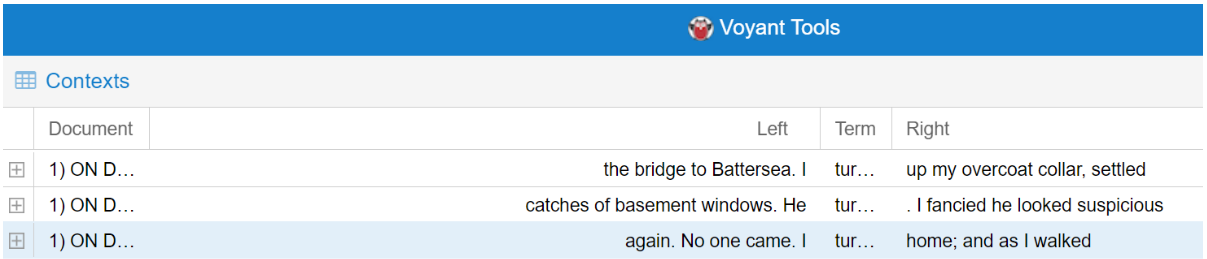


Figure 235 Links, On Destroying Books

Filtering and zooming processes have been done to develop comprehension among different themes. Sometimes, textual evidence is taken for further explanation of a KG, and in some cases, KGs are extended to search connectivity of more themes. This addition is based on Hermeneutica Theory which is “supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166). One triangular KG of "poor, books" reveals that the writer feels pity for the poor books that are lying in the mud of the river. The writer's other feeling about books is that they are terrible, and this has been exhibited in the KG of "thought, bad, books". One more triangular KG joins "thought, books, horrible" to reveal knowledge that the writer considers these books horrible. One more thought has been exposed with KG of "thought, bookish, blaze" because the writer also plans to burn books page by page, though it is an arduous task. Apart from the writer's vision, views of non-bookish people have also been shown that they are not willing to throw any paperbound material whether they read it or not; and it is presented in the KG of "non, bookish, people".

v. Contexts



Document	Left	Term	Right
1) ON D…	the bridge to Battersea. I	tur…	up my overcoat collar, settled
1) ON D…	catches of basement windows. He	tur…	. I fancied he looked suspicious
1) ON D…	again. No one came. I	tur…	home; and as I walked

Figure 236 Contexts, On Destroying Books

The word “turned” as a verb means changing direction, and as a linking verb, it means a change of condition. Both words have different grammatical patterns and semantic shades. The theory has suggested the solution to the polysemic problem. “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). Later, theorization led to the design of text mining tools for showing bidirectional text. Contexts tool disambiguates word sense and shows that the word “turned” is used as a main verb with the meaning of changing direction three times.

6. *The Man Who Was a Hospital* by Jerome K. Jerome

i. Summary

This corpus has 1 document with 1,158 total words and 416 unique word forms. Created about 2 minutes ago (on 13th January 2018).

Vocabulary Density: 0.359

Average Words Per Sentence: 16.8

Most frequent words in the corpus: i (95); got (9); read (8); said (6); housemaid's (5); knee (5); man (5); feel (4); fever(4); li ver (4);

Figure 237 Summary, The Man Who Was a Hospital

Jerome K Jerome writes a unique vocabulary of 416 words, and they are repeated almost three times in this essay until it reaches 1158 total words. By dividing unique words with total words, 0.359 vocabulary density is derived. The length of the sentences is comparatively lesser than the previous essay, and Jerome K Jerome writes almost 16.8 words in a sentence. The most frequent word is “I (95)” which discovers its first-person narrative technique. The most frequent words suggest different names of diseases, for instance, “fever (4)” and “housemaid's knee (5)”; and parts of the body like “liver (4)” tongue and head.

ii. Cirrus

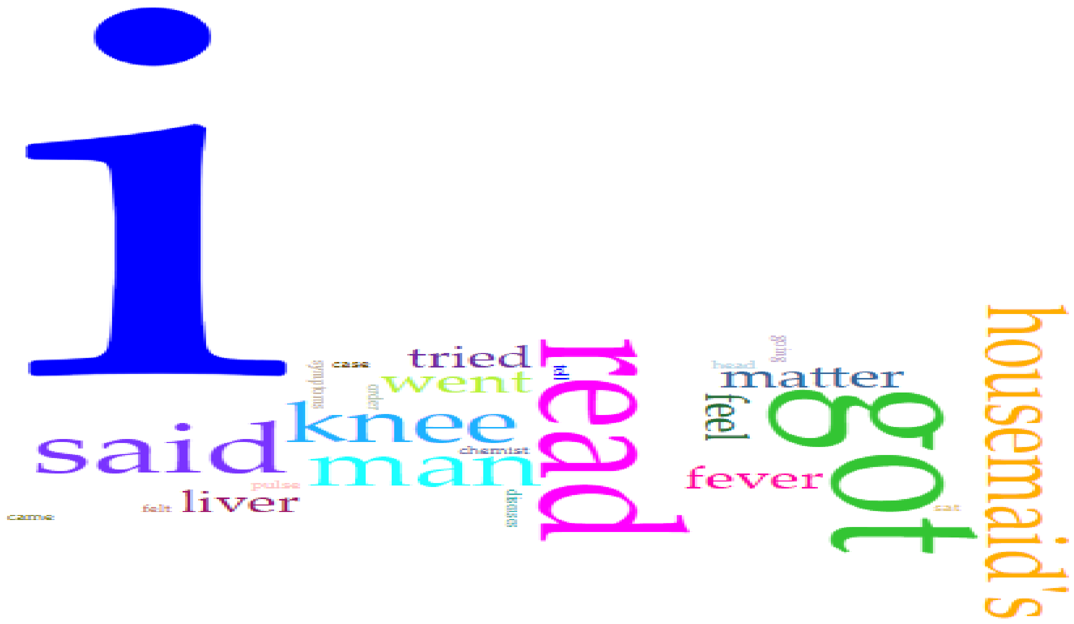


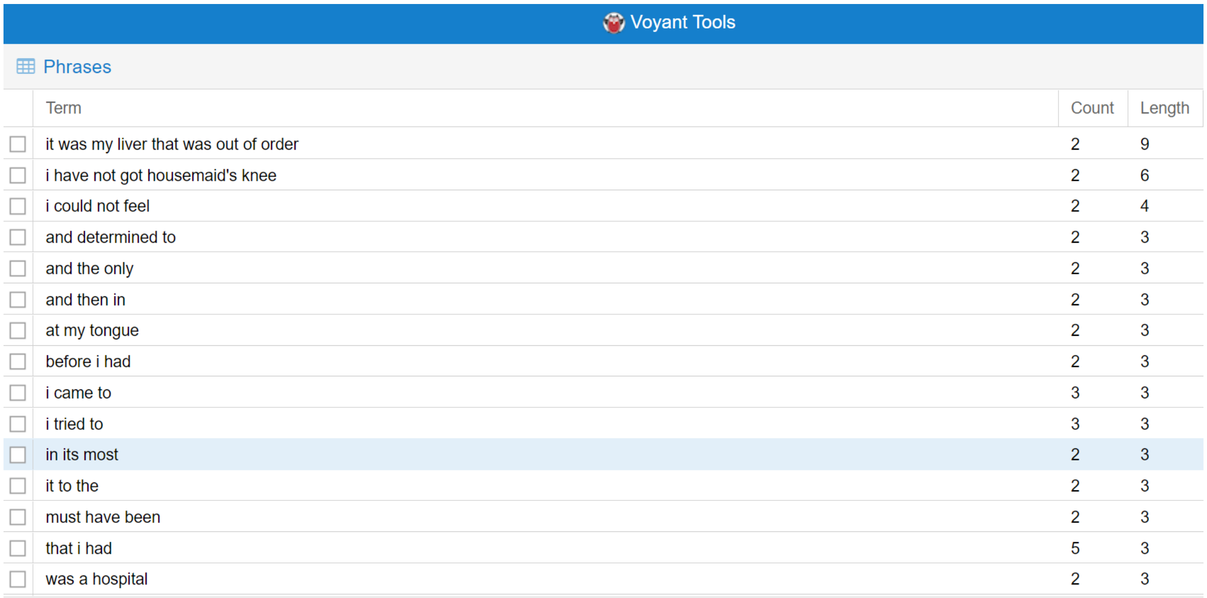
Figure 238 Cirrus, The Man Who Was a Hospital

The topic modelling process facilitates Cirrus generation from any corpus, and the most occurring theme is “I (95)”, which shows the use of the first-person narrative technique; consequently, the entire essay revolves around a self-narrating event. Another prominent theme is "got (9)", and it is associated with names of “diseases (3)” for example “, housemaid's knee (5)”. In fact, “got (9)” means the happening of some negative thing, and it is linked with all diseases. Likewise, the word “out of order (3)” also means the same inefficiency of the body. The word "got" indicates the occurrence of unpleasant things in this essay titled *‘The Man who was a Hospital’*. There is a theme of "tried (4)", and it shows that he strives hard to examine himself whether he is a patient of all those diseases or not. He “feels (4)” that he has all diseases except the housemaid's knee. Moreover, the theme of diseases pertaining to liver, head and fever have also been mentioned to prove him a complete living hospital. The theme of "matter (4)" also refers to different ailments. After self-diagnosis of several diseases, the patient goes to a doctor who writes a funny and judicious prescription. Then he goes to a “chemist (3)” to buy the prescribed medicines, but he is unable to provide those entities.

Another revelation of knowledge is evident with the word "read (8)", that after reading different symptoms of diseases in the pharmacopoeia, he feels that he is the victim of all diseases in the world. Another theme is "said (6)" which shows its dialogic quality that characters directly converse with each other. In this essay, both doctor and chemist treat a whimsical patient as a comic butt.

The human analysis concentrates on only one masculine character and his funny acts (Hussain, 2009, p. 56). Cirrus tool outperforms the main character “I” and “chemist (3)”. Comparing with human analysis, Cirrus tool shows the themes of feel, diseases, housemaid’s knee, tried, and got along with their statistical weight.

iii. Phrases



	Term	Count	Length
□	it was my liver that was out of order	2	9
□	i have not got housemaid's knee	2	6
□	i could not feel	2	4
□	and determined to	2	3
□	and the only	2	3
□	and then in	2	3
□	at my tongue	2	3
□	before i had	2	3
□	i came to	3	3
□	i tried to	3	3
□	in its most	2	3
□	it to the	2	3
□	must have been	2	3
□	that i had	5	3
□	was a hospital	2	3

Figure 239 Phrases, The Man Who Was a Hospital

Phrases tool extracts all collocations from any piece of text. Then human cognition plays its role to separate meaningful collocations from meaningless collocations. One theoretical postulate has suggested the same, “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). Phrases tool extracts bigrams and trigrams, for instance, "out of order" (Id), "got housemaid's knee" (V+N+Apo+N), "and then" (Conj+Adv), "at my tongue" (Prep+Prn+N), "before I had" (Conj+Prn+Aux), "I had to" (Prn+Mod) and "in its most" (Prep+Prn+Adj). They unveil an out of order body system or diseases about the knee and tongue. They occur 2 to 3 times, whereas their length ranges from 2 to 9 words.

iv. Links

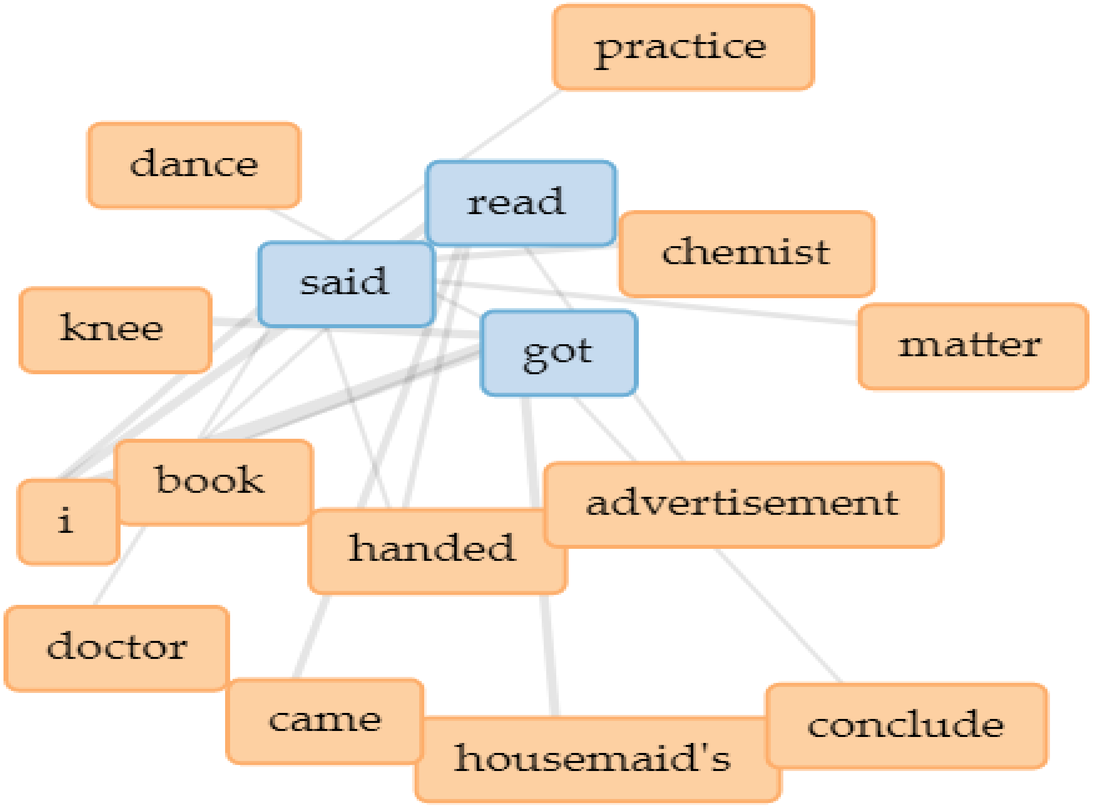
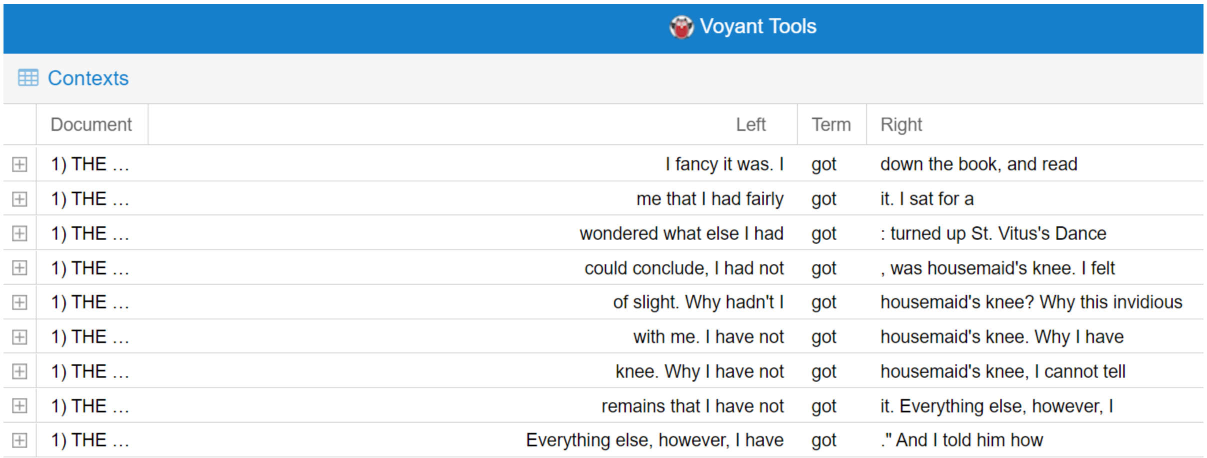


Figure 240 Links, The Man Who Was a Hospital

One KG reveals very interesting knowledge patterns, for example, "I, said, chemist". Another KG shows that “I” (patient) talks to the doctor and chemist. One more KG of “chemist, said, handed” refers to an act of returning prescription to the whimsical patient. Another KG of "conclude, advertisement, read, book" exposes that he reads an advertisement and voluminous book about different diseases; eventually, he concludes that he has become a victim of many diseases. So, reading of pharmacopoeia leads him to be a pseudo victim of all discovered diseases. Another KG of "dance, got, housemaid's knee" unveils a knowledge pattern that the word "got" has been associated with diseases. In figure 239, filtering and zooming processes have been done to access meaningful patterns. They confirm the known textual knowledge and disclose the unknown knowledge patterns; for instance, the node of the word “got” is only linked with diseases of “housemaid” and “knee”.

v. Contexts



Document	Left	Term	Right
1) THE …	I fancy it was. I	got	down the book, and read
1) THE …	me that I had fairly	got	it. I sat for a
1) THE …	wondered what else I had	got	: turned up St. Vitus's Dance
1) THE …	could conclude, I had not	got	, was housemaid's knee. I felt
1) THE …	of slight. Why hadn't I	got	housemaid's knee? Why this invidious
1) THE …	with me. I have not	got	housemaid's knee. Why I have
1) THE …	knee. Why I have not	got	housemaid's knee, I cannot tell
1) THE …	remains that I have not	got	it. Everything else, however, I
1) THE …	Everything else, however, I have	got	." And I told him how

Figure 241 Contexts, The Man Who Was a Hospital

The word “got” has several semantic shades: obtain, understand, become, describe something negative and cause something to happen. In this corpus, the word “got” is used nine times. In the 1st place, “got” means fetched; the 2nd time, it means to understand; and from the 3rd to 9th sentence, it means an attack of several diseases.

7. *My Financial Career* by Stephen Leacock

i. Summary

This corpus has 1 document with 903 total words and 338 unique word forms. Created 5 seconds ago (on 14th January 2018).

Vocabulary Density: 0.374

Average Words Per Sentence: 9.7

Most frequent words in the corpus: said (17); bank (9); manager (8); dollars (7); fifty (7); money (7); accountant(5); looked (5); account (4); asked (4);

Figure 242 Summary, My Financial Career

Technology expedites quantified stylistic analysis of any genre. Stephen Leacock takes 338 unique words, and he repeats them almost three times in this essay to write a total of 903 words. Dividing unique words by total words, 0.374 vocabulary density is calculated, and 9.7 words per sentence have been mentioned as a stylistic feature. Thus, vocabulary density and an average length of sentences are appropriate for basic level readers, and they manifest Stephen Leacock’s stylistic qualities. Such linguistic ease is the requirement of comic essays to comprehend their punch lines. The most frequent words unveil finance-related vocabulary, for example, “bank (9)”, “manager (8)”, “dollars (7)”, “money (7)”, “accountant (5)” and “account (4)”.

ii. Cirrus

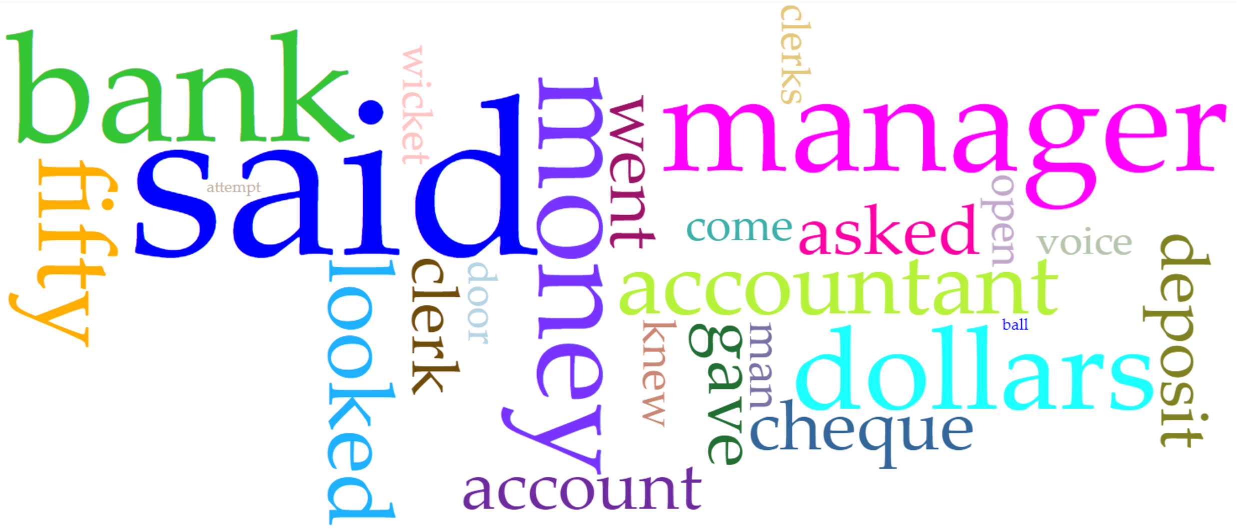
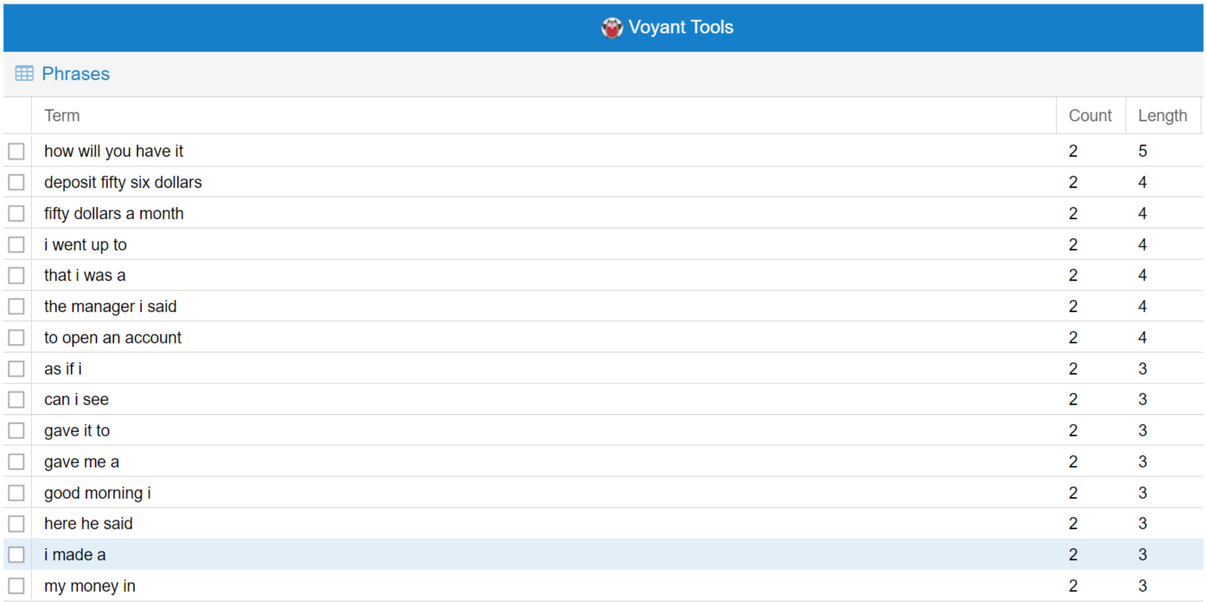


Figure 243 Cirrus, My Financial Career

Figure 242 Cirrus extracts key topics with the topic modelling technique. The most occurring theme, "said (17)" reveals the use of dialogues between the writer and manager as well as the writer and accountant. The second theme of "bank (9)" indicates the geographic setting of the entire essay. The writer, Stephen Leacock, has mentioned his rattling mental condition in the bank. Furthermore, particular banking sector register has been mentioned, for instance, "manager (8)", "fifty (7)", "dollars (7)", "money (7)", "cheque (4)", "deposit (4)", "clerks (3)" and "accountants (5)". These above-mentioned themes support the story of depositing a few rupees in the bank and an odd behaviour that knits the fabric of light comedy.

Human analysis shows that only one male character dominates, whereas female characters are non-existing (Hussain, 2009, p. 56). Comparing with Cirrus tool, Cirrus tool mentions manager, clerks, accountants with their statistical weight, but the human analysis does not quantify them. Main themes of Cirrus are dollars, money, cheque, deposit, while human analysis did not mention them precisely.

iii. Phrases



Voyant Tools

Phrases

	Term	Count	Length
□	how will you have it	2	5
□	deposit fifty six dollars	2	4
□	fifty dollars a month	2	4
□	i went up to	2	4
□	that i was a	2	4
□	the manager i said	2	4
□	to open an account	2	4
□	as if i	2	3
□	can i see	2	3
□	gave it to	2	3
□	gave me a	2	3
□	good morning i	2	3
□	here he said	2	3
□	i made a	2	3
□	my money in	2	3

Figure 244 Phrases, My Financial Career

Derivation of collocation patterns/ n-grams has been based on co-occurrence. So, Phrases tool is employed to extract collocations/ n-grams for example, "how will you have it?" (Prn+Aux+Prn+V+Prn), "went up" (V+Prep), "open an account" (V+Art+N), "as if" (Conj) and "good morning" (Adj+N). They occur twice, and their length ranges from 3 to 5 words.

iv. Links

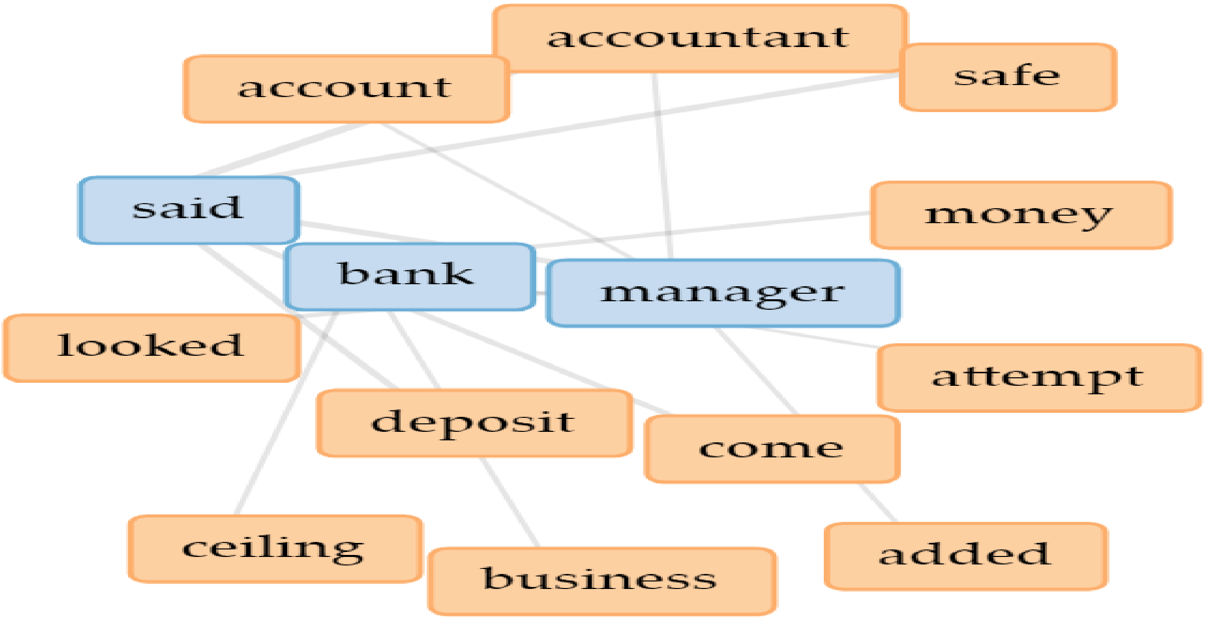
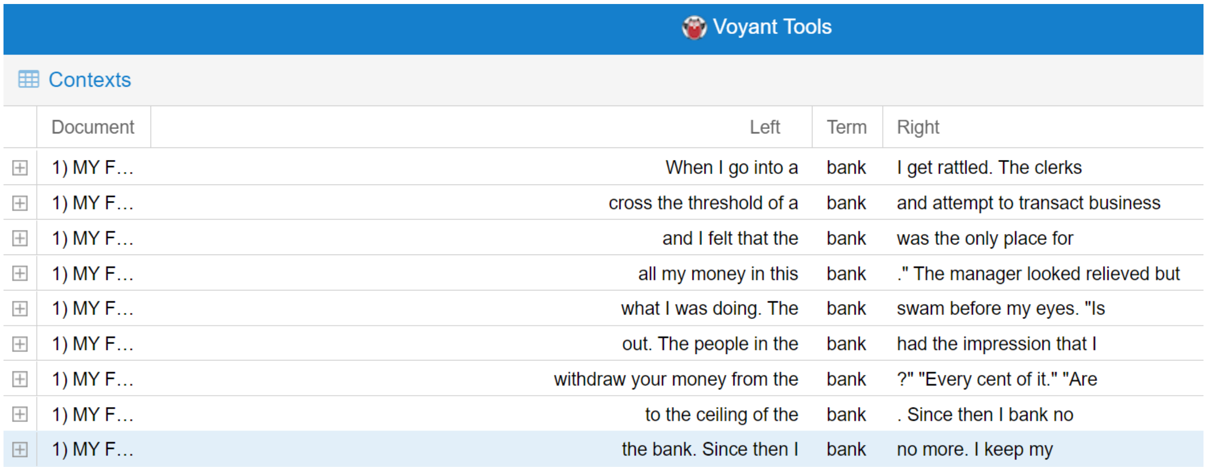


Figure 245 Links, My Financial Career

The KG of "accountant, said, bank, manager" discovers knowledge about the depositor’s conversations with the accountant and manager in this essay. In addition to it, the manager also talks to the accountant and asks him to open an account for the visitor, therefore, this situation has been shown in a multipronged node of the KG, "said, account, accountant, manager". During the meeting of manager and writer, specific points have been discussed, for instance, depositing money, opening an account, amount of money and they have been delineated in this KG "money, bank, deposit". Other KG of "safe, said, bank, manager" reminds us of the event when the writer enters the safe instead of a door, and this comic situation strengthens the manager's point of view that the depositor is an insane person. There is another interesting KG of "attempt, bank, ceiling" which reveals the climax of the essay when he withdraws his money pretending that somebody has mishandled him during the business transaction. In a fit of fury, he closes the recently opened bank account, but the bank staff laughs at him loudly in a mocking tone.

v. Contexts



	Document	Left	Term	Right
⊞	1) MY F…	When I go into a	bank	I get rattled. The clerks
⊞	1) MY F…	cross the threshold of a	bank	and attempt to transact business
⊞	1) MY F…	and I felt that the	bank	was the only place for
⊞	1) MY F…	all my money in this	bank	." The manager looked relieved but
⊞	1) MY F…	what I was doing. The	bank	swam before my eyes. "Is
⊞	1) MY F…	out. The people in the	bank	had the impression that I
⊞	1) MY F…	withdraw your money from the	bank	?" "Every cent of it." "Are
⊞	1) MY F…	to the ceiling of the	bank	. Since then I bank no
⊞	1) MY F…	the bank. Since then I	bank	no more. I keep my

Figure 246 Contexts, My Financial Career

The word "bank" has different meanings as a noun and verb. As a noun, it means a place for depositing and withdrawing money; a sloping verge of a river; a layer of mud; a row of something; a collection of something and to keep money in a financial bank. So, semantic and grammatical ambiguities arise about the most appropriate meaning of the word “bank” in this essay. In this situation, the theoretical framework suggests a solution that “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). On the basis of this idea, contexts tool is designed to show bidirectional context for word sense disambiguation, since clarity of denotative meaning is the first step of hermeneutics. To disambiguate, the word “bank” is retrieved nine times, and the first eight times it is used as a noun with the meaning of a place for the transaction of money, and the 9th time, it is used as a verb with the meaning of a banking process.

8. *China's Way to Progress* by Galeazzo Santini

i. Summary

This corpus has 1 document with 2,244 total words and 934 unique word forms. Created 23 seconds ago (on 17th January 2018).

Vocabulary Density: 0.416

Average Words Per Sentence: 23.1

Most frequent words in the corpus: chinese (19); china (12); world (8); agricultural (6); country (6); day (6); women (6); wor ker (6); workers (6); family (5);

Figure 247 Summary, China's Way to Progress

Computational stylistics finds aspects of the literary style of Galeazzo Santini. The vocabulary density of this essay is 0.416, and it is calculated by the division of unique words by total words. The summary tool shows that 934 unique words are written in it, and they are repeated more than twice, so they become 2244 total words. The most occurring key words are concerned with “Chinese (19)”, “China (12)”, its relationship with “world (8)”, its “agricultural (6)” system and treatment with men, “women (6)” and “workers (6)”. Moreover, without mentioning the remarkable contributions of Mao, the story of Chinese development cannot be completed.

ii. Cirrus

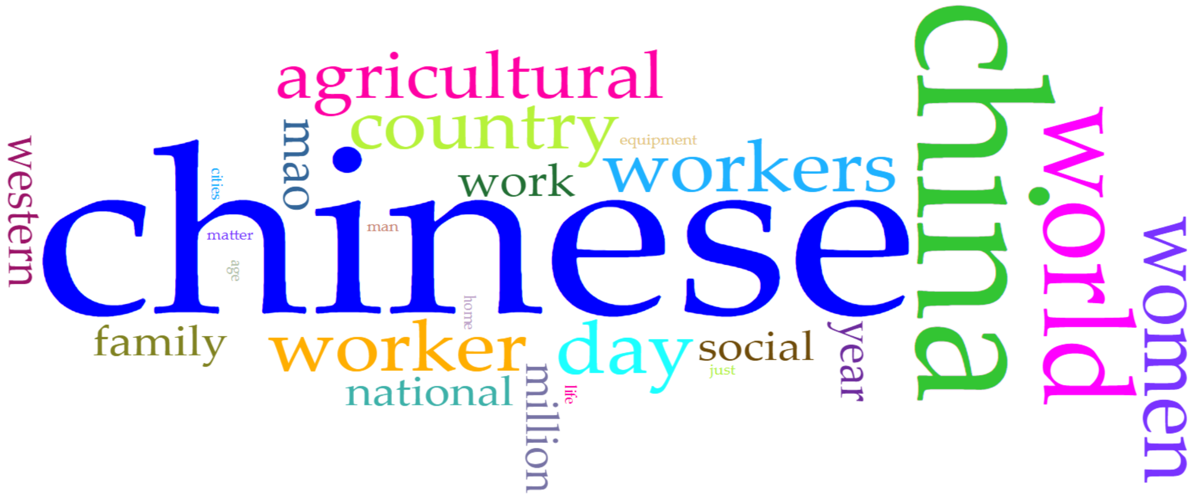


Figure 248 Cirrus, China's Way to Progress

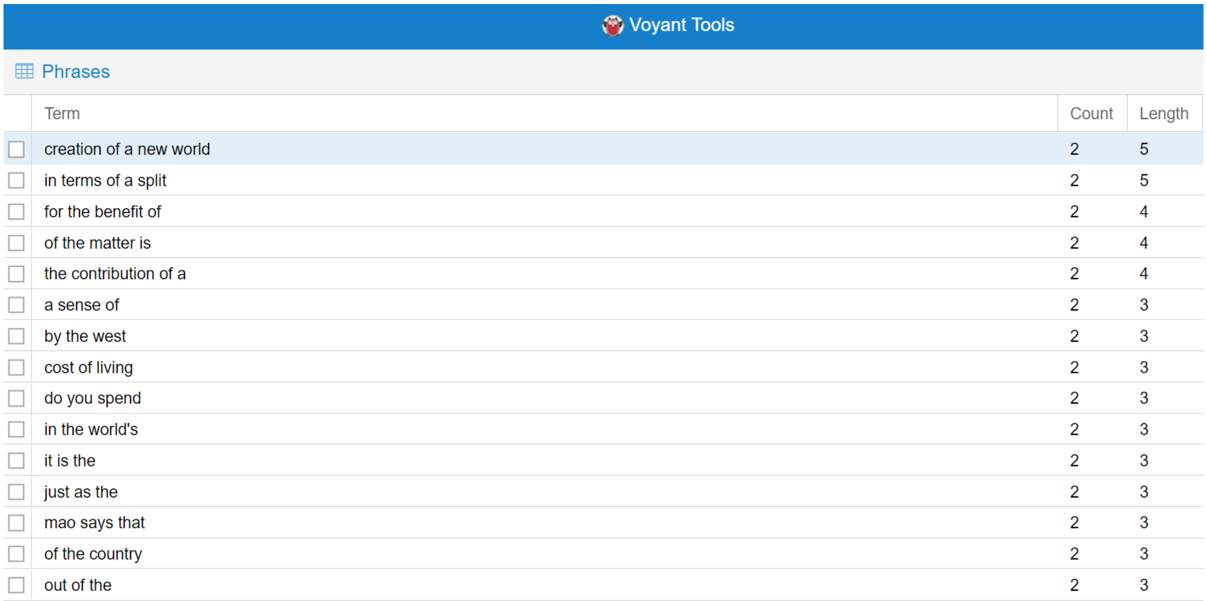
In Cirrus of this essay, the most important theme is "Chinese (19)" but not China. “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166) because Cirrus supports human cognition in the process of knowledge discovery. Again, these tools provide a hermeneutic aid for human multifaceted thinking processes. This Cirrus gives the interesting new knowledge pattern that the Chinese government pays more attention towards welfare projects for Chinese workers and the common masses, so the key policy is humanism; for instance, they provide social security benefits and numerous other advantages to their workers. Two other relevant themes are "family (5)" and "social (5)", which refer to the Chinese social security system that provides social welfare advantages to the Chinese.

Another very realistic knowledge pattern is revealed that "women (6)" and "workers (6)" are preferred to machines, so they include the use of machines carefully lest Chinese workers should face unemployment. Another very interesting knowledge pattern is discovered that women workers are preferred to male workers, and statistics prove that the words "women" and “man” are used six and four times respectively.

Another theme of "agricultural (6)" shows that China does not ignore the agricultural system because they realise that the road of the industry goes through the agricultural fields. If there is no agriculture, there is no industry because the industry requires raw agricultural material. In a nutshell, the industry at the cost of agriculture is a mere loss. The character and theme of "Mao (5)" indicates that they are leading their lives according to the maxims of Mao and his Red Book, that is why he transforms the opium-addicted Chinese nation into an economic superpower.

Human analysis shows that the main characters are Chou En Lai and Mao Tse Tung, whereas unnamed women characters are present. Women work in fields and factories enjoying equal social and professional opportunities and facilities. Other themes are lack of femininity in women, confidence, dignity and undoubted awareness of women about their rights (Hussain, 2009, pp. 58-59). Comparing human and Cirrus tool analysis, characters of Mao, women and workers are the same, but human and machine analysis themes are different.

iii. Phrases



	Term	Count	Length
□	creation of a new world	2	5
□	in terms of a split	2	5
□	for the benefit of	2	4
□	of the matter is	2	4
□	the contribution of a	2	4
□	a sense of	2	3
□	by the west	2	3
□	cost of living	2	3
□	do you spend	2	3
□	in the world's	2	3
□	it is the	2	3
□	just as the	2	3
□	mao says that	2	3
□	of the country	2	3
□	out of the	2	3

Figure 249 Phrases, China's Way to Progress

Standard phraseology is "creation of a new world" (N+Prep+Art+Adj+N), "in terms of" (Prep+N+Prep), "for the benefit of" (Prep+Art+N+Prep), "contribution of" (N+Prep), "a sense of" (Art+N+Prep), "cost of living" (N+Prep+N), "in the world" (Prep+Art+N), "just as" (Id), "out of" (Prep+Prep) and "agricultural Commune" (Adj+N). They occur twice, and their length ranges from 3 to 5 words. Collocations/ n-grams not only teach standard phraseology for fluency but also teach correct grammatical structures which represent the repeated ideology of the essay too, for example, the Chinese ease the life of their people with welfare revolution of the entire society, and this ideation is evident in figure 248.

iv. Links

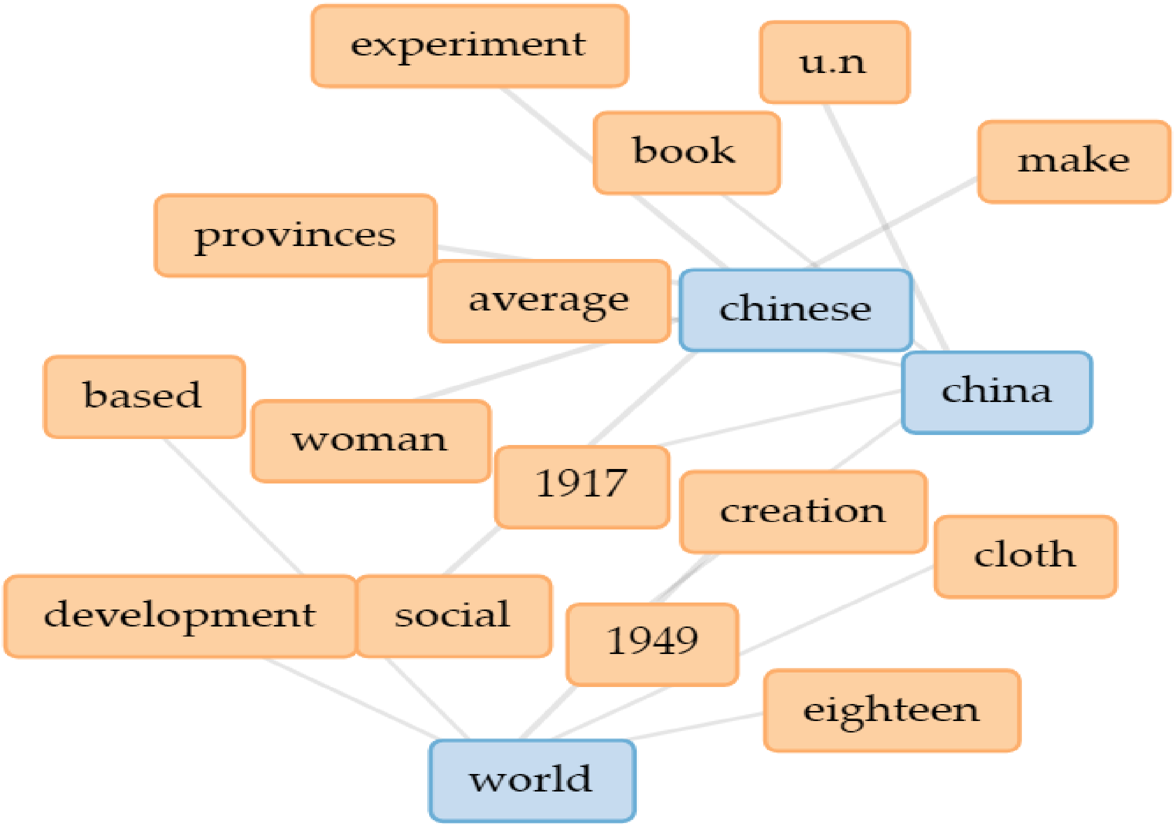
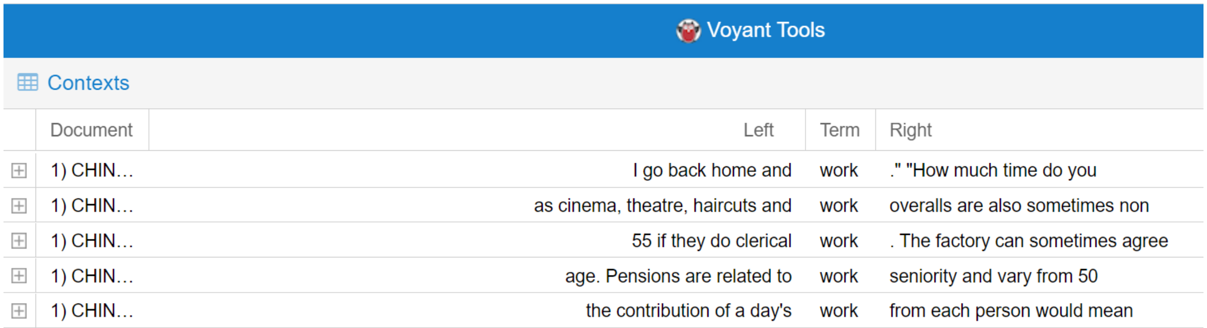


Figure 250 Links, China's Way to Progress

“In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct” (Agrawal, & Psaila, 1995, p. 1). One KG comprises " provinces, Chinese, make" and discovers the knowledge pattern that all Chinese provinces are self-sufficient to survive. Another KG of " social, Chinese, experiment" shows that the Chinese government conducts a social experiment for the welfare of Chinese by giving them several social security benefits. One more KG, "1917, China, creation, 1949" compares the predicament of the Chinese during colonialism and after getting independence from imperialistic powers. It also compares the Chinese condition with the Russian situation during colonial eras. Another KG of " book, Chinese, China " refers to their reading of Mao's visionary Red Book before starting their work; therefore, it motivates them to work for national development. One significant KG of "U.N., China" demonstrates that the United Nations recognizes China after 22 years and regrets this delay. Another interesting KG is "cloth, world, China" that the Chinese nation requires such a long cloth which can be wrapped around the world several times.

v. Contexts



Document	Left	Term	Right
1) CHIN…	I go back home and	work	." "How much time do you
1) CHIN…	as cinema, theatre, haircuts and	work	overalls are also sometimes non
1) CHIN…	55 if they do clerical	work	. The factory can sometimes agree
1) CHIN…	age. Pensions are related to	work	seniority and vary from 50
1) CHIN…	the contribution of a day's	work	from each person would mean

Figure 251 Contexts, China's Way to Progress

The word "work" can be used as a verb, adjective or noun. It can give the semantic shade of physical labour or literary work. To disambiguate word sense, the word "work" is used five times, and it clarifies lexical ambiguity. It is used as a verb in the 1st sentence, in the second sentence, it is used as an adjective, and in the 3rd, 4th and 5th sentences, it is used as a noun.

9. *Hunger and Population Explosion* by Anna Mckenzie

i. Summary

This corpus has 1 document with 1,595 total words and 590 unique word forms. Created about 28 minutes ago (on 17th January 2018).

Vocabulary Density: 0.370

Average Words Per Sentence: 18.1

Most frequent words in the corpus: population (19); famine (15); rate (15); people (14); death (9); food (9); years(9); birth (8); million (8); world (8);

Figure 252 Summary, Hunger and Population Explosion

The summary tool presents the stylometry of Anna Mckenzie who uses 590 unique words in this corpus, and she repeats them almost three times until the total words of this essay reach 1595 words. Dividing unique words by total words, 0.370 vocabulary density has been calculated, and it is suitable for intermediate level readers. Anna Mckenzie writes 18.1 words in a sentence, and it is appropriate for advanced level readers. This essay includes repeated vocabulary and long sentences. The most frequent themes in this essay are related to human "population (19)", "famine (15)”, “death (9)” and “food (9)”.

ii. Cirrus

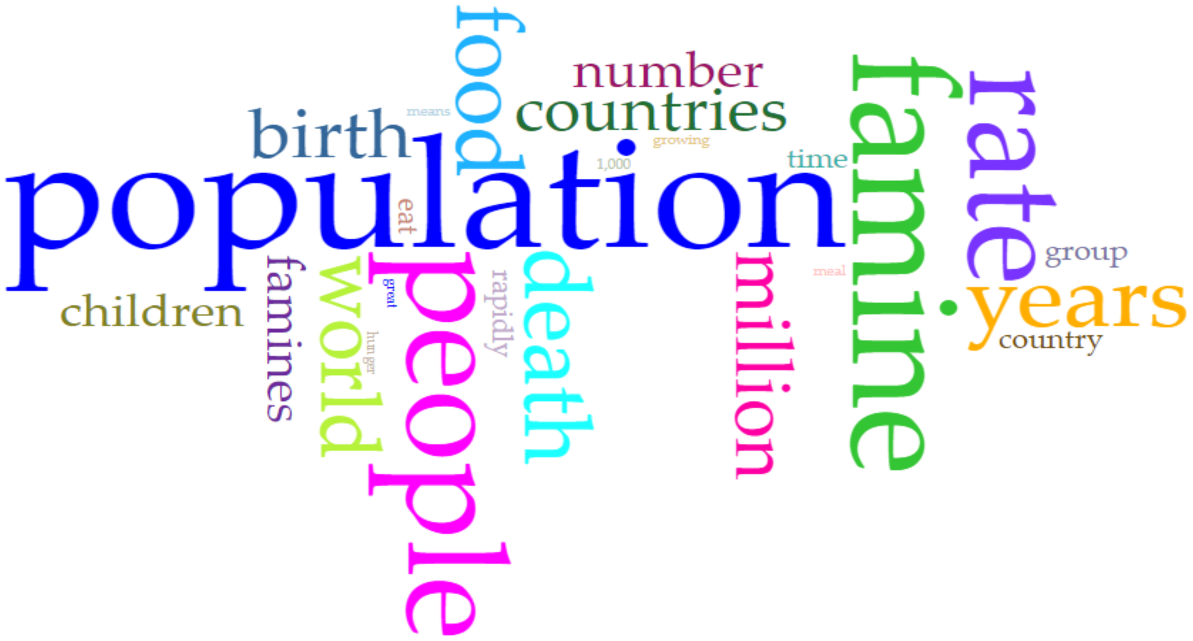


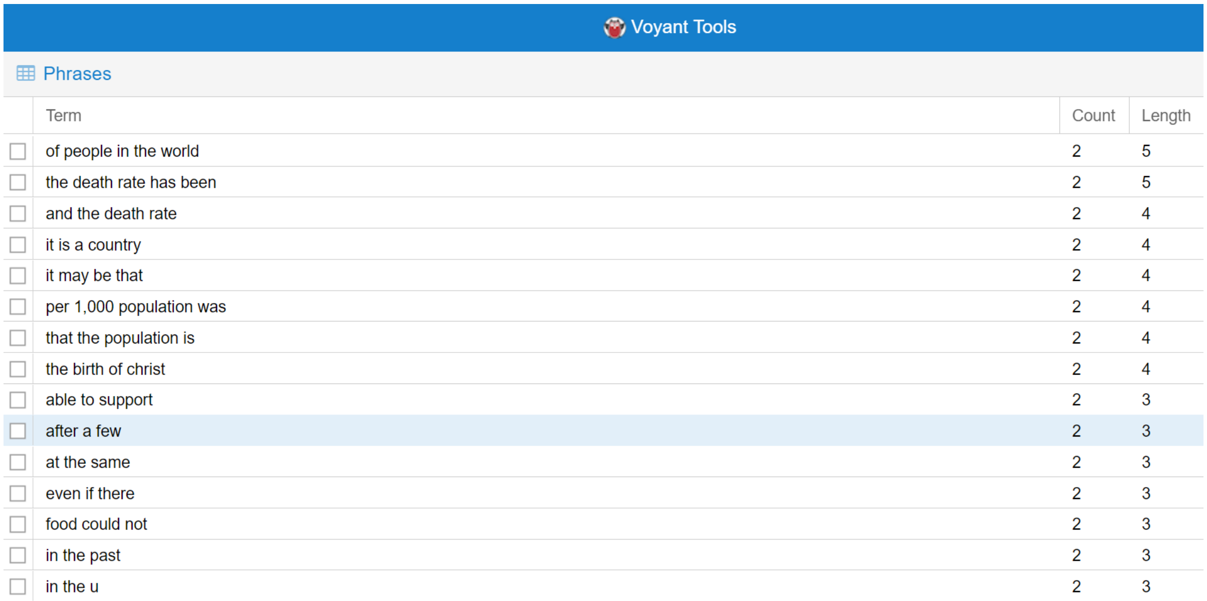
Figure 253 Cirrus, Hunger and Population Explosion

Cirrus is a data visualization technique to sort out dominating and the most occurring themes with topic modelling. The two most occurring themes are "population (19)" and "famine (15)", and they are interlinked since the rise in population leads to famine and scarcity of food. To prevent famine, decreasing the birth rate is essential because it is evident that more people demand more food. On the other hand, sources of developing nations do not grow according to the pace of their population hype. Several historic "famines (6)" and their human losses have been mentioned in this essay. The essayist depicts famine-stricken weakest "children (6)" who are looking like a living skeleton or pieces of liquorice.

Some mathematical themes have also been mentioned, for example, "rate (15), birth (8), death (9), million (8), number (6), 1000 (4)". First of all, the essayist mentions the method to calculate the population growth rate by subtracting the death rate from the birth rate in 1000 persons. Moreover, the population is increasing tremendously in millions, and the scarcity of food is also increasing day by day.

Human analysis shows that no female character has been delineated in this essay, while this literary essay has been written by a female character (Hussain, 2009, p. 59). Cirrus tool deeply analyses different characters of population and children. Key themes are also mentioned precisely about famines, birth control, death, million. To conclude, Cirrus analysis is more profound and detailed as compared to human analysis.

iii. Phrases



	Term	Count	Length
□	of people in the world	2	5
□	the death rate has been	2	5
□	and the death rate	2	4
□	it is a country	2	4
□	it may be that	2	4
□	per 1,000 population was	2	4
□	that the population is	2	4
□	the birth of christ	2	4
□	able to support	2	3
□	after a few	2	3
□	at the same	2	3
□	even if there	2	3
□	food could not	2	3
□	in the past	2	3
□	in the u	2	3

Figure 254 Phrases, Hunger and Population Explosion

Collocation patterns/ n-grams in this corpus are "in the world" (Prep+Art+N), "death rate" (Adj+N), "may be" (Aux), "per 1000 population" (Prep+Nu+N), "birth of Christ" (N+Prep+N), "a few" (Prn), "at the same" (Prep+Art+Adj), "even if" (Phr) and "in the past" (Prep+Art+Adv). They occur twice, and their length ranges from 3 to 5 words. These bigrams and trigrams are co-occurring in the corpus to establish their entity recognition and to enhance fluency in language skills.

iv. Links

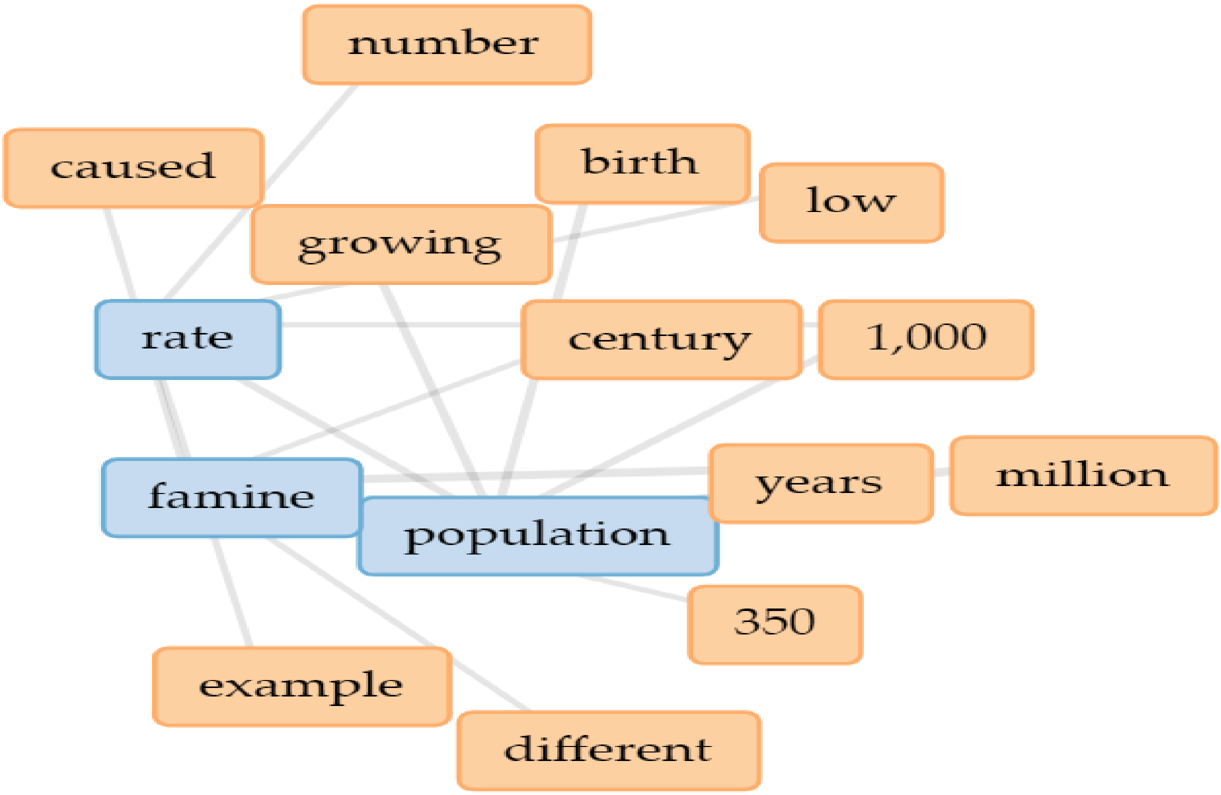
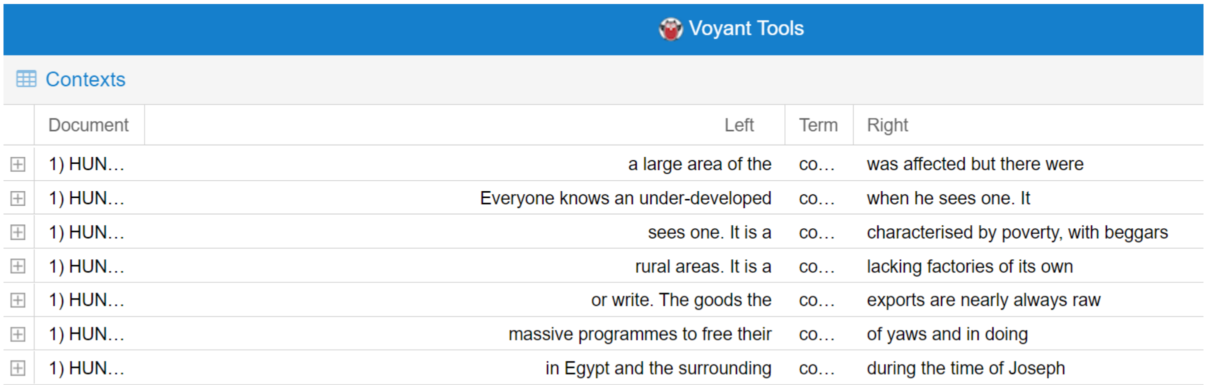


Figure 255 Links, Hunger and Population Explosion

KGs build interrelationships of various entities for multifaceted hermeneutic analysis. The KG of "1000, population, birth" elucidates the method to calculate population growth in any community. Another KG of “rate, famine, population, 350” reveals the factual knowledge that in 18 centuries, other famines in different parts of the world affected humanity for 350 years. Another KG of " low, rate, famine, population" reveals that millions of people died during previous famines, so the human population declined rapidly. To summarize, more knowledge graphs can also be built to discover more hermeneutic patterns. This point has been mentioned in the theoretical framework that “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166).

v. Contexts



	Document	Left	Term	Right
⊞	1) HUN…	a large area of the	co…	was affected but there were
⊞	1) HUN…	Everyone knows an under-developed	co…	when he sees one. It
⊞	1) HUN…	sees one. It is a	co…	characterised by poverty, with beggars
⊞	1) HUN…	rural areas. It is a	co…	lacking factories of its own
⊞	1) HUN…	or write. The goods the	co…	exports are nearly always raw
⊞	1) HUN…	massive programmes to free their	co…	of yaws and in doing
⊞	1) HUN…	in Egypt and the surrounding	co…	during the time of Joseph

Figure 256 Contexts, Hunger and Population Explosion

The word "country" is polysemic, for instance, a land on which a government operates, farming land and land with particular features. To diminish this semantic ambiguity, figure 256 reveals that the word "country" is used seven times as a noun, and every time it means famine-affected third world governments and countries. Since “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166), it resolves all ambiguities for a better understanding of the text. Later, the semantically clear text leads to discover deep and multifaceted hermeneutic patterns.

10. *The Jewel of the World* by Philip K. Hitti

i. Summary

This corpus has 1 document with 2,256 total words and 879 unique word forms. Created 8 seconds ago (on 17th January 2018).

Vocabulary Density: 0.390

Average Words Per Sentence: 22.1

Most frequent words in the corpus: al (22); spain (15); abd (14); rahman (13); cordova (10); caliph (9); muslim (8); world (8) ; years (8); capital (7);

Figure 257 Summary, The Jewel of the World

“Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). Philip K. Hitti employs 879 unique words in this essay, and he repeats them almost a little less than three times, so total words reach the limit of 2256. Dividing unique words by total words, 0.390 vocabulary density is calculated, and it is suitable for intermediate-level readers. Usually, he uses 22.1 words in a sentence, and it is suitable for highly advanced level learners. The most frequent words briefly inform us about central points of this biographical essay which are “Muslim (8) Caliph (9)” “Abd (14)”, “al (22)” “Rahman (13)” ’s life and his socio-political and financial achievements in “Cordova (10)”.

ii. Cirrus

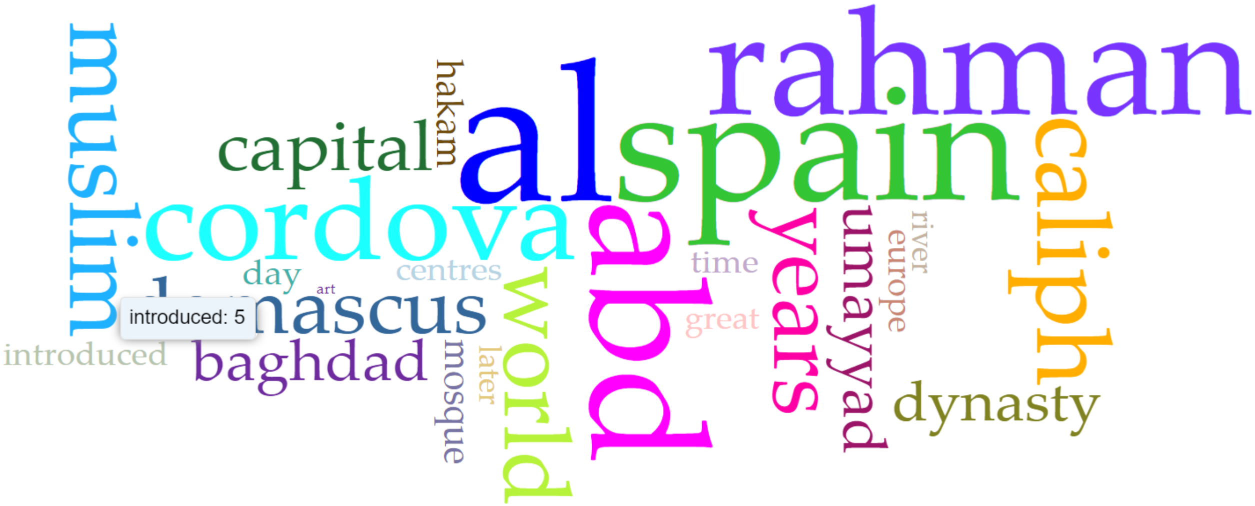


Figure 258 Cirrus, The Jewel of the World

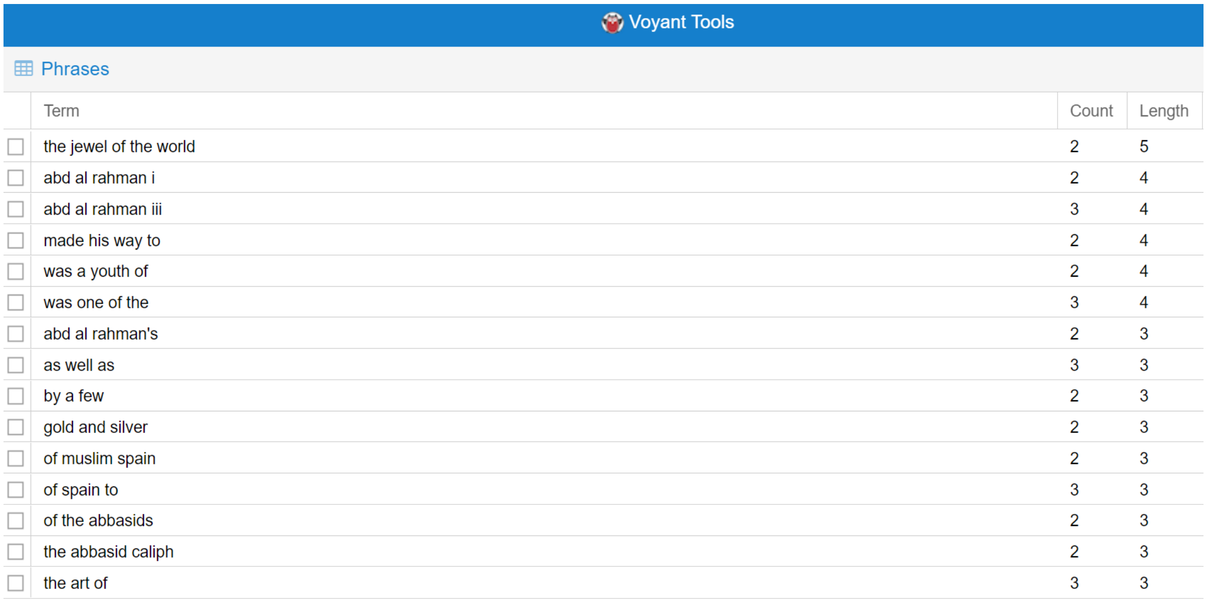
The most common word in this corpus is the Arabic word "Al" which transforms a common noun into a proper noun. It reveals a powerful Arabic linguistic influence on all characters, things and places. Arabic is the official language of Islam, so they establish deep-seated ties, as it is evident in this essay. The theme of "Muslim (8)" also shows that Muslims have occupied Spain for several centuries, and it strengthens the use of Arabic "al" with names and places. Another theme is also linked with the Muslim dynasty because Muslims obey their "caliph (9)" who is a supreme spiritual leader.

The geographical setting of the essay is "Spain (15)" and some other geographical places are "Damascus (7)" and "Baghdad (6)". Under the leadership of Abdul Rahman, "Cordova (10)”, the capital of Spain, became a centre of excellence equal to Damascus and Baghdad which were rivals to Cordova. When Cordova culminated to its zenith in different domains of arts, academia and human welfare, the whole "Europe (5)" was in an abyss of ignorance and darkness.

The most important character is Abdal "Rahman (13)" who is the pioneer of the Muslim dynasty in Spain. There is an old tussle between the "Umayyad (6)" dynasty and the Abbasids dynasty. Abdul Rahman belongs to the Umayyad dynasty; that is why the name of his dynasty is used six times in this corpus. He starts his Spanish rule from scratch and makes it a great welfare Muslim state that surpasses Europe and becomes an illuminating jewel among three hubs of excellence. Then his successors rule with the same title, ‘Amir’ and enhanced the acme of Spain. Another significant character is “Al Hakam” who is a bibliophile and supports students, scholars, libraries and writers generously in Muslim Spain.

Its characters are some males and one nun. Key themes are about male perspectives (Hussain, 2009, pp. 59-60). As Cirrus tool analysis is concerned, it shows "Rahman (13)", Caliph, “Al Hakam” and Muslims separately with their names and statistical weight. It also mentions geographical settings with city names "Damascus (7)" and "Baghdad (6)". To conclude, Cirrus tool presents a more comprehensive analysis in comparison to human analysis.

iii. Phrases



	Term	Count	Length
□	the jewel of the world	2	5
□	abd al rahman i	2	4
□	abd al rahman iii	3	4
□	made his way to	2	4
□	was a youth of	2	4
□	was one of the	3	4
□	abd al rahman's	2	3
□	as well as	3	3
□	by a few	2	3
□	gold and silver	2	3
□	of muslim spain	2	3
□	of spain to	3	3
□	of the abbasids	2	3
□	the abbasid caliph	2	3
□	the art of	3	3

Figure 259 Phrases, The Jewel of the World

Frequent multi-word expressions have been extracted for example "jewel of the world" (N+Prep+Art+N), "Abdal Rahaman" (N+N), "made his way" (V+Prn+N), "one of the" (Prn+Prep+Art), "as well as" (Conj), "a few" (Prn), "gold and silver" (N+Conj+N), "Muslim Spain" (Prop Adj+N), "Abbasid Caliph" (Prop Adj+N), "great mosque" (Adj+N) and "Umayyad dynasty" (Prop Adj+N). These collocations/ n-grams occur 2 to 3 times, whereas their length varies from 3 to 5 words.

iv. Links

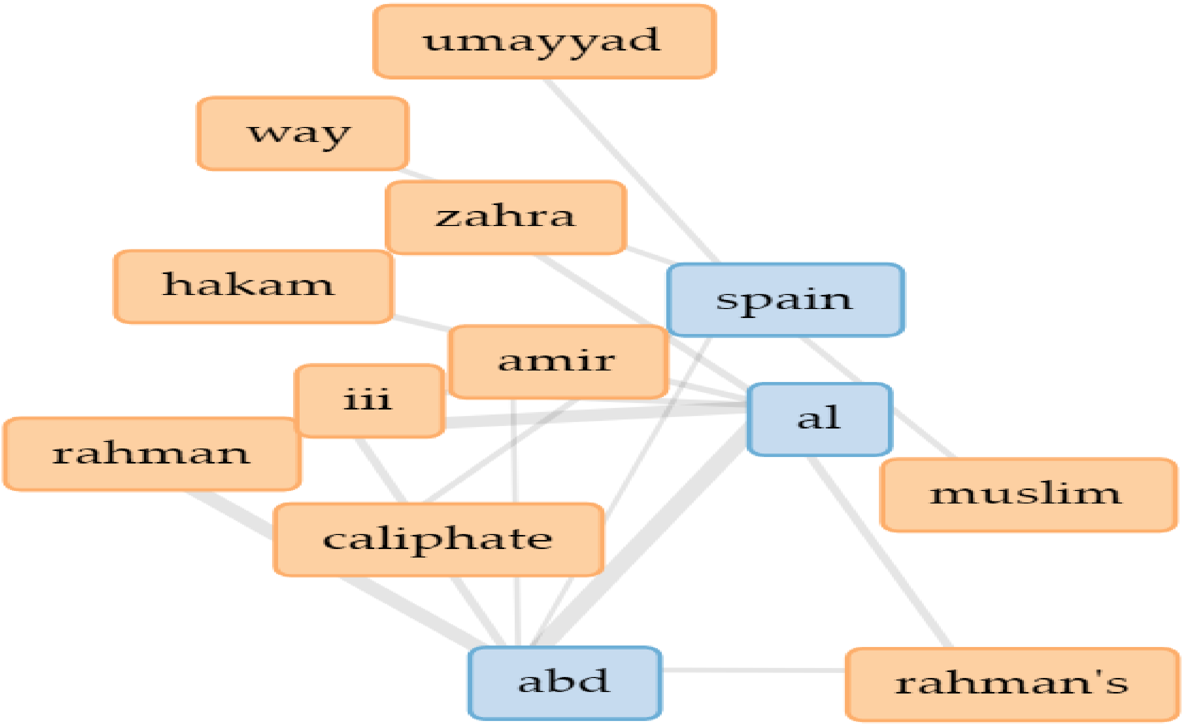
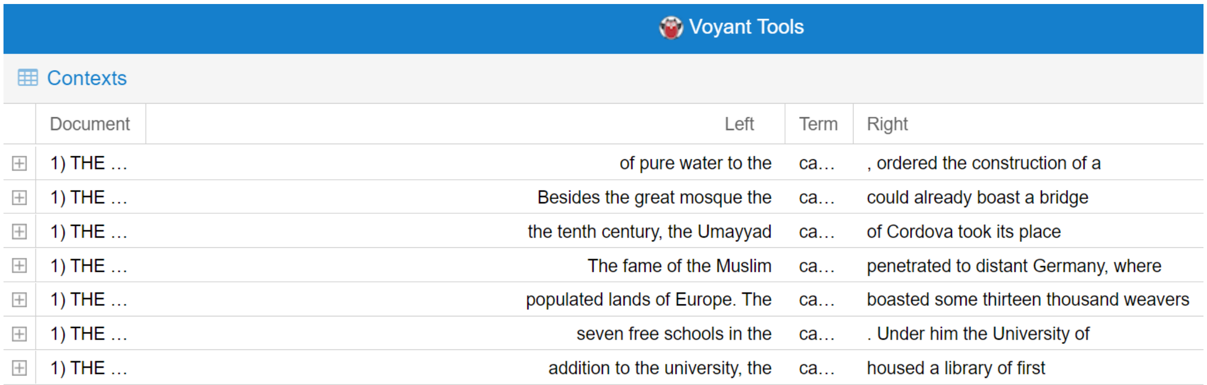


Figure 260 Links, The Jewel of the World

The most frequent word "al" is linked with "Al, Amir, Hakam, iii, Rahman" in the KG to reveal knowledge that Arabic "al" is used with the aforementioned four words. Another KG is built with "al, Abd, Rahman, iii" which informs us that title Amir is given to Abdal Rahman III and his predecessors and successors. The theme of "Umayyad" intersects several other links which show that all these people belong to the Umayyad dynasty. One more KG of "Umayyad, Spain, Muslim" reveals the knowledge pattern that the Umayyad dynasty gets its prestigious position in Spain. One more KG of "al, Amir, Zahra" exposes knowledge that the royal family of Abd Al Rahman lives in the palace. Amir is the title of Abdal Rahman and his successors. One theoretical postulate informs that “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). Human schema, understanding and quest to search knowledge are also required to interpret any KG. In short, these knowledge graphs are used for the exploration of multidimensional hermeneutic patterns.

v. Contexts



	Document	Left	Term	Right
⊞	1) THE …	of pure water to the	ca…	, ordered the construction of a
⊞	1) THE …	Besides the great mosque the	ca…	could already boast a bridge
⊞	1) THE …	the tenth century, the Umayyad	ca…	of Cordova took its place
⊞	1) THE …	The fame of the Muslim	ca…	penetrated to distant Germany, where
⊞	1) THE …	populated lands of Europe. The	ca…	boasted some thirteen thousand weavers
⊞	1) THE …	seven free schools in the	ca…	. Under him the University of
⊞	1) THE …	addition to the university, the	ca…	housed a library of first

Figure 261 Contexts, The Jewel of the World

Word “capital” means huge wealth and the central administrative hub of any country. To disambiguate this semantic shade, it is essential to study the key word in context. The word “capital” is used seven times in this essay, and every time it refers to Cordova, the central administrative hub of Muslim Spain.

4.9 Text Mining of Biographical Essays/ Heroes

A biographical essay is an account of a notable person’s remarkable achievements.

11. *First Year at Harrow* by Sir Winston S. Churchill

i. Summary

This corpus has 1 document with 779 total words and 356 unique word forms. Created about 6 minutes ago (on 16th December 2017).

Vocabulary Density: 0.457

Average Words Per Sentence: 15.9

Most frequent words in the corpus: english (9); latin (6); learn (5); boys (4); mr (4); clauses (3); examinations (3); greek (3); harrow (3); long (3);

Figure 262 Summary, First Year at Harrow

Summarization shows a condensed report of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45). The mined text has been quantified for the stylometry of the text and the essayist. Winston S. Churchill’s essay comprises 356 unique words, and they are used almost twice in this text; thus, the density of its vocabulary counts 0.457 which is suitable for basic level readers. Churchill writes 15.9 words in a sentence, and it is appropriate for intermediate-level readers. The most frequent words indicate that this essay is associated with academia, Latin and English language. This essay has an educational setup in which learning languages, examinations and examiners have been discussed. To conclude, Summary tool shows the computational stylometry of Winston Churchill's essay.

ii. Cirrus

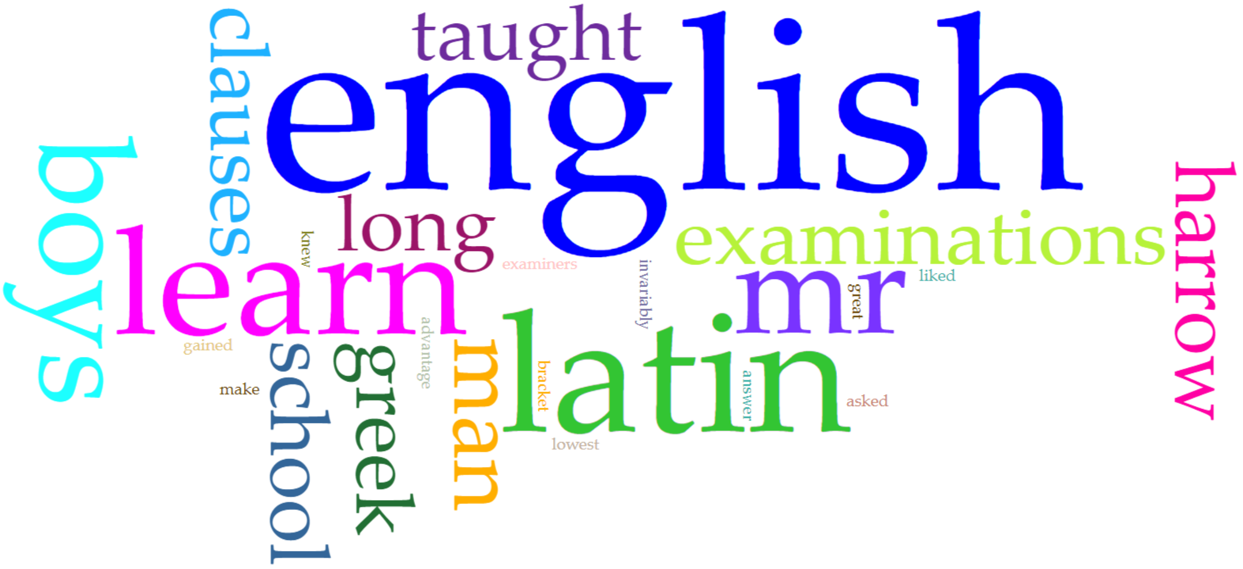


Figure 263 Cirrus, First Year at Harrow

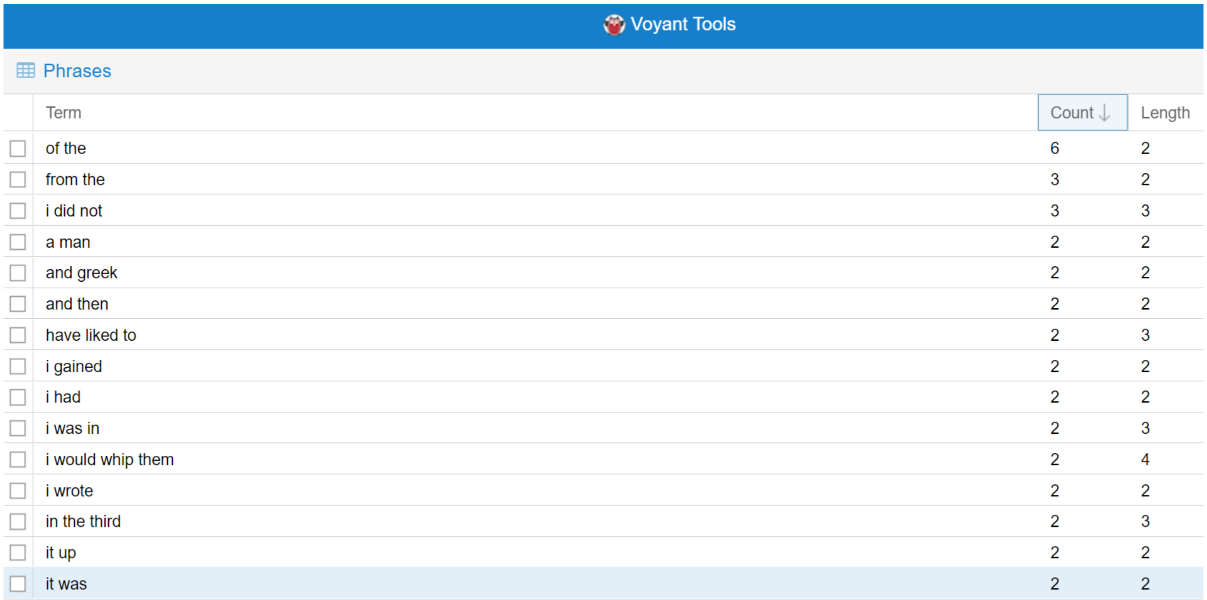
All actions of this essay happen in the geographical setting of "Harrow (3)" "School (3)". This Cirrus shows characters of "man (3)" and "boys (4)", and it also dawns upon us the reality that only boys are taught at Harrow in those days; that is why no feminine name is present in the Cirrus.

There are several themes of "examinations (3)", "examiners (2)", "asked (2)", "answered (2)" for Churchill's entrance exam, and he put "bracket (2)" around the question number. Even then, he was admitted to Harrow, but he was placed in the "lowest (2)" level. Cirrus of this essay reveals major themes of learning "English (9)", "Latin (6)" and "Greek (3)". Previously, brilliant boys learnt Greek and Latin languages, while dunces were educated only in the English language. Churchill was also considered a dunce during his academic career.

Winston Churchill fails at Harrow, and stays in the same class; he has to learn English again and again, so he gets an ultimate "advantage (2)" of learning English structure which penetrates his bone marrow. The rationale for the most occurring word, “English” is that Winston Churchill was the weakest and the most repeatedly failing student at Harrow school, so the major part of the essay covers his academic life and learning styles of English under the guidance of his English teacher. A text mining study of the world-famous 1000 WikiLeaks emails also finds key themes with Cirrus, and its findings are aligned with the findings of the current study (Kemman, 2016).

Human analysis of the text shows that only a male character has been portrayed with a male perspective (Hussain, 2009, p. 56). Cirrus tool reveals characters of the biographical essay in detail and mentions “man (3)”, “boys (4)” and “examiners (2)” as key characters. Comparing human and Cirrus analysis, the former shows a male perspective, while the latter shows learning of “English (9)”, “Latin (6)” and “Greek (3)” languages for different sorts of students. To summarize, Cirrus presents a more detailed and quantified analysis as compared to content analysis.

iii. Phrases



	Term	Count ↓	Length
□	of the	6	2
□	from the	3	2
□	i did not	3	3
□	a man	2	2
□	and greek	2	2
□	and then	2	2
□	have liked to	2	3
□	i gained	2	2
□	i had	2	2
□	i was in	2	3
□	i would whip them	2	4
□	i wrote	2	2
□	in the third	2	3
□	it up	2	2
□	it was	2	2

Figure 264 Phrases, First Year at Harrow

From the most occurring phrases, the standard trigram is “in the third” (Prep+Art+Adj). Besides, these collocations reveal useful information for language learners, for instance, the use of the 3rd form of a verb with the auxiliary verb “have”, for instance, “have liked to” (Aux+V+Prep). Another collocation/ n-gram, “I would whip them” (Prn+Mod+V+Prn) guides the use of the first form of the verb with “would”. Another knowledge pattern shows that collocation/ n-gram “and then" (Conj+Adv) can be used together, and they are not considered redundancy. These collocations/ n-grams occur 2 to 6 times, and their length ranges from 2 to 4 words.

iv. Links

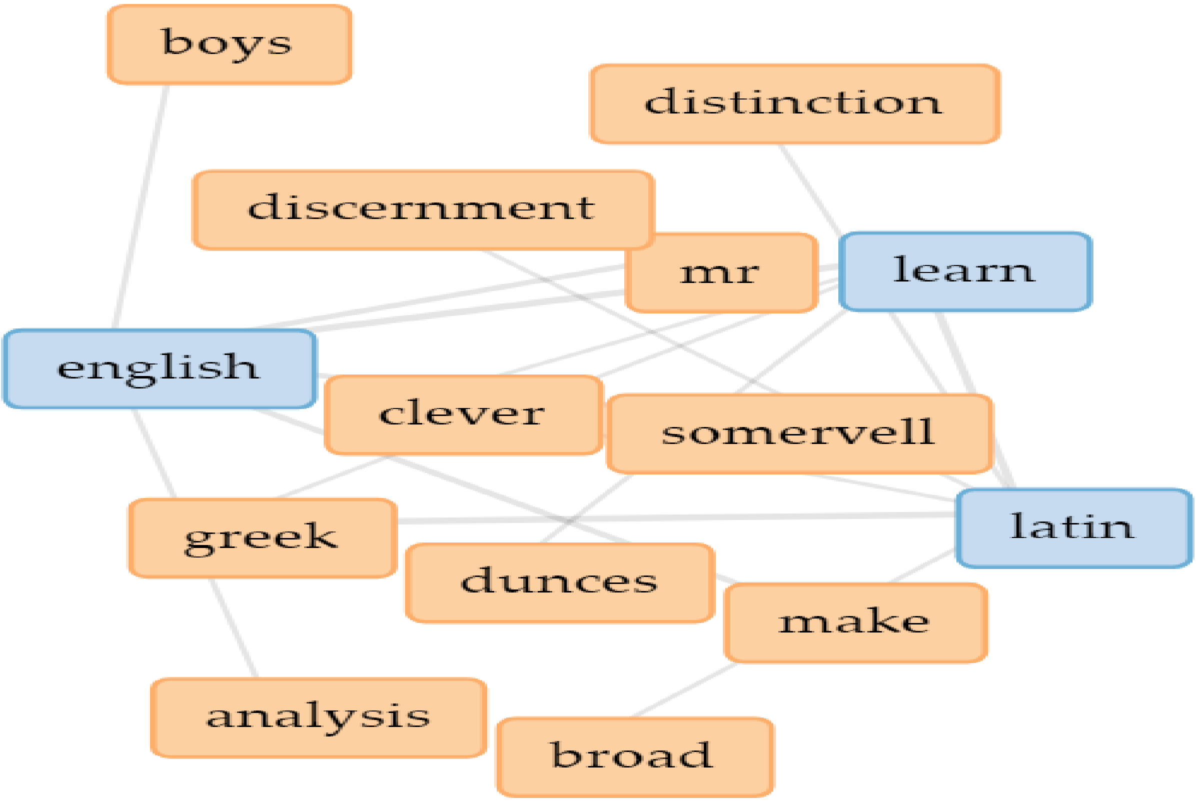
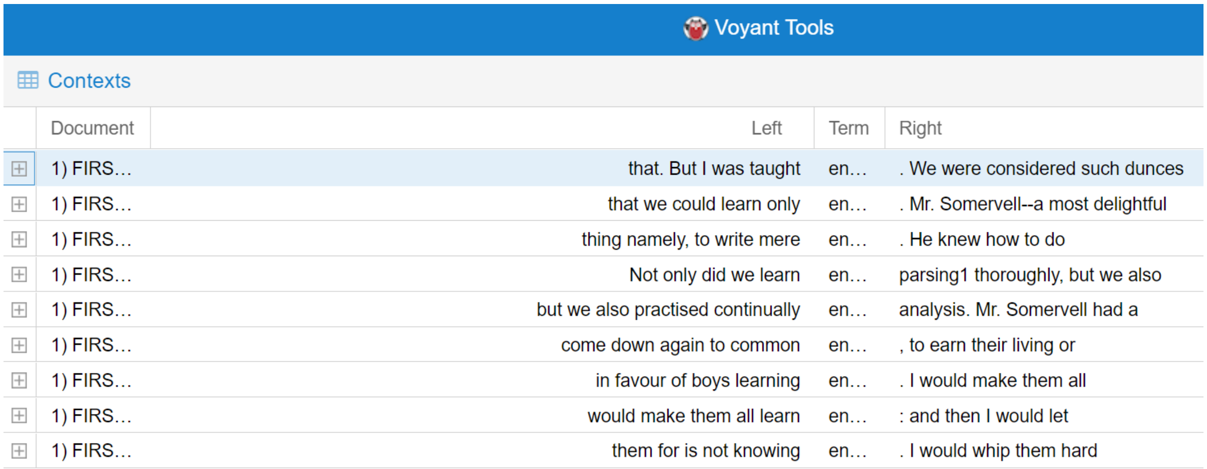


Figure 265 Links, First Year at Harrow

Knowledge Discovery Theory is defined as "the extraction of implicit, previously unknown and potentially useful information from data" (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). The KG reveals an accurate pattern of "clever, Greek, Latin" which shows that clever boys used to learn Greek and Latin languages. The noticeable point is that the words "English" and "dunces" are interconnected because dunces learn English, not Greek or Latin languages. Its evidence is visible with the KG, "learn, dunces, English". Another knowledge pattern of "distinction, learn, Latin," unveils that learning and writing beautiful Latin bears fruit, and they win distinctions. Another KG of "Latin, discernment" reveals that Mr Welldon shows discernment in measuring general competence after looking at Churchill's Latin prose.

In conclusion, Hermeneutica Theory guides that "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166), and it affirms that English is linked with dunces as the textual evidence highlights about the situation of those days. Knowledge bearing quality of text mining tools has been shown in the nodes of knowledge graphs, and human reflection interprets multi-layered themes with the help of KGs. This textual evidence proves the accuracy of the knowledge graphs, as it is mentioned that “It is supplemented by other materials” (Rockwell, & Sinclair, 2016, p. 166).

v. Contexts



Voyant Tools

Contexts

Document	Left	Term	Right
1) FIRS…	that. But I was taught	en…	. We were considered such dunces
1) FIRS…	that we could learn only	en…	. Mr. Somervell--a most delightful
1) FIRS…	thing namely, to write mere	en…	. He knew how to do
1) FIRS…	Not only did we learn	en…	parsing1 thoroughly, but we also
1) FIRS…	but we also practised continually	en…	analysis. Mr. Somervell had a
1) FIRS…	come down again to common	en…	, to earn their living or
1) FIRS…	in favour of boys learning	en…	. I would make them all
1) FIRS…	would make them all learn	en…	: and then I would let
1) FIRS…	them for is not knowing	en…	. I would whip them hard

Figure 266 Contexts, First Year at Harrow

Text analytics is also applied for defining and solving the problem. The most occurring word, “English” has two senses: one is English as a language, and the second is the nationality of a person who lives in England. To solve this apparent WSD, Contexts tool is employed for word sense disambiguation, and this tool is also based on the theoretical postulate which informs that “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). Figure 266 reveals knowledge patterns that the word “English” is used nine times, and every time it refers to English as a subject.

12. *Hitch--- Hiking across the Sahara* by G. F. Lamb

i. Summary

This corpus has 1 document with 3,978 total words and 1,150 unique word forms. Created about 23 minutes ago (on 17th December 2017).

Vocabulary Density: 0.289

Average Words Per Sentence: 18.0

Most frequent words in the corpus: christopher (31); desert (26); water (23); reached (14); sand (14); truck (13); day (12); journey (12); little (12); just (9);

Figure 267 Summary, Hitch--- Hiking across the Sahara

G. F. Lamb's essay comprises a total of 3978 words, while 1150 unique words are repeated almost 3.5 times in it, and it causes ease for readers because readers come across familiar words again and again. After dividing unique words by total words, 0.289 vocabulary density is calculated, and it is suitable for intermediate level readers. On average, each sentence consists of 18 words which are also suitable for intermediate level learners. It has been placed in the category of biographical essays/ heroes according to the classification of the book. Subsequently, it deals with the life of Robert Christopher who dares to go to the Sahara Desert through hitchhiking. G. F. Lamb's most frequent words present its hero and several major happenings, for instance, "Christopher (31)", "desert (26)", "sand (14)".

ii. Cirrus

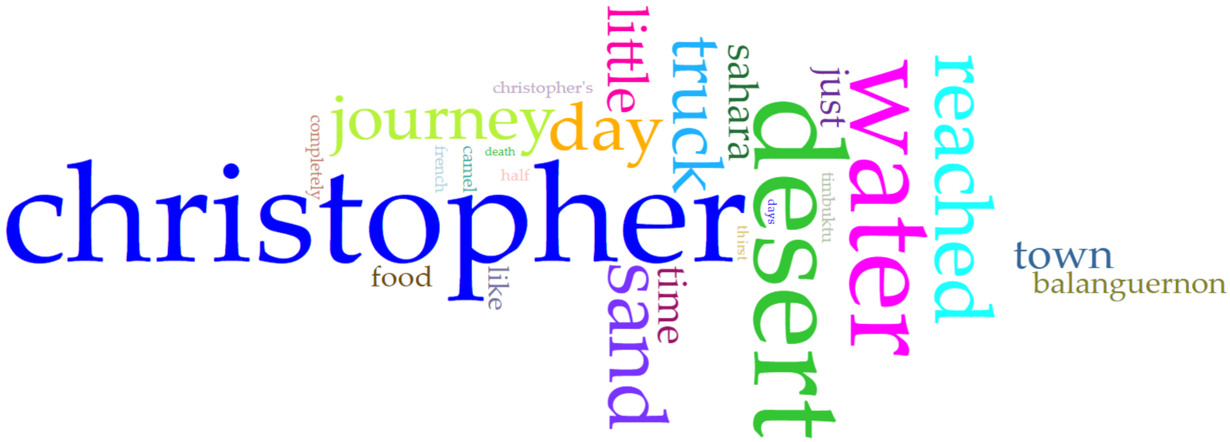


Figure 268 Cirrus, Hitch--- Hiking across the Sahara

This Cirrus shows the most occurring person as "Christopher (31)" in this travelogue. The second most frequent theme is "desert (26)" because he hitchhikes in "Sahara (9)" desert, the geographical setting of the essay. This desert abounds with "sand (14)" and sand dunes. Another theme is "water (23)" because people have to travel sometimes off the track to fetch water, or they even lose their lives in the absence of water. One incident is reported that three car travellers also passed away due to scarcity of water. The second trouble in the desert journey is a scarcity of "food (8)".

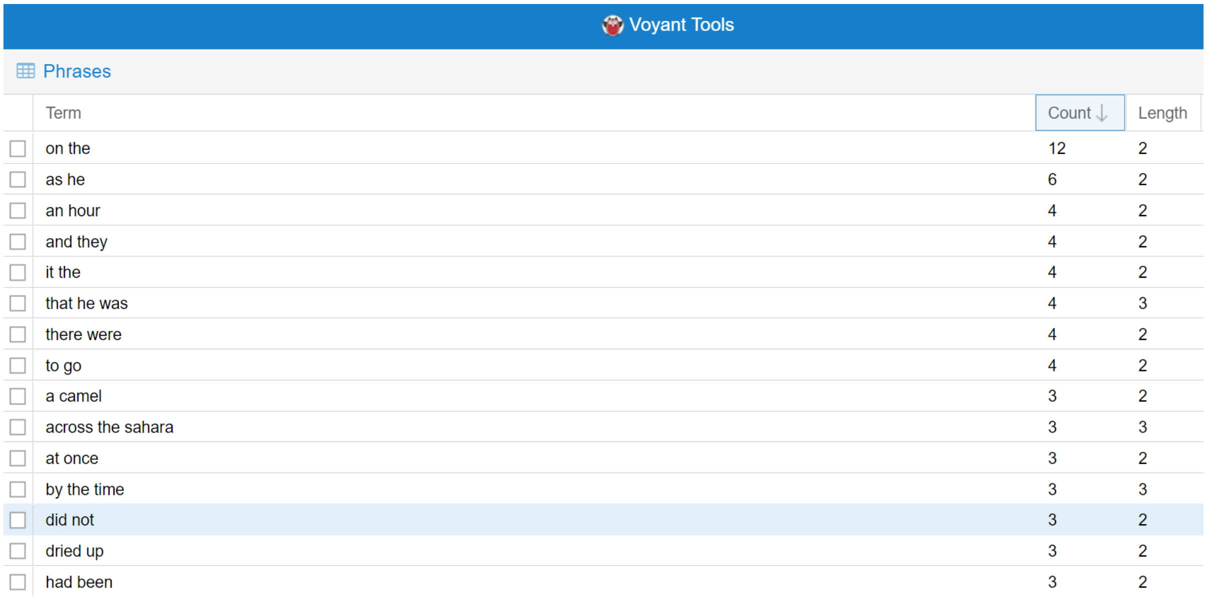
The modern mode of travelling is “truck (13)” which carries pins, daily things, ammunition. Another traditional mode of transportation is “camel (7)” which is also used by Christopher. Most of the “journey (12)” is covered by hitchhiking, while some part is covered on the back of a camel. Christopher covers six places of the Sahara Desert, for instance, Boussada, Ghardia, El Golea, In Salah, Tamanrasset and Timbuktu. The first five areas are covered by taking a lift from trucks, and the last area from Tamanrasset to Timbuktu is covered on the back of a camel.

Another theme of “French (7)” also refers to the domination of French forces on the Sahara Desert that is why French weapon carriers and military men are visible over there. Once Christopher also takes a lift from a weapon carrier by showing an expired permit to the captain.

The second most important human character and redeemer of Christopher is Prof. Claude “Balanguernon (8)” who gives him shelter in Taureg king’s camp; assists him with Taureg guide for further travelling; and eventually searches lost Christopher with desert patrol cars to save his life.

Human analysis reveals the presence of one female character in comparison to 23 male characters. The only female character is the foster mother of Robert Christopher, the adventurous hero of the essay. She threatens to send her son to Timbuktu which leads him to see the distant place (Hussain, 2009, p. 60). Cirrus tool also finds the same hero as “Christopher (31)”, but Cirrus highlights another dominant character “Balanguernon (8)” who has been ignored in human analysis. The character of the foster mother is missing in machine analysis. Furthermore, Cirrus extracts key themes of “desert (26)”, “Sahara (9)” “sand (14)”, “water (23)” and “food (8)”. All of them can be validated through textual evidence.

iii. Phrases



	Term	Count ↓	Length
□	on the	12	2
□	as he	6	2
□	an hour	4	2
□	and they	4	2
□	it the	4	2
□	that he was	4	3
□	there were	4	2
□	to go	4	2
□	a camel	3	2
□	across the sahara	3	3
□	at once	3	2
□	by the time	3	3
□	did not	3	2
□	dried up	3	2
□	had been	3	2

Figure 269 Phrases, Hitch--- Hiking across the Sahara

Among these phrases, the standard collocation pattern/ n-gram is “an hour” (Art+N), while phrases one hour or an hour are called wrong. Moreover, the presence of the article “an” reveals that the “h” letter is silent in the word “hour”. Another phrase, “to go” (Inf V), indicates the correct use of English that after “to”, only 1st form of the verb should be used. Other collocations/ n-grams are “across the Sahara” (Prep+Art+N), “at once” (Prep+Adv), “by the time” (Prep+Art+N) and “dried up” (V+Adv). They occur 3 to 12 times, and their length ranges from 2 to 3 words.

iv. Links

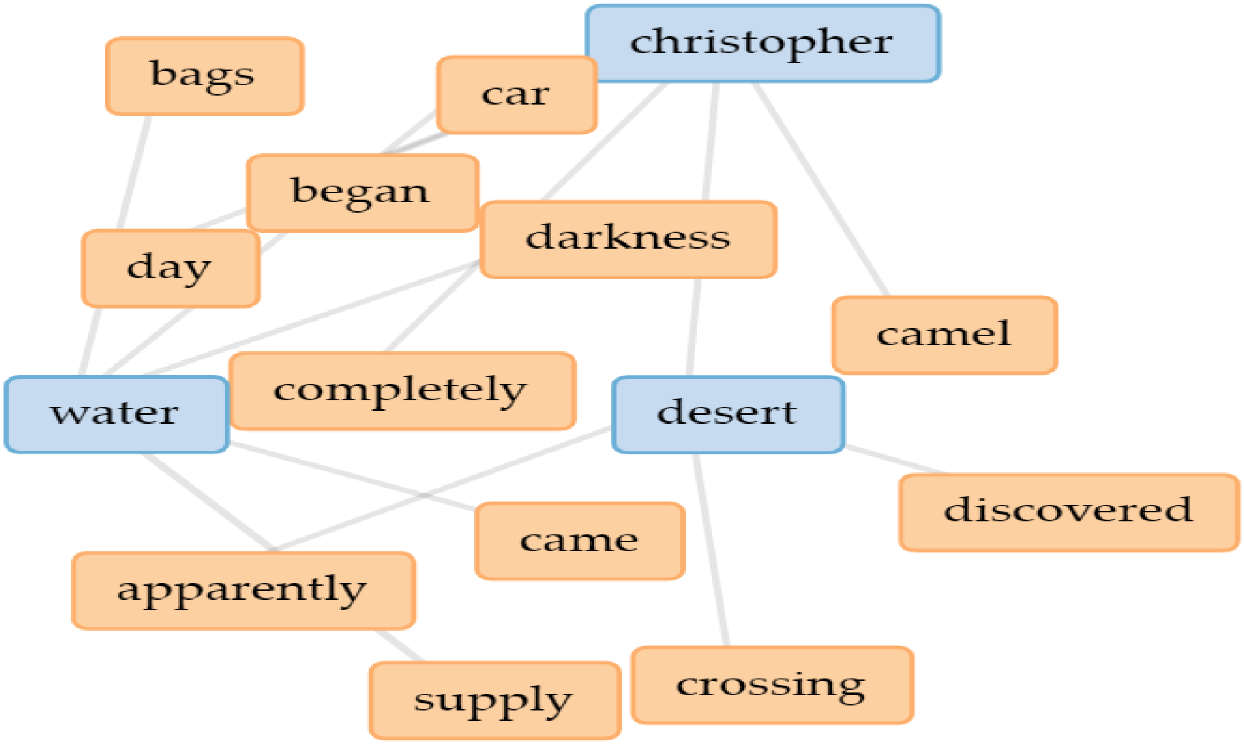
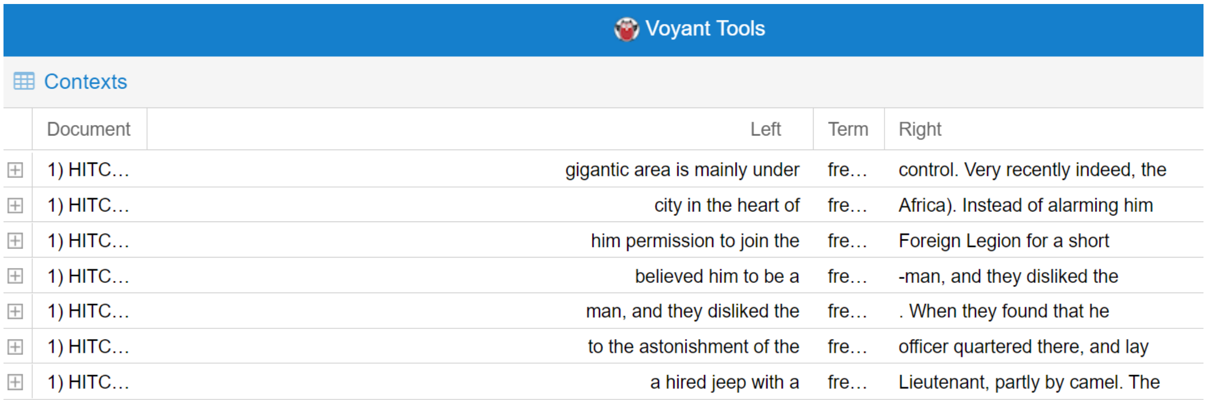


Figure 270 Links, Hitch--- Hiking across the Sahara

The blue coloured KG of "water, Christopher, desert" shows that Christopher faces scarcity of water in the desert, he detours too, and his probable death at In Abbangarit is also imminent due to the shortage of water. He takes water out of the well by making a cord of his recording tape. Another KG of "began, Christopher" points out by the beginning of his journey from one place to another. Another KG of "bag, water, supply" reveals that he loads his camel with the supply of water bags on the back of the camel, and he travels on the camel during his journey from Tamanrasset to Timbuktu. One more KG of "crossing, desert" exposes another story which is narrated by Hantout, a truck driver. He informs that three persons try to cross the Sahara Desert in a car, but their car sticks in the dunes until their dead bodies are found like dried leaves.

There is one knowledge discovery that the word "discovered" has not been connected with anyone because two times it is mentioned with Christopher, and the second time, it is linked with the pronoun "he". It has been used three times with other words; that is why it has not been connected specifically with any word.

v. Contexts



	Document	Left	Term	Right
⊞	1) HITC…	gigantic area is mainly under	fre…	control. Very recently indeed, the
⊞	1) HITC…	city in the heart of	fre…	Africa). Instead of alarming him
⊞	1) HITC…	him permission to join the	fre…	Foreign Legion for a short
⊞	1) HITC…	believed him to be a	fre…	-man, and they disliked the
⊞	1) HITC…	man, and they disliked the	fre…	. When they found that he
⊞	1) HITC…	to the astonishment of the	fre…	officer quartered there, and lay
⊞	1) HITC…	a hired jeep with a	fre…	Lieutenant, partly by camel. The

Figure 271 Contexts, Hitch--- Hiking across the Sahara

Hermeneutica Theory is “embedded in a context” (Rockwell, & Sinclair, 2016, p. 166), so context facilitates readers to determine the true meaning of any word. In this essay, the word “French” is used eight times. There is an ambiguity about its semantic shade whether it refers to the nationality or French language. To resolve this ambiguity, figure 271 reveals a knowledge pattern that the word French is referred to nationality eight times, and it did not discuss the French language.

13. *Sir Alexander Fleming* by Patrick Pringle

i. Summary

This corpus has 1 document with 2,776 total words and 846 unique word forms. Created 23 seconds ago (on 17th December 2017).

Vocabulary Density: 0.305

Average Words Per Sentence: 17.1

Most frequent words in the corpus: fleming (32); germs (27); penicillin (22); discovered (15); medical (13); antiseptic (11); disease (11); leucocytes (10); like (10); problem (10);

Figure 272 Summary, Sir Alexander Fleming

In this biographical essay, Patrick Pringle writes 846 unique words used almost more than three times; hence, they count 2776 total words. Its vocabulary density is 0.305. This lower and repeated vocabulary makes the text easier and familiar to readers. In addition, Patrick Pringle constructs 17.1 words in a sentence. The most frequent word “Fleming (32)” highlights that the entire biographical essay covers his life and his medical achievements. He “discovered (15)” “penicillin (22)”, a natural “antiseptic (11)”, which exterminates harmful germs and protects “leucocytes (10)” naturally.

Furthermore, data compression, quantification and linguistic elements are key features of Information Theory (Shannon, 2009). Stylometric analysis has been performed with the quantification of linguistic data. This precise analysis compresses big data information in a few lines for the comprehension of the data.

ii. Cirrus

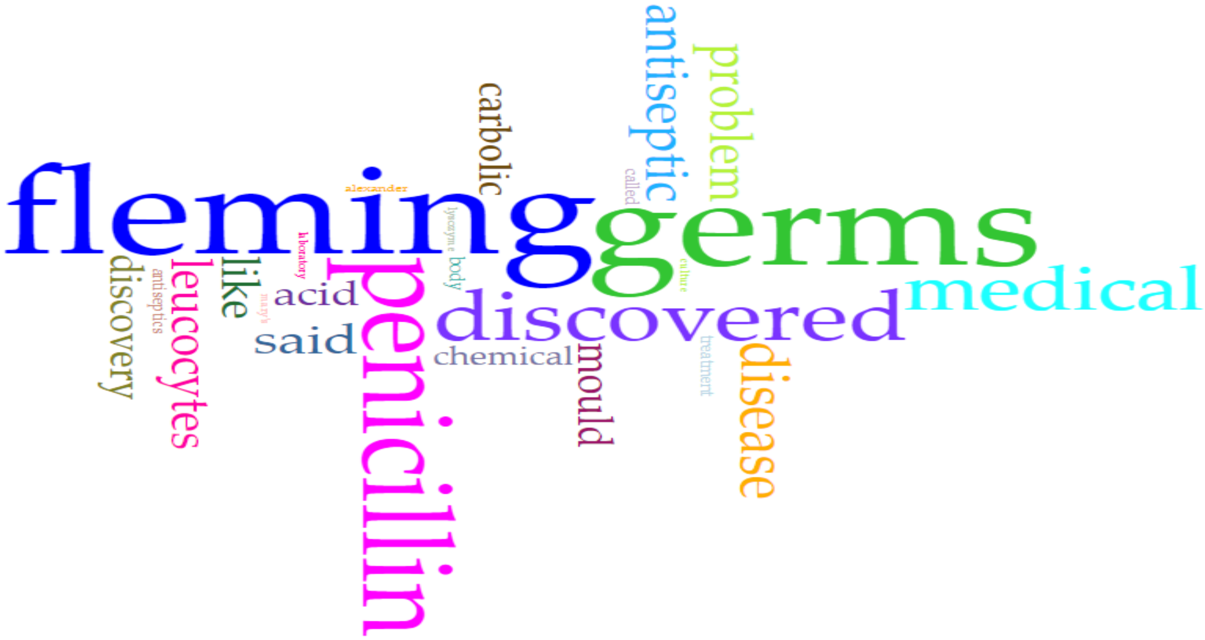
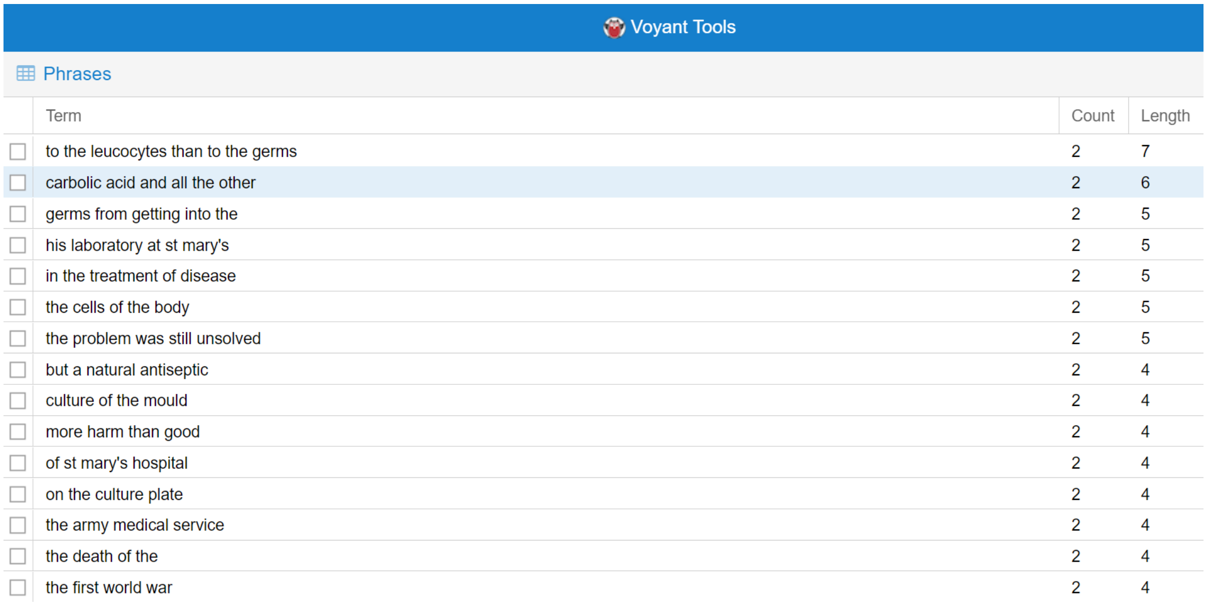


Figure 273 Cirrus, Sir Alexander Fleming

“In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct” (Agrawal, & Psaila, 1995, p. 1). This is a biographical essay, and its central character is “Fleming (32)” who is the most occurring name in this Cirrus. His first discovery is about “lysozyme (6)”, which is a natural antiseptic obtained from nasal secretions. His most famous discovery is “Penicillin (22)”, which kills “germs (27)” naturally, and such material is named “antiseptic (11)”. This discovery is caused by the accidental fall of “mould (9)” from some kitchen to his mini “laboratory (6)”. This is a discovery, so the related words “discovered” and “discovery” occur 15 and 9 times respectively. Besides, penicillin is an “antiseptic (11)” which saves “leucocytes (10)”, and kills harmful germs. It is a radical success in the "medical (13)" field to kill germs without harming lecuocytes. Another theme is the use of "carbolic acid (9)" to kill germs, but soon it is abandoned because it creates more harmful effects than good results. The theme of "said (10)" refers to the talking and exchange of ideas by Fleming and other characters. It also gives evidence of dialogic quality in this biographical essay. Thus, the clustering of Fleming's achievements has been shown in Cirrus 272 precisely.

Human analysis shows that this essay revolves around one male character, namely Fleming and three marginalized women who have been referred without assigning any activity. One key theme is forgetfulness (Hussain, 2009, pp. 60-61). On the other hand, Cirrus tool also finds the same heroic character in the domain of science. Cirrus finds themes of lysozyme (6)", "Penicillin (22)", "germs (27)", "antiseptic (11)" and "mould (9)". In short, human analysis has not discovered these themes which have been extracted by Cirrus.

iii. Phrases



	Term	Count	Length
□	to the leucocytes than to the germs	2	7
□	carbolic acid and all the other	2	6
□	germs from getting into the	2	5
□	his laboratory at st mary's	2	5
□	in the treatment of disease	2	5
□	the cells of the body	2	5
□	the problem was still unsolved	2	5
□	but a natural antiseptic	2	4
□	culture of the mould	2	4
□	more harm than good	2	4
□	of st mary's hospital	2	4
□	on the culture plate	2	4
□	the army medical service	2	4
□	the death of the	2	4
□	the first world war	2	4

Figure 274 Phrases, Sir Alexander Fleming

"Carbolic acid" (Adj+N) is a bigram collocation pattern. Another collocation pattern/ n-gram "getting into" (V+Prep) with the meaning of entering is used. The phrase "at St Mary's" (Prep+Adj+N+Apo) is found because the preposition of place "at" is used before a place name. The phrase "in the treatment of disease" (Prep+Art+N+Prep+N) always links "treatment" with "disease". The phrase "cells of the body" (N+Prep+Art+N) is a quadgram. Two collocation patterns/ n-grams, for instance, "culture of the mould" (N+Prep+Art+N) and "on the culture plate"

(Prep+Art+Adj+N), have a semantic resemblance. These collocations/ n-grams occur two to three times, and their length covers 4 to 7 words.

iv. Links

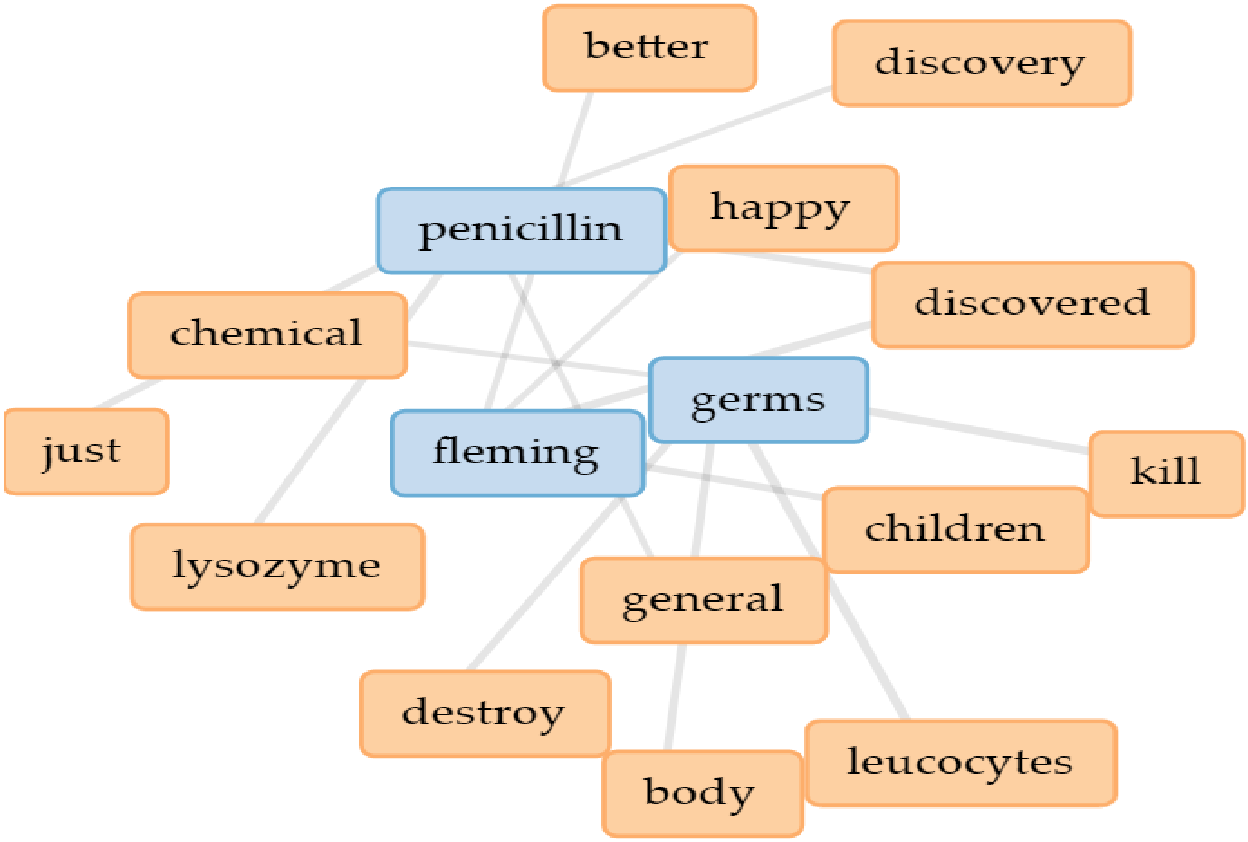
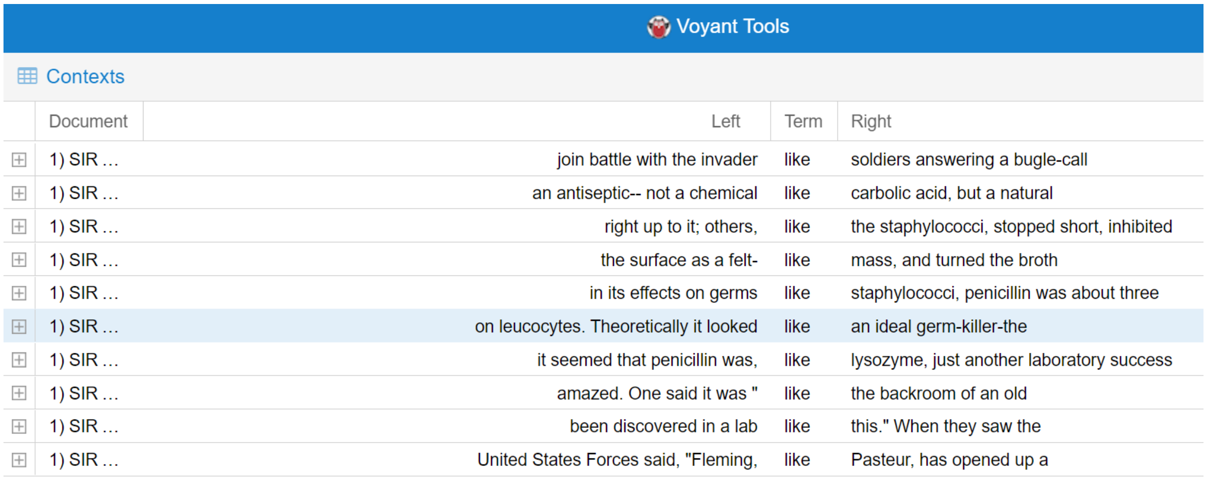


Figure 275 Links, Sir Alexander Fleming

The KG of "discovery, penicillin, lysozyme" refers to the discovery of Lysozyme by Fleming. Another KG explains the function of lysozyme, for instance, "Lysozyme, chemical, germs, kill", and it is the first natural germ killer. The second success is mentioned in the KG of "discovery, penicillin, discovered" because it is found by fortune. One more KG of "chemical, germs, kill" reveals the information that chemicals like carbolic acid kill germs along with the patient. Some other chemical materials cause more harm to the body than good. Another significant incident is exhibited by the KG of "children, Fleming" when a man brings his three children in front of Fleming, and he urges his children to pray for Fleming forever because the lives of his three children are saved because of Fleming's discovered penicillin. To conclude, relationship mining among relevant variables (Barahate, 2012, p. 13) has been done through KGs.

v. Contexts



	Document	Left	Term	Right
⊞	1) SIR ...	join battle with the invader	like	soldiers answering a bugle-call
⊞	1) SIR ...	an antiseptic-- not a chemical	like	carbolic acid, but a natural
⊞	1) SIR ...	right up to it; others,	like	the staphylococci, stopped short, inhibited
⊞	1) SIR ...	the surface as a felt-	like	mass, and turned the broth
⊞	1) SIR ...	in its effects on germs	like	staphylococci, penicillin was about three
⊞	1) SIR ...	on leucocytes. Theoretically it looked	like	an ideal germ-killer-the
⊞	1) SIR ...	it seemed that penicillin was,	like	lysozyme, just another laboratory success
⊞	1) SIR ...	amazed. One said it was "	like	the backroom of an old
⊞	1) SIR ...	been discovered in a lab	like	this." When they saw the
⊞	1) SIR ...	United States Forces said, "Fleming,	like	Pasteur, has opened up a

Figure 276 Contexts, Sir Alexander Fleming

The word “like” can be used as a simile or for extreme adoration. To determine its most appropriate meaning, the study of context is essential because “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166). On the premise of the theory, Contexts tool has been used for word sense disambiguation. So, figure 276 reveals that the word “like” is used 10 times as a simile, however, the word “like” as an action verb has not been used in this essay.

14. *Louis Pasteur* by Margaret Avery

i. Summary

This corpus has 1 document with 3,515 total words and 1,156 unique word forms. Created about 6 minutes ago (on 27th December 2017).

Vocabulary Density: 0.329

Average Words Per Sentence: 30.8

Most frequent words in the corpus: pasteur (36); disease (24); germs (18); work (16); france (14); pasteur's (10); treatment (8); great (7); man (7); war (7);

Figure 277 Summary, Louis Pasteur

“Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). The text has been changed into a quantified and compact description through Summary tool; hence, quantification is a means of knowledge construction. Margaret Avery wrote 1156 unique words which had been repeated almost three times until total words reach the number of 3515. When a reader reads the same word a third time, he/she feels at ease, and his/her reading speed also increases. When unique words are divided by total words, 0.329 vocabulary density is calculated, recommended for intermediate-level readers. Another information pattern is revealed that this biographical essay consists of the longest sentences, and almost 30 words are written in one sentence by Margaret Avery. Sentence length is appropriate for highly advanced level readers. When the most occurring words are analysed, it is found that "Pasteur's (36)" name is centralized because the whole essay covers his lifetime scientific achievements.

ii. Cirrus

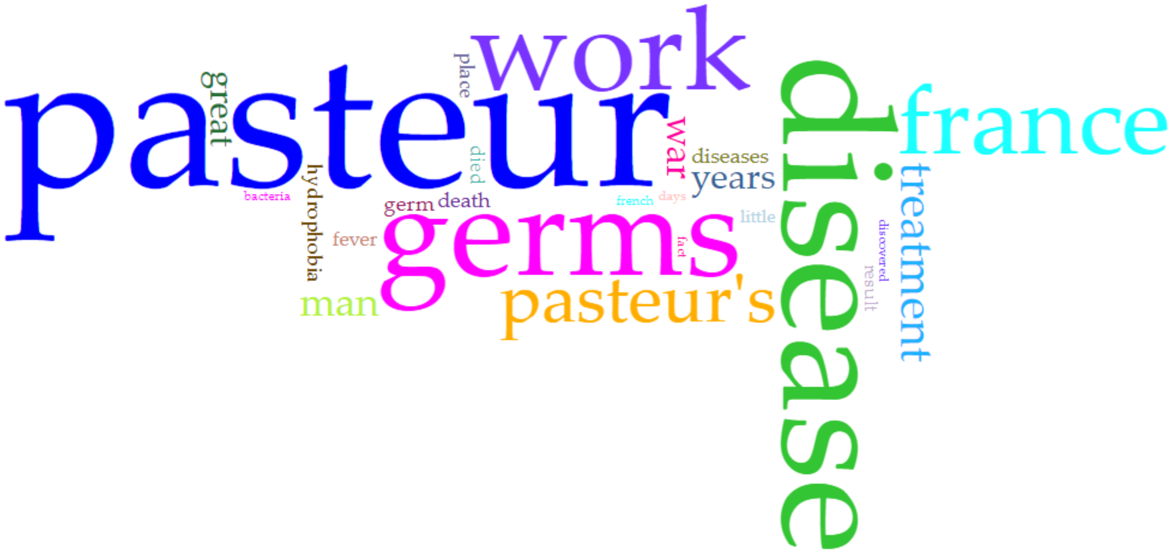


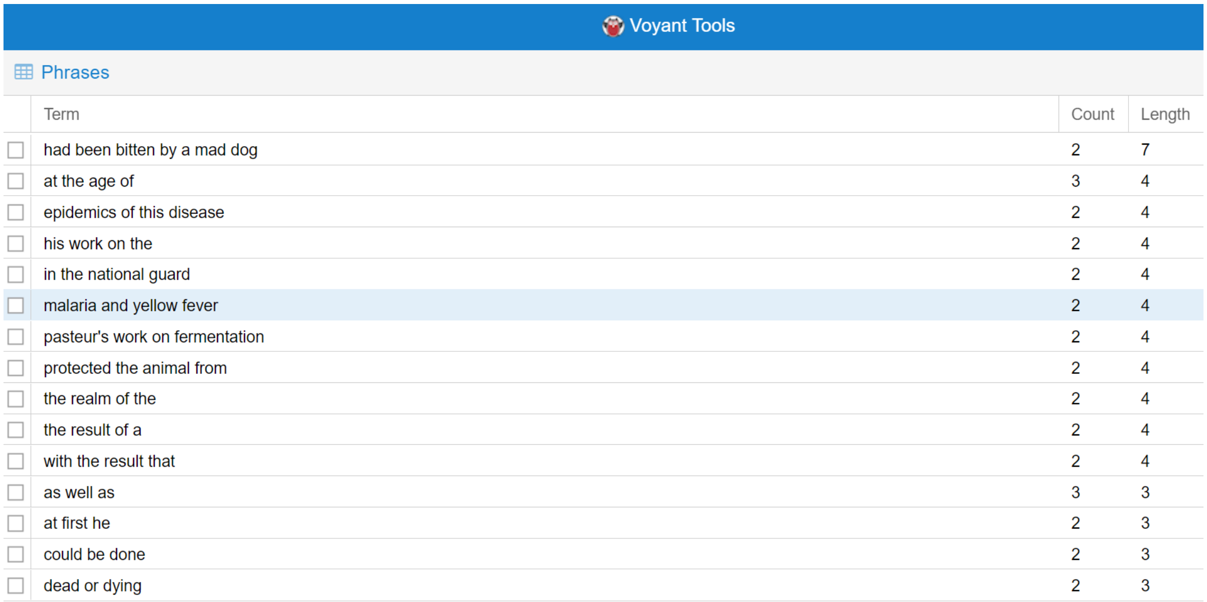
Figure 278 Cirrus, Louis Pasteur

This is a biographical essay which concentrates on Louis Pasteur's scientific achievements in different domains of science, so the name of "Pasteur (36)" occurs most in this corpus. He cures many lethal "diseases (24)" of birds, silk worms, animals, and human beings. In fact, "germs (18)" cause diseases, so he finds those germs and certain ways of their extermination for the welfare of all living beings. He prepares vaccines and cures many "diseases (6)" including anthrax, cholera, "fever (6)" "and hydrophobia (6)", and he lessens the sufferings of diseases on a large scale.

Patriotism is a hallmark in Pasteur's life since his passionate love for his motherland, "France (14)", knows no bounds. When Germany attacks his homeland, he returns his PhD degree, and abandons his professorship in Germany with disgusting feelings. Afterwards, he goes to the soldier recruiting office to fight a man to man "war (7)" against Germany, but he is rejected on medical grounds. To materialize his aims, he takes the sword of science to defeat Germany, and he plays his role well by introducing fermentation, brewing industry, treatment of diseases and preparation of vaccines for birds, animals and human beings. These scientific works recover the health of living beings, generate a considerable amount of revenue, and win friends for France.

Human analysis reveals that this essay discusses the towering personality of Louis Pasteur, whereas feminine characters are caretakers of silkworms in their homes. Pasteur's daughter also uses a microscope (Hussain, 2009, pp. 61-62). Comparing human analysis with machine analysis, Cirrus shows "Pasteur (36)" as a central character and other minor characters. As themes are concerned, Cirrus tool generates comprehensive themes, for instance, "diseases (24)", "fever (6)", "hydrophobia (6)", whereas human analysis did not detect these themes in detail.

iii. Phrases



	Term	Count	Length
□	had been bitten by a mad dog	2	7
□	at the age of	3	4
□	epidemics of this disease	2	4
□	his work on the	2	4
□	in the national guard	2	4
□	malaria and yellow fever	2	4
□	pasteur's work on fermentation	2	4
□	protected the animal from	2	4
□	the realm of the	2	4
□	the result of a	2	4
□	with the result that	2	4
□	as well as	3	3
□	at first he	2	3
□	could be done	2	3
□	dead or dying	2	3

Figure 279 Phrases, Louis Pasteur

One theoretical postulate is "Manipulation is in service of exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166). Here meaningless collocations have been segregated from meaningful collocations with human understanding. Standard phrases, "bitten by a mad dog" (V+Prep+Art+Adj+N), "at the age of" (Prep+Art+N+Prep), "work on" (V+Prep), "as well as" (Conj), "at first" (Prep+Adv) and "disease and death" (N+Conj+N) are standard collocation patterns/ n-grams in this essay. Their presence is counted 2 to 3 times, whereas their length consists of 3 to 7 words.

iv. Links

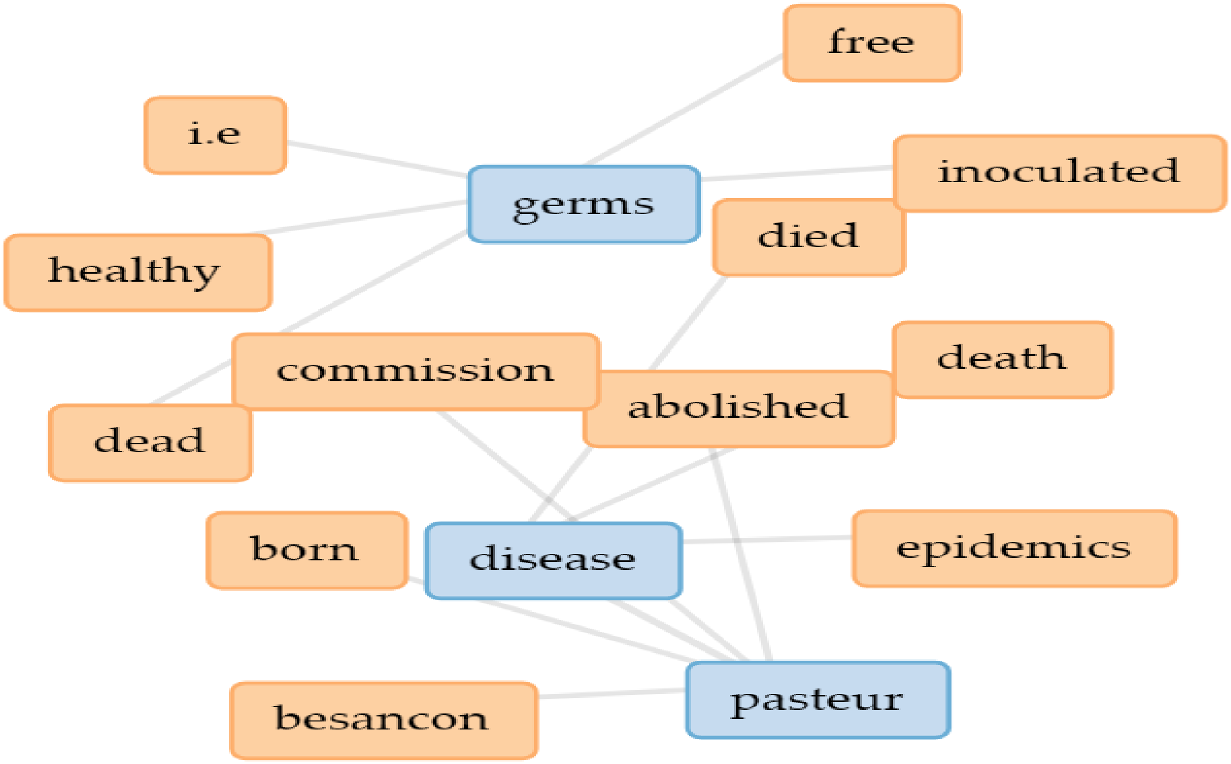
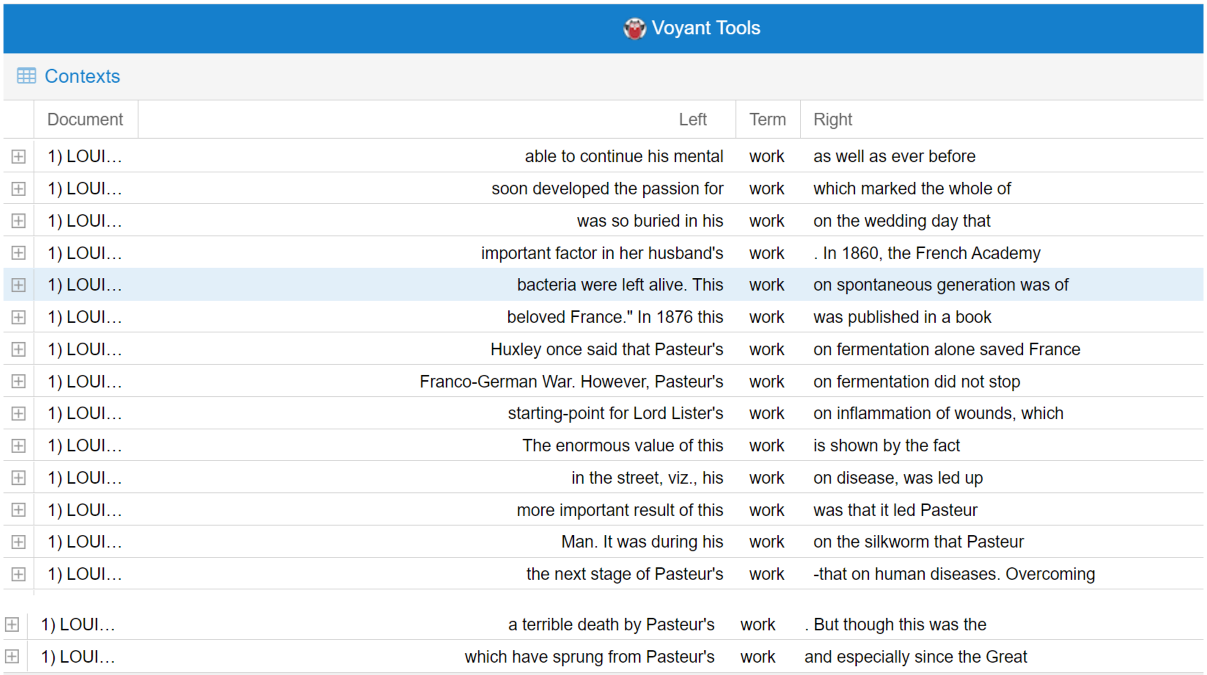


Figure 280 Links, Louis Pasteur

The blue coloured entities are "Pasteur, germs, disease" because germs cause diseases, and Pasteur treats those diseases successfully until he uproots them. Another KG of "dead, germs, inoculated" refers to the process of vaccination in which dead germs are "inoculated" to healthy persons, birds or animals, and they show immunity to the disease. This method is the premise of vaccination, and it has continued successfully by now. One more KG of "Pasteur, abolished, disease" suggests that Pasteur saves the lives of many people, birds, animals and silk worms; and his medical discoveries have abolished lethal epidemics.

One KG of "Pasteur, disease, commission" refers to his seminal work of proving that only living beings give birth to living beings, whereas non-living things cannot put forth living beings. So, the commission decides in favour of Pasteur; thus, he wins this competition and rejects the centuries-old notion. One more KG of "Besancon, Pasteur, disease" indicates that he gets admission to Besancon for better and extensive studies at graduation level, and in the same institute, he meets his future wife.

v. Contexts



Document	Left	Term	Right
1) LOUI…	able to continue his mental	work	as well as ever before
1) LOUI…	soon developed the passion for	work	which marked the whole of
1) LOUI…	was so buried in his	work	on the wedding day that
1) LOUI…	important factor in her husband's	work	. In 1860, the French Academy
1) LOUI…	bacteria were left alive. This	work	on spontaneous generation was of
1) LOUI…	beloved France." In 1876 this	work	was published in a book
1) LOUI…	Huxley once said that Pasteur's	work	on fermentation alone saved France
1) LOUI…	Franco-German War. However, Pasteur's	work	on fermentation did not stop
1) LOUI…	starting-point for Lord Lister's	work	on inflammation of wounds, which
1) LOUI…	The enormous value of this	work	is shown by the fact
1) LOUI…	in the street, viz., his	work	on disease, was led up
1) LOUI…	more important result of this	work	was that it led Pasteur
1) LOUI…	Man. It was during his	work	on the silkworm that Pasteur
1) LOUI…	the next stage of Pasteur's	work	-that on human diseases. Overcoming
1) LOUI…	a terrible death by Pasteur's	work	. But though this was the
1) LOUI…	which have sprung from Pasteur's	work	and especially since the Great

Figure 281 Contexts, Louis Pasteur

The word "work" can be used as a noun and a verb. Owing to different parts of speech, its semantic shades also change in different word senses. To probe the key word "work", Contexts tool illustrates the tabular data in figure 281. The ambiguous word "work" occurs 16 times in this biographical essay, and every time it refers to a common noun. It occurs five times after Proper Noun "Pasteur", and once after Proper noun "Lister", but it is never used as a verb in this corpus. After resolving grammatical ambiguity, semantic shade or translated meaning can be determined easily.

15. *Mustafa Kamal* by Wilfrid F. Castle

i. Summary

This corpus has 1 document with 2,933 total words and 1,021 unique word forms. Created 14 seconds ago (on 30th December 2017).

Vocabulary Density: 0.348

Average Words Per Sentence: 22.6

Most frequent words in the corpus: mustafa (28); kamal (24); turkish (23); anatolia (14); government (14); ottoman (14); istanbul (13); allies (11); greeks (11); turkey (11)

Figure 282 Summary, Mustafa Kamal

The summary tool presents a stylistic pattern of this corpus, for instance, the author's total vocabulary and unique vocabulary. Vocabulary density determines the difficulty level of the text, structure of sentences, whether they are longer or shorter. In this biographical essay, Wilfrid F. Castle uses 1021 unique words, and they are almost repeated three times with 2933 total words. Division of unique words by total words produces 0.348 vocabulary density which is suitable for intermediate level readers. The sentences of this essay are shorter than the previous essay, *'Louis Pasteur'*. Wilfrid F. Castle writes long sentences with average words of 22.6 for advanced level readers. One postulate of the theory is "It is supplemented by other materials" (Rockwell, & Sinclair, 2016, p. 166). Themes of Summary tool match themes of Cirrus, and this process establishes concurrent validity. Likewise, the previous four essays show their protagonist's name; this essay also centres upon the towering personality of "Mustafa (28)" "Kamal (24)", and this is the most occurring name in this corpus.

ii. Cirrus

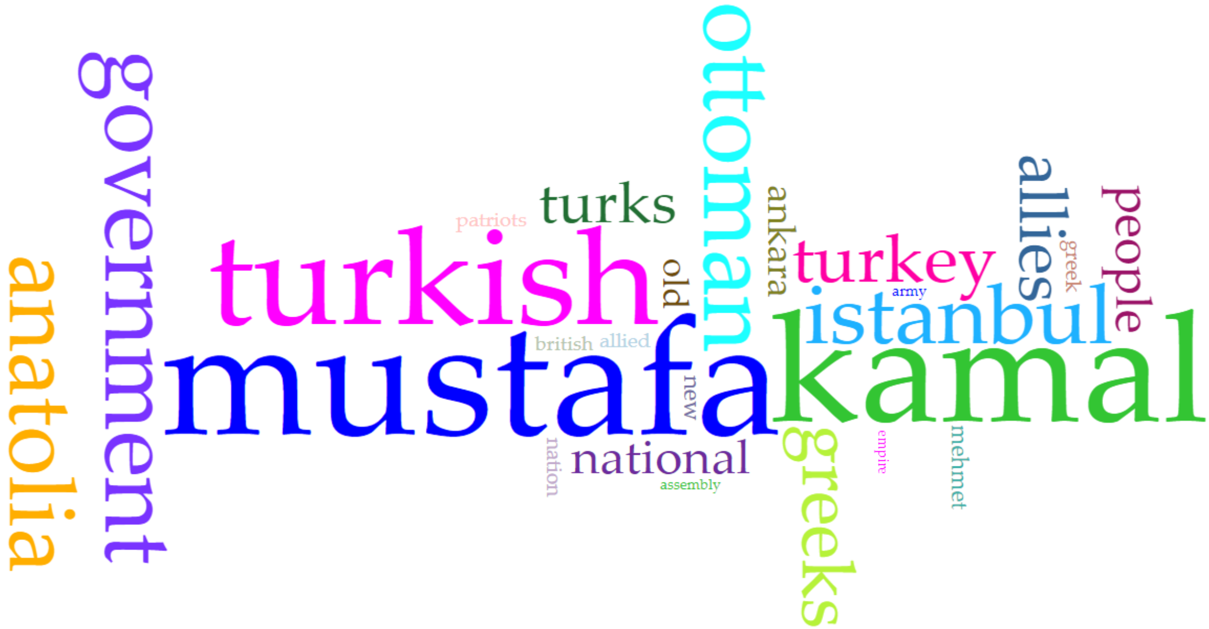


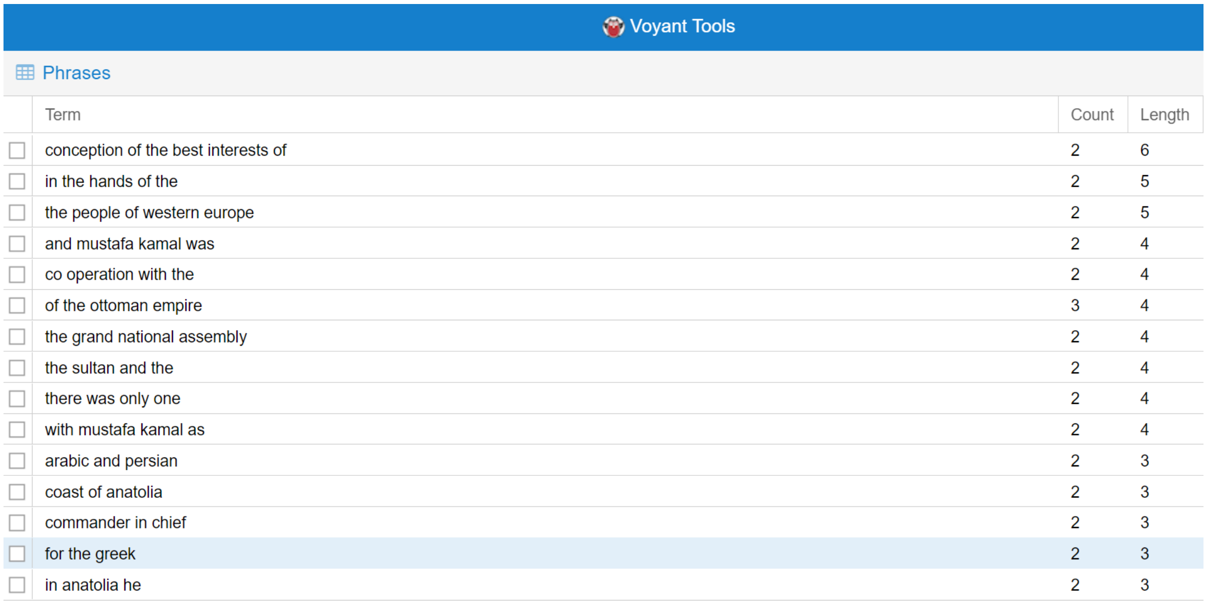
Figure 283 Cirrus, Mustafa Kamal

Knowledge Discovery Theory is defined as “the extraction of implicit, previously unknown and potentially useful information from data” (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998, p. 9). All themes of this Cirrus are potentially useful knowledge patterns. Furthermore, these themes can be used as a previewing technique for readers. This Cirrus has been generated with topic modelling and concept mining techniques. Technically, the statistical method filters topics from a large dataset showing the numeric value of each occurrence. Statistical values change words into a knowledge pattern. This is a biographical essay, and it discusses the military achievements of Mustafa Kamal, hence, in biographical essays, the most occurring character is the protagonist of the essay, “Mustafa (28)” and “Kamal (24)”, and these two names refer to one military leader.

The geographical settings of the entire essay are “Anatolia (14)” and “Turkey (11)”. Mustafa Kamal brings revolution for “Turks (10)” with the support of the valiant “Turkish (23)” nation. In this ordeal, “Anatolia (14)” is the land of resistance against allied forces, and the traitor Mehmet signs a treaty to hand over his all rich provinces to the allied forces. Furthermore, Mehmet resides in “Istanbul (13)”, the capital of the “Ottoman (14)” Empire to facilitate allies and to defeat his resisting valiant soldiers in “Anatolia (14)”. In fact, in 1920, “Turkish (23)” “national (9)” “government (14)” acted as a puppet in the hands of enemies. The “Allies (11)”, for instance, “Greeks (11)”, English, Cicily and other forces practically occupy the capital of Turkey and other areas in fulfilment of Mehmet’s signed treaty with Allied forces. According to the treaty, only the Ottoman capital remains Turkey; thus, Mehmet has stabbed his countrymen at the back by signing this derogatory and destructive treaty. Mustafa Kamal saves the Turkish nation from its opponents; simultaneously, he plays the role of a dictator, democrat, and redeemer for the Turkish nation. The current study conforms to theme extraction from presidential campaigns speeches of Obama and McCain (GitHub, 2014), as mentioned in chapter 2.

Human analysis reveals that this essay primarily concentrates on the life achievements of Mustafa Kamal, a war hero. On the periphery, a female medical practitioner has been shown (Hussain, 2009, p. 62). Cirrus tool mentions the character of “Mustafa (28)” “Kamal (24)”, but the character of lady doctor has not been mentioned. The main themes of Cirrus show different nations, for instance, “Turkish (23)”, “Greeks (11)”. Some geographical settings have also been highlighted, for example, “Istanbul (13)” and “Anatolia (14)”. To conclude, Cirrus-generated themes are more comprehensive and quantified than humanly-generated themes.

iii. Phrases



	Term	Count	Length
□	conception of the best interests of	2	6
□	in the hands of the	2	5
□	the people of western europe	2	5
□	and mustafa kamal was	2	4
□	co operation with the	2	4
□	of the ottoman empire	3	4
□	the grand national assembly	2	4
□	the sultan and the	2	4
□	there was only one	2	4
□	with mustafa kamal as	2	4
□	arabic and persian	2	3
□	coast of anatolia	2	3
□	commander in chief	2	3
□	for the greek	2	3
□	in anatolia he	2	3

Figure 284 Phrases, Mustafa Kamal

Standard collocation patterns/ n-grams of this biographical essay are: “best interest” (Adj+N), “in the hands of” (Prep+Art+N+Prep), “Western Europe” (Adj+N), “Mustafa Kamal” (N+N), “cooperation with” (N+Prep), “the Ottoman Empire” (Art+Adj+N), “national assembly” (Adj+N), “commander in chief” (N+Prep+Adj) and “the Greek” (Art+ N). They occur 2 to 3 times, and their length varies from 3 to 6 words. They also teach the correct use of prepositions and articles before the name of the dynasty.

iv. Links

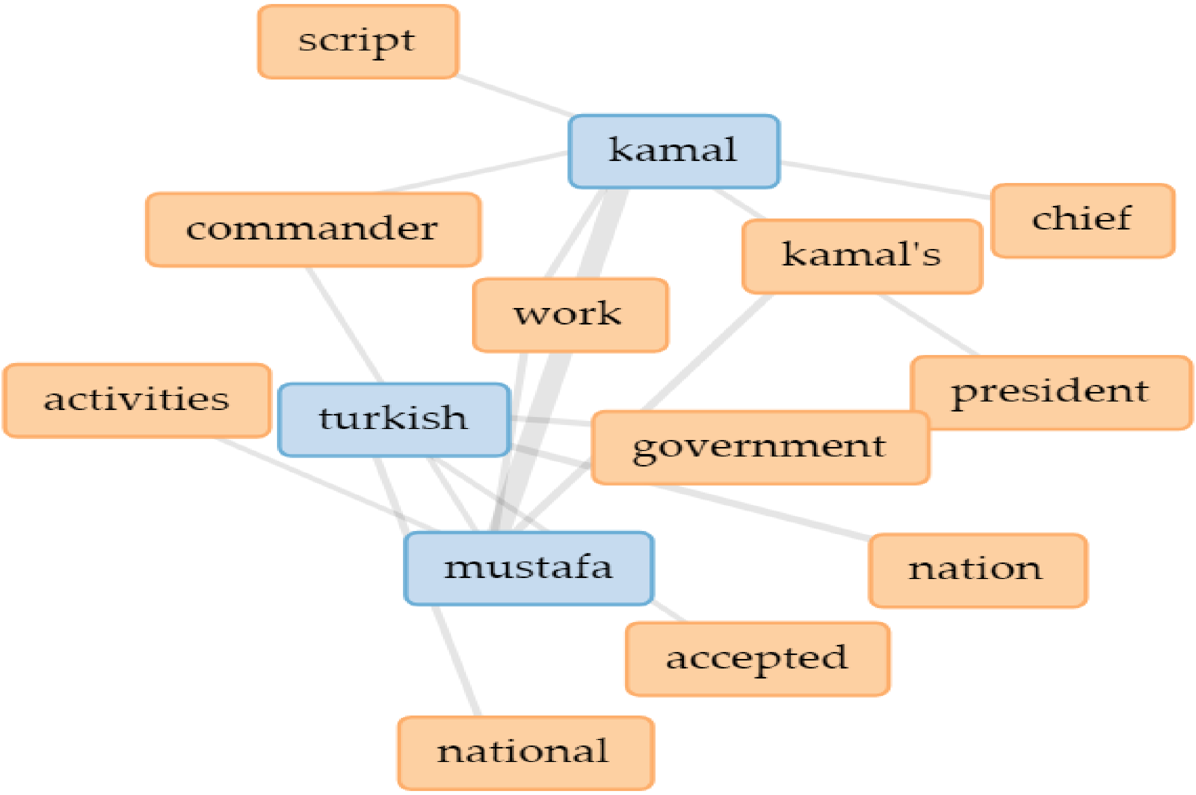
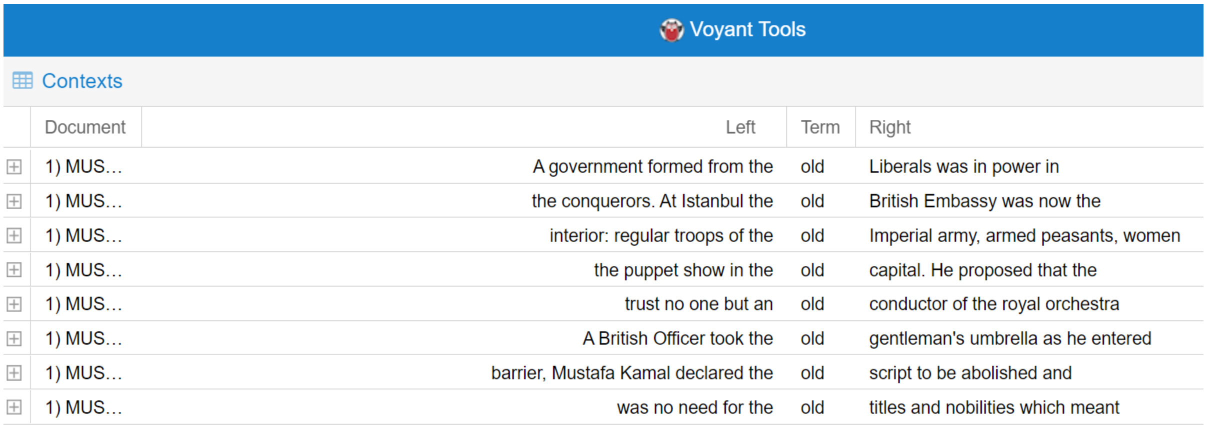


Figure 285 Links, Mustafa Kamal

The blue coloured strongly connected KG of "Mustafa, Kamal, Turkish" expresses his services for Turkish nations. If he were not there to unify his valiant nation to defeat the allied forces, there would be no Turkey now except its small capital. Another KG of "chief, Kamal, commander" refers to his services as a commander in chief of sovereign Turkey. Two KGs of "accepted, Mustafa, Turkish, commander" and "president, Kamal's, Kamal" elaborate one of the central themes that the Turkish nation accepts Mustafa Kamal as its president being a democratic as well as a dictator at the same time. Another KG of "script, Kamal, chief" refers to his efforts to change the old and difficult script of the Ottoman Empire, and he emphasises learning the new script till a certain date, consequently, an academic revolution emerges for the Turkish nation.

v. Contexts



	Document	Left	Term	Right
⊞	1) MUS…	A government formed from the	old	Liberals was in power in
⊞	1) MUS…	the conquerors. At Istanbul the	old	British Embassy was now the
⊞	1) MUS…	interior: regular troops of the	old	Imperial army, armed peasants, women
⊞	1) MUS…	the puppet show in the	old	capital. He proposed that the
⊞	1) MUS…	trust no one but an	old	conductor of the royal orchestra
⊞	1) MUS…	A British Officer took the	old	gentleman's umbrella as he entered
⊞	1) MUS…	barrier, Mustafa Kamal declared the	old	script to be abolished and
⊞	1) MUS…	was no need for the	old	titles and nobilities which meant

Figure 286 Contexts, Mustafa Kamal

The word “old” carries different word senses, for instance, long ago, outdated, and it carries different meanings with living and non-living things, for example, old car and old man, so the mere word “old” exhibits ambiguity. The same discrepancy in machine translation was discussed in chapter 2 by Hutchins (1999). One of the main objectives of text mining is to disambiguate text since knowledge should be free from confusions. So, “Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166), and contextual clues are used to disambiguate the sense of the text. Consequently, its practical expression is given in figure 286. The word ‘old” is used eight times in this corpus, and it is used for human beings only four times, for instance, “old Liberals”, “old imperial army”, “old conductor” and “old gentleman” and it is used four times for non-living things thus: “Old British Embassy”, “old capital”, “old script” and “old titles”. In a nutshell, KWIC resolves word sense ambiguity, as it has been found useful in Fischer (1971) in the 2nd chapter.

4.10 Text Mining of the Novel *Good-Bye Mr Chips* by James Hilton

The novel *‘Good-Bye Mr Chips’* (Appendix D) presents actions through Brookfield school, real-life characters, British history, world wars and a teacher’s devotion to Brookfield school.

i. Summary

This corpus has 1 document with 16,758 total words and 3,160 unique word forms. Created 28 seconds ago (on 17th August, 2017).

Vocabulary Density: 0.189

Average Words Per Sentence: 16.4

Most frequent words in the corpus: chips (155); umph (120); brookfield (86); old (68); school (60); like (49); just (40); boys (39); said (39); boy (38)

Figure 287 Summary, Good-Bye Mr Chips

Summarization shows a condensed report of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45). The novelist James Hilton employs 3160 basic words, and he uses these unique words almost five times in this novel, so the total vocabulary is 16758 words. By dividing unique words with total words, 0.189 vocabulary density is calculated, and it is appropriate for advanced level readers. On average, one sentence consists of 16.4 words which are suitable for advanced level readers. In Hermeneutica Theory, "exploration and understanding" (Rockwell, & Sinclair, 2016, p. 166) have been emphasised and Summary tool also explores corpus features for a better understanding of computational stylistics. Furthermore, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). The whole novel has been transformed into quantified data to represent stylistic features of any text. This is a very compact data summary that a reader can easily comprehend stylometric features of any writer or any text.

Some previous studies of Ch.2, Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O'Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018; also find such stylometric characteristics, and their findings match findings of this study.

ii. Cirrus

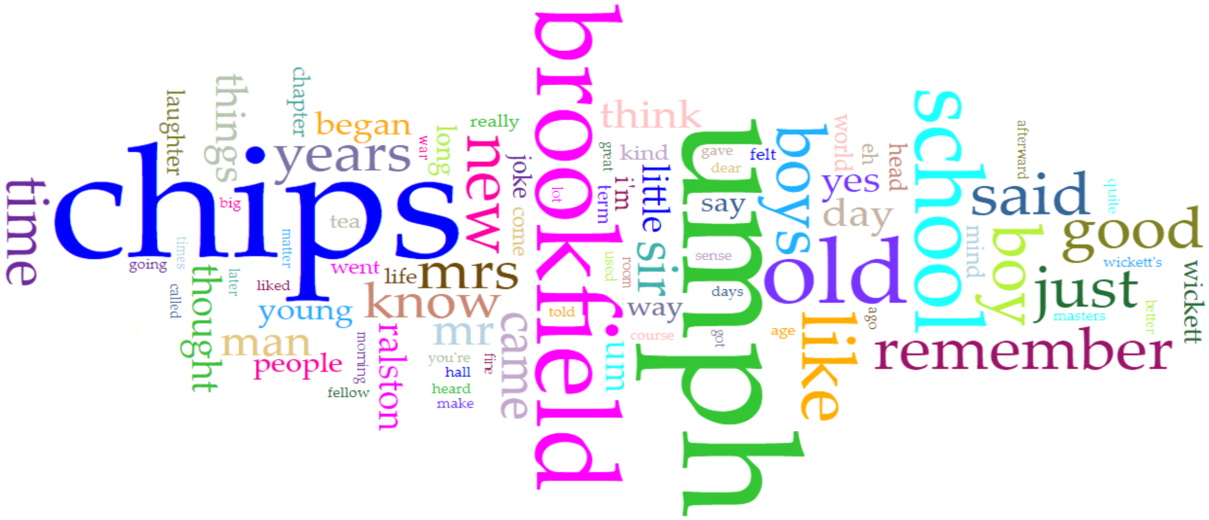


Figure 288 Cirrus, Good-Bye Mr Chips

"In active data mining paradigm,…we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct" (Agrawal, & Psaila, 1995, p. 1). Primarily, word clouds are used for topic modelling and named entity recognition from a large textual dataset. The most prominent theme of the novel is "Chips (155)" who is an endeared school teacher at Brookfield, and the whole novel focuses on his lifetime contributions to the service of students and Brookfield. Consequently, he is the protagonist of the novel, *'Good-Bye Mr Chips'*. The hedge of "Umph (120)" is the repeated habit of Mr Chips during daily conversations and lessons. This is the only most frequently repeated nonverbal expression in this corpus.

Another motivating character is "Katherine (9)" Bridges who is Chips' girlish wife. She is mentioned less because she remains alive about one year after her marriage. The pronoun "she" is used 67 times, and most of the time, it is the deixis for Catherine that Mr Chips thinks of and "remembers (36)" her on many occasions. Before every final decision, he consults the imaginary being of Katherine, her previous comments, and he obeys them unconsciously.

Another feminine character is Mrs. "Wickett (20)" in whose house Mr Chips spends his life after retirement. One reason is that her house is opposite to Brookfield, and the second reason is that she is a very caring and comforting host for Mr Chips. Cirrus also reveals the character of principal "Ralston (25)" with whom Mr Chips has a row, and the latter wins in this controversy.

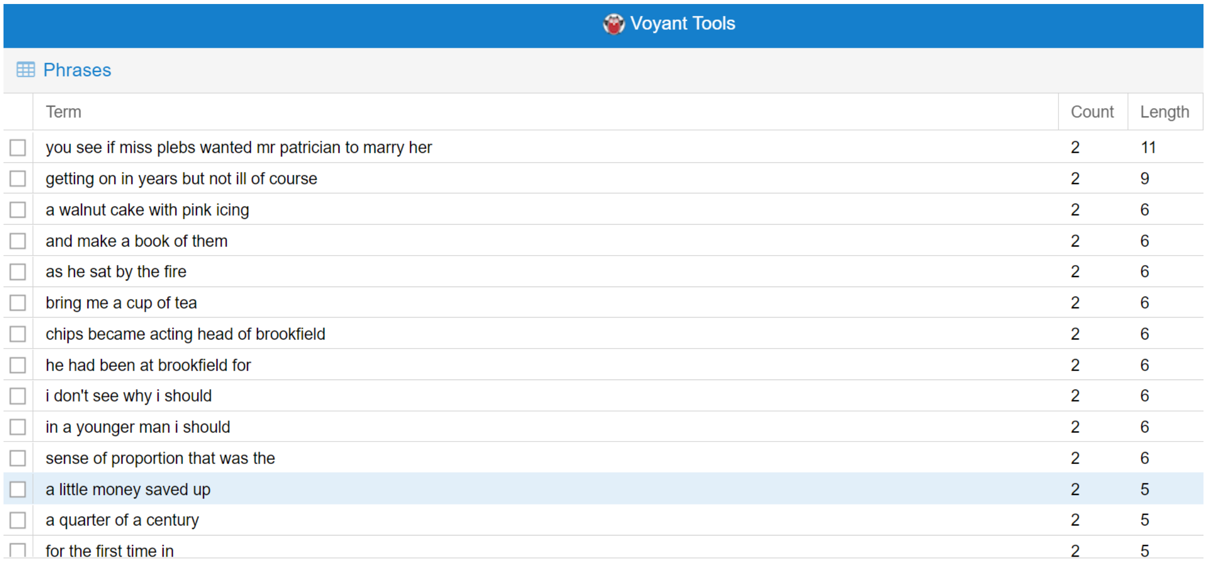
The key theme reveals the setting of the novel, "Brookfield (86)", "school (60)". Mr Chips and Brookfield school tied an unbreakable knot with each other. Defining it further, the theme of "boys (39)" is prevalent during the course of the novel. Mr Chips and boys interact with one another in their whole life, and quite interestingly, Mr Chips knows their fathers and grandfathers because they were also his students. Another very interesting knowledge discovery is that the word "girl" is missing because only boys used to study and reside in Brookfield school. Education of girls or the idea of their boarding school was not popular in Churchill's school days. So, "deviation detection" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45) has been found.

Themes of "time (36)" and "years (34)" proceed with the entire novel and the role of all characters. The cycle of years continues, and boys complete their studies; consequently, they join different professional careers. Some survive and meet him again, whereas some lose their precious lives in world wars, accidents, and the Titanic tragedy.

The word "said" occurs 39 times, and it informs about the frequent use of direct dialogues and conversations in this novel because of its conversational style. The theme of "thought (25)" informs that other characters and Mr Chips think several times on different issues. So, the knowledge discovery process indicates that it is a thought-provoking novel. Findings of the current novel are in harmony with findings of previous novels: Burrows, 2002; Hussain, 2009; Jockers, & Mimno, 2013; Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015; Scrivener, & Davis, 2017; Sinclair, & Rockwell, 2015b; Yeates, 2013; as mentioned in literature review chapter

Human analysis shows that Mr Chips is a central male character while Katherine is a feminine character busy in cycling and rock climbing. After her marriage, she advises her husband in all matters of school administration. Another feminine character is Mrs Wickett, the landlady of Mr Chips after his retirement. Some themes reveal that Mr Chips' habits are slovenly, teaching methods are slack and old fashioned (Hussain, 2009, pp. 74-80). Comparing human and Cirrus analyses, characters are the same, but their extracted themes are different; for instance, Cirrus highlights a repeated nonverbal expression as "umph", but the human analysis does not mention it. Cirrus finds deeper themes, for example, "said", "thought", "time", "years", "school" which have been ignored by human analysis. In conclusion, Cirrus analysis is deeper, quantified and more comprehensive as compared to human analysis.

iii. Phrases



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

Figure 289 Phrases, Good-Bye Mr Chips

Standard collocation patterns/ n-grams are: "to marry her" (Inf V+Prn)," getting on in years" (V+Prep+Prep+N), "walnut cake" (Adj+N), "pink icing" (Adj+N), "make a book" (V+Art+N), "sat by the fire" (V+Prep+Art+N), "bring me a cup of tea" (V+Prn+Art+N+Prep+N), "acting head" (Adj+N), "sense of proportion" (N+Prep+N), "a little money" (Adj+N), "a quarter of a century" (Art+N+Prep+Art+N), "for the first time" (Prep+Art+Adj+N) and "he lived at" (Prn+V+Prep). These collocations/ n-grams occur twice and their length ranges from 5 to 11 words.

“Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). Moreover, narratology is constructed with human reflection and repetition of certain words. As narratology is concerned, "to marry her" highlights the marriage of Chips and Katherine. This marriage changes the entire fabric of his life. Another collocation/ n-gram of "make a book" expresses the desire of Chips to write a book, but this desire was never fulfilled. The hospitality of Chips for students is evident with collocations/ n-grams of "walnut cake" with "pink icing", which he served to his students.

iv. Links

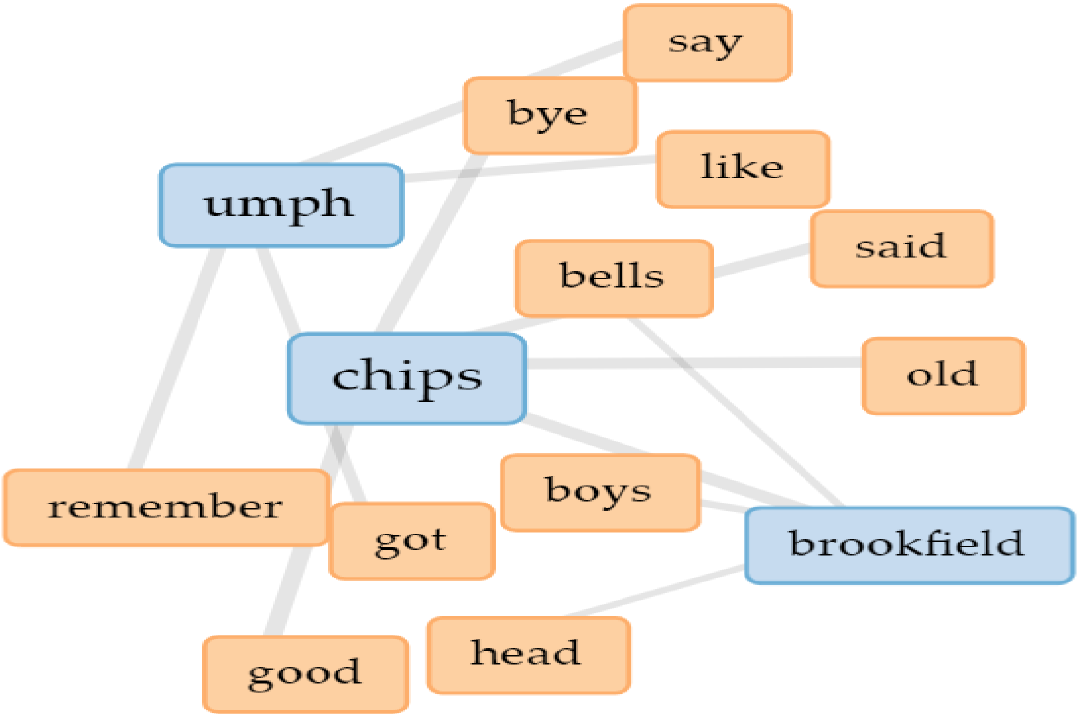


Figure 290 Links, Good-Bye Mr Chips

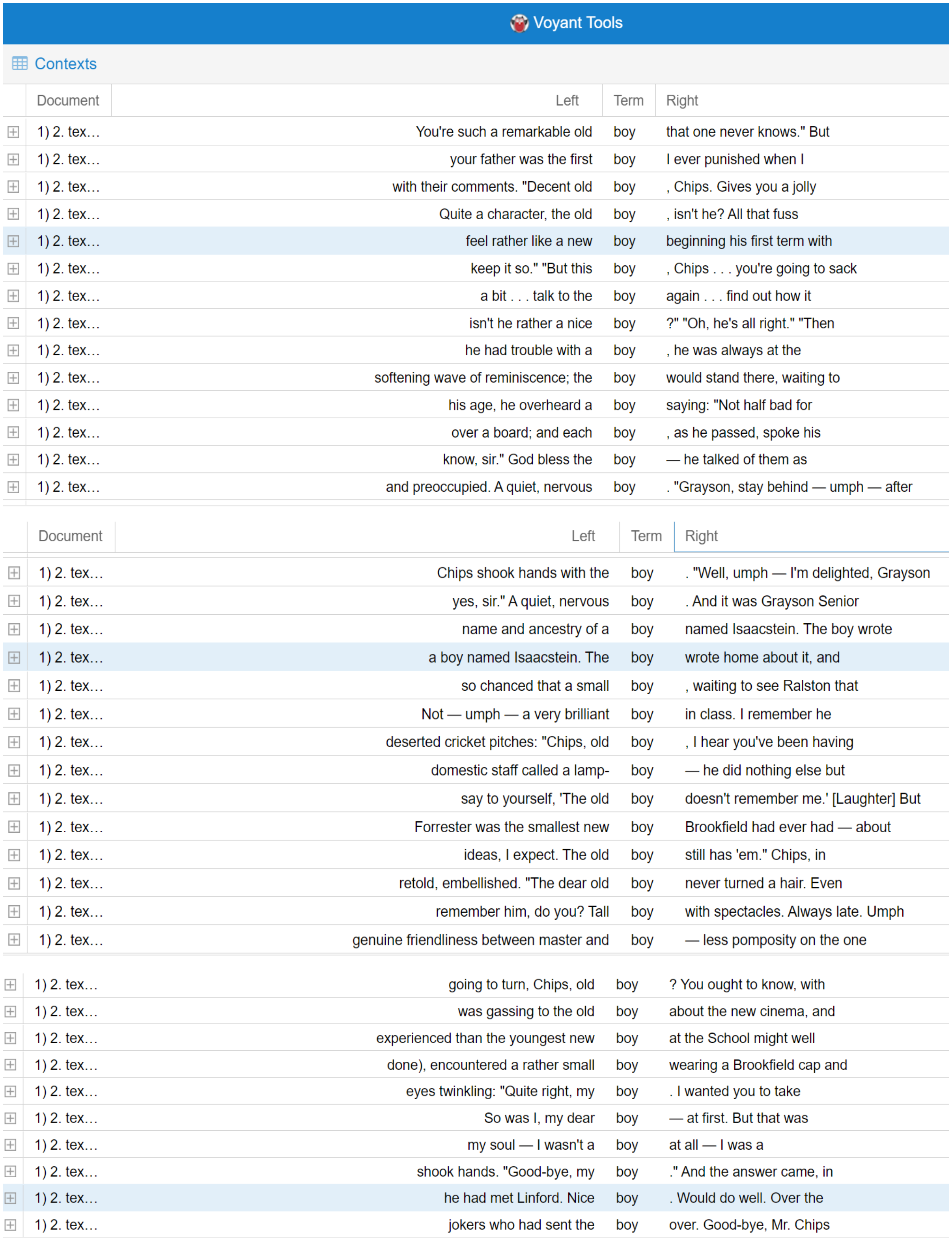
Hermeneutica Theory states that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166), and all knowledge graphs are visual representations of knowledge. Simultaneously, it is a human role to reflect and extract innovative knowledge patterns for better comprehension of the text. Reflection and multidimensional hermeneutic patterns lead to idiosyncratic knowledge patterns. These nodes suggest possible dimensions of interpretation and close association of themes. The following examples strengthen the theoretical underpinnings.

A KG presents the interrelationship of key themes to construct a comprehensive meaning. The KG of "Chips, Brookfield" sustains from beginning to the end of the novel. His whole life is knitted with the fabric of Brookfield school. Furthermore, the KG of "head, Brookfield, Chips" informs that Mr Chips serves as an acting head of Brookfield school even though he is called back to school after retirement. Realising the devotion of Chips for Brookfield, the Board of governors and other heads of Brookfield always pay honour to Chips.

The KG of "say, Umph, Chips" indicates a very interesting pattern that Mr Chips habitually utters this nonverbal expression as a habit. So, this nonverbal expression becomes a hedge in his conversations. Another relevant and interesting KG, "umph, say, remember" reveals that boys also imitate and make fun of his nonverbal expression “umph”.

Another emotional KG of "boys, Chips, remember" refers to his frequent thinking about his wife Catherine Bridges who passed away on 1st April; thereupon, fate fooled him tragically. Her memories hoover and influence his every decision. Extending the same KG, "good, remember, Chips, bye", and they bring to light two memories: Once Catherine Bridges says to him *"Good-Bye Mr Chips"* before her marriage, and the same phrase is uttered by a small boy, Linford, and this phrase takes him back to her sweet and sad memories. The KG of "said, bells, Chips" informs about dialogic style because several dialogues have been used in this novel. To conclude, relationship mining among relevant variables (Barahate, 2012, p. 13) has been done.

v. Contexts

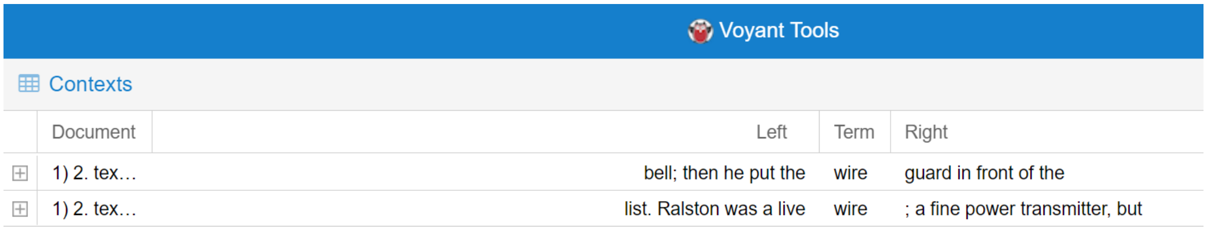


Document	Left	Term	Right
1) 2. tex…	You're such a remarkable old	boy	that one never knows." But
1) 2. tex…	your father was the first	boy	I ever punished when I
1) 2. tex…	with their comments. "Decent old	boy	, Chips. Gives you a jolly
1) 2. tex…	Quite a character, the old	boy	, isn't he? All that fuss
1) 2. tex…	feel rather like a new	boy	beginning his first term with
1) 2. tex…	keep it so." "But this	boy	, Chips . . . you're going to sack
1) 2. tex…	a bit . . . talk to the	boy	again . . . find out how it
1) 2. tex…	isn't he rather a nice	boy	?" "Oh, he's all right." "Then
1) 2. tex…	he had trouble with a	boy	, he was always at the
1) 2. tex…	softening wave of reminiscence; the	boy	would stand there, waiting to
1) 2. tex…	his age, he overheard a	boy	saying: "Not half bad for
1) 2. tex…	over a board; and each	boy	, as he passed, spoke his
1) 2. tex…	know, sir." God bless the	boy	— he talked of them as
1) 2. tex…	and preoccupied. A quiet, nervous	boy	. "Grayson, stay behind — umph — after

Document	Left	Term	Right
1) 2. tex…	Chips shook hands with the	boy	. "Well, umph — I'm delighted, Grayson
1) 2. tex…	yes, sir." A quiet, nervous	boy	. And it was Grayson Senior
1) 2. tex…	name and ancestry of a	boy	named Isaacstein. The boy wrote
1) 2. tex…	a boy named Isaacstein. The	boy	wrote home about it, and
1) 2. tex…	so chanced that a small	boy	, waiting to see Ralston that
1) 2. tex…	Not — umph — a very brilliant	boy	in class. I remember he
1) 2. tex…	deserted cricket pitches: "Chips, old	boy	, I hear you've been having
1) 2. tex…	domestic staff called a lamp-	boy	— he did nothing else but
1) 2. tex…	say to yourself, 'The old	boy	doesn't remember me.' [Laughter] But
1) 2. tex…	Forrester was the smallest new	boy	Brookfield had ever had — about
1) 2. tex…	ideas, I expect. The old	boy	still has 'em." Chips, in
1) 2. tex…	retold, embellished. "The dear old	boy	never turned a hair. Even
1) 2. tex…	remember him, do you? Tall	boy	with spectacles. Always late. Umph
1) 2. tex…	genuine friendliness between master and	boy	— less pomposity on the one
1) 2. tex…	going to turn, Chips, old	boy	? You ought to know, with
1) 2. tex…	was gassing to the old	boy	about the new cinema, and
1) 2. tex…	experienced than the youngest new	boy	at the School might well
1) 2. tex…	done), encountered a rather small	boy	wearing a Brookfield cap and
1) 2. tex…	eyes twinkling: "Quite right, my	boy	. I wanted you to take
1) 2. tex…	So was I, my dear	boy	— at first. But that was
1) 2. tex…	my soul — I wasn't a	boy	at all — I was a
1) 2. tex…	shook hands. "Good-bye, my	boy	." And the answer came, in
1) 2. tex…	he had met Linford. Nice	boy	. Would do well. Over the
1) 2. tex…	jokers who had sent the	boy	over. Good-bye, Mr. Chips

Figure 291 Contexts, Good-Bye Mr Chips

There is an ambiguity with the word “boy” because sometimes it is associated with school boys of Brookfield and sometimes, it refers to Mr Chips, an old school teacher. To disambiguate this semantic and contextual issue, Contexts tool shows that the word “boy” in figure 291 and the word “boy” is used 38 times in this corpus, it refers to young school boys 29 times and it refers to “old boy” or Mr Chips 9 times. Moreover, the word “Chips “ is added before “old boy” twice, and “Chips” is used after the phrase “old boy” once to clarify the referent.



	Document	Left	Term	Right
⊞	1) 2. tex…	bell; then he put the	wire	guard in front of the
⊞	1) 2. tex…	list. Ralston was a live	wire	; a fine power transmitter, but

Figure 292 Contexts, Good-Bye Mr Chips

Another semantic ambiguity of the word “wire” is shown in figure 292 that word "wire" has denotative or connotative meanings. The word “wire” is used twice: The first time, it gives the denotative meaning of an iron wire, and the second time, it means an emotional, dynamic, progressive and aggressive person like Ralston who has a row with Chips.

4.11 Critical Discussion

This segment has been further discussed in five categories: Cirrus, Phrases, Links, Summary, Contexts.

4.11.1 Summary

Stylometry or computational linguistics presents quantified information to exhibit stylistic qualities of any writer or any piece of writing. Summarization shows a condensed report of a subset of mined data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, November, p. 45). Likewise, KDD also mentions that “Summarization involves methods for finding a compact description for a subset of data” (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996). The data generated by Summary tool are reported in table 5. The Summary tool shows total words, unique words, vocabulary density, average words per sentence, the most occurring words and their statistical weight because statistics determines stylistic features (Amancio, 2015), as mentioned in Ch.2.

Table 5 Stylometric Features of Intermediate Books

Serial. No	Short Stories. Book I	Total Words	Unique Words	Vocab Density	Length of Sent
1	Button, Button	2152	660	0.307	9.9
2	Clearing in The Sky	2228	599	0.269	12.8
3	Dark They Were, And Golden Eyed	1858	672	0.362	7.5
4	Thank You, M’am	1361	426	0.313	12.5
5	The Piece of a String	1007	413	0.410	12.9
6	The Reward	1255	446	0.355	17.4
7	The Use of Force	1268	467	0.368	12.2
8	The Gulistan of Sadi	851	365	0.429	15.5
9	The Foolish Quack	833	316	0.379	16.3
10	A Mild Attack of Locusts	990	407	0.411	12.9
11	I Have a Dream	771	255	0.331	18.4
12	The Gift of the Magi	1544	474	0.307	10.0
13	God Be Praised	3181	1017	0.320	12.1
14	The Overcoat	1849	679	0.367	14.0
15	The Angel and the Writer and Others	952	390	0.410	14.6

Serial. No	Plays. Book III	Total Words	Unique Words	Vocab Density	Length of Sent
1	Heat Lightning	3018	638	0.211	8.5
2	Visit to a Small Planet	43151	1004	0.233	7.9
3	The Oyster and the Pearl	5738	964	0.168	8.4

Serial. No.	Poems. Book III	Total Words	Unique Words	Vocab Density	Length of Sent
1	The Rain	63	45	0.714	31.5
2	Night Mail	114	85	0.746	19.0
Serial. No.	**Poems. Book III**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
3	Loveliest of Trees, The Cherry Now	80	56	0.700	26.7
4	O Where are you going?	135	80	0.593	19.3
5	In the Street of Fruit Stalls	92	64	0.696	30.7
6	Sindhi Woman	69	56	0.812	34.5
7	Times	122	44	0.361	122.0
8	Ozymandias	112	85	0.759	28.0
9	The Feed	101	55	0.545	16.8
10	The Hollow Men	87	63	0.724	43.5

11	Leisure	93	56	0.602	31.0
12	Ruba’iyat	87	70	0.805	17.4
13	A Tale of Two Cities	135	90	0.667	27.0
14	My Neighbor Friend Breathing his Last!	97	57	0.588	10.8
15	He came to know himself	76	50	0.658	38.0
16	God’s Attributes	83	50	0.602	20.8
17	The Delight Song	169	68	0.402	84.5
18	Love – An Essence of All Religions	77	41	0.532	77.0
19	A Man of Words and Not of Deeds	121	47	0.388	17.3
20	In Broken Images	112	43	0.384	16.0

Serial. No	**Plays. Book III**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
1	Heat Lightning	3018	638	0.211	8.5
2	Visit to a Small Planet	43151	1004	0.233	7.9
3	The Oyster and the Pearl	5738	964	0.168	8.4

Serial. No.	**Poems. Book III**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
1	The Rain	63	45	0.714	31.5
2	Night Mail	114	85	0.746	19.0
Serial. No.	**Poems. Book III**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
3	Loveliest of Trees, The Cherry Now	80	56	0.700	26.7
4	O Where are you going?	135	80	0.593	19.3
5	In the Street of Fruit Stalls	92	64	0.696	30.7
6	Sindhi Woman	69	56	0.812	34.5
7	Times	122	44	0.361	122.0
8	Ozymandias	112	85	0.759	28.0
9	The Feed	101	55	0.545	16.8
10	The Hollow Men	87	63	0.724	43.5
11	Leisure	93	56	0.602	31.0
12	Ruba’iyat	87	70	0.805	17.4
13	A Tale of Two Cities	135	90	0.667	27.0
14	My Neighbor Friend Breathing his Last!	97	57	0.588	10.8
15	He came to know himself	76	50	0.658	38.0
16	God’s Attributes	83	50	0.602	20.8
17	The Delight Song	169	68	0.402	84.5

18	Love – An Essence of All Religions	77	41	0.532	77.0
19	A Man of Words and Not of Deeds	121	47	0.388	17.3
20	In Broken Images	112	43	0.384	16.0

Sr. No	**Essays. Book II. Part-I**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
1	The Dying Sun	1046	376	0.359	23.8
2	Using The Scientific Method	1276	498	0.390	19.3
3	Why Boys Fail in College	1667	561	0.337	24.9
4	End of Term	897	416	0.464	35.9
5	On Destroying	1207	536	0.444	19.5
6	The Man Who Was a Hospital	1158	416	0.359	16.8
7	My Financial Career	903	338	0.374	9.7
8	China's Way to Progress	2244	934	0.416	23.1
9	Hunger and Population Explosion	1595	596	0.370	18.1
10	The Jewel of The World	2256	879	0.390	22.1
Sr. No	**Biographical Essays/Heroes. Book II. Part II**	**Total Words**	**Unique Words**	**Vocab Density**	**Length of Sent**
1	First Year at Harrow	779	356	0.457	15.9
2	Hitch Hiking Across Sahara Desert	39781	1150	0.289	18.0
3	Sir Alexander Fleming	2776	846	0.305	17.1
4	Louis Pasteur	3515	1156	0.329	30.8
5	Mustafa Kamal	2933	1021	0.348	22.6

Sr. No	Novel	Total Words	Unique Words	Vocab Density	Length of Sent
1	Good Bye Mr Chips	16758	3160	0.189	16.4

While deciding any book for any learner, two things must be kept in mind regarding language difficulty: vocabulary density and average words in a sentence. Higher vocabulary density shows less repetitive vocabulary, and it increases difficulty level in vocabulary, for example, in the poem ‘Ruba’iyat’, unique words are 70, and total words are 81; therefore, only 11 words have been repeated, and its vocabulary density is 0.805 which is very high. On the other hand, lower vocabulary density suggests more repeated vocabulary, and a reader feels at ease during reading the frequently repeated words; for example, the novel *‘Good Bye Mr Chips’* uses 3160 unique words, and they have been repeated almost five times until its total words are calculated at 16758. Its vocabulary density is 0.189 which is very low and easy. So, higher vocabulary density content and books must be recommended for advanced level learners, and lower vocabulary density texts should be included in the syllabus of basic level learners. Besides, theme and genre selection are also key elements for book selection; hence, along with human cognition, Cirrus tool can also be used for theme extraction of books as a preview because digital libraries are showing key words as a preview nowadays.

The data generated by Summary tool are reported in table 6. The following individual vocabulary density has been classified into eight categories. They have been classified from left to right, lower vocabulary density to higher vocabulary density and from easy to difficult vocabulary.

Table 6 Eight Categories of Vocabulary Density

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.100-0.199	**0.200-0.299**	**0.300-0.399**	**0.400-0.499**	**0.500-0.599**	**0.600-0.699**	**0.700-0.799**	**0.800-0.899**
	1 short story	10 short stories	4 short stories				
1 play	2 plays						
		3 poems	1 poem	4 poems	5 poems	5 poems	2 poems
		7 essays	3 essays				
	1 hero	3 heroes	1 hero				
1 novel							

As genre-wise vocabulary density is concerned, short stories have a minimum of 0.269 and a maximum of 0.411 vocabulary density; poems have a minimum 0.361 and the highest 0.812 vocabulary density; One act plays have a minimum 0.168 and the highest 0.211 vocabulary density; literary essays have a minimum of 0.337 and the highest 0.464 vocabulary density; the biographical essays/ heroes have a minimum 0.289 and a maximum 0.457 vocabulary density; and one novel has 0.189 vocabulary density.

In this study, one play and one novel have their vocabulary density from 0.168 to 0.199. One short story, two plays and one biographical essay/ hero have less than 0.299 vocabulary density. 4 short stories, one poem, three essays and one biographical essay/ hero have less than 0.499 vocabulary density. 4 poems have less than 0.599 vocabulary density. Five poems have less than 0.699 vocabulary density. Five poems have less than 0.799 vocabulary density, and two poems comprise less than 0.899 vocabulary density.

Higher vocabulary density suggests less repetitive and difficult vocabulary of the genre, while lower vocabulary density refers to frequently repetitive and easy vocabulary. With the Summary tool, this knowledge pattern is discovered that poems carry the highest vocabulary density; for example, seven poems have more than 0.700 vocabulary density. No other genre including short stories, plays, literary essays, biographical essays/ heroes and a novel have more than 0.500 vocabulary density. Giving a concluding stance, except poems, other genres, for example, short stories, plays, literary essays, biographical essays/ heroes and a novel comprise less than 0.499 vocabulary density. The current study recommends that basic level learners' books should have 0.100 to 0.199 vocabulary density; intermediate level learners' academic content should be between the range of 0.200 to 0.399 vocabulary density; and advanced level learners' books should maintain 0.400 to 0.899 vocabulary density. So,/ with the help of text mining, specific reading material has been recommended for different levels of learners. The same has been mentioned in the KDD theory "In active data mining paradigm,… rules are discovered" (Agrawal, & Psaila, 1995, p. 1). To conclude, the current study has discovered rules and criteria for selection of books according to the level of learners. The data generated by Summary are reported in table 7.

Table 7 Overall Vocabulary Density Classification Chart

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.100-0.199	**0.200-0.299**	**0.300-0.399**	**0.400-0.499**	**0.500-0.599**	**0.600-0.699**	**0.700-0.799**	**0.800-0.899**
	1 short story	10 short stories	4 short stories				
1 play	2 plays						
		3 poems	1 poem	4 poems	5 poems	5 poems	2 poems
		7 essays	3 essays				
	1 hero	3 heroes	1 hero				
1 novel							

Most of the lessons of intermediate English textbooks have 20 words in a sentence, but five poems have more than 36 words in a sentence. Apparently, it is opposite to the existing reality of small poetic lines. In reality, poetic lines are small, but a full stop is marked after one quatrain or sometimes at the end of the entire poem as it happens in the 7th poem, *'Times'*. Such a characteristic of extending a verse beyond two lines is named enjambment.

The current study formulates rules for different levels of learners. The average length of sentences for basic level language learners is 7 to 10 words, 11 to 15 words in a sentence for intermediate level language learners, 16 to 20 words in a sentence for advanced level language learners; and 21 to 25 or more words in a sentence for highly advanced level language learners.

4.11.2 Cirrus

In this dissertation, the discussion centres on Cirrus or word cloud which represents key themes, word clusters and characters from the selected texts, so this topic modelling technique and named entity recognition provide not only statistical information but also aesthetic pleasure; hence, their amalgamation enhances learning and pedagogical processes manifold. The frequent occurrence of a word enlarges its font in Cirrus to exhibit significance and grasp the viewers' attention. The readers can quickly notice major and minor themes as a previewing technique. Without close reading, distant reading presents key motifs of any text through Cirrus. Key characters are easily explored with the help of Cirrus. Furthermore, data compression, quantification and linguistics are key features of Information Theory (Shannon, 2009). Cirrus quantifies all linguistic information in compact form to give a precise idea about the text. So, Cirrus can also be used as a previewing technique, and it can be used before teaching any piece of literature. Some previous Cirrus studies also extract key themes from literary texts of novels. Findings of the current dissertation in Cirrus section match these studies: Burrows, 2002; Hussain, 2009; Jockers, & Mimno, 2013; Lohmann, Heimerl, Bopp, Burch, & Ertl, 2015; Scrivener, & Davis, 2017; Sinclair, & Rockwell, 2015b; Yeates, 2013; elucidated in Ch.2.

4.11.3 Phrases

As the debate of discrepancy between standard and substandard collocations/ n-grams is fostered, those phrases which consist of most of the function words (prepositions, articles, conjunctions, auxiliary verb) for example “on the”, “it the” in *‘Hitch Hiking across Sahara Desert’* are considered as substandard phrases. These two phrases do not convey any meaningful knowledge pattern because “Collocations as instrumentation for meaning is a scientific fact” (Louw, 2010, p. 79). So, two key criteria have been finalised: firstly, a substandard phrase consists of all or most of the function words, and they become almost meaningless. Secondly, standard collocations/ n-grams consist of mostly content words, so they carry meaning, and their learning must be beneficial for learners. The frequent occurrence in corpus proves their established linguistic norm. More precisely, standard collocations/ n-grams consist of more ratio of content words than function words. Content words and their semantic shades supersede function words. Appendix E reveals examples of standard collocations/ n-grams. So, it is also proved in Hermeneutica Theory that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). So, Phrases tool utilises human reflection which differentiates between standard and substandard collocations/ n-grams.

On technical grounds, the machine learning algorithm extracts repeated patterns whether they make sense or not. Therefore, the machine has done its work accurately, and next is the human task to recognize and separate standard and substandard collocations/ n-grams. One theoretical postulate is “Manipulation is in service of exploration and understanding” (Rockwell, & Sinclair, 2016, p. 166). That is why human cognition and understanding have been manipulated for getting refined knowledge patterns. Computers cannot find total sense in repeated phrases, yet 90% accuracy is considered a great success in machine learning models.

This study has extracted the 167 different collocation patterns/ n-grams with their relevant examples from intermediate English textbooks with Phrases tool, and they have been mentioned in Appendix E. In Appendix F, these 167 collocation patterns/ n-grams have been presented with their grammatical classification. Therefore, the extraction of collocation patterns/ n-grams is a knowledge discovery process. Some advantages of standard phraseology are that standard collocation patterns/ n-grams enhance fluency and manifest narratology of each lesson. By learning 167 collocation patterns/ n-grams of Appendix E, learners can easily comprehend their entire intermediate English textbooks and utilise these collocations/ n-grams in their answers. These standard collocation patterns/ n-grams accelerate fluency and accuracy in four language skills. Moreover, they also analyse the input of collocations/ n-grams through intermediate textbooks. One point needs clarification that sometimes a collocation has occurred only once in a corpus, even then it sustains its status as a collocation pattern.

Keeping collocations/ n-grams as a foundation stone, learners can coin some other standard collocations/ n-grams. In fact, language is a chain process that starts from unigram to bigram, trigram and quadgram/ collocations/ n-grams and finally, sentences are spoken or written to produce a complete language code.

Previously 37 collocation patterns/ n-grams (Benson, Benson, & Ilson, 1986), 2000 phrases (Shin, & Nation, 2007), 33% figurative phrases, 33% idiom (Siyanova-Chanturia, Conklin, Caffarra, Kaan, & van Heuven, 2017) have been found and in the same continuation, the current dissertation extracts 167 grammatical collocation patterns/ n-grams mentioned in Appendix E. In Appendix F, 18 grammatical categories of 297 standard collocation instances/ n-grams have been presented. The current study extends previous collocation studies on textbooks (Biber, Conrad, & Cortes, 2004; Hsu, 2008; Lee, 2015; Russell, 2017). The current study extracts lexical bundles from English textbooks like Biber, Johansson, Leech, Conrad, & Finegan’s (1999, pp. 992-993) works. The extraction purpose of these 167 collocations/ n-grams is to teach and learn English language skills fluently. These findings are in harmony with the findings of the previous studies of Seretan, Nerima, & Wehrli, 2004; Shin, & Nation, 2007; Nesselhauf, 2003; as they have been elucidated in the literature review.

4.11.4 Links/ KG

As a human brain functions through the binding and wiring of neurons, KGs also develop their existing and innovative links for exploring multifaceted hermeneutic patterns. Therefore, this study has found interconnectivity of characters, their acts and central themes in each lesson. They verify the text, the interconnectivity of events as well as characters. So, relationship mining among relevant variables (Barahate, 2012, p. 13) has been demonstrated with KGs. The linkage of various themes and characters open new vistas of hermeneutics, and new interpretations are presented. Rarely there is an incomplete KG and it can be made complete and factual by its extension with the slider provided with each tool. All Voyant tools are interactive, so expansion and reduction of data visualization and results can be done easily according to the nature of the query. One postulate of hermeneutica theory also emphasizes this point, “They can be extended to expose new things” (Rockwell, & Sinclair, 2016, p. 166). During interpretation, human beings have to filter and zoom some KGs, and in this regard, Hermeneutica Theory guides that “Knowledge bearing tools provoke reflection” (Rockwell, & Sinclair, 2016, p. 166). Apart from finding links, KGs can be used for determining stylistic qualities (Tweedie, Singh, & Holmes, 1996), as mentioned in Ch. 2.

4.11.5 Contexts

“Hermeneutica Theory is embedded in a context” (Rockwell, & Sinclair, 2016, p. 166); this theoretical underpinning reveals that understanding of context is necessary for word sense disambiguation and interpretation of the text for hermeneutic purposes. Again, hermeneutics also requires compiling all textual evidence for a certain theme to interpret its multidimensional meanings and approaches, as it is done in discursive essays too. It is said that a word is known by the company it keeps. To explore the company of a word from a big dataset is a gigantic task; hence, Contexts tool is capable enough to show the bidirectional context of any word within a few seconds. As an example, in the 12th short story, the word “like” has been disambiguated whether it is a simile or a main verb. Contexts tool clarifies its context, and a reader can easily determine its part of speech as well as its semantic shade. The same was proved in chapter 2 by Bhala, & Abirami (2014). Giving theoretical support, Hermeneutica Theory is “embedded in a context” (Rockwell, & Sinclair, 2016, p. 166) because without context true meaning identification is a big challenge. Moreover, Contexts tool also facilitates the compilation of information and textual evidence regarding one topic from a large dataset.

The next chapter presents the conclusion, major findings and contributions of the current study.

CHAPTER 5

CONCLUSION

5.1 Introduction

This section presents an overview of the current dissertation, key findings, fulfilment of objectives, multifaceted contributions in existing knowledge, certain limitations, the future recommendations and implications of the current study. In addition to them, it also shows glimpses of the present meagre situation of text mining and digital humanities in Pakistan, and it proposes how to improve this predicament in the domains of academia, research and industry.

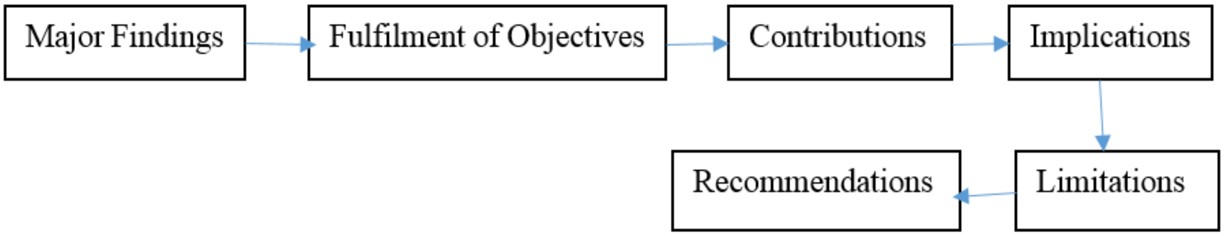


Figure 293 Flowchart of Conclusion Chapter

5.2 Revisiting the Work

Being a study of digital humanities, this research mines the text of intermediate English textbooks to explore innovative and idiosyncratic knowledge patterns. After transforming hard text into digitised text, data of each lesson is visualized into Summary (Corpus summary, stylometry, computational stylistics), Cirrus (word cloud), Phrases (collocations/ n-grams), Links (KGs), and Contexts (KWIC, Word Sense Disambiguation). Furthermore, the triangulation of Knowledge Discovery Theory and Hermeneutica Theory has been applied in the data analysis phase. Then existing knowledge patterns are confirmed, and new knowledge patterns are discovered with valid data visualizations and grids. Five target knowledge patterns have been delineated in the following paragraphs.

Summary tool aptly quantifies all corpus at each short story, play, poem, essay and novel level to present a complete corpus summary of the entire work as total words, unique words, vocabulary density, average words in a sentence and the most occurring themes with their statistical weight. These features represent computational stylistic characteristics or stylometry of the literature and its writer. This study has discovered the stylometry of each literary genre and textbook unit. Roberto Busa spent more than two decades in analysing '*Thomisticus Aquinas'* manually, and Voyant Summary tool extracts stylometric features of any text in a few seconds; therefore, digital humanities performs the manual work of a few decades in a few seconds.

Cirrus brings to light all major characters and key themes with their statistical evidence from each literary work. Most recurring words look bigger and more prominent in each Cirrus, while the least occurring motifs occupy a very small place and size to convey their least occurrence. In addition to it, its interactivity shows statistical weight, and it gives an autonomous choice of selecting 25 to 500 themes. Again, Cirrus transforms a less interesting static text into an interactive, data visual and enjoyable text which motivates an interesting and autonomous way of learning. Thus, digital humanities collaborates with knowledge patterns and aesthetic pleasure to present a compact summary of key themes and characters.

Phrases tool shows collocation patterns/ n-grams in tabular data that are arrangeable in ascending or descending order and in count and length order. This interactivity opens multiple search choices for readers and researchers; however, this study opts for the length of phrases in descending order. The first 15 phrases have been delimited and a total of 167 collocation patterns have been discovered from PIE TCZU (Pakistani Intermediate English Textbook Corpus Zafar Ullah).

KGs have been designed digitally as inter-connected neurons, and they function like human brain neurons which bond multiple ideas and construct new dimensions of hermeneutics. They have linked the main characters and key events of selected literary texts. They develop new relations among different themes and characters to construct new paradigms of knowledge discovery and understanding after zooming in, zooming out and filtering nodes. To summarise, "knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166) and filter nodes after reflection.

The study of semantics, context and grammar are closely linked with one another, and the key word in context is used to disambiguate word sense. To disambiguate any text, Contexts tool facilitates in finding any key word to clarify its context, part of speech and contextual meaning. Another advantage of this tool is that it compiles the required information in a sequence, and the reader easily finds valuable knowledge patterns from a large database; thus, information retrieval has become easier along with information extraction.

5.3 Major Findings of the Current Research

Findings of distant and close readings harmonise each other to a great extent. So, key findings of this research have been mentioned tersely in the light of five research questions. In fact, it is a “dialogical” process of “querying and getting answers back from the computers” (Rockwell, & Sinclair, 2016, p. 159). Firstly, each research question has been written; afterwards, its relevant findings with examples have been delineated:

i. How does text mining summary discover stylometric features from intermediate English textbooks?

The summary tool produces unique words, total words, vocabulary density, average words per sentence and the most occurring words with their statistical weight. It exhibits stylistic qualities of the writers and their works by discovering their unique vocabulary, total words, vocabulary density, and average words in a sentence. Thus, stylistics and literary criticism have been quantified for literary critics, and the same has been proposed by Smith (1978) that literary analysis should be done with computers.

Finding the total vocabulary of any writer in all works and his/her coinage of new words have always been a topic of great debate in stylistics and literary criticism. Furthermore, analysis of the complete work of one writer and the search for his/her word coinage has become convenient with Voyant hermeneutic tools. If a complete list of unique words of any writer is compared with a standard lexicon or corpus of that era, the remaining words express the coinage of the writer. The current study proposes this method, and performs this task practically on single literary works. Consequently, another study can find coinage ability and coined number of words of any writer to enumerate his/her linguistic contributions. Later, they can be incorporated in new lexicons as standard lemmas or as a new meaning.

Different literary genres contain different minimum and maximum vocabulary density. This study discovers that the lowest vocabulary density is 0.269 in the short stories, and the highest vocabulary density is 0.411. The poems display 0.361 as the lowest vocabulary density and 0.812

as the highest vocabulary density. The One-Act Plays contain a minimum 0.168 and a maximum 0.211 vocabulary density. Literary essays comprise a minimum of 0.337 and a maximum of 0.464 vocabulary density. Biographical essays/ heroes have a minimum of 0.289 and a maximum 0.457 vocabulary density, and the novel *'Good Bye Mr Chips'* has 0.189 vocabulary density.

The vocabulary density of Summary tool determines the difficulty level of the text. So, vocabulary density should be classified for each age group and class before writing textbooks or giving them reading material. This criterion facilitates educators and content developers to select material based on vocabulary density. The results lend strong support to propose classified criteria for different levels of reading material based on vocabulary density and average words per sentence. The book selection for each level of class and age should be determined through its vocabulary density. The data yielded by this study provides strong convincing evidence that if vocabulary density is higher, for instance, 0.759, it should be recommended for higher-level learners, and if vocabulary density is lower for example, 0.189, it should be recommended for basic level learners. The current research proposes that basic level learners' textbooks should have 0.100 to 0.199 vocabulary density; intermediate level learners' textbooks should have 0.200 to 0.399 vocabulary density; and advanced level learners' textbooks should have 0.400 to 0.899 vocabulary density.

Average sentence length also decides the suitability of content for a different level of learners. The current study proposes a hierarchical average word per sentence for different levels of learners. For basic level English students, English books should have 7 to 10 words in a sentence. Those books with 11 to 15 words in a sentence are suitable for intermediate-level learners. Advanced level learners' books must have an average of 16 to 20 words in a sentence. Highly advanced level learners' books should contain 21 to 25 or more words in a sentence. Textbook writers and syllabus designers ought to follow this classification for different levels of learners.

ii. How does an interactive word cloud/ Cirrus reveal major themes and characters from intermediate English textbooks?

Cirrus highlights major themes and main characters of the text; for instance, 2nd biographical essay, *'Hitch-Hiking across the Sahara'* shows Christopher as a hero, and Cirrus also shows "Christopher (31)" as the most occurring character. Its whole text discusses the journey in the desert, so "desert (26)" is the geographical setting of the essay. Moreover, in the 1st essay, *'Dying Sun',* the most prominent theme is "life" and its sustenance in this universe. Thus, there is ample support for the claim that Cirrus correctly explores major characters, themes and significant events of the text.

Comparing human and machine extracted themes and characters, Cirrus tool quantifies text, while human beings give different names/ synonyms to the extracted themes according to their perceptions. It is not necessary that human beings take exact words from the text; rather human beings tend to use synonyms or other terms to declare them as themes of the text. Furthermore, human beings extract themes with the lens of feminism, Marxism or with their personal inclinations/ interests; consequently, their results would vary after every reading or after an analysis of a text by different persons, but machine extracted themes remain the same after every analysis; therefore, digital tool analysis enhances validity and reliability of the text mining.

iii. What types of collocation patterns/ n-grams have been unveiled to extract the standard phraseology with its parts of speech?

Phrases tool presents the most occurring 15 phrases from each short story/poem/novel/essay/biographical essay. Then standard phrases have been selected and substandard phrases have been ignored for example, in the 5th short story, the phrase "about his" has been omitted. They have been analysed according to their grammatical collocation pattern/ n-grams, for instance, (Adjective+Verb). This study develops a claim that it compiles a list of 167 standard collocation patterns/ n-grams along with examples from intermediate English textbooks. They have been further classified into 18 grammatical categories. Apart from it, those collocations enhance fluency in four language skills.

The most common pattern of collocations/ n-grams is Adj+N. An innovative collocation pattern is also found that bigrams (two words) co-occur, but they are one part of speech for example, "may be" (Aux) in 1st play *'Heat Lightning'*; "of course" (Adv) in 2nd play *'A Visit to Small Planet'*; "to stand" (Inf V) in the poem *'Leisure'* and "out of order" (Id) in the essay *'The Man Who was a Hospital'.* Trigram "as well as" has been used as a conjunction. In a nutshell, two or three words may represent one grammatical category.

The construction of new collocation patterns/ n-grams is based on existing collocations/ n-grams for example, the 5th poem, *'Street of Fruit Stalls'* shows a bigram "fruit stalls" which can be replicated with newly coined bigrams of vegetable stalls and crockery stalls.

As a fact, correct English can be learnt by the use of standard collocations as they have been discussed in the analysis and holistic conclusion segments. In the 11th poem, *'Leisure'*, bigram "to stand" informs us about the use of the first form of the verb after "to". Besides, prepositions also frequently cause difficulties for English as a Second Language (ESL) learners. So, a bigram "full of" (Adj+Prep) in the 11th poem *'Leisure'* also teaches the correct use of the preposition.

Collocation patterns/ n-grams also reflect philosophical stances, ideologies, narratologies and hedges of the text and its writer; for instance, "I have a dream" expresses the equality-based vision of Martin Luther King in the 11th short story, '*I Have a Dream*'.

iv. How do knowledge graphs present the interrelationship of various key themes and characters for digital hermeneutics?

Links tool finds interrelationship of various themes and characters; therefore, textual examples prove them factual, valid and reliable. In 1st biographical essay *'First Year at Harrow'*, Links tool illustrates the connection of Knowledge Graph (KG) "Greek, Latin, clever, learn" and the KG of "learn, English, dunces" and it explains that only clever boys learnt Greek and Latin, while dunces learnt only English during Churchill's childhood. The word "clever" is not linked with the word "English", and the word "English" is not connected with the KG of "Greek, Latin, clever". Consequently, Links tool presents factual visualization to validate the text. Another evidence has been presented in the 2nd play, *'Visit to a Small Planet'*, a node of character Aide has been linked with only General Powers because Aide just obeys and reports to General Powers, and Aide does not communicate with other characters. I put forward the claim that textual evidence suggests that Links tool generates valid, knowledge bearing and interesting KGs in most of the cases.

v. How does the context of certain problematized words disambiguate the word sense by showing interactive bidirectional context?

Contexts tool finds true word sense of any word along with its context; for example, the word "entered" is present in the 15th short story named *'The Angel and the Author-and others'*. It can have two meanings: the first meaning is "step into," and the second meaning can be "to enter data". So, Contexts tool guides to the latter semantic shade (enter data) which is entirely appropriate in this short story.

Many words are the same in their orthographic patterns, but their grammatical parts of speech are different. If parts of speech are different, their meanings also differ. To disambiguate word sense, Contexts tool searches interactive bidirectional context to understand its part of speech and meanings. Determining parts of speech for any word facilitates readers to assign contextual meaning. In the 19th poem, Contexts tool disambiguates the word “like” in this way that it is used seven times as a simile; therefore, the difference between a simile and a main verb can be easily discriminated. Besides, deictic pronouns are also disambiguated with Context tool, as it has been done in 4th poem. Deictic pronoun “you” have been clarified that who are addresser and who are addressees in the whole poem.

Relevant adjective for a particular noun has been searched with Contexts tool; for example, the 20th poem unveils adjectives with noun “images” as “broken images”, “clear images”.

To find any reference or exact textual quote from a big dataset has become convenient and efficient through Contexts tool. Furthermore, it can be used in designing of objective type test or in finding the answer to any objective type question paper without a close reading.

Contexts tool also practically instructs Voyant users on collecting material concerning one topic or about one character without close reading of the whole text. Consequently, it saves time for researchers and students.

5.4 Addressing Statement of the Problem

Some major problems of static, time consuming, inadaptable, unstructured text were mentioned in the problem statement in the 1st chapter. After the data analysis section, it becomes evident that all intermediate English textbooks have become digitised, machine readable, and interactive for digital natives. Multi-coloured data visualisation and interactivity can diminish the boredom of distant readers, save learners’ time and enhance their motivation as well as interest to some extent. Besides, this study employs only five tools to show their reading and learning styles, but the reader is free to use 25 tools of Voyant, and all digitised texts are adaptable into any data visual. Those who want to study in the traditional mode can also use Reader tool to study from top to bottom. Those who want to apply smart study skills, they should take advantage of 25 Voyant tools. This ease facilitates technology addicted youngsters; consequently, learners learn new knowledge patterns until their interest culminates to the zenith with the support of Voyant text mining tools. This distant reading creates interactive data visualization, aesthetic pleasures, derivation of collocations/ n-grams, stylistic analysis, interlink of themes with characters, and contextual reading. To conclude, the current research may replace close reading with distant reading technique, especially if textual data are big. Distant reading facilitates readers to derive conclusions hermeneutically. The same outcome has also been mentioned by the designers of Voyant tools "to replace thorough reading and let the readers jump to conclusions" (Rockwell, & Sinclair, 2016, p. 49). To conclude, the current study matches the vision of Voyant tool designers.

This is the age of technology, and the need to learn programming and coding is increasing for humanities students. Recently, autoML has diminished the need of learning coding and programming. Voyant tools have incorporated NLP and Python libraries in it, that even a technophobe can easily do text mining tasks without any knowledge of coding, programming and Python libraries.

5.5 Fulfilment of Objectives

In this section, each research objective along with its accomplishment has been delineated.

i. To produce a summary of text mining to extricate quantified information about stylometry, vocabulary density, the average length of sentences and the most frequent words in the corpus.

All short stories, essays, poems, biographical essays, and the novel show total words, unique words, vocabulary density, average words in a sentence, and the most occurring words that unveil the writer's stylometry and his/her literary work.

In the essay, *'My Financial Career'*, Stephen Leacock writes 338 unique words from 903 total words. Therefore, his writings have a 0.374 vocabulary density, and his sentences consist of an average of 9.7 words per sentence. To comprehend comedy, small sentences and easy vocabulary have been used by Stephen Leacock. The most frequent words, for example, "bank (9)", "manager (8)", "dollars (7)", "money (7)", "accountant (5)" and "account (4)" elucidate the most frequent characters and aspects of the story.

ii. To generate Cirrus/ word clouds to unveil the prominent motifs and characters.

All short stories, literary essays, biographical essays and the novel perfectly extract key motifs and major characters from their source texts in the data visualization of Cirrus/ word clouds. In the 3rd biographical essay, *'Alexander Fleming'*, the protagonist "Fleming" is the most significant and frequent theme in it. In the 2nd short story, *'Clearing in the Sky'*, only two major characters are “I, he” and both of them have been prominently displayed with large fonts and different colours in figure 17. Mostly Cirrus/word cloud correctly mines text and highlights prominent motifs and characters from any text.

iii. To point out collocation patterns/ n-grams to extract the most frequent standard phraseology.

Bigrams and trigrams are extracted to study phrases, whereas substandard phrases (for example, “and I”) have been excluded. For more comprehension, their grammatical parts have also been written to classify them as certain collocation patterns/ n-grams, for example, Adj+N collocation pattern/ n-gram which is the most common pattern in this corpus. Thereupon, standard phraseology is sorted out from Pakistani Intermediate English Textbook Corpus Zafar Ullah (PIE TCZU). The current study presents 167 collocations patterns belonging to 18 grammatical categories.

iv. To create knowledge graphs to explain the interconnectivity of various themes and characters in digital hermeneutics.

In most of the cases, a KG ties a rational knot among characters, their relevant ideas and themes; for example, in the novel *‘Good Bye Mr Chips’*, one KG associates with “Chips, head, Brookfield” which discloses that Mr Chips also serves as the head of Brookfield; school is the geographical setting of the novel because most of the actions develop in it, and all characters of the novel belong to Brookfield directly or indirectly. Mr Chips, being a central character of Brookfield, gets the status of an institution and authority to resolve several conflicts.

v. To explore the bidirectional context of ambiguous words to comprehend the contextual word sense.

If there is a word “land”, the Contexts tool shows bidirectional context, and the reader easily differentiates whether “land” in the 2nd short story*, ‘Clearing in the Sky’* means a piece of land, or it is about the landing of an aeroplane. It is found three times in the corpus, and three times it refers to a cleared piece of land on the mountain top. To conclude, Contexts tool disambiguates selected words and conveys the true meanings of every word along with its bidirectional context.

5.6 Current Text Mining Research Situation in Pakistan

Pakistan is just at its initial stage of text mining, data mining, computational linguistics, digital humanities and corpus linguistics. Recently, data mining subjects and degrees have been started. There are a few data science labs, and one of them is in Information Technology University, Lahore. There are three computational centres: Firstly, Centre for Language Engineering in UET, Lahore which has designed Urdu Screen Reader, Urdu POS Tagger, English, Urdu, Punjabi and Sindhi online dictionaries (Centre for Language Engineering, 2016). Secondly, Language Engineering Centre in the University of Punjab, Lahore is just in its infancy period. Thirdly, Corpus Centre in Air University, Islamabad, has compiled a Pakistani English variety corpus. Fourthly, Center for Research in Urdu Language Processing (CRULP) is present at FAST-NU, Lahore.

5.7 Contributions of the Current Research

Edison said, “The value of an idea lies in the using of it” (Edison, 2001). So, an idea becomes practicable and expedient for the masses after its materialization for general goodness. The current study mainly contributes to the domains of academia and pedagogy, but, it can also be utilized in forensic linguistics, library science and cyber security. Furthermore, it bridges gaps between distant and close reading to a great extent by finding their homogeneity.

5.7.1 Comprehensive Contribution

A few papers, one master thesis with one tool, lecture notes and some blogs have been written with the use of Voyant tools, but the current study is the first comprehensive work on five Voyant tools, namely Summary, Cirrus, Phrases, Links, and Contexts. The current study interconnects domains of digital humanities, educational data mining, text mining and English textbooks for knowledge discovery and hermeneutic analysis. Moreover, this study has explored criteria for selecting and writing textbooks for basic, intermediate and advanced level learners.

5.7.2 Academic Contributions in Pakistani Society

The current study contributes to the following aspects:

i. University students in advanced countries are studying, conducting research projects and developing industry linkages with text analytic tools, but Pakistan is ignorant of these productive techniques. The current study fills this niche by elevating pedagogy and learning practices in academia and saleable research through digital study methods. The researcher has started teaching and conducting research with Voyant tools to his BS English students at National University of

Modern Languages, Islamabad, and Capital University of Science and Technology, Islamabad, Pakistan. With distant reading, it enables learners to learn knowledge patterns interestingly in the shortest possible time.

ii. Data visualization of the intermediate textbooks has been done in five different interactive visuals. Firstly, from Corpus tools, Summary tool summarizes the entire corpus from every short story, essay, poem, novel and play. Secondly, from Documents tools, Cirrus (word cloud) is generated for each textual unit. Thirdly, from Grid tools, Phrases tool generates phrases/ collocations/ n-grams in a tabular form with their counts in terms of occurrence and length. Fourthly, from Visualization tools, Links tool generates a multi-coloured visual which develops a KG among several themes. Fifthly, from Grid tools, Contexts tool extracts a table of interactive context for the ambiguous word.

iii. Presently, only scanned books are available online, and they cannot be used interactively for text mining. The academic contribution of the current study is that digital reformatting has been done in the way that printed intermediate English textbooks have been transformed into digitised textbooks, and their corpus has been built. These digitized books have been shared on different digital platforms for the benefit of intermediate students. It can facilitate more than 1 million students from nine Boards of Intermediate and Secondary Education in Punjab, Diploma of Associate Engineering (DAE) students in Punjab Board of Technical Education. These learners can study interactive and digital intermediate English textbooks for their self-paced learning.

iv. The current study has experimented with five specified Voyant tools to reveal their strong and weak areas. Voyant analyses big data or large texts better than small texts, and they show erroneous results during the analysis of short poems.

v. Corpus, text analytics, data visualization, tabular data and digital interactive methods have become study skills for students.

vi. Like foreign advanced level universities, academic growth and authenticity of knowledge patterns have been established with Voyant tools.

vii. The current study ignites a wave of preparation of digital material for Pakistani students.

viii. Interpretations of the generated visuals in the light of DH bridges humanities and computing to discover new data visuals and knowledge patterns.

5.7.3 One Word Ubiquitous Library Concept Manifestation

Recently knowledge laden mini silicon chips have been introduced, and big libraries are being transformed into smart digital libraries; that is why the concept of e-learning has been initiated. Following the same trend, e-learning libraries in TMA (Tehsil Municipal Administration) parks have been established in Punjab, Pakistan. Extending this digital trend, the current study transforms one book into one word and by just clicking that word, the entire book opens online with key features of text mining and data visualization. Encompassing the same strategy, the next step is just to click one word and a complete corpus of all books within one domain will open, and the distant readers can get the desired information from all books with just one word. Thus, multi-layered one-word libraries can be accessed online, and anybody can carry his/her shelves of the library with one word. The current research proposes and proves an innovative concept of a smart ubiquitous one-word library concept in library science.

Appendix G exhibits the one-word library concept of the entire "Pakistani Intermediate English Textbook Corpus Zafar Ullah (PIE CTZU)" at micro and macro levels, for example, the corpus of each unit, complete 1st year books, complete 2nd year books and a complete corpus of PIE CTZU. The most occurring word from each corpus has been specified for the one-word library concept. Just click on the yellow coloured one word, the reader will access the relevant corpus. Now 25 tools for close reading purposes have been introduced. In the Documents category, a close reader can also select Reader tool for close reading.

A common student is confused to observe the discrepancy between the number of words in Voyant tool and the number of words in MS word file as shown in Appendix G. In 2nd year corpus, words in Voyant tool are more than the MS word file. This difference appears when Voyant tool counts some specific one word as two; for instance, Voyant tools count the word "Ruba'iyat" as two words, whereas the MS Word file counts the word "Ruba'iyat" as one word. Thus, Voyant tools differ in the counting of some single words. Secondly, the automatic stop words tool may also decrease the number of some function words in the data of Voyant tools.

Earlier studies in Ch.2 (Chakraborty, 2012; Eder, Rybicki, & Kestemont, 2016; Li, Ji, & Xu, 2017; O'Sullivan, Bazarnik, Eder, & Rybicki, 2018; Sundberg, & Nilsson, 2018) harmonise with the current study in the elucidation of stylometric qualities. The current study builds intermediate English textbook corpus like Burnard and McEnery (2000), Sinclair (2004), Connor and Upton (2004) whose works are on the use of corpora in TEFL. As the current study applies automated Voyant text mining analysis of textbooks, similarly the following six types of research applied automated methods for textbook analyses (Anping, 2005; Biber et al., 2004; Chujo, 2004; Gouverneur, 2008; Meunier, & Gouverneur, 2007; Romer, 2004b, 2006).

Biber et al. (2002) built a 27 million words first textbook corpora named the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus) of discourses in American universities. Secondly, in Germany, Romer (2004a) prepared 100,000 words of German English as a Foreign Language Textbook Corpus (GEFL TC). Thirdly, Meunier and Gouverneur (2007) prepared TeMa textbooks corpora of 700 0000 words. Extending the list of new textbooks corpora, the current study builds "Pakistani Intermediate English Textbook Corpus Zafar Ullah (PIE CTZU)" with 82,487 words.

5.7.4 Linguistic Contribution for Lexicography

New words are also coined by changing a Proper Noun into a common noun or some other parts of speech; for example, Google (Proper Noun) has been changed into a verb, for instance, "googling" and googler. Similarly, Voyant (Proper Noun) has been proposed as a verb (voyant, voyanted, voyanted, voyanting), adjective (voyanting), adverb (voyantly) and noun (voyanter) through this study, and the researcher has used these proposed grammatical forms several times in this study and one conference paper. Since it is a new word, its meaning has also been written in parenthesis, such as using Voyant tools to mine text, study, teach and conduct research hermeneutically. Besides, the same will be emailed to notable lexicography publishing houses.

5.7.5 Exploration of New Dimensions in Voyant Tools

The researcher finds some other beneficial things which have not been mentioned by its designers, for instance, Voyant tool is equally useful for Urdu, Punjabi, Arabic, Turkish and other code-mixed languages. Besides, flaws in tools lead to an improvement in tools. There is a flaw of Inpage file acceptance in Voyant tools, and it necessitates improvements in its future design. The researcher has also informed tool designers about some tool design recommendations mentioned in 5.10.4.

5.7.6 Theoretical Contribution: Text Mining Insight Theory

"Modern theory is usually developed through a series of steps by academics" (Skills You Need, 2018), and the following modern theory has delineated certain steps for text mining. The researcher of the current dissertation humbly proposes Text Mining Insight Theory which has the following steps and features.

1. Hard Text to Digitisation: Paperbound texts should be transformed into digitised text or already available digitized text, or databases can be selected for this purpose.

2. Pruning: Textual data requires pruning or cleansing; for example, names and captions of advertisements, images, titbits, references and printing press are given in web texts. It is vital to prune and purify text and to prepare it for the text mining process.

3. Processing: To extract information patterns according to aims or to solve any problem, the selected text/ texts should be uploaded on the most relevant tools as the current study has done through Voyant tools. Processing also involves coding, preparing certain file type or arranging text to fulfil the prerequisites of the tool.

4. Statistical Information: Statistics are powerful locomotives to tug its knowledge bogies to their safe destinations speedily. In another way, statistics are foundations on which an edifice of text mining and visualization is built. A visual is transformed into a knowledge pattern with statistics which shapes and modifies all visuals because interactivity is the key element of the voyanting process.

5. Textual Information: Extraction of textual information is based on statistics as the Summary tool does in Voyant. The researcher can choose particular segments to strengthen his/ her views, or text can be presented as evidence of visuals and statistics.

6. Visual Information: Certain types of visuals (Cirrus, Trends, Links) have been generated by tools which are replete with explicit and implicit hermeneutic information.

7. Tabular Information: Information is also provided in table form, and a researcher can adjust it according to ascending or descending order. Usually unigrams, bigrams are mentioned in tables along with their occurrence and length.

8. Hermeneutic Knowledge: Information extraction, recognition of knowledge patterns and converting visuals, tables and texts into harmonized knowledge is necessary to make it a result-oriented hermeneutic activity.

9. Segregation of Known and Unknown Information: Text mining testifies and consolidates already known knowledge with visuals, tables, texts and statistics. The remaining unknown or new knowledge patterns are used for discovering new dimensions and insights.

10. Insight Deduction: Delving deep into knowledge and bringing a deep understanding of the text is the hallmark of Knowledge Discovery Theory. This insight is manifold, unique, in-depth and harmonious with visible information patterns.

11. Forecasting Intelligence: The judicious organizations, persons and nations foresee the future and prepare themselves at their best. They build their capacity to meet forthcoming challenges. So, text mining forecasts the future trends, happenings and business tendencies to prepare for the futuristic challenges. Therefore, knowledge discoveries and inventions are outcomes of forecasting.

12. Transapplication of Sublime Insight: One text does not serve only one purpose or one field of study; rather, sublime insight is applicable to many situations, subjects, hardware and software. Voyant tools are used for reading textbooks and generating sublime insights which lead to cybersecurity, prediction, pinpointing of extremists and their hate crimes, one-word library concept, data mining, business intelligence, language engineering, pedagogic uses and the publishing industry. Transapplication validates sublime insight, and makes it beneficial for many diverse fields. To summarise, wide-ranging applications and multifaceted advantages elevate sublime insight.

13. Progressive Modification and Recycling: Finding its faults critically and modifying its tools, visuals, tables, statistical formulae and insights are considered progressive signs. With each cycle of Text Mining Insight Theory, new possible modifications should be essentially done. In this way, this progressive cycle of unveiling knowledge keeps on moving forward to its indefinite intellectual dimensions and fields.

5.8 Pedagogical Implications of This Research

Digital pedagogy is the practical and academic result of this study, though conducting a user study with pre-test and post-test of Voyant tools is beyond the scope of this study. Such studies have been recommended for future studies.

5.8.1 Educational Context and Potential Application

The present study centres upon educational context since intermediate English textbooks have been selected for text mining, hermeneutics and knowledge discovery process. Five Voyant tools, namely Summary, Cirrus, Phrases, Links, and Contexts have been used in an educational context because more than fifteen renowned universities, including Michigan State University, University of North Texas, University of Toronto and libraries also apply them to an educational context. The brief overview of Voyant tools in an educational context has been presented in the following paragraph:

Firstly, Summary tool presents stylistic features of any text in the form of unique words, total words, average words per sentence and vocabulary density. Thus, this tool shows a person’s total vocabulary and this knowledge pattern can also assist in analysing any subjective type answer in an educational context. Secondly, Cirrus tool extracts key themes of any type of text, books, websites as a preview or an executive visual summary. Therefore, the distant reading process is done before close reading through Cirrus tool. Furthermore, a reader also develops a link between the most frequently occurring words and the title or genre of the text. Thirdly, Phrases tool extracts collocation patterns of any text, and they are used to enhance language learners' fluency. In addition to it, these collocations manifest narratology and hedges. Fourthly, Links tool shows the interrelationship of associated themes and characters to elucidate different hermeneutic layers. Furthermore, multidimensional interpretive and discursive essays can also be developed with the help of knowledge graphs. Fifthly, Contexts tool disambiguates the word sense of any word by showing bidirectional context. Moreover, context helps to differentiate between two homographs, for instance, verb or noun, adverb or adjective. Besides, if a person aims to compile relevant information on any topic from big data, this tool can accumulate all relevant text in a sequence in the shortest possible time.

5.8.2 Digital Humanities and Digital Pedagogy

Digital pedagogy started in the 1990s, since our classrooms are multimodal and all senses play their role in absorbing knowledge, imparting knowledge in academic settings. Thus, digital tools, human beings and institutions collaborate for the process of learning and teaching (Walker, 1999). Digital humanities unveils knowledge patterns used for learning and teaching purposes, especially it promotes autonomous learning according to each learner's interest, learning style, and IQ level. Another level of comparison is that traditional reader consumes text, while interactive digital reader experiences digital tools and creates digital knowledge patterns.

5.8.3 Digital Pedagogy and TPACK Model

Digital pedagogy and TPACK (Technological Pedagogical Content Knowledge) model emphasize teaching educational content with technological tools and perspectives. Practically, Voyant users take text or other educational content to study with technological tools such as Voyant suite, and they learn new dimensions, for instance, repeated themes, characters, collocation patterns/ n-grams, linkage of one theme with other associated themes and characters through Links tool. In addition to it, Summary tool quantifies stylistic features, unique words, total words, vocabulary density. With the traditional method, it can take several years to analyse quantitative features of the whole text. Furthermore, disambiguation of the word sense process can be done with Contexts tool.

5.8.4 Digital Pedagogy and LearningWheel Model

To learn and develop digital learning and digital literacy skills, LearningWheel has been emphasized because it engages learners. Moreover, it reshapes the content as Voyant tools reshape the text into various grids, data visualizations and other forms to present the crucial information patterns aesthetically. As PowerPoint tools support lectures, technology also provides intellectual support to digital learners. LearningWheel model has the following characteristics.

1. Visuals inform for pedagogical purposes
2. Data visuals are created by practitioners and for the practitioners
3. It is categorised into four pedagogic 'modes of engagement'
4. It can be 'Resource' specific
5. It can be 'Contextualised' to a subject area
6. It can be 'level' specific

7. It promotes free and accessible digital resources
8. It shows a collaborative approach
9. It manifests an international perspective
10. It is scalable: flexible and adaptable (Kellsey, & Taylor, 2016).

5.8.5 Pedagogy and Educational Data Mining

Educational data mining assists pedagogical practices, as the current study has provided pedagogical support (Baker, 2010) with Voyant text mining tools. The first tool is Summary which counts total words, unique words, the average length of sentences, vocabulary density and the most repeated words. These knowledge patterns facilitate the elaboration of stylistic features of any writer or any piece of writing. Moreover, these knowledge patterns support quantified criticism of any literary piece of writing. This corpus summary also guides the criterion to select reading material for basic, intermediate and advanced level learners because EDM enables “data-driven decision making for improving the current educational practice and learning material’’ (Calders, & Pechenizkiy, 2012). Preview of the text helps learners and teachers to get a brief introduction of the text; hence, Cirrus tool generates word clouds with their statistical weight. Furthermore, the most repeated themes and characters are also extracted to develop a background and genre of the text. TEFL teachers should use this technique before teaching any text. Another tool, namely Phrases extracts collocation patterns/ n-grams and they can be used for enhancing linguistic fluency. In addition to it, the extracted collocations/ n-grams also indicate hedges and narratology of the selected text. In advanced countries, textbooks give a list of collocations in the textbooks, but Pakistani textbooks lack this quality. I propose the inclusion of a collocation list in every English textbook. Moreover, students should also be trained to make their own list of collocations with the help of Phrases tool. Another tool, for example, Links tool builds various knowledge graphs to show the interconnectivity of themes and characters, and they are based on the textual evidence of the text, for instance, nodes of Greek and Latin are linked with bright students; and English node is linked with only dull students in the biographical essay *‘First Year at Harrow’*. Thus, it provides hermeneutic support by widening multidimensional layers of interpretation. Besides, EDM promotes learners’ critical thinking about any phenomenon (Romero, & Ventura, 2012), and learners interpret different knowledge graph nodes. Language and literature teachers should use these tools for language learning and language teaching purposes. Moreover, educationists should explore knowledge patterns and insights into the data. The fifth tool is Contexts which supports the process of word sense disambiguation by showing bidirectional context. It can also be used to compile information on one topic to write a subjective answer, and it can also be used to set or check objective type papers.

Technological Pedagogical Content Knowledge (TPACK) combines the use of technology, pedagogy and content knowledge (Koehler, Mishra, 2009). Modern education cannot be imparted effectively without following the TPACK model because it engages maximum senses in the learning process. During the challenge of the pandemic, TPACK model has been followed successfully in the developing and developed world simultaneously. Furthermore, autonomous learning has been introduced, even modern MOOC and Coursera online courses have been designed by following TPACK model. To conclude, educational data mining and pedagogy have developed inseparable ties.

5.8.6 Contemporary Practices of Teaching with Voyant Tools

Academically advanced countries have applied Voyant tools in academia, teaching, learning, research activities and library settings to prove an effective nexus between teaching and Voyant tools, for instance, 22 famous universities of the world, including the University of North Texas, USA, Stanford University, USA, Michigan State University, USA, University of Toronto, Canada and others have employed Voyant tools for teaching to their graduate students. Moreover, scholars of PhD digital humanities studied use of Voyant tools at the University of Melbourne, Australia. As far as research activities are concerned, seven international digital humanities and data mining conferences concentrated on Voyant tools. More than 28 literary kinds of research projects and websites employed Voyant tools. In advanced countries, libraries play a very conducive role in constructing knowledge at international paradigms; hence, more than six universities, including the University of California, Indiana University, Western Michigan University, USA, Carnegie Mellon University, USA, mentioned Voyant tools on their library pages (Sinclair, & Rockwell, 2017).

5.8.7 Digital Pedagogy and Voyant Tools

Craig Saper mentioned implications of DH thus “A field of study, research and teaching” (Terras, Nyhan, & Vanhoutte, 2016, p. 286) and the current DH research displays its implications for learning, teaching, publishing and research purposes. The current study suggests the application of Voyant tools in Pakistani academic institutes for learning, teaching, research and development of industry linkages. The current study proposes Voyant tools as the most appropriate pedagogic tools for knowledge discovery pursuits because Technological Pedagogical Content Knowledge (TPACK) combines technology, pedagogy and content knowledge (Koehler, Mishra, 2009).

i. Pakistani Universities should voyant (Verb. the use of Voyant tools to mine text, to study, to teach, to conduct research hermeneutically) for reading and comparison of voluminous and small texts. Moreover, new big data works or databases can be studied with Voyant tools.

ii. Literary style and quantified criticism of different novelists, dramatists, short story writers and poets should be studied quantitatively and qualitatively through Summary tool. Consequently, digital criticism can generate better and precise discussion and comparison of different literary texts.

iii. Collocation patterns/ n-grams and standard phraseology lists should be compiled to enhance fluency in reading and writing. Moreover, these phrases reflect the ideology and narratology of the writer. Apart from it, the collocational study produces lists of the most common phrases for basic language learners.

iv. Comparison and contrast of different complete works can become easy with digital interactive Voyant tools; for instance, complete works of Jane Austen can be compared with Shakespeare's complete works to find similar and contrasting elements. Thus, they can be used for learning and teaching purposes in the shortest possible time.

v. When an educationist aims to select a storybook or a novel for different age group learners, usually, he/she selects just by intuition or by just reading a few passages, but there is no distinct criterion. The current study presents text analytics with the Summary tool which shows unique words, total words, vocabulary density, the average length of sentences and the most frequent words with their statistics. Higher vocabulary density words show less repetitive and rather difficult language. On the other hand, lower vocabulary density texts show more repetitive words, hence, it indicates the use of easy language in the text. A rational educationist should select lower vocabulary density and small sentence books for children, while higher vocabulary density and long sentences must be recommended for advanced level learners. The current research proposes that basic level learners' textbooks should have 0.100 to 0.199 vocabulary density; intermediate level learners' textbooks should have 0.200 to 0.399 vocabulary density; and advanced level learners' textbooks should have 0.400 to 0.899 vocabulary density.

Average sentence length also decides the suitability of content for a different level of learners. The current study proposes a hierarchical average word per sentence for different levels of learners. For basic level English students, English books should have 7 to 10 words in a sentence. Those books with 11 to 15 words in a sentence are suitable for intermediate-level learners. Advanced level learners' books must have an average of 16 to 20 words in a sentence. Highly advanced level learners' books should contain 21 to 25 or more words in a sentence.

v. Usually, subjective and objective type questions are asked in all examinations, Contexts tool is quite appropriate to design objective type question papers. To find key words, answers to objective type questions and to collect information on one topic have also become easier with Contexts tool.

vii. Quality of subjective answers or essay paper of CSS exam can be evaluated with the Summary tool which shows vocabulary power, unique words, the average length of sentences and key themes of the written work.

viii. A writer can easily assess his/her writings, trends, and repeated words through Cirrus tool (University of Victoria, 2019). Nowadays, SEO (Search Engine Optimization) is done on websites to accelerate web traffic and make the website most accessible.

ix. Textual repeated concepts can be traced in the whole documents through Trend and Bubbles tools in Voyant suite (University of Victoria, 2019).

x. Corpus building from any type of file or text or web source has become easier with Voyant tools. Then the learner can easily explore various interactive knowledge patterns from the text.

xi. Voyant tools can be recommended for pedagogy, finding textual examples or quotes, setting objective type question papers and evaluating subjective papers.

5.9 Limitations

Although the current research has accomplished most of the aims and satisfactorily answered five research questions, yet there are a few rare unavoidable limitations. Research integrity requires to mention them, since there should be some endeavours to overcome them in future studies. This study has highlighted several problem areas.

i. Some writers use synonyms in their works, so, the same word repetition and deriving themes based on statistical value may fail or mislead. Sometimes, a writer indirectly points out a theme, but he does not use the key thematic word repeatedly. It is considered that the reader is wise enough to understand the implied central theme. If any such shortcoming is found, it can be addressed easily with close reading or Contexts tool. Such limitation is rarely found especially in the selected English textbooks. One specimen of this misconception is shown in the 17th poem, *'The Delight Song'* in which the most recurring words appear as "alive (4)", "good (4)", "relations (4)" and "stand (4)", but they do not represent the most important theme and title words "delight (1)", "song (1)" which occur once in the text.

ii. Phrases Voyant tool shows all collocation patterns/ n-grams, but all patterns are not linguistically important. So, the researcher has to differentiate between standard and substandard phrases. "Knowledge bearing tools provoke reflection" (Rockwell, & Sinclair, 2016, p. 166). This reflective ability helps to differentiate between standard and substandard collocations/ n-grams. Meaningful or standard collocation is "seek justice", while substandard and less meaningful collocation/ n-gram is "and the".

iii. Sometimes, a short poem does not have any phraseology or collocation pattern/ n-gram because of its diversified and unique vocabulary; for example, the 10th poem, *'Hollow Men'* by T.S. Eliot does not show any repeated collocation pattern. No phraseology is also meaningful because it reveals unique vocabulary, and it is a great scholarly effort and stylistic quality of any writer.

iv. Sometimes, Voyant tools show error and incompleteness during Links or KG development of some poems. Moreover, some poems cannot present key themes because of the non-repetition of themes, their brevity, figurative language, implied and symbolic meanings. This error message is caused by a small poem, less or rare repeated words and terse poetic style. The first rule of data mining is that it supports big data (Adriaans, & Zantinge, 2009, p. 102), and large texts produce better results than small texts. That is why small poems show an error, but the text of the novel does not show any anomaly. Exception to some small poems, all tools produce accurate results in prose. It is also a fact, that no text mining tool can perform better cognitive tasks than human researcher. Text mining tools provide an aid in hermeneutics. Finding such faults in tools lead to improvements and academic contributions. To conclude, no human being can count text mining patterns faster than computers, but human cognition and wisdom cannot be ignored in the presence of text mining tools.

v. Data can be "extended to expose new things" (Rockwell, & Sinclair 2016, p. 166). Voyant tools are interactive, and their interactivity sustains during the online mode, but to present them in this dissertation, it is necessary to convert interactive data into a static image. Firstly, deficiencies start in a static image; for example, the statistical weight of each theme in Cirrus disappears. Secondly, a KG shows the path of its relationship with other themes and nodes on Voyant website, but as soon as it changes into a static image, its coloured path disappears.

vi. In the 17th poem, only the last sestet has been changed into the KG, and the first 14 lines have not been transformed into the KG, so it is the flaw of Links tool in poetic analysis. A KG connects only those repeated themes which have interrelationship in the text, and this poetic text is neither repeated nor connected in the form of the KG.

vii. Sometimes a KG misses some themes; for instance, in the 12th short story, the theme "buy" is missed with Jim, while he buys a gift for Della. Primarily, its solution is the extension of the KG and a larger KG incorporates more themes. To seek more accurate results, a larger KG should be structured.

viii. Voyant tools take "apostrophe sign" as a separator; for instance, the word "Ruba'iyat" is displayed as two words, "Ruba" and "iyat" in the data visualization of the 12th poem, whereas "iyat" is a meaningless word. Similarly, Voyant cannot comprehend "came to know" as one theme of "know". Voyant takes "came", "to" and "know" as three separate themes which do not match textual and semantic reality.

ix. Contexts tool is unable to display capitalization of proper noun "Morning" in 15th short story.

x. Another limitation is that only a few previous research projects have been found relevant to Voyant tools. Others have been mentioned in blogs or lecture notes of different professors.

xi. As technological deficiency is concerned, Voyant tools do not accept and upload InPage text, while they accept MS word files, pdf files, text files and web links.

The above-mentioned limitations have been transformed into future recommendations.

5.10 Recommendations

Recommendations lead to future domains to progress the existing knowledge in multidimensional branches. The results of this study suggest several new avenues for future research.

5.10.1 Voyant Related Recommendations for Futuristic Research Works

There are 25 Voyant tools, and the current study delimits only five tools, namely Summary, Cirrus, Phrases, Links and Contexts. These recommendations are specifically concerned with the five aforementioned tools.

i. Stylometry should be used for forensic analysis; for instance, wills can be studied to judge their genuineness or fictitiousness.

ii. The total and unique words of any writer should be compared with the corpus or dictionary of that time to find out newly coined words by the writer.

iii. Lexicography and text analytics should collaborate in producing new dictionaries/ word lists/ phrase lists of each subject and profession, for instance, a dictionary/ wordlist for doctors. This lexicographic work should be done with Phrases tool in Voyant.

iv. Code mixing and code-switching should be analysed with Cirrus, Phrases and Summary tools. Furthermore, it will find linguistic tendency to replace language items, as it is being done with Urdu that many English words are replacing Urdu words.

v. With text analytics, translation and original work should be compared at several levels with Summary, Cirrus and Phrases tools. When any book or work is translated into another language, it is necessary to measure its equality in terms of the most occurring themes with Cirrus tool, statistical weight, total words, unique words with the Summary tool; phrase comparison with Phrases tool and interrelationship of themes with Links tool.

vi. The current research has discovered knowledge patterns from intermediate English textbooks. The next step is to practise these knowledge patterns in the classrooms. Further steps should be to conduct user studies on all Voyant tools with pre-test and post-test to measure the efficacy of these tools.

5.10.2 Recommendations for Future Text Mining Research Projects

Taking this research as a pioneering research in digital humanities (DH) in Pakistan, further research projects keeping in view quantum humanities can be generated on the following grounds:

i. Ubiquitous data mining will analyse various aspects of our daily lives since data mining has affected academia, shopping choices, leisure activities, health, work, search options.

ii. Invisible data mining, for instance, software, search engines, emails can be mined in the future.

iii. Digital mining with a privacy preservation option may be used without exposing personal information. To accomplish this task, social media sites can be mined to find target customers.

iv. Current data stream mining can be used in future research.

v. Web mining of academic websites, university sites should be done to search their key themes in the future.

vi. Multimedia data (audio, video, image, text) mining can be done for deep learning. Audio and video mining projects from telephonic conversations, webinars, podcasts and audiobooks should be done to convert speech into the text, to search phonemes, to search particular word meanings and their classification.

vii. Image mining can also be done in future research projects because it discovers interesting knowledge patterns from images. After all, image data are increasing on Google Image, ELTpics, Instagram, social media. Furthermore, bias detection from images can also be done.

viii. Social media mining and emotion mining will be done to explore trends, sentiments, user flow, and prediction patterns. Social media mining has been started for cybersecurity and Donald Trump has issued orders to mine social media accounts of USA visa applicants (Chan, 2018).

ix. Table mining projects will mine unstructured, ungrammatical, scattered and heterogeneous data which are usually ignored because some research projects are being conducted on plain and structured texts.

x. Topic mining, topic clouding, concept mining, and opinion mining can be done for topic detection from a large dataset.

xi. Multilingual web mining or cross-lingual mining can be done because digital natives use their mother tongue or code meshed language, especially in social media posts.

xii. Mind reading of digital content users or judging tendencies of students can be done through text mining. Such information can be beneficial for career counsellors.

xiii. Cross comparison of mined books in the same discipline and different disciplines will initiate new interdisciplinary and cross-disciplinary research dimensions.

xiv. Cybersecurity is significant for every country, whereas Pakistan is lagging behind in this field. Text analytics can be used for prediction trends, so it finds a person whose inclinations are destructive. Terrorist activities and hate crimes will be monitored by text analytics; thus, it would mingle forensic linguistics, cybersecurity, text analytics, statistics and computation.

xv. Nowadays, emails show readymade answers or digital auto-generated cues, and they are the results of text mining. Studies can be conducted for analysis of their utility and correctness.

xvi. Text mining can be employed to explore gender biases, gender-specific hedges and other linguistic specifications.

xvii. Librarians and researchers should practically use the one-word library concept which is propounded by the current research.

5.10.3 Recommendations for the Nexus between Voyant Tools and Industry

The following aspects promote industry linkages with Voyant tools.

i. The latest Oxford and Cambridge books use a specific corpus sticker to verify the inclusion of the relevant corpus. British books benefit from British National Corpus

(BNC) and other corpora to make them authentic and harmonious to the extensive and true store of language. Voyant tools, for example, Summary and Phrases tools can analyse any corpus. New English and Urdu books must consult already built corpora, and new corpora should also be built with Voyant tools. Therefore, corpus experts will be able to work in collaboration with publishing houses.

ii. Data Visualization is an emerging trend in publishing houses. Cirrus in Voyant produces an interactive word cloud, and each chapter can be summarised with Cirrus (word cloud) and Links (KG).

iii. Now demands for works in ESP (English for Specific Purpose) and genre-focused lexicons are increasing because of super specialization in every field, so Phrases tool in Voyant can generate dictionaries/ word lists/ phrases for learners. It is said that a learner can learn any language with the most occurring 3000 vocabulary items. Voyant tool can generate this vocabulary list within no time. So, Voyant tool can be used to prepare vocabulary and Phraseology books for different domains.

iv. Intermediate book writers and PTB administration should include collocation list in intermediate English textbooks to upgrade English textbooks.

5.10.4 Recommendations for Voyant Tool Designers

Voyant is an open access code that can be taken from the given link: https://github.com/sgsinclair/Voyant/find/master It can be modified for the integration of new recommended features. With the advent of quantum computers and quantum humanities, the following dimensions can be added to upgrade Voyant tools.

i. Meanings should be visible by clicking on word cloud entities and Links. To accomplish this task, Oxford Advanced Learner’s Dictionary should be linked with this tool.

ii. Oxford Thesaurus or www.thesaurus.com website should be linked with word cloud because Pakistani students of intermediate have to select the most appropriate English synonym in the objective paper.

iii. An English to Urdu Dictionary by Shan Ul Haq Haqqi should be linked with Voyant tools, so that Pakistani students can easily understand English words in their national language. Furthermore, the demand for English to Urdu, and Urdu to

English translations will also be fulfilled for the translation industry and intermediate Punjab Board exams.

iv. Voyant does not accept uploaded data in the form of an InPage document. It would be upgraded to upload an InPage document.

v. New ways of text visualizations should be devised for better and effective comprehension.

vi. Stanford Parts of Speech (POS) tagger should be linked with Voyant tools to facilitate the voyanting process (adj. the use of Voyant tools to mine text, study, teach, and conduct research).

vii. Options of sharing on social media sites should be added on each panel of Voyant tools.

viii. The one-click print option should also be added to each skin of Voyant tools.

ix. The sentiment analysis option should be added in Voyant tools.

x. Hyperlink in the Summary tool should not be ineffective after a year; rather it should be permanent to retrieve the data. It will make the one-word library concept permanent.

5.11 Conclusion

Concluding this chapter, the current research has answered all research questions, accomplished its prime objectives and highlighted its valuable contributions in the domains of academia, pedagogy, research, industry and library science. In addition to it, the implications of five Voyant tools in academic settings have been highlighted. The current study finds some limitations and nonexistence of text mining centres in Pakistan; hence, they necessitate future recommendations regarding Voyant tools, pedagogy, the establishment of research and industry linkages with academia.

REFERENCES

Abdul-Rahman, A., Lein, J., Coles, K., Maguire, E., Meyer, M., Wynne, M., & Johnson, C. R. (2013). Rule-based visual mappings–with a case study on poetry visualization. *Computer Graphics Forum*, *32,* 381–390

Adamic, L. A., & Glance, N. (2005, August). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36-43). ACM.

Adriaans, P., & Zantinge, D. (2009). *Data mining*. New Delhi: Pearson Education.

Agrawal, R., & Psaila, G. (1995, August). Active Data Mining. In *KDD* (pp. 3-8). Menlo Park, California: American Association for Artificial Intelligence.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, *12*(1), 307-328. Menlo Park, California: AAAI Press.

Aggarwal, C. C., & Zhai, C. (2012). An introduction to text mining. In *Mining text data* (pp. 1-10). NY: Springer.

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2012). Data mining for improving textbooks. *ACM SIGKDD Explorations Newsletter*, *13*(2), 7-19.

Agrawal, R. (2013). Reimaging Textbooks Through the Data Lens. *Stanford Infoseminar*, 15.

Alexander, C. (1979). *The timeless way of building*. NY: Oxford University Press USA.

Altun, T., & Akyildiz, S. (2017). Investigating student teachers' technological pedagogical content knowledge (TPACK) levels based on some variables. *European Journal of*

Education Studies. 3(5), 467-485. Retrieved from http://oapub.org/edu/index.php/ejes/article/download/681/1916&hl

Amancio, D. R. (2015). A complex network approach to stylometry. *PloS one, 10*(8), e0136076.

Anping, H. (2005). *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.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383-398. doi:10.1007/s11423-012-9235-8

Argamon, S., Šarić, M., & 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.

Australian National University. (2016, December 7). *Research projects - Centre for digital humanities research. ANU college of arts and social sciences*. Retrieved July 27, 2017, from http://cdhr.anu.edu.au/research-projects

Avidov-Ungar, O., & Shamir-Inbal, T. (2017). ICT coordinators' TPACK-based leadership knowledge in their roles as agents of change. *Journal of Information Technology Education:* Research, 16, 169-188. Retrieved from http://www.informingscience.org/Publications/3699

Baalen, R. V. (2012). *The singing narrators of fictional lies: A close and distant reading of Dutch mendacious songs* (Master's thesis, Utrecht University, Utrecht, Netherland). Retrieved from https://dspace.library.uu.nl/handle/1874/254725

Baker, P., Hardie, A., & Mcenery, T. (2006). *Glossary of corpus linguistics* (1st ed.). Edinburgh: Edinburgh University Press.

Baker, R. S. (2014). Educational data mining: An advance for intelligent systems in education. *AI and education*, 78-82. Retrieved from http://ieeexplore.ieee.org/document/6871689/

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.

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

Banko, M., Cafarella, M. J., Michael, J., Soderland, S., Stephen, J., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* (pp. 2670–2676).

Barabási, & Lászlo, A. (2002). *Linked: The New Science of Networks*.

Barahate, S. R. (2012). Educational data mining as a trend of data mining in educational system. In *IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology (ICWET 2012) icwet (9): March* (pp. 11-16).

Barlow, M. (1996). Corpora for theory and practice. *International Journal of Corpus linguistics, 1*(1), 1–37.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental Social Psychology*. Cambridge: Cambridge University Press.

Barzen, J., & Leymann, F. (2020). Quantum humanities: a vision for quantum computing in digital humanities. SICS Software-Intensive Cyber-Physical Systems, 35(1), 153-158.

Basharat, S. U. (2004). *Teaching language skills through literature use at intermediate level* (Unpublished master's thesis). Allama Iqbal Open University, Islamabad, Pakistan.

Beaugrande, R. (2001). If I were you…: Language Standards and Corpus Data in EFL. *TESOL Quarterly*. Retrieved from http://beaugrande.bizland.com/Ifwereyou.htm

Bembenik, R., Skonieczny, L., Rybinski, H., Kryszkiewicz, M., & Niezgodka, M. (2013). *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions* (Vol. 467): Springer.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations.* Amsterdam: John Benjamins.

Berry, D. M. (2012). Introduction: Understanding the digital humanities. In *Understanding digital humanities* (pp. 1-20). Palgrave Macmillan, London.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 225–234). Frankfurt am Main: Peter Lang.

Bernardini, S. (2002). Serendipity expanded: Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis. Papers from the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July 2000* (pp. 165–182). Amsterdam: Rodopi.

Bertalanffy, L.v. (1969). *General System Theory*. New York: George Brazillier.

Berthon, P., Nairn, A., & Money, A. (2003). Through the paradigm funnel: a conceptual tool for literature analysis. *Marketing Education Review*, *13*(2), 55-66.

Bhala, R. V., & Abirami, S. (2014). Trends in word sense disambiguation. *Artificial Intelligence Review*, *42*(2), 159-171.

Bhattacharyya, D. K., & Hazarika, S. M. (2006). *Networks, data mining and artificial intelligence: trends and future directions*. New Delhi: Narosa Publishing House.

Biber, D. (1993). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, *19*(3), 531-538.

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.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on interactive presentation sessions* (pp. 69–72). Association for Computational Linguistics.

Bird, S., Klein, E., & Loper, E. (2014). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Beijing: O'Reilly.

Blackburn, S. (2016). The Oxford dictionary of philosophy. doi:10.1093/acref/9780198735304.001.0001

Blanco, R. & Zaragoza, H. (2010). Finding support sentences for entities. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. pp 339–346. ACM.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine learning research*, 993–1022,

Boitshwarelo, B. (2011). Proposing an integrated research framework for connectivism: Utilising theoretical synergies. *International Review of Research in Open and Distributed Learning*, *12*(3), 161-179.

Borgmann, A. (2000). *Holding on to reality: The nature of information at the turn of the millenium*. Chicago: The University of Chicago Press.

Botley, S., McEnery, T. & Wilson, A. (eds). (2000). *Multilingual Corpora in Teaching and Research.* Amsterdam: Rodopi.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. Routledge.

Brown, S., Fisher, S. Clements, P., Binhammer, K., Butler, T. Carter, K. Grundy, I. & Hockey, S. (1997). *SGML and the Orlando Project: Descriptive Markup for an Electronic History of Women's Writing*. *Computers and the Humanities* 31: 271–84.

Brown, S., Ruecker, S., Radzikowska, M., Milena, Patey, M., Sinclair, S., Antoniuk, J. (2009). Visualizing varieties of association in Orlando. *JDHCS*, *1*(1), 1-5. Retrieved from https://letterpress.uchicago.edu/index.php/jdhcs/article/view/7

Burch, R. (2010, September 21). Charles Sanders Peirce. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2010/ entries/peirce/

Buurma, R. S., & Heffernan, L. (2018, April 11). *Search and replace: Josephine Miles and the origins of distant reading | Modernism / Modernity print*. Retrieved January 10, 2019, from https://modernismmodernity.org/forums/posts/search-and-replace

Burdick, A., Drucker, J., Lunenefeld, P., Presner, T., & Schnapp, J. (2012). *Digital humanities* (1st ed.). MA: MIT Press.

Burdick, A. (2016). *Digital humanities*. Cambridge, Mass: MIT Press.

Burnard, L. (1988). *Report of Workshop on Text Encoding Guidelines*. *Literary and Linguistic Computing* 3: 131–3.

Burnard, L. & McEnery, T. (eds). 2000. *Rethinking language pedagogy from a corpus perspective.* Paper presented at the Third International Conference on Teaching and Language Corpora. Frankfurt am Main: Peter Lang.

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267-287. doi:10.1093/llc/17.3.267

Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, *14*(2), 83-90. Retrieved from http://www.jstor.org/stable/30207304

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

Califf, M. E., & Mooney, R. J. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference* (pp. 328–334).

Campbell, D. & Fiskel, D. (1959). Convergent and discriminant validation by the multitrait-multimethod Matrix. *Psychological Bulletin*, 56: 81-105.

Capurro, R. (2006). Towards an ontological foundation of information ethics. *Ethics and Information Technology*, *8*(4), 175-186. doi:10.1007/s10676-006-9108-0

Capurro, R. (2010). Digital hermeneutics: an outline. *AI & SOCIETY*, *25*(1), 35-42. doi:10.1007/s00146-009-0255-9

Carter, R. & M. McCarthy (1995a). Grammar and the Spoken Language. *Applied Linguistics* 16(2): 141-158.

Carter, R. & M. McCarthy (1995b). Spoken grammar: what is it and how can we teach it? *ELT Journal* 49(3): 207-217.

Castellà, Q., & Sutton, C. (2014). Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. *International World Wide Web Conference Committee (IW3C2)*. Retrieved from homepages.inf.ed.ac.uk/ scutton/ publications/castella14word.pdf

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Studies in Computational Intelligence, 62*, 183-221.

Centre for Language Engineering. (2016, January 19). *Center for Language Engineering.* Retrieved February 23, 2018, from http://www.cle.org.pk/

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010, May). Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 4, No. 1). *Icwsm, 10*(10-17), 30.

Chakraborty, T. (2012). Authorship identification in Bengali literature: A comparative analysis. *Proceedings of COLING 2012: Demonstration Papers.* pp. 41-50

Chalker, S. (1994). *Pedagogical grammar: Principles and problems.* Hemel Hempstead, Prentice Hall: 31-44.

Chambers, F. (1997). Seeking consensus in coursebook evaluation. *ELT Journal* 51 (1): 29–35.

Chan, S. (2018, August 7). *14 Million Visitors to U.S. Face Social Media Screening.* Retrieved January 4, 2019, from https://www.nytimes.com/2018/03/30/world/ americas/travelers-visa-social-media.html

Chattamvelli, R. (2016). *Data mining methods*. Oxford: Alpha Science International Ltd.

Chaturvedi, M., Gannod, L, G., Mandell, L., Armstrong, H., & Hodgson, E. (2012). Myopia: A visualization tool in support of close reading. *Digital Humanities*, *2*.

Chen, J., Li, Q., Wang, L., & Jia, W. (2004). *Automatically generating an e-textbook on the web.* Paper presented at the International conference on advances in web based learning (pp. 35–42).

Chen, Y. (2017). Dictionary use for collocation production and retention: A CALL-based study. International Journal of Lexicography, 30(2), 225-251.

Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2008). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics, 30*(2), 236-252.

Choi, H. Y., & Chon, Y. V. (2012). A corpus-based analysis of collocations in tenth-grade high school English textbooks. *Multimedia Assisted Language Learning, 15*(2), 41-73.

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.

Churchland, P. S., & Sejnowski, T. J. (1992). *The Computational Brain.* MIT Press.

Clement, T., Plaisant, C., & Vuillemot, R. (2008). *The story of one: Humanity scholarship with visualization and text analysis.* Retrieved from http://www.cs.umd.edu/hcil/trs/2008-33/2008-33.pdf

Clement, T. (2012). Distant listening or playing visualizations pleasantly with the eyes and ears. *Digital Studies*, *3*(2).

Cobb, T., Greaves, C., & Horst, M. (2000). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la Didactique des Langues Secondes*. Montréal: Éditions Logique. Online: http://132.208.224.131/ResearchWeb/ consulted: 25.11.2003.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. Routledge.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, C. (2006). Docuburst: Document content visualization using language structure. In *Proceedings of IEEE Symposium on Information Visualization, Poster Compendium*.

Collins, C., Viegas, F. B., & Wattenberg, M. (2009). Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology, 2009* (pp. 91–98). IEEE Symposium.

Connor, U. & Upton, T. (eds). (2004). *Applied Corpus Linguistics: A Multidimensional Perspective.* Amsterdam: Rodopi.

Conrad, S. (2000). Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly 34*(3): 548-560.

Conrad, S. & D. Biber. (2000). Adverbial marking of stance in speech and writing. *Evaluation in text: Authorial stance and the construction of discourse*. S. Hunston & G. Thompson. Oxford, Oxford University Press: 56-73.

Corbett, A. T., & Anderson, J. R. (1995). *Knowledge tracing: Modeling the acquisition of procedural knowledge*. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.

Crane, G. (2006). What Do You Do with a Million Books? *D-Lib Magazine*, *12*(3). doi:10.1045/march2006-crane

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cristobal, R., & Sebastian, V. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Crystal, D. (2001). *Language and the internet*. Cambridge, England: Cambridge University Press.

Cunningsworth, A. (1984). *Evaluating and Selecting EFL Teaching Materials.* Oxford: Heinemann.

Cunningsworth, A. (1995). *Choosing Your Coursebook.* Oxford: Heinemann.

da Silva, E. B., Orenha-Ottaiano, A., & Babini, M. (2017). < b> Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora. *Acta Scientiarum. Language and Culture,* 39(4), 345-353.

Daily Times. (2016, December 15). Interactive whiteboards at KP schools. *Daily Times*: [Peshawar]. Retrieved from http://dailytimes.com.pk/khyber-pakhtunkhwa/15-Dec-16/interactive-whiteboards-at-kp-schools

DeCamp, P., Frid-Jimenez, A., Guiness, J., & Roy, D. (2005). Gist icons: Seeing meaning in large bodies of literature. In *Proceedings of IEEE Symposium on Information Visualization*, IEEE.

Dellar, H. & D. Hocking. (2000). *Innovations*. Hove, Language Teaching Publications.

Denzin, N. (1970). *The research act in sociology.* Chicago: Aldine.

Denzin, N. (1978). *The research act: A theoretical introduction to sociological methods.* New York: McGraw-Hill.

Digital Preservation Management. (2013). *Timeline: Digital technology and preservation*. Retrieved July 25, 2017, from http://www.dpworkshop.org/dpm-eng/timeline/viewall.html

Dixon, D. (2012). Analysis tool or research methodology: Is there an epistemology for patterns? In *Understanding digital humanities*. Houndmills, Basingstoke: Palgrave Macmillan.

DMK Force. (2008). *Data presentation architecture-Something to Learn (DMK).* Retrieved December 22, 2018, from https://sites.google.com/site/somethingtolearndmk/dadata-presentation-architecture

Don B., C. Plaisant., A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., & Shneiderman, B. (2007). Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on information and knowledge management* (pp. 213–222). ACM.

Drucker, J. (2009). *How do you define Humanities Computing / Digital Humanities?* - Taporwiki. Retrieved May 19, 2018, from http://www.artsrn.ualberta.ca/taporwiki/index.php/How_do_you_define_Humanities_Computing_/_Digital_Humanities%3F

Du, H. (2010). *Data mining techniques and applications: An introduction*. Delhi, India: Cengage Learning.

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

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

Edison, T. A. (2001). *BrainyQuote.* Retrieved May 6, 2018, from https://www.brainyquote.com/quotes/thomas_a_edison_136633

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS), 48*.

Ellegård, A. (1962). *A Statistical Method for Determining Authorship: The Junius Letters 1769–1772*. Gothenburg: Gothenburg Studies in English.

Enns, J. T. (2005). Perception, Gestalt Principles of In Lynn Nadel (ed.). *Encyclopedia of Cognitive Science*. Hoboken, NJ: Wiley.

ETANA Electronic Tools and Ancient Near East Archives. (2019). *Center for computer analysis of texts (CCAT) public archive | ETANA*. Retrieved February 17, 2019, from http://www.etana.org/node/4495

European Association for Digital Humanities. (2017). Projects | EADH - The European Association for Digital Humanities. Retrieved July 27, 2017, from https://eadh.org/projects#block-views-project-list-block-1

Everitt, B. (2009). *Chance rules: An informal guide to probability, risk and statistics*. Springer Science & Business Media.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1535–1545).

Fang, L., Sarma, A. D., Yu, C., & Bohannon, P. (2011). Rex: explaining relationships between entity pairs. *Proceedings of the VLDB Endowment*. 5(3):241–252.

Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy. (1996). *An outline of the steps of the KDD process* [Graph]. Retrieved from http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Fayyad, U. M., Shapiro, G. P., Smyth, P., & Uthurusamy, R. (Eds.). (1996). *Advances in knowledge discovery and data mining*. Menlo Park, Canada: AAAI Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a unifying framework. In *KDD-96 Proceedings*. (Vol. 96, pp. 82-88).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, November). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fekete, J. D., & Dufournaud, N. (2000). Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *Proceedings of the fifth ACM conference on Digital libraries* (pp. 47–55). ACM.

Fiala, D. (2007). *Web mining methods for the detection of authoritative sources* (Unpublished doctoral dissertation). University Louis Pasteur Strasbourg I, Strasbourg, France.

Fischer, M. (1971). The KWIC index concept; A retrospective view. In *Key papers in information science*. Washington DC: American Society for Information Science.

Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Ed.), *The Power of CALL* (pp. 97–113). Houston, TX: Athelstan.

Fox, G. (1998). Using corpus data in the classroom. *Materials Development in Language Teaching*. B. Tomlinson. Cambridge, Cambridge University Press: 25-43.

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.

Gacitua, R., & Sawyer, P. (2008, May). Ensemble methods for ontology learning-an empirical experiment to evaluate combinations of concept acquisition techniques. In *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)* (pp. 328-333). IEEE.

Gamerman, E. (2015). Data miners dig into ‘Watchman’. *Wall Street Journal*. Page. 107, Gavioli. L. (2000). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (Ed.), *Learning with Corpora* (pp. 108–137). Houston, TX: Athelstan / Bologna: CLUEB.

Gee, J. P. (2004). *Situated Language and Learning: A Critique of Traditional Schooling*. Florence: Taylor and Francis.

Gelman, A., Carlin, J.B, Stern H. S., & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall.

George, P., Vangelis, K., Anastasia, K., Georgios, P., & Constantine, S. D. (2009). Semi-automated ontology learning: The boemie approach. In *Proceedings of the first ESWC workshop on inductive reasoning and machine learning on the semantic web.* Heraklion, Greece.

Gibbs, F. (2016). Digital humanities definitions by type. In *Defining digital humanities: A reader* (pp. 289-300). London: Routledge.

Gillani, S. A. (2015). *From text mining to knowledge mining: An integrated framework of concept extraction and categorization for domain ontology* (Unpublished doctoral dissertation). Corvinus University of Budapest, Budapest, Hungary.

GitHub. (2014). *Semantic word cloud visualization- Description.* Retrieved January 23, 2018, from http://wordcloud.cs.arizona.edu/description.html

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. *Phraseology in foreign language learning and teaching*, 223-243. Amsterdam: John Benjamins.

Graham, S., Milligan, I., & Weingart, S. (2013). *Voyant Tools | The Historian's Macroscope: Big Digital History*. Retrieved August 11, 2017, from http://www. The macroscope. org/?page_id=639

Granger, S., Hung, J. & Petch-Tyson, S. (eds). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Green, B. A. (2018). Corpora in Language Learning. *The TESOL Encyclopedia of English Language Teaching*, 1-9.

Grier, M. (2014, October 29). Voyant Tools | DH and Literary Studies. Retrieved March 25, 2017, from http://modernist-magazines.org/?q=taxonomy/term/782

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition, 5*(2), 199-220.

Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In IEEE international conference on management of innovation and technology (pp. 715–719).

Ham, F. V., Wattenberg, M., & Viegas, F. B. (2009). Mapping text with phrase nets. In *Visualization and Computer Graphics* (p. 1169–1176). IEEE Transactions.

Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In World conference on open learning and distance education, Hong Kong.

Hammouda, K., & Kamel, M. (2010). Data mining in e-learning. In *E-learning networked environments and architectures: A knowledge processing perspective* (pp. 374-404).

Han, J., Kamber, M., & Pei, J. (2012). *Data mining connections and techniques* (3rd ed.). MA: Morgan Kaufmann.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Massachusetts: MIT Press.

Hanna, M. (2004). Data mining in the e-learning domain. *Computers & Education Journal*, 42(3), 267–287.

Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla *Advances and Innovations in Systems, Computing Sciences and Software Engineering* (pp. 121-126): Springer.

Hayes, S. (2008). Toolkit: Wordle. *Voices from the Middle, 16*(2), 66-68.

Healey, A. (1989). *The corpus of the dictionary of old English: Its delimitation, compilation and application*. Paper presented at the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary. Oxford, September, 1989.

Hearst, M. A. (1999, June). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). Association for Computational Linguistics.

Heidegger, M. (1954). *Die Frage nach der Technik (The question concerning technology)* (1st ed.). Retrieved from http://www.psyp.org/ question_ concerning _technology.pdf

Heidegger, M. (2010). *Being and time*. Suny Press.

Heydenreich, L. H. (2019, January 25). Leonardo da Vinci | biography, art, & facts. Retrieved February 2, 2019, from https://www.britannica.com/biography/Leonardo-da-Vinci

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Oklahoma: StatSoft.

Hockey, S. & I. Marriott (1979a). *The Oxford Concordance Project (OCP) – Part 1. ALLC Bulletin* 7: 35–43.

Hockey, S. & I. Marriott (1979b). *The Oxford Concordance Project (OCP) – Part 2. ALLC Bulletin* 7: 155–64.

Hockey, S. & I. Marriott (1979c). *The Oxford Concordance Project (OCP) – Part 3. ALLC Bulletin* 7: 268–75.

Hockey, S. & I. Marriott (1980). *The Oxford Concordance Project (OCP) – Part 4. ALLC Bulletin* 8: 28–35.

Hockey, S. (2001). *Concordance programs for corpus linguistics* (pp. 76-97). University of Michigan Press.

Hofstee, D. E. (2006). *Constructing a good dissertation: A practical guide to finishing a Master's, MBA or PhD on schedule* (1st ed.). Retrieved from http://www.exactica.co.za/book-chapters.php

Holton, W. Coffeen (2021, August 17). quantum computer. Encyclopedia Britannica. https://www.britannica.com/technology/quantum-computer

Hotho, A., Nürnberger, A., & Paaß, G. (2005). *A brief survey of text mining.* Paper presented at the Ldv Forum.

Hsu, J. (2008). Role of the multi-word lexical units in current EFL/ESL textbooks. *US-China Foreign Language,6* (7), 27-39.

Hutchins, J. (1999). The historical development of machine translation. *Submission for the Degree of Doctor of Philosophy by Publication at the University of East Anglia url: http://www. hutchinsweb. me. uk/PhD-1999. pdf,(Accessed: 23/10/2011).*

Hunston, S. & G. Francis. (1998). Verbs observed: a corpus-driven pedagogic grammar. *Applied Linguistics* 19(1): 45-72.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: CUP.

Huntston, S. (2010). The usefulness of corpus-based descriptions of English for learners: The case of relative frequency: Collecting, Annotating and exploiting a corpus of textbook material. In *Corpora and language teaching* (pp. 179-202). Amsterdam: John Benjamins.

Hussain, M. N. (2009). *Gender representation in English language textbooks at HSSC level: An analytical study* (Unpublished master's thesis). International Islamic University, Islamabad, Pakistan.

Hwang, M., Choi, D., Ko, B., Choi, J., & Kim, P. (2011). An automatic method for wordnet concept enrichment using Wikipedia titles. In *Reliable and Autonomous Computational Science* (pp. 347-365). Springer, Basel.

Ibsen, H. (1890). *The Wild Duck: A Drama in Five Acts*. Christiania, Norway.

Ingram, A. L. (1999). Using web server logs in evaluating instructional web sites. *Journal of educational technology systems*, *28*(2), 137-157.

Iqbal, A. M. (1924). (Bang-e-Dra-007) Aik Pahar Aur Gulehri. Retrieved March 2, 2019, from http://iqbalurdu.blogspot.com/2011/02/bang-e-dra-7-aik-pahar-aur-gulehri.html

Iqbal, M. (2011). *An evaluative study of English textbooks* (Unpublished master's thesis). Allama Iqbal Open University, Islamabad, Pakistan.

Jacobs, G. M., & Ball, J. (1996). An investigation of the structure of group activities in ELT coursebooks. *ELT Journal*, *50*(2), 99-107.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)* (pp. 1–21).

Jeffreys, H. (1973). *Scientific inference* (3rd ed.). Cambridge, England: Cambridge University Press.

JISC. (2006). *Text mining.* Retrieved May 18, 2017, from http://jisc.ac.UK/publications

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769. doi:10.1016/j.poetic.2013.08.005

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).

Johnson, S., Arago, S., Shaik, N., & Palma-Rivas, N. (2000). Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments. *Journal of Interactive Learning Research, 11*(1), 29–49.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher, 33*(7), 14-26.

Kadushin, C., Hecht, S., Sasson, T., & Saxe, L. (2008). Triangulation and Mixed Methods Designs: Practicing What We Preach in the Evaluation of an Israel Experience Educational Program. *Field Methods*, *20*(1), 46-65. doi:10.1177/1525822x07307426

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

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms.* (2nd ed.). New Jersey: IEEE Press.

Karegar, M., Isazadeh, A., Fartash, F., Saderi, T., & Navin, A. H. (2008). Data-mining by probability-based patterns. In *Proceedings of the 30th international conference on information technology interfaces*. pp. 353–360.

Kasneci, G., Suchanek, F. M., Ifrim, G., Ramanath, M., & Weikum, G. (2008, April). Naga: Searching and ranking knowledge. In *Data Engineering, 2008.* ICDE 2008. IEEE 24th International Conference. pp. 953-962. IEEE.

Kellsey, D., & Taylor, A. (2016). *The LearningWheel: A model of digital pedagogy* (1st ed.). England: Critical Publishing.

Kemman, M. (2016, November 24). *A Republic of Emails: What are the contents?* Retrieved March 25, 2017, from http://www.maxkemman.nl/2016/11/a-republic-of-emails-what-are-the-contents/

Kenny, A. (1992). *Computers and the Humanities*. Ninth British Library Research Lecture. London: British Library.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, *31*(2), 91-113.

Kirschenbaum, M. (2016). What is Digital Humanities and what's it doing in English departments? In *Debates in the digital humanities 2016* (pp. 3-11). London: University of Minnesota Press.

Kleinberg, J., Papadimitriou, C., & Raghavan, P. (1998). A microeconomic view of data mining. *Data mining and knowledge discovery*, *2*(1), 311-324. Retrieved from https://www.cs.cornell.edu/home/kleinber/dmkd98-seg.pdf

Klosgen, W., & Zytkow, J. (2002). *Handbook of data mining and knowledge discovery.* New York: Oxford University Press.

Knobloch, E. (2004). Mathesis – The Idea of a Universal Science. In R. Seising, M. Folkerts & U. Hashagen (Eds), *Form, Zahl, Ordnung. Studien zur Wissenschafts- und Technianneschichte.* Stuttgart: Steiner, 77–90.

Knorr-Cetina, K. D. (1992). Science as practice and culture. In A. Pickering (Ed.), *The couch, the cathedral, and the laboratory: On the relationship between experiment and laboratory in science.* Chicago: Chicago University Press, 113–38.

Knowledge Platform. (2017). *Blended learning solutions. Asia. Adaptive learning*. Retrieved May 20, 2017, from https://www.knowledgeplatform.com/

Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education. 9*(1), 60-70.

Komprise. (2017). *What Is Data Archiving? Data Archiving Definition.* Retrieved August 8, 2017, from https://www.komprise.com/glossary_terms/data-archiving/

Kongthon, A. (2004). *A text mining framework for discovering technological intelligence to support science and technology management* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, *59*(4), 322-32. http://dx.doi.org/ 10.1093/elt/cci061

Kosmos, H. (2014). *Alexander von Humboldt-foundation- A brief history of digital humanities.* Retrieved May 20, 2017, from https://www.humboldt-foundation.de/web/kosmos-cover-story-102-5.html

Koutri, M., Avouris, N., & Daskalaki, S. (2004). A survey on web usage mining techniques for web-based adaptive hypermedia systems. Adaptable and Adaptative Hypermedia Systems. *Idea Inc. Hershey*.

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

Kwary, D. A. (2018). A corpus and a concordancer of academic journal articles. *Data in brief*, *16*, 94-100.

Lankshear, C., & Knobel, M. (2011). *New literacies: Everyday practices and classroom learning*. Maidenhead: Open University Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 1–23). London: Longman.

Li, D., & Wang, S. (2005). Concepts, principles and applications of spatial data mining and knowledge discovery. *ISSTM*, *36*(25), 1-12.

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

Liu, A. Y. (2012). *Where is cultural criticism in the digital humanities?* (pp. 490-509). eScholarship, University of California.

Littlejohn, A. (1998). The analysis of language teaching materials: Inside the Trojan Horse. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp.179-211), Cambridge: Cambridge University Press.

Lloyd, V. (2016, August 8). Turning data into insight | theHRD. Retrieved January 30, 2021, from https://www.thehrdirector.com/features/hr-in-business/turning-data-into-insight/

Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015). ConcentriCloud: Word Cloud Visualization for Multiple Text Documents. *19th International Conference on Information Visualisation*, 114-120. Retrieved from https://puma.ub.uni stuttgart.de/bibtex/2fb567abbd01d10cb6efc8a59105f8114/visus

Louw, B. (2010). Collocation as instrumentation for meaning: A scientific fact. In *Literary education and digital learning* (pp. 79-101). Zimbabwe: IGI Global.

LSP. (2016). *Learn Smart Pakistan | Pakistan's Online Learning Challenge.* Retrieved May 20, 2017, from http://www.learnsmartpakistan.org/Lsp/Index

Maedche, A., & Staab, S. (2004). Ontology learning *Handbook on ontologies.* (pp. 173-190): Springer.

Malpas, J., & Gander, H. H. (Eds.). (2014). *The Routledge companion to hermeneutics*. Routledge.

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.

Mathiak, B., & Eckstein, S. (2004, September). Five steps to text mining in biomedical literature. In *Proceedings of the second European workshop on data mining and text mining in bioinformatics* (Vol. 24, pp. 47-50).

McCarthy, P. M., & Boonthum-Denecke, C. (Eds.). (2011). *Applied Natural Language Processing: Identification, Investigation and Resolution: Identification, Investigation and Resolution*. Pennsylvania: International Science Reference.

McCarty, W. (2005). *Humanities Computing*. New York: Palgrave Macmillan.

McCarty, W. (2009). Attending from and to the machine. *Inaugural lecture. Kings College London*.

McCurdy, N., Lein, J., Coles, K., & Mayer, M. (2016). Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on visualization and computer graphics*, *221*, 439-448.

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

McNaught, C., & Lam, P. (2010). Using Wordle as a supplementary research tool. *The Qualitative Report Volume*, *15*(3). Retrieved from http://nsuworks.nova.edu/cgi/viewcontent.cgi?article=1167&context=tqr

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.

Meunier, F & Gouverneur, C. (2010). New types of corpora for new educational challenges: Collecting, Annotating and exploiting a corpus of textbook material. In *Corpora and language teaching* (pp. 141-156). Amsterdam: John Benjamins.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *science, 331*(6014), 176-182.

Mihalcea, R. (2007, April). Using wikipedia for automatic word sense disambiguation. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (pp.196-203).

Miley, F., & Read, A. (2011). Using word clouds to develop proactive learners. *Journal of the Scholarship of Teaching and Learning*, *11*(2), 91-110. Retrieved from https://www.iupui.edu/~josotl/archive/vol_11/no_2/v11n2miley.pdf

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications* (1st ed.). Elsevier.

Mitra, S., & Acharya, T. (2003). *Data mining: multimedia, soft computing, and bioinformatics*. New Jersey: John Wiley & Sons, Inc.

Mohs, C. (2013, April 22). *Voyant Tools | DH and Literary Studies.* Retrieved March 25, 2017, from http://modernist-magazines.org/?q=node/687

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, *53*(4), 730-741. doi:10.1016/j.dss.2012.05.030

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Moretti, F. (2013). *Distant reading*. London: Verso.

Morton, A. Q. (1965). *The Authorship of the Pauline Epistles: A Scientific Solution*. Saskatoon: University of Saskatchewan.

Mosteller, F. & D. L. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Mukherjee, J., & Rohrbach, J. M. (2006). Rethinking applied corpus linguistics from a languagepedagogical 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 am Main: Peter Lang.

Muralidharan, A., & Hearst, M. A. (2012). Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing*, *28*(2), 283-295. doi:10.1093/llc/fqs044

Nahm, U. Y. (2001). *Text mining with information extraction: Mining prediction rules from unstructured text*. AL: University of Texas.

Naser, M. A. (2012). *Analyzing the effectiveness of poetry in teaching English grammar at the intermediate level* (Unpublished master's thesis). Allama Iqbal Open University, Islamabad, Pakistan.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. http://dx.doi.org/10.1093/applin/24.2.223

Neisser, U. (2005). Pattern Recognition. In D. A. Balota & E. J. Marsh (eds.), *Cognitive Psychology: Key Readings*. New York: Psychology Press.

Nguyen, T. T., Chang, K., & Hui, S. C. (2011). Word cloud model for text categorization. In *11th IEEE International Conference on Data Mining* (pp. 487-496). Singapore: IEEE.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11-33.

Nie, X., & Zhou, J. (2008, April). A domain adaptive ontology learning framework. In *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on* (pp. 1726-1729). IEEE.

Nilakant, K., & Mitrovic, A. (2005). Application of data mining in constraint-based intelligent tutoring systems. In *Proceedings of the artificial intelligence in education, AIED* (pp. 896–898).

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books.

Norris, D., Baer, L., Leonard, J., Pugliese, L., & Lefrere, P. (2008). Action analytics. *Educause Review, 43*(1), 42-67.

O’Keeffe, A., McCarthy, M. & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: CUP.

Olsen &Wendy. (2004). Triangulation in Social Research: Qualitative and Quantitative Method Can Really be Mixed. In Holborn M. (Ed.), *Developments in Sociology*. Ormskirk: Causeway Press.

O’Neill, R. (1993). Are textbooks symptoms of disease? *Practical English Teaching* 14(1): 11–13.

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

Oracle. (2017). *What Is Data Mining?* Retrieved July 25, 2017, from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCJ

Oxford University Press. (2021). Oxford advanced learner's dictionary. Retrieved April 7, 2021, from https://www.oxfordlearnersdictionaries.com/definition/english/cirrus?q=Cirrus

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the LREC*, 1320–1326.

Palace, B. (1996). Data mining. technology note. Prepared for Management 274A. *Technology Note, Anderson Graduate School of Business at UCLA.*

Paley, W. B. (2002). *Textarc: Showing word frequency and distribution in text.* Poster presented at IEEE Symposium on Information Visualization.

Paltridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes* 21: 125–143.

Parrish, S. M. (1962). *Problems in the Making of Computer Concordances*. *Studies in Bibliography* 15: 1–14

Parsehub. (2019). *Free web scraping - Download the most powerful web scraper | ParseHub*. Retrieved February 16, 2019, from https://www.parsehub.com/

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, *8*(3), 489-508.

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.

Peña-Ayala, A. (2014a). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432-1462. doi:10.1016/j.eswa.2013.08.042

Peña-Ayala, A. (2014b). *Educational data mining: Applications and trends*. Cham: Springer International Publishing.

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making*, 7(4), 639–682.

Pirrò, G. (2015, October). Explaining and suggesting relatedness in knowledge graphs. In *International Semantic Web Conference* (pp. 622-639). Springer, Cham.

Plato. (2014). *Theaetetus*. Trajectory, Inc.

Ponzetto, S. P., & Navigli, R. (2010, July). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1522-1531).

Powers, R. (1996). *Galatea 2.2*. New York: Harper Perennial.

Prensky, M. (2001). Digital Natives, Digital Immigrants. *On the horizon*, *9*(5), 1-6. doi:10.4135/9781483387765.n6

Prensky, M. (2012). *From digital natives to digital wisdom: Hopeful essays for 21st century learning*. Thousand Oaks, CA: Corwin.

Price-Wilkin, J. (1994). *Using the World Wide Web to Deliver Complex Electronic Documents: Implications for Libraries*. *The Public-Access Computer Systems Review* 5:5–21. Retrieved July 21, 2017, from http://jpw.umdl.umich.edu/pubs/yale.html.

Prieur, C., Cardon, D., Beuscart, J. S., Pissard, N., & Pons, P. (2008). The strength of weak cooperation: A case study on flickr. *arXiv preprint arXiv:0802.2317*.

Punjab Information Technology Board Government of the Punjab. (2017, May 10). e.Learn.Punjab. *Express* [Islamabad], p. ix.

Radzikowska, M., Ruecker, S., & Sinclair, S. (2011). *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-prospect Browsing*. Ashgate Publishing Group.

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.

Ramsay, S. (2014). The hermeneutics of screwing around; or what you do with a million books. In *Pastplay: Teaching and learning history with technology*. MI: University of Michigan Press.

Ramsden, A., & Bate, A. (2008). *Using word clouds in teaching and learning.* Retrieved January 24, 2018, from http://opus.bath.ac.uk/474/1/using%2520 word%2520clouds%2520in%2520teaching%2520and%2520learning.pdf

Ranalli, J. M. (2003). *ELT coursebooks in the age of corpus linguistics: constraints and possibilities*. Retrieved from University of Birmingham website: https://www.birmingham.ac.uk/Documents/collegeartslaw/cels/essays/corpuslinguistics/ Ranalli6.pdf

Ratner, B. (2017). *Statistical and machine learning data mining: Techniques for better predictive modeling* (2nd ed.). USA: CRC Press.

Readings, B. (1996). *The university in ruins*. MA: Harvard University Press.

Rech, J., Feldmann, R., & Ras, E. (2012). Knowledge patterns. In *Organizational Learning and Knowledge: Concepts, Methodologies, Tools and Applications* (pp. 578-586). IGI Global.

Rieder, B., & Rohle, T. (2012). Digital methods: Five challenges. In *Understanding digital humanities* (pp. 67-84). Houndmills, Basingstoke: Palgrave Macmillan.

Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, *18*(2), 209-219. doi:10.1093/llc/18.2.209

Rockwell, G. (2016). Is humanities computing an academic discipline? In *Defining digital humanities: A reader* (pp. 13-34). London: Routledge.

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., & Bra, P. D. (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14(5), 425–464.

Romero, C., Ventura, S., Delgado, J. A., & De Bra, P. (2007, September). Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In *European conference on technology enhanced learning* (pp. 292-306). Springer, Berlin, Heidelberg.

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.

Romero, C., & Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27. doi:10.1002/widm.1075

Romero, C., Romero, J. R., & Ventura, S. (2013). *Educational data mining: Applications and trends* (pp. 29-64). Cham, NY: Springer International Publishing.

Russell, D. B. (1965). *COCOA-A Word-Count and Concordance Generator*. Retrieved March 14, 2017, from http://www.chilton-computing.org.uk/acl/applications/cocoa/p001.htm

Russell, D. B. (1967). *COCOA - A Word Count and Concordance Generator for Atlas*. Chilton: Atlas Computer Laboratory.

Russell, T. S. (2017). *Finding the formula: formulaic language use in Hong Kong primary school English textbooks* (Doctoral dissertation, University of Birmingham).

Rychlý, P., & Kovář, V. (2007). Displaying Bidirectional Text Concordances in KWIC format. Czech Republic: Masaryk University

Ryle, G. (1945). Knowing how and knowing that: The presidential address. *Proceedings of the Aristotelian Society*. New Series, 46 (1945–1946): 1–16.

Salkind, N. J. (2010). *Encyclopedia of research design*. California, CA: SAGE.

Saunders, M., Lewis, P., & Thornhill, A. (2012). *Research methods for business students* (6th ed.). Pearson Education Limited.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (p. 44–49). Manchester, UK.

Schöch, C., & Schoech, C. (2012). Topic modeling genre: An exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*, *11*(2). Retrieved from http://digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Schreibman, S. (2013). The Digital Humanities and Humanities Computing: An Introduction. *A companion to digital humanities* (pp. 15-39). Malden, MA: Blackwell Pub.

Scott, M. (1996). *WordSmith Tools*. Ver. 3.0. Oxford: Oxford University Press.

Scrivner, O., & Davis, J. (2017). Interactive Text Mining Suite: Data visualization for literary studies. *Visualization method,* 29-38. Retrieved from http://ceur-ws.org/Vol-1786/scrivner.pdf

Seidlhofer, B. (2000). Operationalizing intertextuality: Using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 207–223). Frankfurt am Main: Peter Lang.

Seretan, V., Nerima, L., & Wehrli, E. (2004). A Tool for Multi-Word CoUocation Extraction and Visualization in MultUingual Corpora In *Proceedings of the 11th EURALEX International Congress*. Universite´ de Bretagne-Sud, Faculte´ des lettres et des sciences humaines, 2004. pp. 755-766.

Seyfert, M., & Viola, I. (2017). Dynamic word clouds. *Proceedings of the 33rd Spring Conference on Computer Graphics - SCCG '17*. doi:10.1145/ 3154353.3154358

Shalizi, C. R., Shalizi, K. L., & Crutchfiel, J. P. (2002). Pattern discovery in time series, part I: Theory, algorithm, analysis, and convergence. *Journal of Machine Learning Research*. Retrieved from https://www.santafe.edu/media/workingpapers/02-10-060.ps.gz

Shannon, C. E. (2009). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi:10.1109/9780470544242.ch1

Shapin, S. & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life.* Princeton, NJ: Princeton University Press.

Shaw, R. (2012, September 20). *Text mining as a research tool* [ppt]. Retrieved from https://aeshin.org/textmining/

Shaw, B. (1930). The Collected Works of Bernard Shaw: *The intelligent woman's guide to socialism and capitalism* (Vol. 20). WH Wise.

Sheldon, L. (1988). Evaluating ELT textbooks and materials. *ELT Journal* 42(4): 237–246.

Shin, D., & Nation, P. (2007). Beyond single words: The most frequent collocations in spoken English. *ELT journal*, *62*(4), 339-348.

Shu-Hsien, L., Pei-Hui, C., & Pei-Yuan, H. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12), 11303–11311.

Siegmund, D. O. (2014). Probability theory | mathematics | Britannica.com. In *Encyclopedia Britannica*. Retrieved June 14, 2017, from https://www.britannica.com/topic/probability-theory

Simpson, Z. B. (2000, May). *Project Guttenberg Vocabulary Analysis.* Retrieved August 18, 2017, from http://www.mine-control.com/zack/guttenberg/

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In Carter, R. & M. Mc-Carthy (eds.) Vocabulary and language teaching, 140–158. Harlow: Longman.

Sinclair, J.M. (1996). The search for units of meaning. *Textus, 9*(1), 75–106.

Sinclair, J. M. (Ed.). (2004). *How to use corpora in language teaching* (Vol. 12). Amsterdam: John Benjamins Publishing.

Sinclair, J. M. (2007). Collocation reviewed. *Manuscript.* Tuscan Word Centre, Italy.

Sinclair, J. McH. (1987). Collocation: A progress report In R. Steele & T. Threadgold (eds): *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam: John Benjamins. pp. 319–331.

Sinclair, S., & Rockwell, G. (2015a). *Principles of Voyant Tools | Voyant Tools Documentation.* Retrieved May 29, 2017, from http://DOCS.VOYANT-TOOLS.ORG/CONTEXT/ PRINCIPLES/

Sinclair, S., & Rockwell, G. (2015b). *Examples Gallery | Voyant Tools Documentation.* Retrieved May 29, 2017, from http://docs.voyant-tools.org/about/examples-gallery/

Sinclair, S., & Rockwell, G. (2017). Voyant tools help. Retrieved April 9, 2021, from https://voyant-tools.org/docs/#!/guide/gallery

Sion, A. (2010). Talmudic Hermeneutics. *Logic in Religious Discourse*, 104. Ontos Verlag

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. (2017). Representation and processing of multi-word expressions in the brain. *Brain and language*, *175*, 111-122.

Skills You Need. (2018). *What is theory?* Retrieved August 10, 2018, from https://www.skillsyouneed.com/learn/theory.html

Smith, J. B. (1978). Computer criticism. *Style Fayetteville, Ark.*, *12*(4), 326-356.

Smith, M. & Kleine, P. (1986). Qualitative Research and Evaluation: Triangulation and Multimethods Reconsidered. In Williams D. (Ed.). *Naturalistic Evaluation* (New Directions for Program Evaluation). San Francisco: Jossey-Bass.

Soderland, S. (1999). Learning information extraction rules for semistructured and free text. *Machine Learning*, *34*(1-3), 233–272.

Stamatatos, E., Fakotakis, N., & 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.

Stephen, H. (2015, October 28). Google books: A complex and controversial experiment. *New York times* [New York]. Retrieved from https://www.nytimes.com/2015/10/29/ arts/international/google-books-a-complex-and-controversial-experiment.html

Sundberg, D., & Nilsson, J. (2018). *Papa Revisited: A Corpus-Stylistic Perspective on the Style and Gender Representation of Ernest Hemingway's Fiction.* (Master's thesis, Linnaeus University, Sweden). Retrieved 6th January 2019 from http://lnu.diva-portal.org/smash/record.jsf?pid=diva2%3A1175700&dswid=-4751

Surbhi, S. (2016, December 8). *Difference Between Research Method and Research Methodology (with Comparison Chart)- Key Differences.* Retrieved August 9, 2017, from http://keydifferences.com/difference-between-research-method-and-research-methodology.html

Swales, J. M. (1995). The role of the textbook in EAP writing research. *English for Specific Purposes 14*(1): 3–18.

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).

Tang, Z., & MacLennan, J. (2005). *Data mining with SQL Server 2005*. Indianapolis, Indiana: J. Wiley.

TAPoR, McMaster University, University of Alberta. (2015). *TAPoR.* Retrieved March 20, 2017, from http://tapor.ca/home

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 Dawn. (2019, September 5). BISEs across Punjab announce Inter results. *The Dawn* [Lahore]. Retrieved from https://www.dawn.com/news/1503542

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

The BioText Project. (2016, August 31). *The BioText Project.* Retrieved March 20, 2017, from http://biotext.berkeley.edu/

The National Centre for Text Mining. (2016, July 25). *National Centre for Text Mining — Text Mining Tools and Text Mining Services*. Retrieved March 20, 2017, from http://www.nactem.ac.uk/

The Editors of Encyclopædia Britannica. (2017). Computer simulation | Britannica.com. In *Encyclopedia Britannica*. Retrieved from https://www.britannica.com/technology/ computer-simulation

Thompson, P. (2016, July 25). *National Centre for Text Mining — NaCTeM — Frequently Asked Questions*. Retrieved March 20, 2017, from http://www.nactem.ac.uk/faq.php?faq=12

Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S., & Gurbuz, M. (2011). *Analysis of preprocessing methods on classification of Turkish texts.* Paper presented at the Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on.

Tribble, C. (1997). Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk & J. P. Melia (Eds.), *PALC ’97. Practical Applications in Language Corpora* (pp. 106–118). Łód´z: Łód´z University Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information.* 2nd ed. Cheshire, CT: Graphics Press.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1-67.

Turdakov, D. Y. (2010). Word sense disambiguation methods. Programming and Computer Software, 36(6), 309-326.

Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, *30*(1), 1-10.

Two Crows Corporation. (1999). *Introduction to Data Mining and Knowledge Discovery* (3rd ed.). MD: Author.

Ueno, M. (2004a, August). Data mining and text mining technologies for collaborative learning in an ILMS" Ssamurai". In *Advanced Learning Technologies, 2004. Proceedings. IEEE International Conference on* (pp. 1052-1053). IEEE.

Ullah, Z., & Mahmood, A. (2019). Stylometry of short stories through Voyant corpus summary tool: A text mining study. *kashmir Journal of Language Research*, *22*(1), 1-17.

Ullah, Z., Uzair, M., & Mahmood, A. (2019). Extraction of key motifs as a preview from 2017 Nobel prize winning novel, 'Never Let Me Go': An interactive word cloud study. *Journal of Research in Social Sciences*, *7*(2), 83-98.

United Nations. (2018, November 19). YouthStats: Information and communication technology. Retrieved March 25, 2021, from https://www.un.org/youthenvoy/information-communication-technology/

University of Florida. (2017a, June 21). U.S.A. - French and Francophone Digital Humanities Projects - Guides @ UF at University of Florida. Retrieved July 20, 2017, from http://guides.uflib.ufl.edu/c.php?g=592869&p=4100319

University of Florida. (2017b, June 21). Australia - French and Francophone Digital Humanities Projects - Guides @ UF at University of Florida. Retrieved July 24, 2017, from http://guides.uflib.ufl.edu/c.php?g=592869&p=4100266

University of Florida. (2017c, June 21). U.K. - French and Francophone Digital Humanities Projects - Guides @ UF at University of Florida. Retrieved July 24, 2017, from http://guides.uflib.ufl.edu/c.php?g=592869&p=4100318

University of California, Davis. (2009). Introduction: Reliability and validity. Retrieved August 8, 2017, from http://psc.dss.ucdavis.edu/sommerb/ sommerdemo/ intro/validity.htm

University of Victoria. (2019). Voyant tools. Retrieved February 13, 2021, from https://pedagogy-toolkit.org/tools/VoyantTools.html

Uszkoreit, H., & Xu, F. (2013, January). From Strings to Things SAR-Graphs: A New Type of Resource for Connecting Knowledge and Language. In *NLP-DBPEDIA@ ISWC*.

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,

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 JLTonline Conference Proceedings.* Retrieved March 25, 2017, from http://www.jltonline.de/ index.php/ conferences/ article/view/502/1306

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).

Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with wordle. Visualization and Computer Graphics. In *IEEE Transactionson* (pp. 1137–1144).

Virtual University of Pakistan. (2015). Virtual University of Pakistan-Introduction. Retrieved June 14, 2017, from http://vu.edu.pk/AboutUs/AboutVU.aspx

Voskarides, N., Meij, E., Tsagkias, M., De Rijke, M., & Weerkamp, W. (2015). Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (Vol. 1, pp. 564-574).

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219-232.

Voyant tools. (2018, January 1). Retrieved from https://voyant-tools.org/?corpus=fc6f9a58e2f79b4fab7dbcb26febbbfa

Wales, K. (2011). Stylometry- Stylometrics. In *A dictionary of stylistics* (3rd ed., p. 402). London, UK: Routledge.

Walker, J. R. (1999). Digital pedagogy in the humanities. Retrieved February 13, 2021, from https://digitalpedagogy.mla.hcommons.org/keywords/classroom/

Wang, H. (1963). Toward Mechanical Mathematics In K. M. Sayre & F. J. Crosson (Eds), *The Modeling of Mind*. South Bend, IN: Notre Dame University Press.

Wang, J., & Good, R. (2007). *The repetition of collocations in EFL textbooks: A corpus study*. Paper presented at The Sixteenth International Symposium and Book Fair on English Teaching in the Republic of China, Taipei.

Wang, Y. (2004). Various approaches in text pre-processing. *TM Work Paper No. 2* (5).

Wang, Y. (2012). Novel approaches to pre-processing document base in text classification: *CiteSeer*. Retrieved January 6, 2016, from *https://pdfs.semanticscholar.org/f0f5/ 0de6a98fa39588671cd8ed4d2fcacfcd2215.pdf*

Wang, J., Zhao, J., Guo, S., North, C., & Ramakrishnan, N. (2014, May). ReCloud: Semantics-based word cloud visualization of user reviews. In *Proceedings of Graphics Interface 2014* (pp. 151-158). Canadian Information Processing Society.

Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In *INFOVIS 2002* (pp. 110–116). IEEE.

Wattenberg, M., & Viegas, F. B. (2008). The word tree, an interactive visual concordance. In *Visualization and Computer Graphics* (p. 1221–1228). IEEE Transactions.

Waugh, S., Adams, A., Tweedie, F., & Waugh, D. (2000). Computational stylistics using artificial neural networks. *Literary and Linguistic Computing*, *15*(2), 187-198. doi:10.1093/llc/15.2.187

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences* (Vol. 111). Chicago: Rand McNally.

Weir, G.; (2007) *The posit text profiling toolset.* In: Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics. UNSPECIFIED.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, *16*, 180-205.

Weir, G. R., & Ozasa, T. (2010). Learning from Analysis of Japanese EFL Texts. *Educational Perspectives*, *43*(1), 56-66. Retrieved from https://eric.ed.gov/?id=EJ912116

Widdowson, H. G. (1991). The description and prescription of language. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1991* (pp. 11–24). Washington, DC: Georgetown University Press.

Widdowson, H. G. (1992). Communication, community and the problem of appropriate use. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1992* (pp. 305–315). Washington, DC: Georgetown University Press.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3–25.

Williams, D. (1983). Developing criteria for textbook evaluation. *ELT Journal* 37 (3): 251–255.

Williams, B. (2008). *Ancient Egyptian war and weapons* (3rd ed.). Chicago, IL: Heinemann Library.

Willis, D. (1990). *The lexical syllabus*. London, Collins.

Willis, D. (1994). A Lexical Approach. *Grammar and the Language Teacher*. M. Bygate, A. Tonkyn & E. Williams. Hemel Hempstead, Prentice Hall. 56-66.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching.* Cambridge: Cambridge University Press.

Wisbey, R. (1963). The analysis of Middle High German texts by computer—some lexicographical aspects. *Transactions of the Philological Society*, *62*(1), 28-48.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization 1995* (pp. 51–58). IEEE.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). California: Morgan Kaufmann.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of the play. *Annual Review of Applied Linguistics, 32*, 231 254.http://dx.doi.org/10.1017/S026719051200013X

Wray, A. (2013). Formulaic language. *Language Teaching, 46*(3), 316-334.

Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (p. 118–127).

Yang, Y, Y. Akers, L. Klose, T, & Yang, B. (2008). Text mining and visualization tools-Impressions of emerging capabilities. *World Patent Information.* Pp. 280-293.USA. Elsevier

Yeates, R. (2013, May 2). Voyant Tools | Post-Apocalyptic Cities. Retrieved March 25, 2017, from https://postapocalypticcities.wordpress.com/2013/05/02/voyanttools/

Zaiane, O., Xin, M., & Han, J. (1998). Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in Digital Libraries* (pp. 19–29).

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

Zhou, X., & Han, H. (2005, May). Survey of Word Sense Disambiguation Approaches. In *FLAIRS conference* (pp. 307-313).

Zhong, N., & Skowron, A. (1999). *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing: 7th International Workshop, RSFDGrC'99, Yamaguchi, Japan, November 9-11, 1999 Proceedings* (Vol. 1711). Springer Science & Business Media.

Zhu, Y. (2018). Knowledge discovery and data mining - IBM. Retrieved January 31, 2021, from https://researcher.watson.ibm.com/researcher/view_group.php?id=144

Zongyao, S., & Fuling, B. (2003). Method and application of comprehensive knowledge discovery. *Geo-spatial Information Science*, *6*(3), 48-55. doi:10.1007/bf02826893

APPENDIX A

Book I Intermediate Punjab Textbook, Punjab, Pakistan

Due to space limitation, file is available from online resources. To access the file, please visit the following link:

https://pctb.punjab.gov.pk/download_books

APPENDIX B

Book III Intermediate Punjab Textbook, Punjab, Pakistan
Due to space limitation, the scanned file is available from online resources. To access the file, please visit the following link:
https://pctb.punjab.gov.pk/download_books

APPENDIX C

Book II Intermediate Punjab Textbook, Punjab, Pakistan

Due to space limitation, the scanned file is available from online resources. To access the file, please visit the following link:

https://pctb.punjab.gov.pk/download_books

APPENDIX D

Novel ‘*Good-Bye Mr. Chips’* by James Hilton

Due to space limitation, the scanned file is available from online resources. To access the file, please visit the following link:

http://gutenberg.net.au/ebooks05/0500111h.html

APPENDIX E

Phrases/ Collocations of Intermediate Books

Grammar Pattern	Examples
1. Adj+Adj+N	blue suede shoes
2. Adj+N	noble deeds, Christmas time, a few minutes, night mail, best interest, Western Europe, national assembly, walnut cake, pink icing, acting head, a little money, fruit stalls, bright boy, many boys. summer holidays, early childhood, household use, good morning, eating habits, agricultural Commune, death rate, great mosque,
3. Adj+N+N	fifty thousand dollars
4. Adj+N+Prep+N	this sort of thing, sweet land of liberty
5. Adj+N+Adv	a few days later
6. Adj+N+Prep	total number of
7. Adj+Prep	afraid of, 2. full of,
8. Adj+Prep+Art	2. one of the, (one. Adj. Merriam Webster)
9. Adj+Prep+Prn	Most of them
10. Adj+Prep+N+Art+N+Adv	2. loveliest of trees the cherry now
11. Adv	a lot of, of course
12. Adv+Adj	too hot, not mere
13. Adv+Art+N	before the king, when the fact fails, as a result,
14. Adv+Prep	just to
15. Adv+Prep+Prn	ahead of him
16. Adv+Prn+Aux	before I had
17. Adv+Prn+V	as she left
18. Adv+Adj+N	when all god’s children
19. Adv+Art+N+Prep+N	as a matter of fact,

20. Art+N	the sun, the Greek, a man, a spaceship, the house, a bit, a dollar, the grain, an hour,
21. Art+N+V+Prep	the years went by
22. Art+N+Prep+N	a pair of shoes, a piece of string,
23. Art+N+Prep	a list of, a short of, a piece of, a sense of,
24. Art+N+Prep+Art+N	a quarter of a century
25. Art+N+Prep+Adj+N	the post of court acrobat
26. Art+N+Inf V+Prep	a time to cast away
27. Art+N+Conj+Art+N+Prep+Art+N	a book and a pearl in the oyster
28. Art+N+Prep+N+Conj+Adv+Prep+N	a man of words and not of deeds
29. Art+N+Inf V	a time to keep
30. Art+Adj	a little, a whole,
31. Art+Adj+N+Prep	a new kind of
32. Art+N+Prep+Adj	the sons of former
33. Art+N+Aux+V+Prep+Art+Adj+N	the package was lying by the front door
34. Aux	ought to, 2. may be
35. Aux+Adj+N+Aux	may be some eccentric millionaire is
36. Aux+V	have to look, did come,
37. Aux+V+Prep	have liked to
38. Aux+V+Prn	doesn’t intrigue you
39. Aux+Adj+Inf V+N	will be able to join hands
40. Aux+Prn+Adj	aren’t you ashamed
41. Conj	as if, 2. as well as, as soon as (Subordinate Conjunction)
42. Conj+Adv	2. and then
43. Conj+Prn	as she
44. Conj+Prn+V+Conj	and you wonder that
45. Conj+V+Prn+V	and let him go
46. Conj+Prn+Mod	if we can
47. Det+N	no time

48. Id	out of order, just as (Merriam Webster)
49. Inf V+Prn+N	to wash your face, to dig her grave,
50. Inf V+Art+N+Prep	to shed the blood of
51. Inf V	2. to stand
52. Inf V+N	to seek justice
53. Inf V+Inf V+Art	to try to burn a
54. Int+Adv	oh yes
55. Int+Adv+Aux+Prn+V	O where are you going
56. Int Prn+Aux+Prn,+N	what is it, son
57. Int Prn+Aux+Adj	who are stupid
58. Mod	used to
59. N	Carbolic acid, commander in chief
60. N+N	Mustafa Kamal, Abdal Rahaman,
61. N+N+N	melon, guava, mandarin
62. N+Apo+N	God’s attributes
63. N+Aux	faith is
64. N+Conj+N	gold and silver, disease and death, communication and transportation,
65. N+Prep	cooperation with, beak with, contribution of,
66. N+Prep+N	degrees of frost, grain of sand, sense of proportion, kinds of food, use of science, birth of Christ, cost of living,
67. N+Prep+Art+N	story of the string, cells of the body, jewel of the world, end of the week, culture of the mould, surface of the sun, clearing in the sky
68. N+Prep+Art+Adj+N	creation of a new world
69. N+V+Prep	applause broke out
70. N+Prep+Art+N+N	land of the pilgrims’ pride,
71. N+V+Prep+Prn	Arthur stared at her

72. N+Prep+Art	letters for the,
73. N+V	Leaves drinking, calculation shows,
74. N+V+Prn+N	friend breathing his last
75. N+V+Ref Prn	God calls himself
76. N+Prn+V	rain I hear
77. Nu+Adj+N	three hundred dollars
78. Prep+Prep	out of
79. Prep+Art+N	down the hall, at the hills, on the daybed, to the village, on the road, in the world, to the ground, for a moment, for a moment, in the presence, for a while, across the street, in a row, into the cold, on a Friday, in the world, in the world, across the Sahara, in the broiler
80. Prep+Art+N+Prep	in the street of, in the hands of, for the benefit of
81. Prep+Art+N+N	in the hills sir
82. Prep+Art+N+Prep+N	in the treatment of disease
83. Prep+Art+Adj+N	toward the deep valley
84. Prep+Adj	of two
85. Prep+Adj+N	in other words, in broken images
86. Prep+Adj+N+Apo	at St Mary’s
87. Prep+Det+N	in such cases
88. Prep+Art+Adj+N	to the next village, for a long time, in a low voice, on the culture plate, for the first time
89. Prep+Art+N+Conj+Prn+Mod	to the end that you may
90. Prep+Art+Adj	at the same, in the third
91. Prep+Art+Adj+N	on the lower side
92. Prep+Art+Adv	in the past
93. Prep+Art+N+Adj+Prep+N	like a garden full of weeds
94. Prep+N+Art	through love the

95. Prep+N+Prep+Art	on top of the
96. Prep+Aux+V	of being drowned
97. Prep+Adv	by now, at first,
98. Prep+Nu+N	per 1000 population
99. Prep+N	at last
100. Prep+N+Prep	in terms of
101. Prep+Prn	like that
102. Prep+Prn+Adj	in its most
103. Prep+Prn+N	in my experience, at my tongue, by his shirt front (shirt front-N)
104. Prep+Prn+N+Prep+N	by my word of honour
105. Prep+Prn+N+Prn+Aux+V+Prep	with this faith we will be able to
106. Prn	no one, 2. a few,
107. Prn+Adv+V	he almost whispered
108. Prn+Aux+V+Prep	I'm going to, you have learnt to
109. Prn+Aux+N+Prep+Art+N	it was 97 in the shade
110. Prn+Aux+V	you were seen,
111. Prn+Aux+Art	I am the, I am a,
112. Prn+Aux+Prn+V+Inf V	what are you going to do
113. Prn+Aux+Prep+V	I had to smell,
114. Prn+V+Adv+Prn	he said as he
115. Prn+Mod	I might, I had to,
116. Prn+Mod+Mod	it would be,
117. Prn+Mod+V	I can cure, it may be,
118. Prn+Mod+V+Prn	I would whip them
119. Prn+Aux+Adj	it was like
120. Prn+V+Art+N+N+Prep+Prn+N	she took the card halves from her purse
121. Prn+V+Prn	he told her
122. Prn+V+Prep	She put on, he lived at,
123. Prn+V+Prep+Prep+Art+N	she went back into the kitchen

124.	Prn+V+Prep+Art+N	she picked up the receiver
125.	Prn+V+Prn	I followed him
126.	Prn+N+Aux	her hair was
127.	Prn+N	their relevance
128.	Prn+N+Conj+N+Aux+V	my father and mother had cleared
129.	Prn+Prep	none to
130.	Prn+V+Prep	he went into, he came by
131.	Prn+V+Prn+Aux+V+Inf V	you thought I was going to say
132.	Prn+V+Inf V+Ref Prn	He came to know himself
133.	Prn+Adj	6. my dear
134.	Prn+Adj+N	his clear images
135.	Prn+V	he assumes, I question,
136.	Prn+Aux	he was
137.	Prn+Aux+Prn+V+Prn	how will you have it?
138.	Prn+V+Prep+Adj+N+	I stand in good relation to all that is
139.	Prep+Adv+Prn+Aux	
140.	Prn+V+Prep+Adj+N+Prep+Art	I stand in good relation to the
141.	Prn+Prep+Prn+N	one of those mysteries
142.	Prn+Prep+Art	one of the
143.	Prn+Aux+Prep	I had to
144.	Prn+Rel Prn	those who
145.	Phr	even if
146.	Prop Adj+N	Muslim Spain, Abbasid Caliph, Umayyad dynasty
147.	PP	in front of
148.	V+Adv	dried up
149.	V+Art+N	open an account, make a book
150.	V+N+Apo+N	got housemaid's knee
151.	V+Prep+Prn	looked at him
152.	V+Prep	got up, belonged to, comes in, hung with, begin to, getting into, work on, went up,

153.	V+Prep+Prep+N	getting on in years
154.	V+Prep+Art+N	pick up the pocket book, looked at the door, hesitated for a moment,
155.	V+Prn+N	snatch my pocket book, cleared his throat, open your mouth, made his way, cleared this land
156.	V+Prep+Prn	sticks to it
157.	V+Prep+N	go to college,
158.	V+Prep+Prep+Art	sat down on a
159.	V+Prep+Art+N	sat by the fire, went into the living room
160.	V+Prep+Art+Adj+N	bitten by a mad dog, made up his mind
161.	V+Prep+Prn+N	look at your throat,
162.	V+Prep+Prn+Inf V+Art	like for us to have a
163.	V+Art+Adj+N	becoming the first customer
164.	V+Prn+Adv	placed it before
165.	V+Prn+Aux+Art+Adj+N	suppose it’s a genuine offer
166.	V+Prn+Art+N+Prep+N	bring me a cup of tea
167.	V+Adv+Prep	2.go back to

APPENDIX F

Evaluation Chart of Collocations

Serial. No	**Sequence No.**	**Grammatical Category**	**Collocation Instances**
1	1-10	Adjective	33
2	11-19	Adverb	13
3	20-33	Article	27
4	34-40	Auxiliary Verb	10
5	41-46	Conjunction	10
6	47	Determiner	1
7	48	Idiomatic expression	2
8	49-53	Infinite	7
9	54-57	Interjection	5
10	58	Modal	1
11	59-76	Noun	37
12	77	Number	1
13	78-105	Preposition	58
14	106-143	Pronoun	52
15	144	Phrase	1
16	145	Proper Adjective	3
17	146	Prepositional Phrase	1
18	147-167	Verb	37
	Total Collocation Patterns= 167	**Total Grammatical Categories= 18**	**Total Instances= 297**

APPENDIX G

One Word Library Concept with Hyperlinks

Serial. No	**Short Stories. Book I**	**One Word Library**
1	Button, Button	norma
2	Clearing in The Sky	i
3	Dark They Were, And Golden Eyed	said (41)
4	Thank You, M’am	said (28);
5	The Piece of a String	mr
6	The Reward	jorkens
7	The Use of Force	throat
8	The Gulistan of Sadi	king
9	The Foolish Quack	camel
10	A Mild Attack of Locusts	margaret
11	I Have a Dream	freedom
12	The Gift of the Magi	jim
13	God Be Praised	maulvi
14	The Overcoat	young
15	The Angel and the Writer & Others	good

Serial. No	**Plays. Book III**	**One-word Library**
1	Heat Lightning	man
2	Visit to a Small Planet	spelding
3	The Oyster and the Pearl	harry

Serial. No	**Poems. Book III**	**One-word Library**
1	The Rain	drop
2	Night Mail	letters
3	Loveliest of Trees, The Cherry Now	cherry
4	O Where are you going?	said (6)
5	In the Street of Fruit Stalls	dark
6	Sindhi Woman	bare
7	Times	time (21)
8	Ozymandias	ozymandias
9	The Feed	grain
10	The Hollow Men	men
11	Leisure	time (6);
12	Ruba’iyat	faith
13	A Tale of Two Cities	burnt

14	My Neighbor Friend Breathing his Last!	aghast
15	He came to know himself	came
16	God’s Attributes	god
17	The Delight Song	alive
18	Love – An Essence of All Religions	love
19	A Man of Words and Not of Deeds	like
20	In Broken Images	images

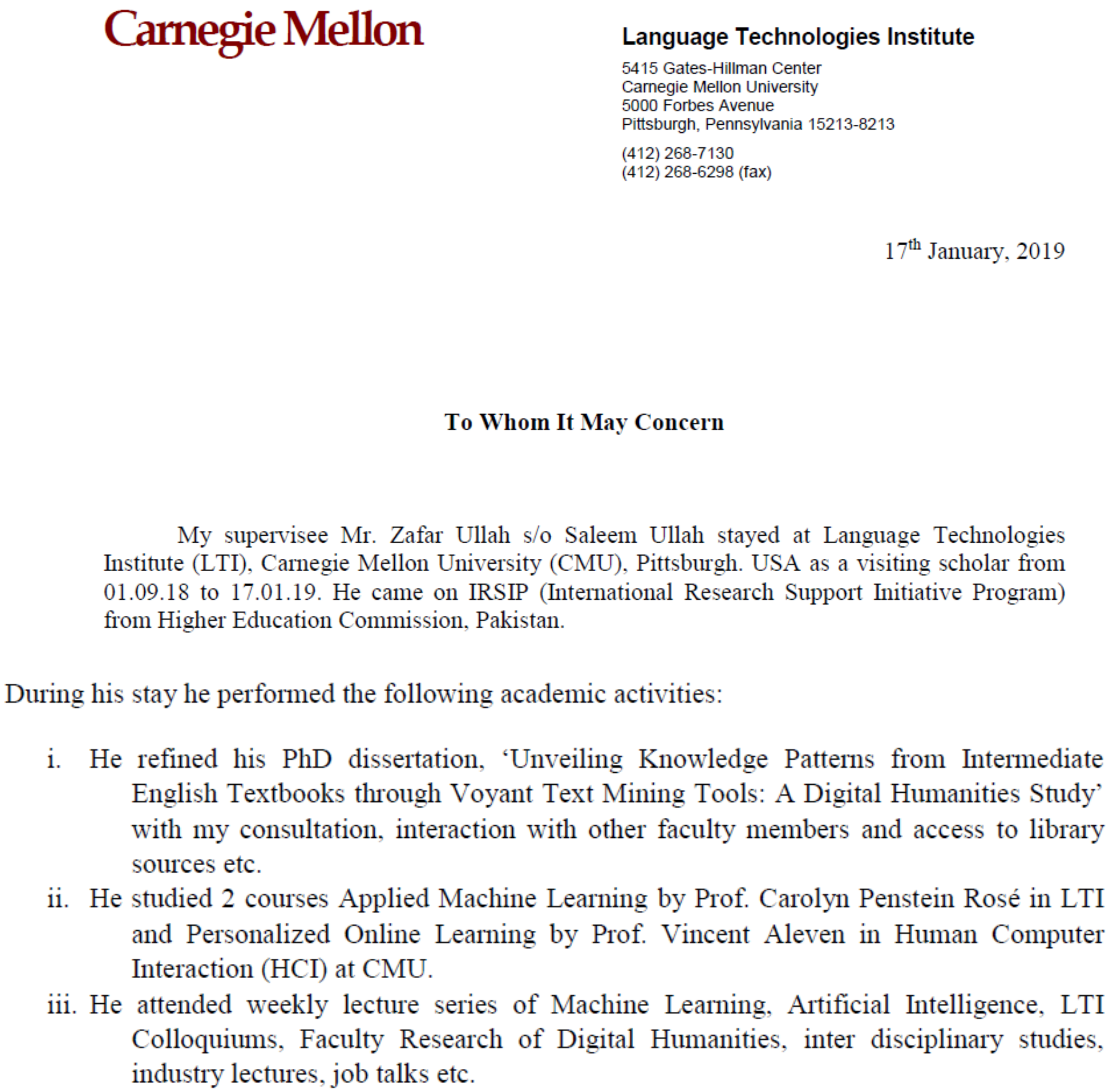
Serial. No	**Essays. Book II. Part-I**	**One-word Library**
1	The Dying Sun	life
2	Using the Scientific Method	people
3	Why Boys Fail in College	boy
4	End of Term	school
5	On Destroying	books
6	The Man Who Was a Hospital	i (95)
7	My Financial Career	said (17);
8	China's Way to Progress	chinese
9	Hunger and Population Explosion	population
10	The Jewel of The World	al
Serial. No.	**Heroes. Book II. Part II**	
11	First Year at Harrow	english
12	Hitch Hiking Across Sahara Desert	christopher
13	Sir Alexander Fleming	fleming
14	Louis Pasteur	pasteur
15	Mustafa Kamal	mustafa

Sr. No	**Novel**	**One-word Library**
1	Good Bye Mr. Chips	chips

Corpus Name	**Total Words in Voyant**	**Total Words in Word File**	**One-word Library**
Complete 1st year Corpus	37,402	37,452	man (214)
Complete 2nd Year Corpus	45,092	45,038	chips (156)
Complete PIE CTZU	82,487	82,495	said (290)

APPENDIX H

Fellowship Letter of Co-Supervisor from Prof. Carolyn Penstein Rose, Carnegie Mellon University, USA



Carnegie Mellon

Language Technologies Institute

5415 Gates-Hillman Center
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213-8213

(412) 268-7130
(412) 268-6298 (fax)

17th January, 2019

To Whom It May Concern

My supervisee Mr. Zafar Ullah s/o Saleem Ullah stayed at Language Technologies Institute (LTI), Carnegie Mellon University (CMU), Pittsburgh. USA as a visiting scholar from 01.09.18 to 17.01.19. He came on IRSIP (International Research Support Initiative Program) from Higher Education Commission, Pakistan.

During his stay he performed the following academic activities:

i. He refined his PhD dissertation, 'Unveiling Knowledge Patterns from Intermediate English Textbooks through Voyant Text Mining Tools: A Digital Humanities Study' with my consultation, interaction with other faculty members and access to library sources etc.

ii. He studied 2 courses Applied Machine Learning by Prof. Carolyn Penstein Rosé in LTI and Personalized Online Learning by Prof. Vincent Aleven in Human Computer Interaction (HCI) at CMU.

iii. He attended weekly lecture series of Machine Learning, Artificial Intelligence, LTI Colloquiums, Faculty Research of Digital Humanities, inter disciplinary studies, industry lectures, job talks etc.

iv. He attended and participated in research group meetings and individual meetings to share his own work and to discuss others' work.

v. He attended several proposal and dissertation defenses.

vi. He interacted and discussed with several faculty members in LTI, Modern Language Department and HCI.

vii. He worked as a team member on a project related to Rhetorical Structure Theory (RST) in collaboration with Turnitin.com and LTI.

viii. He worked on a research paper on collocations and discussed some future research projects.

I wish him best for his future research, studies and career.

Sincerely,

Carolyn Rosé

Dr. Carolyn Penstein Rosé
Professor
Language Technologies Institute and Human-Computer Interaction Institute
School of Computer Science, Carnegie Mellon University