Exploring Stylistic Features in Agatha Christie's Works through Stylometric Approaches

A Thesis Submitted for the Degree of Doctor of Philosophy,
Studies in Language and Culture,
Graduate School of Language and Culture,
Osaka University

by

TSUCHIMURA, Narumi
（土村　成美）

November　2020

Acknowledgements

I owe my deepest gratitude to my supervisors Professor Tabata Tomoji and Associate Professor Bor Hodošček for their guidance, support and professional advice on my research. Without their help, this dissertation would not have been possible. In addition, I would like to thank the members of the research project on Digital Humanities at Osaka University for their insightful comments on my research.

Contents

Chapter 1

Introduction

1.1 Background and aim

Agatha Christie (1890–1976) is a renowned English crime writer. She is well-known as the creator of detectives Hercule Poirot, Miss Marple, Tommy Beresford, and Tuppence Beresford. Christie is considered to be the best selling writer of all time. Her collective sales have been estimated to be more than two billion copies, and only the Bible and the works of William Shakespeare outstrip her sales[1]. Her novels have been translated into many languages and read by many people all over the world.

As the preeminent translator of Christie's works into Japanese, Nakamura (1986) says, "When the truth is hidden in a subtle figure of speech as Christie often does, we translators have to translate the sentence very carefully so as not to go too far or not to say too little as if we are united with the author . (Nakamura, 1986; 23, 24)" Inaki (2013) notes in her book's conclusion that Christie is an author whose words have a striking fascination for readers. What attracts readers to her novels? Qualitative studies have been carried out using close reading, but quantitative methods are rarely used. This study makes aims to shed light on some of the reasons Christie's works fascinate a number of people, using quantitative methods that compare her with other authors.

This study focuses on not only the differences from other authors but also on the diachronic

[1]Agatha Christie: Mystery of crime writer's disappearance tackled in Kate Mosse story (https://www.independent.co.uk/arts-entertainment/books/news/agatha-christie-mystery-of-crime-writers-disappearance-tackled-in-kate-mosse-story-10500837.html) (Accessed on June 15th, 2018)

[1]Author's translation

changes in Christie's style. Jin (2016) states that it is possible that one's style can be affected by his or her environment, thoughts, and emotions, and one's writing is related to his or her mental state. He cites the study on the relationship between Akutagawa Ryunosuke's suicide and the stylistic changes in his works. He also proposes a relationship between Uno Koji's brain disease and the changes in his writing style. Jin notes that studies on the relationship between one's style and his or her mental state is a meaningful consideration in today's society (Murakami et al., 2016; 101, 102). It is also possible that an author's style changes as they age. This study focuses on style changes in relation to aging and the environment, and aims to analyse whether any such diachronic changes appear in Christie's works.

Along with the development of computer technology in the last few decades, digital humanities have progressed; the linguistic resources that have been previously investigated by qualitative methods are beginning to be analysed using quantitative, statistical methods. Such linguistic resources include literary works, press, political discourses and so on. In this study, we approach Christie's works from the perspective of such quantitative, stylometric analyses.

1.2 Agatha Christie's life

Before going on to the analysis, I would like to briefly introduce Christie's life.

Christie was born Agatha Mary Clarissa Miller in Torquay, Devon in 1890. Her parents did not allow her to go to school as her brother and sister did, so she was educated at home. During her childhood, Christie's sister encouraged her to read detective novels, so she read crime novels such as the Sherlock Holmes series. She began to write poems and short stories in her teens. She studied opera in Paris, but did not have adequate talent to pursue it as a career.

In 1912, Christie met military officer Archibald Christie at a dance party. Two years later they were married. While Archibald was in France during World War I, Christie worked as a voluntary nurse at a hospital in Torquay. Following that she worked in a dispensary. The murder methods in her stories were inspired by the knowledge of poisons she acquired in the dispensary and appeared in her first crime novel.

In 1920, after being rejected by several publishers, Christie's first crime novel, The Mysterious Affair at Styles was published by The Bodley Head. Her sixth novel, The Murder

of Roger Ackroyd, published in 1926, aroused criticism for its unconventional use of the protagonist as a murderer. She became a popular writer due to this novel and signed over publishing rights to Collins.

That same year her mother died and Archibald asked for a divorce because he had an affair. On December 3, Christie disappeared after driving away from her home, which caused a great stir in the UK. Eleven days later, she was found in a hotel in Yorkshire staying under a false name. This disappearance contributed to her book sales. The truth behind her disappearance was never established. Archibald and Christie divorced in 1928.

Christie became increasingly popular as she aged. Her new novels were published every Christmas; her publishing company Collins established the slogan, 'A Christie for Christmas'.

In 1930, during her travels in Mesopotamia, she met an archaeologist named Max Mallowan. She married him that same year. She often travelled to excavation sites in the Middle East with Max, and these experiences later appeared in her novels. After her marriage to Mallowan, Christie began to write love stories under the pen name Mary Westmacott. Christie did not make it public that Mary Westmacott was her pen name until the Sunday Times broke the news. She wrote only six romance novels.

Christie also wrote plays, the most outstanding of which is The Mousetrap, which was first performed at London's Ambassadors Theatre in 1952. It is the longest-running show in the world. She was made Commander of the Order of the British Empire in 1956 and promoted to Dame Commander in 1971 for her great contribution to literature. Christie's health began to deteriorate in the 1970s. She could no longer write new novels after the Postern of Fate was published in 1973. Christie died at her home in 1976 at age 85 from natural causes (said to be a cold).

1.3 Structure of this thesis

The structure of this thesis is the following: In Chapter 2, we briefly review previous studies on stylometry and studies on Christie's style. In Chapter 3, we provide an overview of the authors whose works are used for the data in this analysis. Then, an explanation of the data is given. In Chapter 4, we compare Christie's style with three other authors using stylometric analysis to examine the features characteristic of her work. In Chapter 5, we

analyse diachronic changes in Christie's style, and we will see these changes in her works. In Chapter 6, we compare Christie's style with other authors through topic modelling and investigate topics that are specific to Christie. In Chapter 7 we will make a conclusion.

Chapter 2

Previous studies

2.1 Stylometry

Crystal (1987) states, 'style is viewed as the set of language features that make people distinctive – the basis of their personal linguistic identity'. Stylometry is also called 'stylostatistics' or 'stylometrics'; this field investigates the characteristics of authors' styles and texts using quantitative variables such as word frequency through statistical methods. There is a great deal of literature and religious writing where author attribution is unknown or disputed, or whose date of creation is not clear. In stylometry, researchers try to reveal authors' identities along with the date of the writing, or capture characteristic features of the texts using quantitative analysis.

Early studies of quantitative analysis of literary texts were almost all on authorship attribution. The beginning of stylometry dates from Augustus de Morgan's suggestion in his letter to a friend in 1851 that authors of the Bible might be able to be distinguished by using their mean word lengths. Inspired by the idea, in 1887 T.C. Mendenhall investigated the distributions of the lengths of words in Oliver Twist by Charles Dickens, Vanity Fair by William Thackeray, Political Economy and Essay on Liberty by John Mill and so on. He described line graphs of word length distributions, and called them 'word-spectrum' or 'characteristic curve'. He proposed that the word length which occurs most frequently would match if the authors of the writings were the same (Mendenhall, 1887; Murakami, 2004).

In 1901, Mendenhall investigated the distributions of the length of words in a play written by William Shakespeare and a work written by Francis Bacon. This was due to the suspicion

that Shakespeare's works were written by Bacon. As a result of the analysis, Shakespeare was found to have used four-letter words most frequently, while Bacon favored three-letter words. Mendenhall concluded that it was unlikely Bacon had written Shakespeare's plays.

In the 1960s, most outstanding researches on linguistic resources through a quantitative analysis were on authorship attribution.

Ellegård (1962) analysed the *Junius Letters*, a series of letters criticising the government that appeared in Public Advertiser, a London newspaper published from 1769 to 1772. These letters were called *Junius Letters* because their author used the pseudonym Junius. The actual author of the letters was not identified at the time of publication. Ellegård compared the frequencies of what he called Junian plus words (the words that the author of the letters preferred to use), Junian minus words, and synonyms between *Junius Letters* and texts written by suspected authors of the letters. Ellegård concluded that the author of the letters was likely Sir Phillip Francis.

Mosteller & Wallace (1963) researched *The Federalist Papers*. *The Federalist Papers* were written by Alexander Hamilton (1755–1804), John Jay (1745–1829) and James Madison (1751–1836) to promote the ratification of the Constitution. They were published during 1787–1788. *The Federalist Papers* consisted of 77 essays: 43 written by Hamilton, 5 by Jay, 14 by Madison, and 3 by Hamilton and Madison. The authors of the remaining 12 essays were in dispute, between Hamilton and Madison, so Mosteller & Wallace attempted to identify the authorship of those disputed papers using the rate of the occurrence of words through statistical approaches. They identified 28 words whose frequencies in Hamilton's and Madison's texts were significantly different. Using those words as the object of the analysis, they investigated who wrote the 12 essays through discriminant analysis and Bayes' theorem. As a result of the analysis, they argued that all 12 essays were written by Madison.

Since the 1980s, technology developed, multivariate analysis began to be used. The application of this technique to stylometry was pioneered by the studies on the style of Jane Austen by Burrows. Burrows (1987) analysed the novels of Jane Austen using statistical methods. He analysed the frequencies of the top 30 most frequent words (most of them were not content words but function words) using principal component analysis (PCA). He showed that the patterns of occurrences of the most frequent words were strongly connected

to idiolects of the characters in Austen's novels and the changes of idiolects within a novel. Burrows' research is worthy of special mention in that he showed that although the frequencies of the most common words such as function words seemed not to be useful as a variable of quantitative analysis, the pattern of the occurrence of these words was helpful in stylometry.

2.2 Researches on Christie's style

Christie is suspected to have suffered from Alzheimer's disease in her later life, so researches on her style using stylometric approaches has mainly focused on the decline in her linguistic ability and vocabulary in her later novels.

Lancashire & Hirst (2009) analysed the vocabulary size and richness in Christie's 14 novels through word types, repeated phrase-types, and occurrences of indefinite words such as *thing*, *anything*, and *something*. As a result, they showed that Christie's linguistic ability declined in her later novels. The vocabulary richness declined as she aged, and the number of repeated phrases and occurrences of indefinite words increased in her later novels. They concluded that these were significant markers of her dementia.

Le et al. (2011) analysed lexical and syntactic changes in the language of Alzheimer's disease using the novels of Christie, Iris Murdoch, and P. D. James through statistical methods. Iris Murdoch was a writer in the UK who was diagnosed with Alzheimer's disease in her later life and died of it. P. D. James was selected for comparison as a writer who has aged healthily. At the lexical level, Le et al. analysed the vocabulary size using type/token ratio (TTR) and word-type introduction rate (WTIR), *i.e.* the cumulative number of unique lemmatised types counted at every 10,000-token interval, n-gram repetitions, lexical specificity by computing the proportions of indefinite nouns (*thing(s)*, *something*, *anything*, and *nothing*), 35 high-frequency, low-imageability verbs, proportions of word class and proportion of interjections and fillers. At the syntactic level, they analysed syntactic complexity (mean length of utterance and mean number of clauses per sentence, parse tree depth, D-level, *i.e.*, a psycholinguistics-based ranking of sentence types), and passive voice. Christie's and Murdoch's later works showed a decline in linguistic ability such as loss of vocabulary, repetition of fixed phrases, decrease in noun-token proportion with a compensatory increase in verb-token proportion, increase in fillers, and so on. In some analyses, Christie showed a

more significant decline in linguistic ability than Murdoch. Lancashire & Hirst (2009) and Le et al. (2011) are studies that focused on the decline in Christie's linguistic ability.

Inaki (2013) assembled a collection of researches using a small corpus; each analysis treats one of Christie's novels. It is a more qualitative study than Lancashire & Hirst (2009) and Le et al. (2011). She analysed *The Murder of Roger Ackroyd*, *And Then There Were None* and *Endless Night* using one novel for each analysis and extracted keywords using WordSmith Tools[1]. She then qualitatively investigated each work in detail . In her analysis of *The Murder of Roger Ackroyd*, a novel in the first person narrative whose narrator is the criminal, she extracted keywords using WordSmith Tools; from these keywords she focused on *-ly* adverbs and one of the most frequent collocates with them, *said*. These *-ly* adverbs have two interpretations (subject-oriented and manner adverbs); Inaki showed that Christie allows readers to misunderstand *-ly* adverbs not as subject-oriented, but as manner adverbs. She insisted that by using this technique, Christie delayed the identification of the criminal. In her analysis on *And Then There Were None*, out of the list of keywords Inaki focused on the expressions of thought of the characters. She clarified that Christie used different expressions of thought purposely to mislead readers. In the analysis of *Endless Night*, Inaki focused on the noun *house*, which was extracted as a keyword with the highest keyness among nouns. She showed that Christie used possessive pronouns (*my* and *our*) modified the noun *house* purposely to mislead readers. Throughout the series of these studies, Inaki showed that Christie led readers not to identify a criminal until the end of the novel, providing them with necessary and sufficient information. She concluded that Christie used words cleverly to misdirect readers keeping fairness as a detective novel writer (not to tell a lie) and providing readers with necessary and sufficient information to identify the criminal.

These studies on Christie's style have limitations in that only a limited number of novels are used for analysis. The studies on dementia, such as Lancashire & Hirst (2009) and Le et al. (2011) that put an emphasis on Christie's decline in linguistic ability treat more works in her later life than works in her earlier life. Therefore the data used in these studies seem to be a little unbalanced. The same applies to Inaki (2013), where she treats one novel for each

[1]WordSmith Tools is a Windows software for finding word patterns. It enables us to make concordance lines and word lists, and to extract keywords. (https://www.lexically.net/wordsmith/) (Accessed on February 20th, 2019)

analysis. In this study, we try to analyse all of her long detective novels.

Chapter 3

Data

In this paper, we chose the writings of Dorothy L. Sayers, Margery Allingham, and Ngaio Marsh for comparison with Christie's writing. As mentioned previously, all three authors were Christie's contemporaries and published many novels during the interwar period. Joannou (2013, 144) noted the following:

> Agatha Christie (1890–1976), Ngaio Marsh (1895–1982), Margery Allingham (1904–1966), Dorothy L. Sayers (1893–1957), Gladys Mitchell (1901–1983), Elizabeth Mackintosh (Josephine Tey) (1896–1952) – all of whose careers began in the interwar years. The first four are often dubbed the so-called 'queens of crime' of Britain's 'Golden Age' of crime fiction.

We collected the novels of the 'queens of crime' of Britain's golden age of crime fiction: Christie, Sayers, Allingham and Marsh. In the following section, we provide brief introductions of the authors, with the exception of Christie, whose writings are used for the analysis. Then, the overview of the data is shown.

3.1 Authors

3.1.1 Dorothy L. Sayers

Dorothy Leigh Sayers was a famous English mystery writer. She is best known for a detective in her novels, Lord Peter Wimsey. She was versatile: she is well-known not only for her novels, but also for essays and translations, particularly her translation of Dante's *Divine Comedy*.

Sayers was born in Oxford in 1893. She learned Latin, French and German during her childhood, and entered Somerville College, Oxford. She was one of the first women to receive a degree. After graduating from college, she taught at a high school, but soon quit teaching and began to work at bookshop and publisher Blackwell. In 1919 she went to France with her lover, where she stayed for one year. After returning to the UK, she started teaching school again. To supplement her income, she began to write mystery novels. She became a popular mystery writer in the 1930s; however, she stopped publishing crime novels after *Busman's Honeymoon* in 1937, instead putting her efforts into writing Christian plays and translating Dante's *Divine Comedy* until her death in 1957.

Christie was the most popular writer signed to Collins publishers and Sayers was the most popular writer signed to Gollancz publisher, so the two were good rivals as well as good friends. In the 1930s and 40s, Christie and Sayers competed against each other for the sales of their detective novels (Malling and Peters, 1997), so they are often compared to each other. According to Mori (ed.) (1998), Sayers had a better writing ability and was better at describing characters than Christie. Today's leading female mystery writers such as P. D. James and Ruth Rendell, say that Sayers, not Christie, is the ideal writer. On the other hand, Christie has been recently criticised by contemporary mystery writers, who say the characters in her novels are in a fixed form, and her style is too plain. On this point, Sayers forms a sharp contrast to Christie.

3.1.2 Margery Allingham

Margery Allingham is also a British detective-story writer. According to Mori (1998), when Rampo Edogawa, a Japanese writer in the Taisho and Showa periods, introduced Allingham in his review, she was not yet a well-respected author. However, she is now highly esteemed in Europe and America and considered to be one of the most famous crime fiction writers during Britain's golden age of crime fiction. Christie also had a high opinion of Allingham.

Allingham was born in Ealing, London. Both of her parents were writers, so she was encouraged to write at an early age. When she was eight years old, her writing appeared in a magazine published by her aunt. Allingham published her first long novel, (which was not a

crime novel) when she was 19.

Allingham's first crime novel was *The White Cottage Mystery*, published in 1928. However, she felt the writing was awkward and was not proud of it. In her second novel *The Crime at Black Dudley*, she introduced Mr. Campion, the detective who would appear in her crime fiction series. Allingham was at the height of her popularity in the 1930s. Her strength was in creating strong characters rather than murder tricks or story plots.

3.1.3 Ngaio Marsh

Ngaio Marsh did most of her writing in New Zealand, where she was born. She wrote both novels and plays, mainly in New Zealand. She began writing plays in her teens, but paused when she attended the Canterbury College School of Art. After graduating from college, she was asked to join a theatrical company in London, where she resumed playwriting. She traveled to London in 1930, where she stayed until 1932, when she returned to New Zealand after her mother died. Marsh wrote many successful plays and novels. She was appointed a Dame Commander of the Order of the British Empire in 1966 in recognition of her great contribution to writing novels and plays.

Mori (1998) says that she didn't show her true skill in her first novel *A Man Lay Dead*, published in 1934. However, she succeeded in her second novel, *Enter a Murderer*, and after her third novel, *The Nursing Home Murder*, was published in 1935, she became one of the most popular crime novel writers. In 1939 she joined Christie in being represented by Collins.

Marsh is famous for the detective in her novels, Inspector Roderick Alleyn. Mori (1998) notes that Marsh put more emphasis on describing characters than on plots and murder tricks, so her novels are more comparable to Sayers' than Christie's.

3.2 Novels

The corpus used for the analysis in this paper contains 130 novels: 66 texts of Christie, 11 of Sayers, 20 of Allingham, and 33 of Marsh. We use only long novels because most of the short stories do not have enough tokens. We can indeed regard a collection of short stories as a long novel, but some of the short stories are included in multiple collections. For example, the short story *The Market Basing Mystery* was first published in 1951, in a short

story collection called *The Underdog and Other Stories*. It was also included in *13 for Luck!* in 1961 and *Poirot's Early Cases* in 1974. The novels of each author that are used in this paper are listed in Table 3.1 to Table 3.4. The word types and tokens of each author's corpus are shown in Table 3.5. The total word tokens of all corpora add up to 9,343,357 words.

All of the four authors wrote not only crime novels but also romances, plays, essays and so on. We only use detective novels because if the genres of data are different, the style and vocabulary are also different between genres.

Table 3.1: Christie's novels used in this study

No.	Title	Code	Year	Tokens	No.	Title	Code	Year	Tokens
1	*The Mysterious Affair at Styles*	C1	1920	56,938	34	*Towards Zero*	C34	1944	56,953
2	*The Secret Adversary*	C2	1922	75,647	35	*Death Comes as the End*	C35	1944	60,907
3	*The Murder on the Links*	C3	1923	59,751	36	*Sparkling Cyanide*	C36	1945	62,928
4	*The Man in the Brown Suit*	C4	1924	75,558	37	*The Hollow*	C37	1946	71,982
5	*The Secret of Chimneys*	C5	1925	74,804	38	*Taken at the Flood*	C38	1948	62,567
6	*The Murder of Roger Ackroyd*	C6	1926	70,086	39	*Crooked House*	C39	1949	55,956
7	*The Big Four*	C7	1927	55,801	40	*A Murder Is Announced*	C40	1950	72,739
8	*The Mystery of the Blue Train*	C8	1928	70,365	41	*They Came to Baghdad*	C41	1951	70,150
9	*The Seven Dials Mystery*	C9	1929	67,251	42	*Mrs. McGinty's Dead*	C42	1952	60,556
10	*The Murder at the Vicarage*	C10	1930	69,995	43	*They Do It with Mirrors*	C43	1952	51,364
11	*The Sittaford Mystery*	C11	1931	64,193	44	*After the Funeral*	C44	1953	71,306
12	*Peril at End House*	C12	1932	55,098	45	*A Pocket Full of Rye*	C45	1953	58,125
13	*Lord Edware Dies*	C13	1933	65,046	46	*Destination Unknown*	C46	1954	60,878
14	*Murder on the Orient Express*	C14	1934	58,896	47	*Hickory, Dickory, Dock*	C47	1955	56,968
15	*Why Didn't They Ask Evans?*	C15	1934	63,761	48	*Dead Man's Folly*	C48	1956	56,675
16	*Three-Act Tragedy*	C16	1935	57,430	49	*4.50 from Paddington*	C49	1957	66,063
17	*Death in the Clouds*	C17	1935	60,208	50	*Ordeal by Innocence*	C50	1958	68,830
18	*The ABC Murders*	C18	1936	59,043	51	*Cat Among the Pigeons*	C51	1959	68,380
19	*Murder in Mesopotamia*	C19	1936	67,436	52	*The Pale Horse*	C52	1961	63,099
20	*Cards on the Table*	C20	1936	56,466	53	*The Mirror Crack'd from Side to Side*	C53	1962	69,425
21	*Dumb Witness*	C21	1937	74,817	54	*The Clocks*	C54	1963	74,243
22	*Death on the Nile*	C22	1937	78,283	55	*A Caribbean Mystery*	C55	1964	53,136
23	*Appointment with Death*	C23	1938	54,508	56	*At Bertram's Hotel*	C56	1965	62,438
24	*Hercule Poirot's Christmas*	C24	1938	59,649	57	*Third Girl*	C57	1966	72,092
25	*Murder Is Easy*	C25	1939	58,010	58	*Endless Night*	C58	1967	62,077
26	*And Then There Were None*	C26	1939	53,002	59	*By the Pricking of My Thumbs*	C59	1968	70,619
27	*Sad Cypress*	C27	1940	55,594	60	*Hallowe'en Party*	C60	1969	67,372
28	*One, Two, Buckle My Shoe*	C28	1940	53,440	61	*Passenger to Frankfurt*	C61	1970	69,317
29	*Evil Under the Sun*	C29	1941	57,834	62	*Nemesis*	C62	1971	77,341
30	*N or M?*	C30	1941	55,836	63	*Elephants Can Remember*	C63	1972	60,448
31	*The Body in the Library*	C31	1942	50,089	64	*Postern of Fate*	C64	1973	75,675
32	*Five Little Pigs*	C32	1942	65,631	65	*Curtain*	C65	1975	56,694
33	*The Moving Finger*	C33	1942	56,669	66	*Sleeping Murder*	C66	1976	59,047

Table 3.2: Sayers' novels

No.	Title	Code	Year	Tokens
1	*Whose Body?*	S1	1923	59,363
2	*Clouds of Witness*	S2	1926	82,203
3	*Unnatural Death*	S3	1927	81,146
4	*The Unpleasantness at the Bellona Club*	S4	1928	72,201
5	*Strong Poison*	S5	1930	77,691
6	*The Five Red Herrings*	S6	1930	111,728
7	*Have His Carcase*	S7	1932	138,018
8	*Murder Must Advertise*	S8	1933	108,599
9	*The Nine Tailors*	S9	1934	107,924
10	*Gaudy Night*	S10	1935	159,577
11	*Busman's Honeymoon*	S11	1937	116,569

Table 3.3: Allingham's novels

No.	Title	Code	Year	Tokens
1	*The White Cottage Mystery*	A1	1928	41,077
2	*The Crime at Black Dudley*	A2	1929	68,272
3	*Mystery Mile*	A3	1930	68,595
4	*Look to the Lady*	A4	1931	77,560
5	*Police at the Funeral*	A5	1931	88,436
6	*Sweet Danger*	A6	1933	74,114
7	*Death of a Ghost*	A7	1934	75,832
8	*Flowers for the Judge*	A8	1936	83,091
9	*Dancers in Mourning*	A9	1937	97,997
10	*The Case of the Late Pig*	A10	1937	38,084
11	*The Fashion in Shrouds*	A11	1938	104,798
12	*Black Plumes*	A12	1940	69,689
13	*Traitor's Purse*	A13	1941	70,812
14	*Coroner's Pidgin*	A14	1945	80,420
15	*More Work for the Undertaker*	A15	1948	74,816
16	*The Tiger in the Smoke*	A16	1952	93,597
17	*The Beckoning Lady*	A17	1955	79,356
18	*Hide My Eyes*	A18	1958	72,517
19	*The China Governess*	A19	1962	89,163
20	*The Mind Readers*	A20	1965	86,236

Table 3.4: Marsh's novels

No.	Title	Code	Year	Tokens
1	*A Man Lay Dead*	M1	1934	49,328
2	*Enter a Murderer*	M2	1935	56,703
3	*The Nursing Home Murder*	M3	1935	61,041
4	*Death in Ecstasy*	M4	1936	72,902
5	*Vintage Murder*	M5	1937	74,395
6	*Artists in Crime*	M6	1938	99,172
7	*Death in a White Tie*	M7	1938	96,943
8	*Overture to Death*	M8	1939	91,334
9	*Death at the Bar*	M9	1940	83,697
10	*A Surfeit of Lampreys*	M10	1941	96,334
11	*Death and the Dancing Footman*	M11	1942	105,552
12	*Colour Scheme*	M12	1943	94,216
13	*Died in the Wool*	M13	1945	84,119
14	*Final Curtain*	M14	1947	91,545
15	*Swing, Brother, Swing*	M15	1949	86,050
16	*Opening Night*	M16	1951	72,794
17	*Spinsters in Jeopardy*	M17	1954	85,932
18	*Scales of Justice*	M18	1955	78,644
19	*Off With His Head*	M19	1957	84,114
20	*Singing in the Shrouds*	M20	1959	68,765
21	*False Scent*	M21	1960	67,226
22	*Hand in Glove*	M22	1962	62,358
23	*Dead Water*	M23	1964	68,688
24	*Death at the Dolphin*	M24	1967	81,102
25	*Clutch of Constables*	M25	1968	71,662
26	*When in Rome*	M26	1970	70,846
27	*Tied Up in Tinsel*	M27	1972	73,568
28	*Black As He's Painted*	M28	1974	75,144
29	*Last Ditch*	M29	1977	80,280
30	*Grave Mistake*	M30	1978	79,935
31	*Photo Finish*	M31	1980	77,578
32	*Light Thickens*	M32	1982	68,424

Table 3.5: Basic summary of each author's data

Author	No of novels	Types	Tokens	STTR
Christie	66	46,009	4,183,485	42.25
Sayers	11	36,003	1,115,019	44.40
Allingham	20	34,912	1,534,462	45.61
Marsh	32	45,242	2,510,391	44.23

In this paper, all the texts are tagged with part-of-speech (POS) tags. POS tagging is conducted using Tree Tagger. Tree Tagger is a tool for annotating text with part-of-speech and lemma information developed by Helmut Schmid at the Institute for Computational Linguistics at the University of Stuttgart [1]."

Table 3.6: Tree Tagger tag set

POS tag	Description	POS tag	Description
CC	coordinating conjunction	PRP$	possessive pronoun
CD	cardinal number	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential there	RBS	adverb, superlative
FW	foreign word	RP	particle
IN	preposition/subord. conj.	SYM	symbol
JJ	adjective	TO	*to*
JJR	adjective, comparative	UH	interjection
JJS	adjective, superlative	VB	verb, base form
LS	list marker	VBD	verb, past
MD	modal	VBG	verb, gerund or present participle
NN	noun, singular or mass	VBN	verb, past participle
NNS	noun plural	VBP	verb, non-3rd person singular present
NNP	proper noun, singular	VBZ	verb, 3rd person singular present
NNPS	proper noun, plural	WDT	wh-determiner
PDT	predeterminer	WP	wh-pronoun
POS	possessive ending	WP$	possessive wh-pronoun
PRP	personal pronoun	WRB	wh-adverb

[1]TreeTagger – a part-of speech tagger for many languages (http://www.cis.uni-muenchen.de/ schmid/tools/ TreeTagger/) (Accessed on June 15th, 2018)

Chapter 4

Comparing Christie's style with other authors' style

In this chapter, we compare Christie's style with three other authors using stylometric approaches: unsupervised approaches and supervised approaches. We will compare Christie's style with the other three authors' styles using unsupervised methods first, and we will examine if we can distinguish Christie's works from other authors' works using a supervised method. Özgür (2004) says that machine learning approaches can be grouped in two main categories as supervised (document classification) and unsupervised (document clustering). She notes, 'text classification, also known as text categorization or topic spotting, is a supervised learning task, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents,' and 'document clustering is an unsupervised learning task, which does not require pre-defined categories and labelled documents.' The research questions in this chapter are the following:

1. Can we find stylistic differences between Christie's works and other authors' works using statistical methods?

2. What are the stylistic differences between the works of Christie and the other authors?

4.1 Methods

To compare Christie's style with the styles of the other authors we use both unsupervised methods and a supervised method. As unsupervised methods, we use cluster analysis and

principal component analysis. As supervised methods, we use Random Forests and compare the results with unsupervised methods. In the following section, we introduce Random Forests.

4.1.1 Random Forests

Proposed by Breiman (2001), Random Forests (RF) is a machine learning method for classification and regression. In this study, RF is applied for classification. RF constructs a multitude of decision trees using a bootstrap training set. About one-third of the data are left out of the bootstrap sample and are not used in the construction of the trees. These data were used for testing the constructed model, thus there is no need for cross-validation. According to Jin (2007), RF has the following merits :

1. The accuracy of classifying/regression is high.
2. It works on big data efficiently, and we can treat hundreds of thousands of variables.
3. It estimates the importance of the variables used in the analysis.
4. It is valid in inferring missing values and in sustaining the accuracy of the data holding a lot of missing values.
5. The balance of the errors is maintained when using data in which the number of individuals is unbalanced among groups in classifying.
6. It calculates the information related to the relationship between classifying and variables.
7. It can calculate the degree of approximation among groups.
8. It can be adapted to the data without external criteria, and so on.

Random Forests has been used for classification of texts and authorship attribution. Jin & Murakami (2007) employed RF for authorship identification within three different types of texts (novels, compositions and diaries), and showed that this method is more effective than other classifiers such as Nearest Neighbour, Learning Vector Quantization, Support Vector Machine, Bagging and Boosting.

Kobayashi, Tanaka & Tomiura (2012) employed RF to classify English scientific papers into two groups: papers by native speakers and those by non-native speakers. They discussed characteristic expressions of each group.

Tabata (2014) used RF to extract marker words that distinguish Charles Dickens from Wilkie Collins. Tabata argues that RF overcomes common problems in keyword measures such as Log-Likelihood or Chi-squared scores, creating an attractive alternative. In this study, we use RF to extract keywords that distinguish Christie's works from the other authors' works.

4.2 Result 1: Unsupervised methods

We compare Christie's style with the other three authors' styles using unsupervised methods. The variables used in these analyses are the top 700 common words present in all four authors' novels. The number 700 is the number of variables at which the RF classification was found to be the most accurate. The most common words are reported to be helpful for authorship identification in the series of Burrows' studies on the style of Jane Austen (such as Burrows 1987, 1992), as well as many other studies on stylistics and authorship attribution. The top 700 words included protagonists' proper nouns such as Poirot, Wimsey, Alleyn and Campion, and were thus excluded from the analysis. This is because Christie wrote novels with different protagonists (Hercule Poirot, Miss Marple, Tommy and Tuppence Beresford and so on), while each of the other authors' novels featured only a single protagonist. Therefore, in Allingham's novels, for example, the word Campion occurs in all of her novels. The results of both unsupervised methods and RF are influenced. There is a possibility that these words would be extracted as characteristic words.

Figure 4.1 is a dendrogram of the cluster analysis. We used Euclidean Distance for interval and Ward's Method for cluster method. We can see that although some of Christie's earlier works are included in other authors' clusters, most of Christie's works form a distinct cluster.

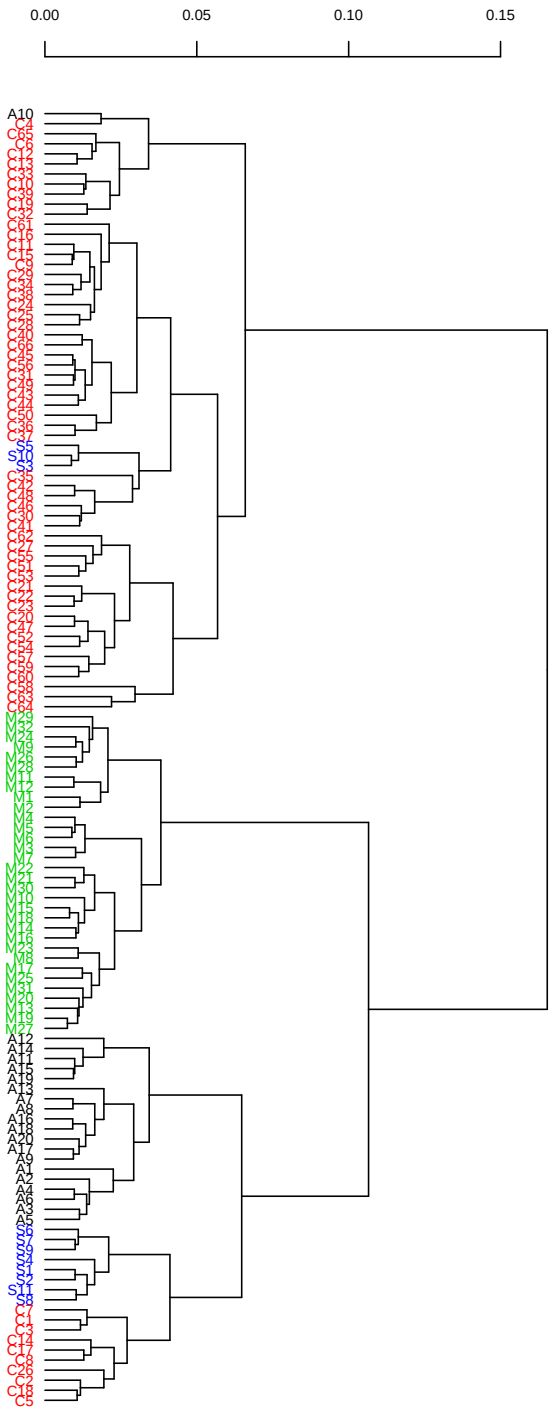

Figure 4.1: A dendrogram of cluster analysis

After overviewing the data briefly through the dendrogram of cluster analysis, we would like to look at what words are characteristic in Christie's texts. Figure 4.2 is the resulting plot of principal component analysis (PCA). The horizontal axis represents principal component one, and the vertical axis represents principal component two. The cumulative loadings are about 52%. We can see most of Christie's works on the right side of the graph and other authors' works on the left. Some works of Christie and other authors overlap, but they are all Christie's earlier works.

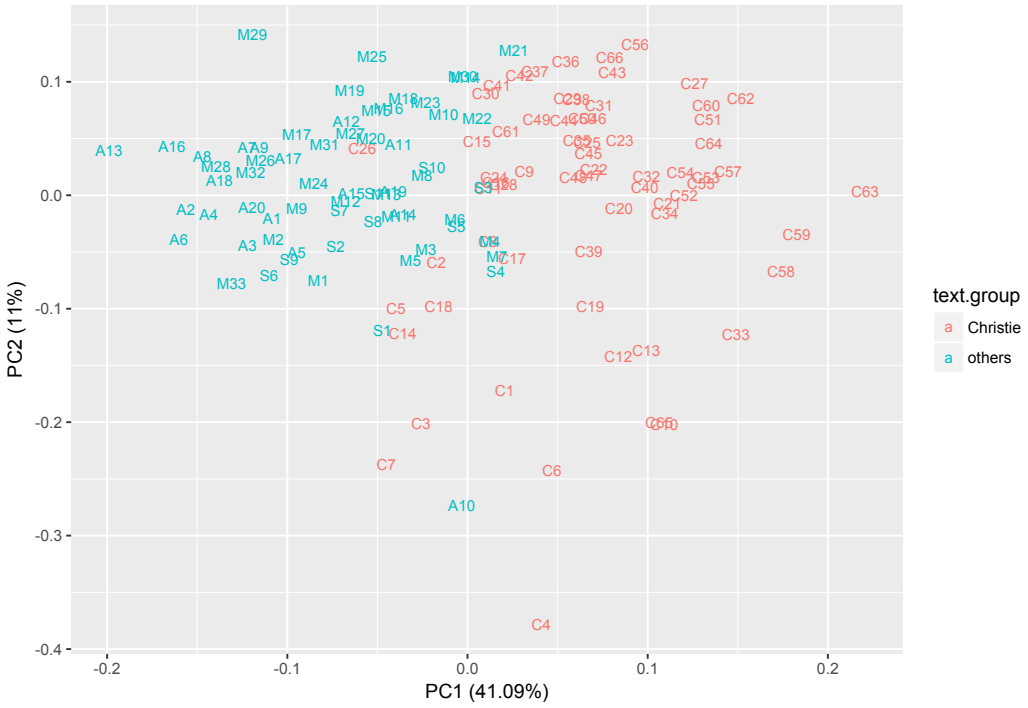

Figure 4.2: The results of principal component analysis

To see what words contribute to the first and second principal components, we draw two barplots that show the top 20 words with the largest and the smallest values which contribute to each principal component, principal component one and two (See Figure 4.3 and 4.4).

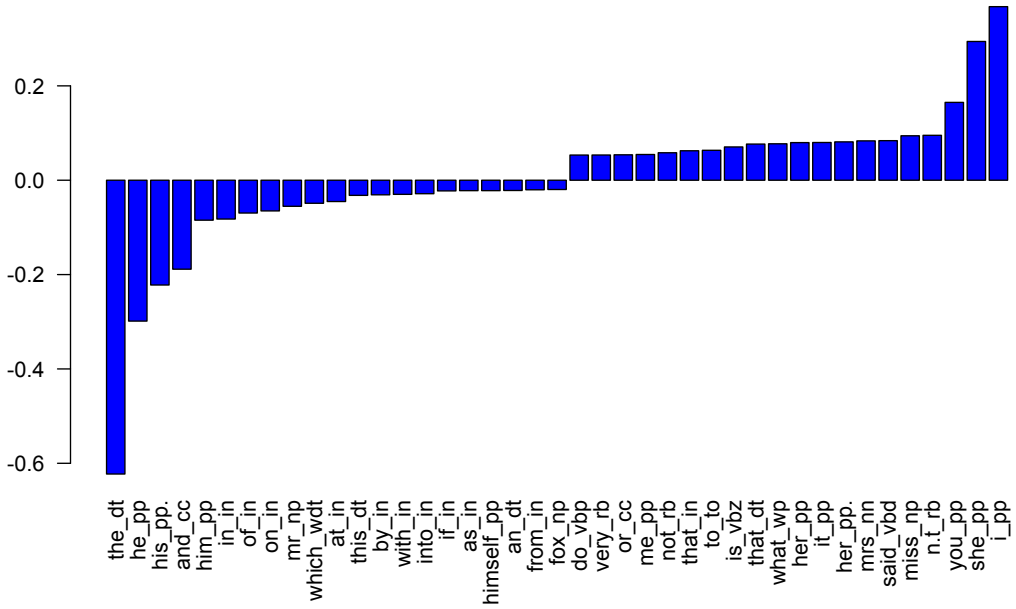

Figure 4.3: Barplot of the words with the biggest/smallest 1st principal component

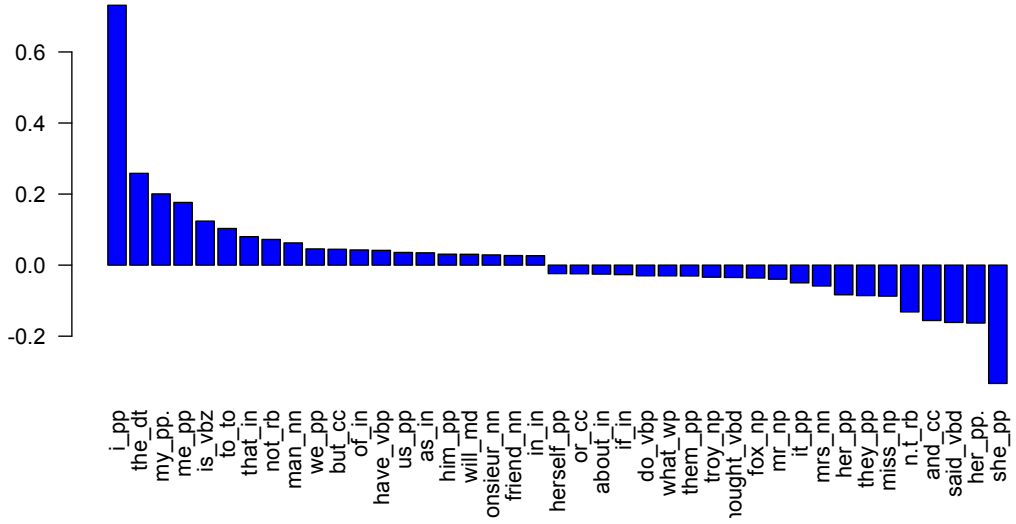

Figure 4.4: Barplot of the words with the biggest/smallest 2nd principal component

In Figure 4.3, we can see a contrast between female and male words: the positive side

contains words related to the female gender such as *she*, *her*, *miss*, and *mrs*, and the negative side contains words related to the male gender such as *he*, *his*, *him* and *himself*. This means that Christie's works relate to female words more than the other three authors' works, and other authors' works relate to male words more than Christie's.

Moreover, we can find that the top 20 words with the highest values contain some first and second person pronouns, such as *I*, *you*, and *me*.

From the words contained in Figures 4.3 and 4.4, at first it seems that this result of the PCA is influenced by the detectives of each story and whether the narrative of the story is first person narrative or third person narrative.

In Sayers', Allingham's and Marsh's works, the detectives are men. Christie's works feature both male (Poirot, Tommy) and female detectives (Miss Marple, Tuppence). Readers tend to think that the protagonist detective should make the most appearances in a crime novel; therefore the pronouns of the protagonist's gender should appear frequently. However, in C63 (*Elephants Can Remember*), which is plotted on the right-most side of the plot, the detective is Poirot. Thus, there seems to be little relationship between the result of the PCA and the presence of detectives in the story. The reason why C63 is on the eastern-most side of the plot might be that Mrs. Oliver, a recurring character in Christie's novels, is the main character in this particular work.

In the result of the PCA, there is a contrast between female and male words. In order to see whether these words are used differently in Christie's and the other authors' novels, we would like to see the tables of the top pentagrams including the pronouns *she* and *he* as an example. Christie's works are featured by more female-leaning female pronouns although other authors' works are featured by more male-leaning pronouns. Therefore we focus on *she* in Christie's novels and *he* in the other authors' novels.

Table 4.1: Top pentagrams with *she* in the Christie corpus (Left) and those with *he* in the other authors' corpus (Right)

No	Cluster	Freq	No	Cluster	Freq
1	she went out of the	37	1	he waited for a moment	45
2	she said in a low	31	2	there you are he said	38
3	she was silent for a	25	3	i don't know he said	36
4	she shook her head no	24	4	at the moment he was	35
5	or two then she said	21	5	for the first time he	34
6	a minute or two she	20	6	a matter of fact he	31
6	she paused and then said	20	6	as if he had been	31
8	that she was going to	19	6	as soon as he had	31
9	but i don't think she	17	6	he said at last i	31
9	she was very fond of	17	6	he said i don't know	31
9	shook her head she said	17	11	he rose to his feet	26
12	a matter of fact she	16	12	he walked over to the	24
12	for the first time she	16	12	if you don't mind he	24
12	i asked her if she	16	14	as far as he could	23
12	i do not think she	16	15	he hung up the receiver	22
12	she shook her head i	16	15	he was one of the	22
12	she was going to be	16	17	he hesitated for a moment	20
18	it was as though she	15	17	that he was going to	20
18	moment or two then she	15	19	he caught a glimpse of	19
18	she paused a minute and	15	19	he said over his shoulder	19
18	she rose to her feet	15	19	he seems to have been	19

Table 4.1 is the list of pentagrams including *she* in the Christie corpus and those with *he* in other authors' corpora. There are many expressions related to conversations, such as *she said in a low*, *there you are he said*, and so on. In the expression of the subjunctive mood, we can see the contrast between Christie and the other authors: although the other three authors prefer to use *as if* over *as though*, Christie prefers *as though* over *as if*. Christie hardly ever uses *as if* in her novels. She uses *as if* only 91 times in her works though she uses *as though* more than 1500 times (See Figure 4.5).

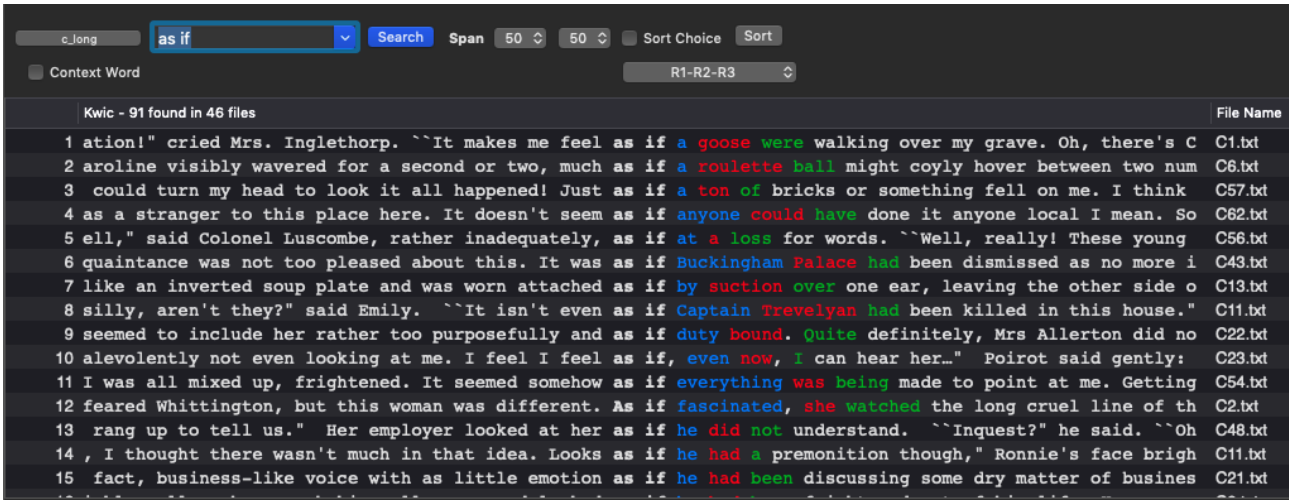

Figure 4.5: An example of the concordance lines including *as if* in Christie's corpus

To see the difference between the use of the female pronoun and that of the male pronoun more clearly, we would like to see the collocates of *she* and *he* using the Mutual Information (MI) score. MI score is one of the measures of the significance of a co-occurrence. According to Hunston (2002), MI score depends on 'two calculations: how many instances of the co-occurring word are found in the designated span of the node word (the Observed), and how many instances might be expected in that span, given the frequency of the co-occurring word in the corpus as a whole (the Expected).' Hunston also notes, 'the MI score is the Observed divided by the Expected, converted to a base-2 logarithm.' The feature of MI score is described as the following:

> If a word occurs rarely, but in most of its few occurrences appears in the proximity of another word, the collocation between those words will obtain a high MI score. MI, then, is a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of the two words (Hunston 2002: 72).

In this paper, MI score is selected in order to find collocates which do not occur frequently but are significant.

Since Sinclair and Jones (1974) state that 'for any node, a very high proportion of relevant information could be obtained by examining collocates at positions N-4 to N+4,' this study

deals only with collocates at position N-4 to N+4. The minimum frequency of collocates was set at 10, and the threshold of the MI score was set at 3.00.

Table 4.2: Top collocates of *she* in the Christie corpus (left) and *he* in other authors' corpore (right) (MI score >=3.00)

No	Word	MI	Word	MI	No	Word	MI	Word	MI
1	fainted (VBD)	4.11	hit	3.65	24	breathlessly	3.25	thinks	3.14
2	stricken	4.01	sweated	3.64	25	hated	3.22	related	3.10
3	wept	3.87	states	3.58	26	burst	3.20	resignedly	3.07
4	bit	3.86	swears	3.55	27	hates	3.18	deserved	3.06
5	frightens	3.80	dopey	3.47	28	realizes	3.18	illusions	3.06
6	choked	3.69	admits	3.41	29	accusingly	3.17	supposed	3.05
7	fainted (VBN)	3.65	cherubims	3.4	30	owns	3.16	straightened	3.04
8	minded	3.62	pulls	3.33	31	clung	3.13	kills	3.04
9	sobbed	3.60	feared (VBN)	3.31	32	appealingly	3.11	confessed	3.04
10	lied	3.60	shaved	3.31	33	unreliable	3.11	winked	3.03
11	sensed	3.55	realizes	3.29	34	added	3.10	fanatic	3.01
12	unwrapped	3.55	fancied	3.28	35	says	3.06		
13	fragile	3.53	reckons	3.27	36	collided	3.05		
14	shrank	3.51	conversationally	3.26	37	walks	3.03		
15	tartar	3.47	short-sighted	3.24	38	exhausted	3.03		
16	riding	3.46	stripped	3.24	39	wistfully	3.03		
17	talks	3.40	feared (VBD)	3.23	40	bustled	3.02		
18	herself	3.33	drinks	3.19	41	clenched	3.02		
19	supposed	3.31	rambled	3.19	42	screamed	3.01		
20	appealed	3.31	reflectively	3.19	43	scared	3.01		
21	insists	3.31	hates	3.17	44	likes	3.00		
22	boasted	3.31	enquired	3.15					
23	shivered	3.26	ground	3.15					

Table 4.2 shows the collocates of *she* in Christie's works and those of *he* in other authors' works whose MI scores are over 3.00. The collocates of *she* in Christie's works consist of words related to crying such as *wept*, *sobbed* and *screamed*. There are also several words related to actions indicative of fear or sadness, such as *fainted*, *shrank*, *shivered*, *clung* and *scared*. Moreover, there is a word related to weakness: *fragile*. It is interesting that the collocates of *he* in other authors' works contain both words related to fears (*feared*) and words

related to killers (*kills*).

Table 4.3: Top collocates of *she* in other authors corpore (MI score >=3.00)

No	Word	MI	No	Word	MI
1	refuses	4.02	30	loved	3.22
2	herself	3.84	30	courageous	3.22
3	breathlessly	3.73	30	sobbed	3.22
4	migraine	3.71	30	crying	3.22
5	huskily	3.69	34	scream	3.21
6	frightened	3.62	35	disliked	3.20
7	stammered	3.61	36	upset	3.15
8	shivered	3.60	37	governess	3.14
9	realised	3.59	37	blushed	3.14
10	suspects	3.58	37	confided	3.14
11	hendy	3.54	37	remembers	3.14
12	hates	3.52	37	feared	3.14
13	faint	3.50	42	lives	3.12
14	resolutely	3.47	42	respond	3.12
15	burst	3.43	42	shivering	3.12
15	supposed	3.43	42	rouse	3.12
17	married	3.42	46	carstairs	3.11
17	resented	3.42	47	imagined	3.10
19	distractedly	3.39	48	fainted	3.08
19	defiantly	3.39	48	patted	3.08
21	pathetically	3.34	50	loves	3.06
21	unsteadily	3.34	50	suffered	3.06
23	adored	3.30	52	slept	3.05
24	emphatic	3.29	53	wished	3.04
24	impulsively	3.29	53	accused	3.04
24	whispered	3.29	53	hopes	3.04
27	excitable	3.28	56	flatly	3.01
28	composedly	3.27	57	realized	3.00
29	murderess	3.26			

(1) "She's as romantic as a valentine. As far as I can gather she *adored* her husband and wants every girl to have the same glorious experience." (*Hide My Eyes* by Margery Allingham)

(2) "You're not allowed to snub me, Fabian, or talk over my head or go intellectual at me.

I loved her. She was my friend." (*Died in the Wool* by Ngaio Marsh)

In order to see whether the use of *she* in Christie's works is different from that of the other authors' works, we would like to see the collocates of *she* in their works. Table 4.3 is the list of the collocates of *she* in the other authors' novels. Like the collocates in Christie's works, there are words related to crying or weakness such as *crying*, *shivered* and so on. What is different from the collocates in Christie's novels is that those in other authors' works contain some words related to love or marriage, such as *married*, *adored*, *loved* and *loves*. In Christie's novels, we can see stereotypical femininity from the collocates of *she*, but in other authors' novels, we see not only that women have stereotypical femininity but also that they *love/loved/adored* someone/something, or they are *loved/adored* by someone.

The result of the analysis above was effected by the third person pronouns. What become if we exclude the third person pronouns from the data? Figure 4.6 is a resulting plot of PCA. We excluded third pronouns (*he/his/him/himself*, *she/her/herself*, *they/their/them/themselves*) from the data used in PCA.

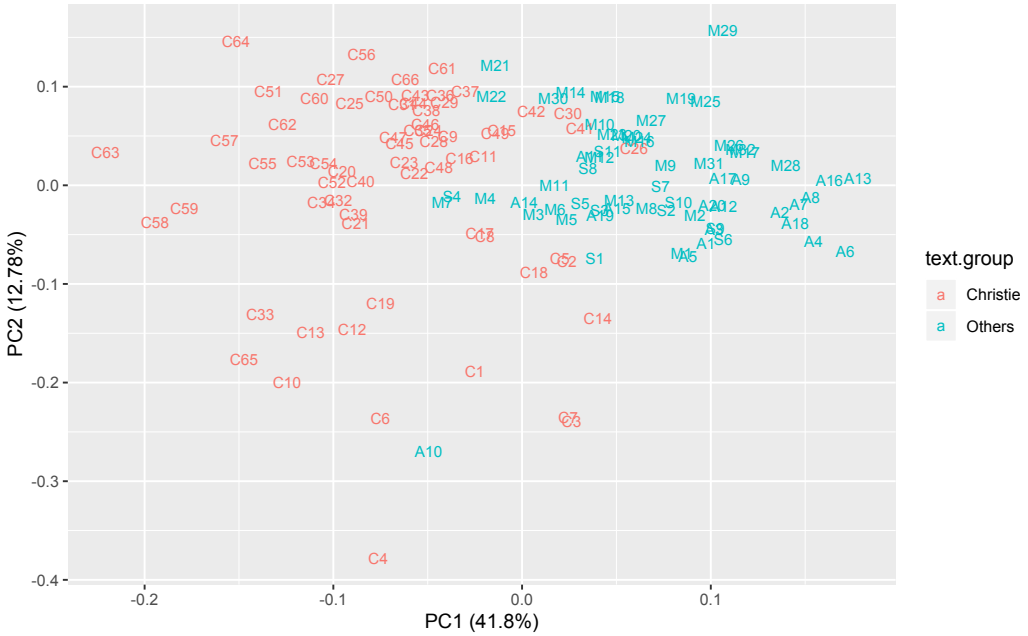

Figure 4.6: The result of principal component analysis (third person pronouns are excluded)

Likewise the result of PCA in Figure , most of Christie's works are plotted on the left side and other authors' works are on the right side of the graph.

In the barplots in Figures 4.3 and 4.4, we can see the word *said*. It is possible that *said* is extracted as a characteristic word because Christie's novels consist of many conversations. Therefore, proportion of conversations in Christie's works was investigated. Figure 4.7 shows the proportion of conversations and narratives in each Christie's work.

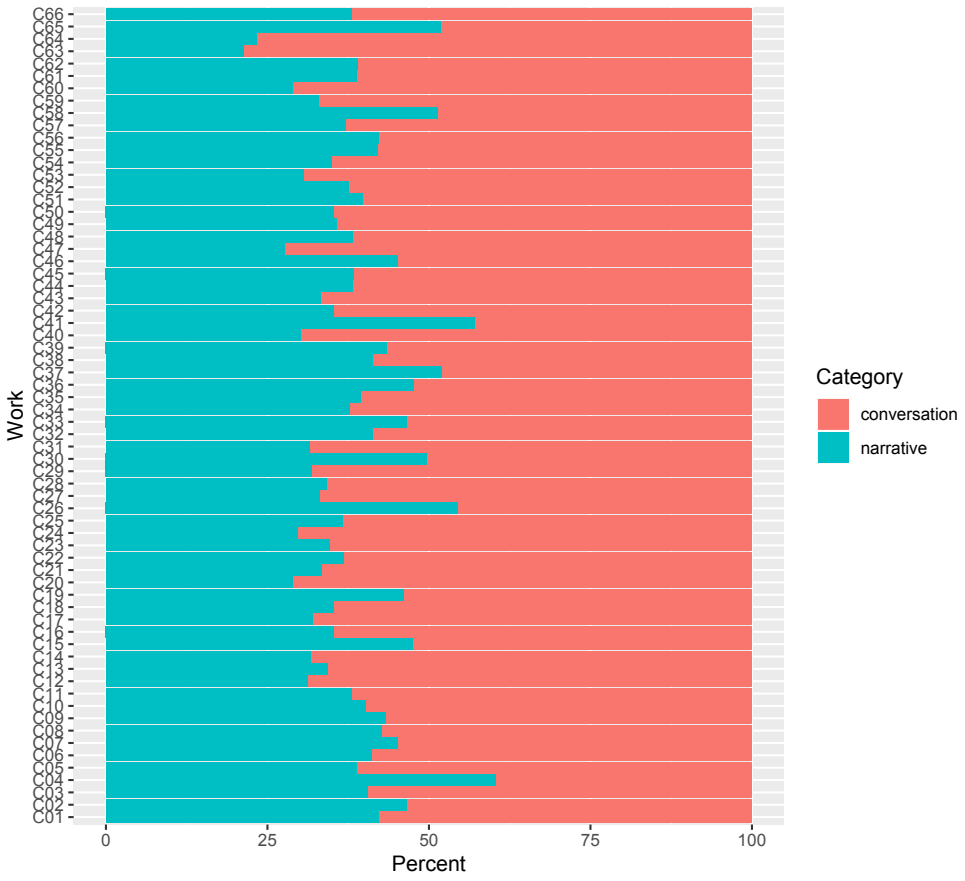

Figure 4.7: The proportion of conversations and narratives in Christie's novels

In most of Christie's novels, the proportion of conversations is larger than that of narratives. This figure indicates that her works are mainly conversations. This feature is outstanding in her later works, especially in her last two novels (*Elephants Can Remember* and *Postern of Fate*).

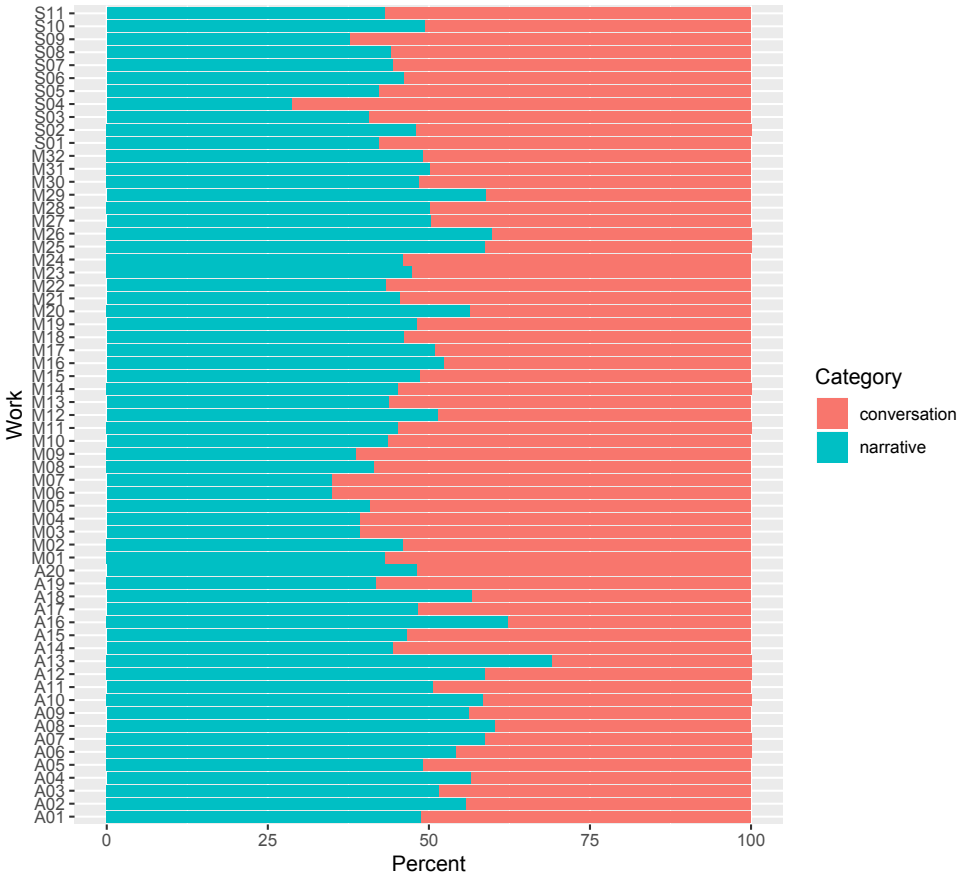

Figure 4.8: The proportion of conversations and narratives in other three authors' novels

Figure 4.8 shows the proportion of conversations and narratives in the works of Sayers, Allingham and Marsh. Compared with Figure 4.7, the proportion of conversations in some of Sayers' and Marsh's novels is large, but the proportion of conversations in most of other authors' works is smaller than that of Christie's novels. Marsh's earlier works have larger proportion of conversations than her later works. This might be because she was good at writing both novels and plays, and her novels were like plays. According to Figures 4.7 and 4.8, Christie's works seem to have more conversations than other three authors.

4.3 Result 2: Supervised method

In this section, we employ RF to distinguish Christie's texts from those of the other authors and extract keywords from Christie's works. RF is trained and validated on the 129 texts with differing numbers of most common words ranging from 1000 to 100 in 100 words steps. The four authors' texts were classified into four different groups–the most accurately with the top 700 words. The average accuracy was 100%. R ver. 3.5.0 is employed to conduct RF, and the package RF is used. Table 4.4 shows the result of RF.

Table 4.4: An example of the result of Random Forests

Call:			
randomForest(formula = author ~.,			
data=csam.all, proximity=T, importance=T, ntree=10000)			
Type of random forest: classification			
Number of trees:	10000		
No. of variables tried at each split:	26		
OOB estimate of error rate:	0%		
Confusion matrix:			
	Christie	Sayers,Allingham,Marsh	class.error
Christie	66	0	0.0000000
Sayers, Allingham, Marsh	0	63	0.0000000

Figure 4.9 is a Multi-Dimensional Scaling Plot of the result of RF. Christie's works are plotted on the left side and the other authors works are on the right side. They are completely separated.

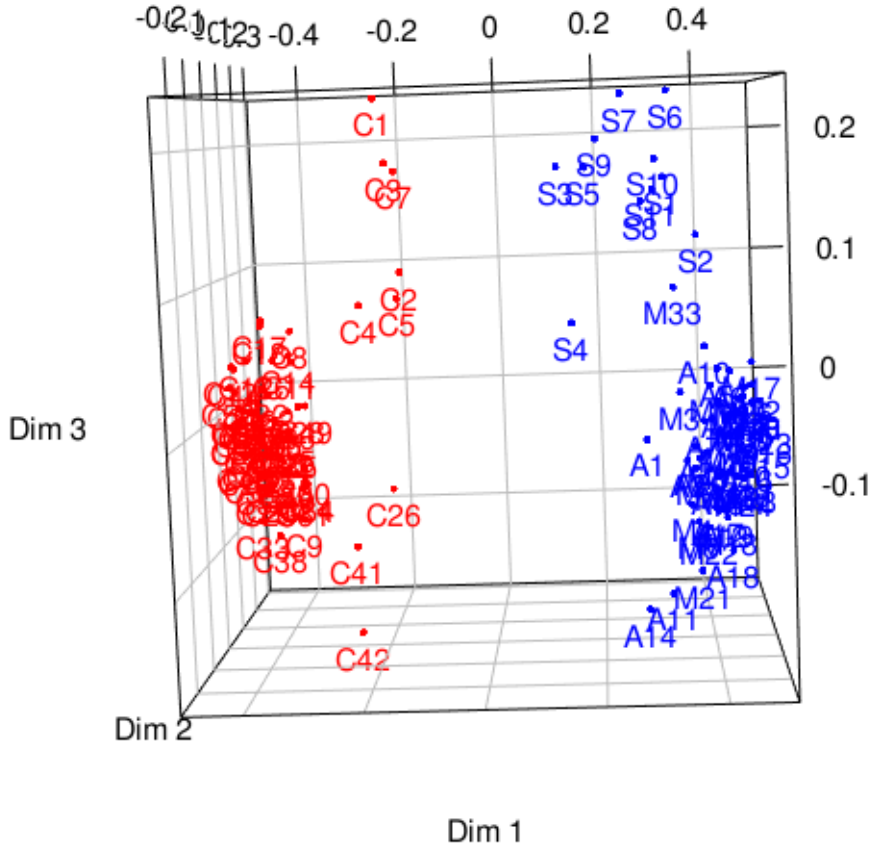

Figure 4.9: MDS plot of the result of Random Forests

The word cloud in Figure 4.10 shows the characteristic words that contributed to the classification the most, based on the mean decrease in the GINI importance.

Figure 4.10: Wordcloud of the characteristic words that contributed to the Random Forest classification the most

Table 4.5: The top 100 key words of each group extracted by Random Forests

Christie
thoughtfully_rb, chapter_np, different_jj, what_wp, true_jj, shook_vbd, always_rb, quite_rb, idea_nn, sometimes_rb, really_rb, things_nns, wrong_jj, kind_nn, that_wdt, anyone_nn, no_uh, that_dt, mean_vbp, though_in, just_rb, reason_nn, nodded_vbd, not_rb, easy_jj, life_nn, became_vbd, deal_nn, real_jj, then_rb, minute_nn, something_nn, seems_vbz, very_rb, seem_vb, think_vbp, is_vbz, sure_jj, everything_nn, why_wrb, perhaps_rb, exactly_rb, wanted_vbd, interesting_jj, killed_vbn, slightly_rb, slowly_rb, difficult_jj, possible_jj, yes_uh, woman_nn, say_vbp, someone_nn, nice_jj, know_vb, death_nn, once_rb, likely_jj, had_vbn, women_nns, day_nn, people_nns, anybody_nn, young_jj, sharply_rb
Sayers, Allingham and Marsh
while_in, above_in, appeared_vbd, if_in, its_pp, under_in, light_nn, god_np, and_cc, began_vbd, the_dt, mouth_nn, into_in, wait_vb, back_nn, that_in, before_in, when_wrb, his_pp$$, sort_nn, on_in, but_cc, found_vbd, made_vbd, look_vb, glanced_vbd, hands_nns, been_vbn, there_ex, rest_nn, stood_vbd, which_wdt, was_vbd, off_in, upon_in

The top 100 keywords are identified and classified into two groups: Christie or the other three authors. The result of the classification of keywords is shown in Table 4.5.

In the characteristic words of Christie, we can see many *-ly* adverbs such as *thoughtfully*, *exactly, slightly*, *slowly* and *sharply*. No *-ly* adverb was extracted as the characteristic word of the other three authors.

First, take a look at *thoughtfully*. Table 4.6 is the list of top three-word clusters of *thoughtfully* in the Christie corpus and Table 4.7 is the list of top four-word clusters of *thoughtfully*.

Table 4.6: Top three-word clusters with *thoughtfully* in the Christie corpus

No	Cluster	Freq
1	said_vbd poirot_np thoughtfully_rb	85
2	he_pp said_vbd thoughtfully_rb	41
3	i_pp said_vbd thoughtfully_rb	38
3	miss_np marple_np thoughtfully_rb	38
5	at_in her_pp$ thoughtfully_rb	33
6	looked_vbd thoughtfully_rb at_in	30
7	at_in him_pp thoughtfully_rb	25
7	said_vbd thoughtfully_rb i_pp	25
9	poirot_np nodded_vbd thoughtfully_rb	24
10	poirot_np said_vbd thoughtfully_rb	22
11	she_pp added_vbd thoughtfully_rb	21
12	he_np added_vbd thoughtfully_rb	18
12	his_pp$$ head_nn thoughtfully_rb	18
14	she_pp said_vbd thoughtfully_rb	16
14	thoughtfully_rb at_in the_dt	16

As we can see from these tables, *thoughtfully* often appears with the verbs *said* (*say*) and *looked* (*look*). *Thoughtfully* is often used with these words to describe the situation of utterance or character action.

(3) "His voice?" *said Poirot thoughtfully*. "Ah, his voice! Would you know it again, Mademoiselle Zia?" (*The Mystery of the Blue Train*)

(4) "I just wondered if it might have been the wrong murder," *said Miss Marple thoughtfully*. (*The Mirror Crack'd from Side to Side*)

(5) She spoke impatiently. Poirot *looked at her thoughtfully* as he asked: "Has she threatened you in actual words in public? Used insulting language? Attempted any bodily harm?" (*Death on the Nile*)

Table 4.7: Top four-word clusters with *thoughtfully* in the Christie corpus

No	Cluster	Freq
1	looked_vbd at_in her_pp$ thoughtfully_rb	33
1	said_vbd miss_np marple_np thoughtfully_rb	33
3	looked_vbd at_in him_pp thoughtfully_rb	23
4	yes_np said_vbd poirot_np thoughtfully_rb	19
5	thoughtfully_rb i_pp do_vbp n't_rb	12
6	looked_vbd at_in me_pp thoughtfully_rb	10
6	shook_vbd his_pp$ head_nn thoughtfully_rb	10
8	looked_vbd thoughtfully_rb at_in the_dt	9
8	said_vbd poirot_np thoughtfully_rb i_pp	9
10	said_vbd poirot_np thoughtfully_rb it_pp	8
10	yes_np i_pp said_vbd thoughtfully_rb	8
12	eyes_nns rested_vbd thoughtfully_rb on_in	7
12	he_pp looked_vbd thoughtfully_rb at_in	7
12	looked_vbd thoughtfully_rb at_in her_pp$	7
12	poirot_np looked_vbd thoughtfully_rb at_in	7
12	said_vbd poirot_np thoughtfully_rb he_pp	7

Moving on to *exactly*, Table 4.8 is the list of three-word clusters, and Table 4.9 is the list of four-word clusters in the Christie corpus.

(6) "Hastings, mon ami, that woman is either *exactly what she* seems or else she is a very good actress." (*Dumb Witness*)

(7) "*What exactly do you* want to know, Dr Calgary?" asked Superintendent Huish, but before Calgary could speak the telephone rang on Huish's desk and the superintendent picked it up. (*Ordeal by Innocence*)

This word is used mostly in conversations, but it is also used in narratives when accompanied by *the same*.

(8) In the doorway, dressed *exactly the same* as she had been at Sunny Ridge, and smiling the same way with that air of vague amiability, was Mrs Lancaster in person. (*By the Pricking of My Thumbs*)

(9) George, Rosemary and Iris all called her Ruth and she often came to Elvaston Square to lunch. She was now twenty-nine and looked *exactly the same* as she had looked at twenty-three. (*Sparkling Cyanide*)

Table 4.8: Top three-word clusters with *exactly* in the Christie corpus

No	Cluster	Freq
1	exactly_rb the_dt same_jj	48
2	exactly_rb what_wp i_pp	44
3	n't_rb know_vb exactly_rb	38
4	exactly_rb what_wp you_pp	34
5	exactly_rb what_wp he_pp	30
5	know_vb exactly_rb what_wp	30
7	exactly_rb do_vbp you_pp	29
8	exactly_rb what_wp she_pp	25
9	what_wp exactly_rb do_vbp	23
10	to_to know_vb exactly_rb	22
11	me_pp exactly_rb what_wp	21
12	tell_vb me_pp exactly_rb	20
13	is_vbz exactly_rb what_wp	19
13	it_pp was_vbd exactly_rb	19
13	just_rb exactly_rb what_wp	19
16	exactly_rb as_in it_pp	18
16	exactly_rb like_in a_dt	18
18	exactly_rb as_in you_pp	17
18	exactly_rb what_wp it_pp	17
20	exactly_rb said_vbd miss_np	16
20	n't_rb remember_vb exactly_rb	16
20	was_vbd n't_rb exactly_rb	16
20	what_wp exactly_rb is_vbz	16

Table 4.9: Top four-word clusters with *exactly* in the Christie corpus

No	Cluster	Freq
1	do_vbp n't_rb know_vb exactly_rb	35
2	what_wp exactly_rb do_vbp you_pp	23
3	tell_vb me_pp exactly_rb what_wp	14
4	exactly_rb said_vbd miss_np marple_np	12
5	exactly_rb what_wp he_pp was_vbd	11
5	n't_rb know_vb exactly_rb what_wp	11
5	that_dt is_vbz exactly_rb what_wp	11
8	ca_md n't_rb remember_vb exactly_rb	10
8	exactly_rb what_wp it_pp was_vbd	10
10	exactly_rb do_vbp you_pp mean_vb	9
10	is_vbz exactly_rb what_wp i_pp	9
10	to_to know_vb exactly_rb what_wp	9
13	it_pp was_vbd n't_rb exactly_rb	8
13	you_pp tell_vb me_pp exactly_rb	8
15	difficult_jj to_to say_vb exactly_rb	7
15	exactly_rb as_in it_pp had_vbd	7
15	exactly_rb the_dt same_jj as_in	7
15	exactly_rb what_wp i_pp was_vbd	7
15	exactly_rb what_wp you_pp mean_vbp	7
15	i_pp did_vbd n't_rb exactly_rb	7
15	it_pp 's_vbz not_rb exactly_rb	7
15	want_vbp to_to know_vb exactly_rb	7

Next, consider *slightly*. Table 4.10 is the list of top three-word clusters and Table 4.11 is the list of top four-word clusters in the Christie corpus. We can see that Christie used this word frequently with fixed phrases such as *out of breath* and *taken aback*.

Table 4.10: Top three-word clusters with *slightly* in the Christie corpus

No	Cluster	Freq
1	in_in a_dt slightly_rb	46
2	slightly_rb taken_vbn aback_rb	32
2	with_in a_dt slightly_rb	32
4	her_pp$ head_nn slightly_rb	16
4	looked_vbd slightly_rb surprised_vbn	16
4	slightly_rb out_rb of_in	16
7	ever_rb so_rb slightly_rb	13
8	looked_vbd slightly_rb taken_vbn	12
9	slightly_rb as_in he_pp	11
10	his_pp$ head_nn slightly_rb	10
10	slightly_rb at_in the_dt	10
12	a_dt slightly_rb different_jj	9
13	was_vbd slightly_rb taken_vbn	8
14	she_pp felt_vbd slightly_rb	7
15	he_pp looked_vbd slightly_rb	6
15	he_pp was_vbd slightly_rb	6
15	looked_vbd slightly_rb puzzled_vbn	6
15	she_pp was_vbd slightly_rb	6
15	slightly_rb surprised_vbn i_pp	6
15	was_vbd slightly_rb out_rb	6

(10) *Slightly out of breath*, she came to a halt outside the ground glass door with the legend painted across it: "Esthonia Glassware Co." (*The Secret Adversary*)

(11) When Miss Marple, *slightly out of breath* and rather tired, got back to the Golden Boar, the receptionist came out from her pen and across to greet her. (*Nemesis*)

(12) "Enjoy yourself, darling." Mr Enderby looked *slightly taken aback*. (*The Sittaford Mystery*)

(13) "Oh," said Tuppence, *slightly taken aback*, then added quickly, "What fun." (*By the Pricking of My Thumbs*)

What should also be specifically mentioned is that these fixed expressions are used more frequently in Christie's later novels than her earlier works. Figures 4.11 and 4.12 are the concordance lines of *slightly out of breath* and *slightly taken aback*, which are made using

CasualConc, a concordance program developed by Dr. Imao Yasuhiro at Osaka University. We can see that Christie's later works appear more frequently than her earlier works. This result matches those of previous studies such as Lancashire & Hirst (2009) and Le et al. (2011), which conclude that the number of repeated phrases increased in her later works.

Table 4.11: Top four-word clusters with *slightly* in the Christie corpus

No	Cluster	Freq
1	slightly_rb out_rb of_in breath_nn	14
2	looked_vbd slightly_rb taken_vbn aback_rb	12
3	said_vbd in_in a_dt slightly_rb	9
4	was_vbd slightly_rb taken_vbn aback_rb	8
5	was_vbd slightly_rb out_rb of_in	6
6	looked_vbd slightly_rb surprised_vbn i_pp	5
6	seemed_vbd slightly_rb taken_vbn aback_rb	5
6	shrugged_vbd her_pp$$ shoulders_nns slightly_rb	5
6	slightly_rb on_in one_cd side_nn	5
10	head_nn slightly_rb on_in one_cd	4
10	in_in a_dt slightly_rb different_jj	4
10	raised_vbd his_pp$$ eyebrows_nns slightly_rb	4
10	turned_vbd her_pp$$ head_nn slightly_rb	4
10	with_in slightly_rb stooping_vbg shoulders_nns	4
15	a_dt slightly_rb breathless_jj voice_nn	3
15	a_dt slightly_rb embarrassed_jj manner_nn	3
15	in_in a_dt slightly_rb breathless_jj	3
15	in_in a_dt slightly_rb embarrassed_jj	3
15	shook_vbd his_pp$ head_nn slightly_rb	3
15	slightly_rb taken_vbn aback_rb and_cc	3
15	speaking_vbg in_in a_dt slightly_rb	3
15	with_in slightly_rb raised_vbn eyebrows_nns	3

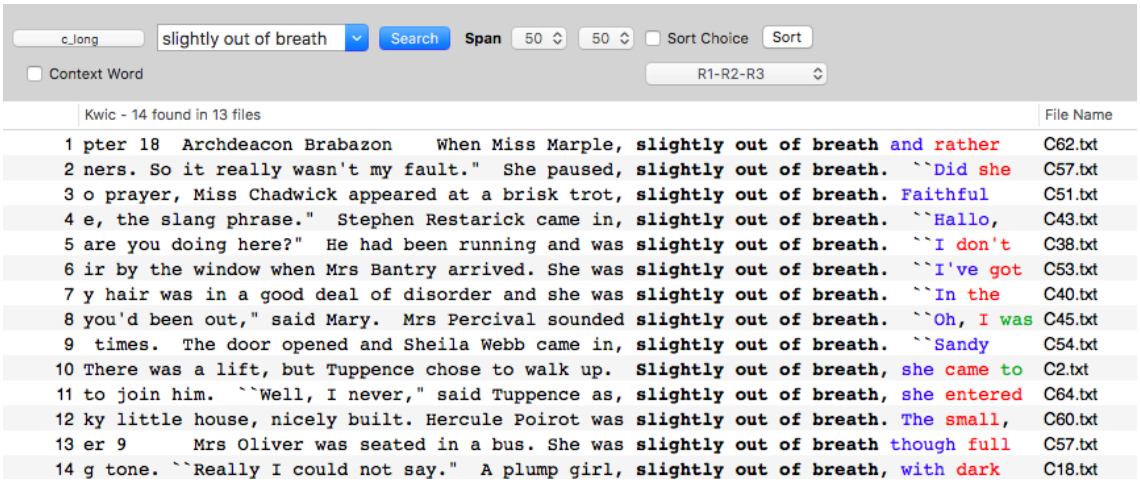

Figure 4.11: Concordance lines of *slightly out of breath*

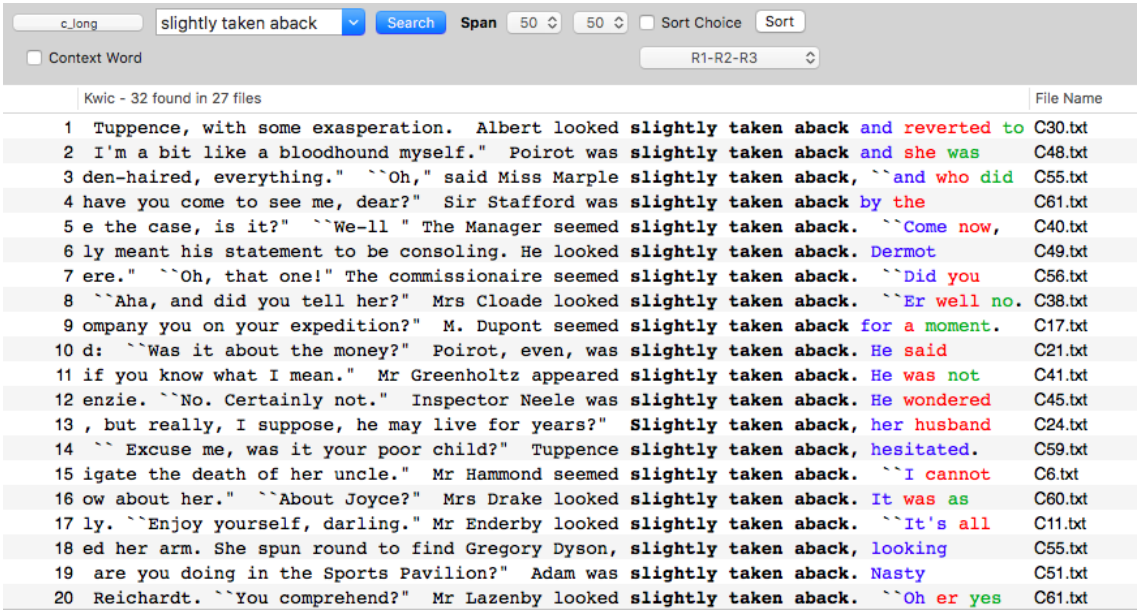

Figure 4.12: Concordance lines of *slightly taken aback*

As for *slowly*, Table 4.12 is the list of top three-word clusters and Table 4.13 is that of four-word clusters in the Christie corpus.

Table 4.12: Top three-word clusters with *slowly* in the Christie corpus

No	Cluster	Freq
1	said_vbd slowly_rb i_pp	130
2	i_pp said_vbd slowly_rb	95
3	he_pp said_vbd slowly_rb	76
4	she_pp said_vbd slowly_rb	63
5	poirot_np said_vbd slowly_rb	61
6	said_vbd slowly_rb it_pp	45
7	his_pp$ head_nn slowly_rb	41
8	slowly_rb i_pp do_vbp	34
9	said_vbd slowly_rb you_pp	30
10	her_pp$ head_nn slowly_rb	22
11	said_vbd slowly_rb and_cc	19
11	slowly_rb along_in the_dt	19
13	slowly_rb it_pp 's_vbz	18
14	said_vbd poirot_np slowly_rb	17
15	slowly_rb up_in the_dt	16
16	said_vbd slowly_rb that_dt	15
17	said_vbd slowly_rb yes_uh	14
17	slowly_rb back_rb to_to	14
17	slowly_rb i_pp think_vbp	14
17	slowly_rb it_pp is_vbz	14
17	slowly_rb it_pp was_vbd	14
17	slowly_rb shook_vbd his_pp$	14

Table 4.13: Top four-word clusters with *slowly* in the Christie corpus

No	Cluster	Freq
1	i_pp said_vbd slowly_rb i_pp	28
1	slowly_rb i_pp do_vbp n't_rb	28
3	said_vbd slowly_rb i_pp do_vbp	26
4	nodded_vbd his_pp$ head_nn slowly_rb	21
5	then_rb he_pp said_vbd slowly_rb	16
6	slowly_rb shook_vbd his_pp$ head_nn	14
7	she_pp said_vbd slowly_rb i_pp	13
8	he_pp said_vbd slowly_rb i_pp	12
8	poirot_np said_vbd slowly_rb i_pp	12
8	poirot_np said_vbd slowly_rb it_pp	12
8	said_vbd slowly_rb it_pp is_vbz	12
8	then_rb she_pp said_vbd slowly_rb	12
13	shook_vbd his_pp$ head_nn slowly_rb	11
13	slowly_rb out_rb of_in the_dt	11
15	said_vbd slowly_rb i_pp think_vbp	10
15	said_vbd slowly_rb it_pp 's_vbz	10
17	said_vbd slowly_rb i_pp have_vbp	9
17	said_vbd slowly_rb i_pp suppose_vbp	9
17	said_vbd slowly_rb it_pp was_vbd	9
17	slowly_rb up_in the_dt hill_nn	9

From Tables 4.12 and 4.13, we can find two patterns Christie tends to use frequently. First is the use of *slowly* with verbs of utterance (*said*). Second is the expression describing shaking one's head.

(14) *Poirot said slowly*:

"An old woman of the name of Ascher who keeps a little tobacco and newspaper shop has been found murdered." (*The ABC Murders*)

(15) "One thing that worries me," *said Poirot slowly*, "is the place where the blowpipe was found." (*Death in the Clouds*)

(16) Miss Marple nodded *her head slowly* and gravely.

"May God have mercy on her soul," she said. (*At Bertram's Hotel*)

(17) "Oh, M. Poirot," I exclaimed. "What do you think really happened?"

He *shook his head slowly* and thoughtfully.

"Tell me," he said. "You are not afraid to go back there tonight?" (*Murder in Mesopotamia*)

Moving on to *sharply*, Table 4.14 is the list of three-word clusters and Table 4.15 is the list of four-word clusters in the Christie corpus.

Table 4.14: Top three-word clusters with *sharply* in the Christie corpus

No	Cluster	Freq
1	he_pp said_vbd sharply_rb	44
2	looked_vbd up_rp sharply_rb	37
3	said_vbd sharply_rb what_wp	36
4	she_pp said_vbd sharply_rb	34
5	at_in him_pp sharply_rb	25
5	i_pp said_vbd sharply_rb	25
7	said_vbd sharply_rb i_pp	21
7	said_vbd sharply_rb you_pp	21
9	at_in her_pp$ sharply_rb	20
10	her_pp$ head_nn sharply_rb	18
10	sharply_rb what_wp do_vbp	18
12	her_pp$ breath_nn sharply_rb	16
13	looked_vbd sharply_rb at_in	14
14	he_pp turned_vbd sharply_rb	13
14	his_pp$ head_nn sharply_rb	13
16	lydia_np said_vbd sharply_rb	11
16	sharply_rb i_pp do_vbp	11
16	sharply_rb what_wp 's_vbz	11
19	sharply_rb that_dt 's_vbz	10
20	said_vbd sharply_rb of_in	9

We can see that there are mainly two patterns in the surroundings of *sharply*: one is used with *said*; the other is used with *looked*.

(18) Symmington put my thoughts into words. *He said sharply*:

"But that narrows it down to about half a dozen to a dozen people in the whole place!" (*The Moving Finger*)

(19) Rosamund, shooting a quick look at him, followed the direction of his eyes. *She said sharply*:

"What are you looking at my hands for? Do you think do you think?" (*Evil Under the Sun*)

(20) Something in the doctor's tone caught Poirot's attention. He *looked at him sharply*. The little Greek was standing staring down at the body with a puzzled frown. (*Murder on the Orient Express*)

Table 4.15: Top four-word clusters with *sharply* in the Christie corpus

No	Cluster	Freq
1	looked_vbd at_in him_pp sharply_rb	20
2	looked_vbd at_in her_pp$ sharply_rb	16
2	sharply_rb what_wp do_vbp you_pp	16
4	turned_vbd her_pp$ head_nn sharply_rb	13
5	in_in her_pp$ breath_nn sharply_rb	11
6	sharply_rb i_pp do_vbp n't_rb	10
7	he_pp said_vbd sharply_rb what_wp	9
8	he_pp looked_vbd up_rp sharply_rb	8
8	said_vbd sharply_rb what_wp 's_vbz	8
10	said_vbd sharply_rb i_pp do_vbp	7
10	said_vbd sharply_rb of_in course_nn	7
10	she_pp looked_vbd up_rp sharply_rb	7
13	in_in his_pp$ breath_nn sharply_rb	6
13	said_vbd sharply_rb _ i_pp	6
13	said_vbd sharply_rb _ what_wp	6
13	said_vbd sharply_rb _ you_pp	6
13	said_vbd sharply_rb what_wp do_vbp	6
13	sharply_rb what_wp is_vbz it_pp	6
13	turned_vbd his_pp$ head_nn sharply_rb	6

Furthermore, *sharply* is used with *in his/her breath* to describe one's breathing and with *turned his/her head.*

(21) "Murder?" Sarah drew *in her breath sharply*. "But what evidence of that is there? The flimsiest imaginable! Dr Gerard himself cannot be sure!" (*Appointment with*

Death)

(22) He paused his large handsome face suddenly altered he *turned his head sharply*. “Hallo," he said. “Now then, Poirot, what are you driving at?" (*Five Little Pigs*)

We have focused on five *-ly* adverbs extracted by RF and investigated their use in Christie’s novels. It is clear that Christie tends to employ fixed expressions using *-ly* adverbs to describe how characters speak and behave.

4.4 Summary of this chapter

In this chapter, we applied unsupervised and supervised methods to investigate the characteristic words of Christie compared with three other authors. As unsupervised methods, we applied cluster analysis and PCA. From the result of the cluster analysis, most of Christie’s works made a distinct cluster from other authors’ works. We also investigated the difference of styles between Christie and the other three authors by using PCA, and Christie’s works were plotted distinctively from other authors’, although some of their works overlapped. We calculated the top 20 words with the biggest and the smallest value contributions to the first and second principal components. With the first principal component, we could see a contrast between female and male words: the words with the smallest value contained words related to the male gender, but those with the largest value contained words related to the female gender. Then, we saw the collocates of the pronouns *she* in Christie’s novels and those of *he* in other authors’ novels as an example by using MI score. We could see the stereotypical femininity in the collocates of *she* in Christie’s works, such as words related to crying, or words related to actions indicative of fear or sadness. This tendency was seen in the collocates of *she* in other authors’ novels, but they also contained words related to love.

Moreover, we investigated the proportion of conversations and narratives because the word “*said*" was extracted as a characteristic word in the PCA. The proportion of conversations in Christie’s works is lager than that of other three authors’ novels. Christie’s novels tend to have more conversations than other authors’ ones.

As for the supervised method, we applied RF. As a result of RF, Christie’s texts were classified into a different group from the other authors’ texts. We identified characteristic

words that contributed to the classification the most. There were many *-ly* adverbs in the characteristic words for Christie's texts, such as *thoughtfully*, *exactly, slightly*, *slowly* and *sharply*. There were no *-ly* adverbs in the characteristic words for other authors. From investigating the use of these *-ly* adverbs in Christie's novels, we could see that Christie tended to use fixed expressions using these words to describe the manner of characters' utterances and behaviour.

Chapter 5

Diachronic changes of Christie's style

In this chapter, we investigate whether there is any diachronic change of style in Christie's works. As noted in Chapter 2, after Christie broke her wrist in 1952, she changed her method of writing novels from typing with a typewriter to using a Dictaphone. It is possible that her writing style changed after this change of writing medium. In this chapter, we analyse whether Christie's style changed as her method of writing changed. The research questions in this chapter are the following:

1. Does Christie's style change after breaking her wrist?

2. What are the differences between Christie's earlier works and later works?

5.1 Data

All of the texts written by Christie are labeled from C1 to C66 in the chronological order of their publishing years. The last two novels, *Curtain* (C65) and *Sleeping Murder* (C66), while published in the 1970s, were written in the 1940s. We assume that the rest of the novels were written just before their year of publication. Out of these 66 texts, we removed 12 works which were published during 1947 and 1957 (five years before and five years after she broke her wrist in 1952 and changed the way she wrote). This is because removing the novels published around 1952 might clarify change of her style. We grouped texts from 1920 to 1946 (C1–C37, C65, C66) into the earlier group, and texts from 1958 to 1973 (C50–C64) into the later group. The number of the texts in the earlier group is much larger than that in the later group, so we reduced the number of texts in the earlier group. The selection of

texts in the earlier group is random. We analyse 35 texts. The tokens of these texts add up to 2,330,584.

5.2 Methods

In this chapter, we also use both unsupervised methods and supervised methods: cluster analysis and PCA as unsupervised methods, and RF as a supervised method. The method of analysis is almost the same as in the previous chapter. We will discuss the results of each method.

5.3 Result 1: Unsupervised methods

The parameters are the same as those in the unsupervised analysis in Chapter 4; that is, using 700 most common words in of the Christy Corpus without excluding protagonists' proper nouns, as this analysis only focuses only on Christie. Figure 5.2 is the result of PCA on all Christie works and Figure 5.1 is the result of their cluster analysis. We used Euclidean Distance for interval and Ward's method as the clustering method.

In Figure 5.1, almost all of Christie's later works make the first smallest cluster, and some of her earlier works and later works make mixed clusters. In Figure 5.2, we can see most of Christie's earlier texts on the left side of the plot and later works on the right. These figures show that it is possible Christie's style changed in her later life from her style in her earlier works.

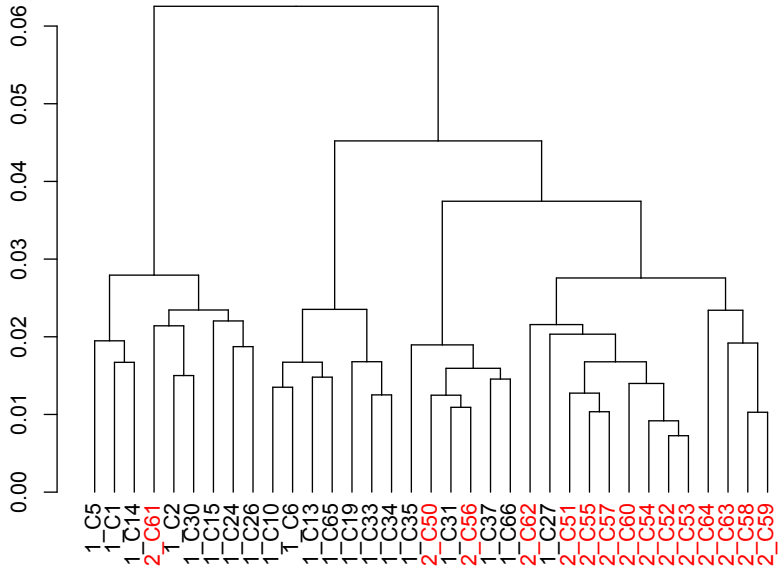

Figure 5.1: A dendrogram of the cluster analysis

Figure 5.2: A plot of the result of principal component analysis

Figure 5.3 and 5.4 show the top 20 words with the largest and the smallest values in the first and second principal components.

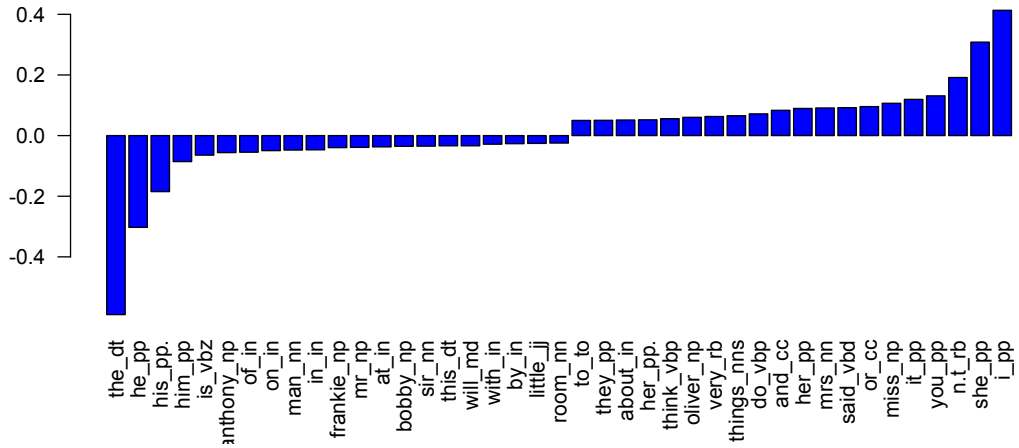

Figure 5.3: Barplot of the words with the biggest/smallest 1st principal component

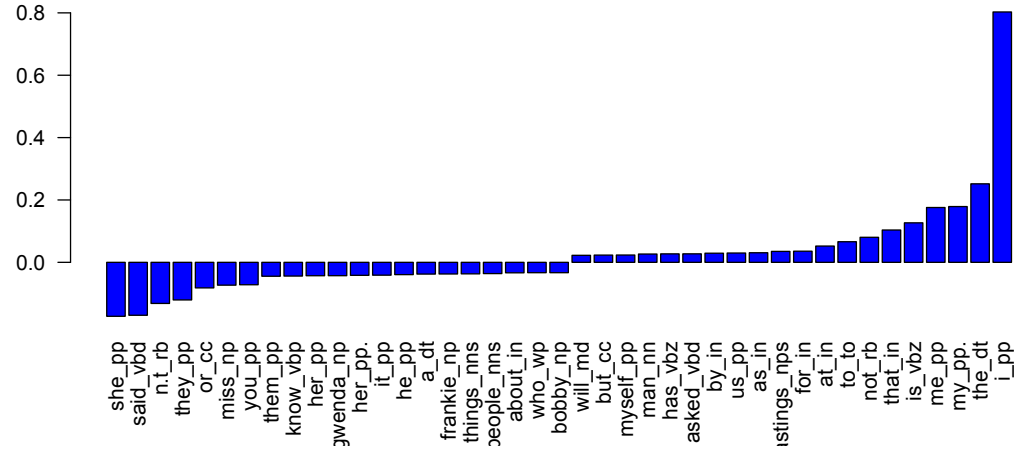

Figure 5.4: Barplot of the words with the biggest/smallest 2nd principal component

From Figure 5.3, we can see the contrast between female words and male words: the words with the biggest principal components contain the words related to female gender (*she*, *mrs* and *her*; the words with the smallest principal components contain the words related to male gender (*he*, *his*, *him* and *man*).

From Figures 5.3 and 5.4, it seems at first glance that this result of the PCA is influenced by the detectives of each story–and whether the narrative of the story is first-person or third-person, just like the previous chapter.

The work which is plotted on the rightmost hand is C63 (*Elephants Can Remember*). The detective in this novel is Hercule Poirot. The novel is written in the third person narrative. Near C63, there is C59 (*By the Pricking of My Thumbs*). The detectives in this novel are Tommy and Tuppence Beresford. This novel is also written in the third person narrative novel. C58 (*Endless Night*) is also plotted on the most eastern side of the plot. In this novel no famous detectives appears (such as Poirot or Miss Marple); it is a third person narrative novel as well.

On the other hand, the works plotted on the western side among Christie's works are C26 (*And Then There Were None*) and C14 (*Murder on the Orient Express*). In C26 no famous detective appears, and it is a third person narrative novel. In C14, the detective is Hercule Poirot and the novel is written in the third person. We can see C5 (*The Secret of Chimneys*) and C2 (*The Secret Adversary*) near C26 and C14. The detective in C5 is Inspector Battle, and the novel presents a third person narrative. The detectives in C2 are Tommy and Tuppence Beresford; it is written in the third person. There are more works written in the third person on the western side of the plot than on the eastern side, so it seems that there is little relationship between the result of PCA and the narrative of the story. Moreover, there are many works written in the first person among Christie's works on the western side, and many works whose protagonist is Poirot on the eastern side of the plot, so it seems that the narrative of the story and the protagonists of the novels have little effect on the result of the PCA.

Paralleling the results of the previous chapter, there is a contrast between female and male words in this result of the PCA. To see whether these words are used differently in Christie's novels, we would like to see the top pentagrams including the pronouns *she* and *he* (Table 5.1).

Table 5.1: Top pentagrams with *she* (left) and *he* (right) in the Christie corpus

No	Cluster	Freq	No	Cluster	Freq
1	she went out of the	37	1	minute or two then he	52
2	she said in a low	31	1	or two then he said	52
3	she was silent for a	25	3	he was silent for a	44
4	she shook her head no	24	4	he paused and then said	37
5	or two then she said	21	4	he rose to his feet	37
6	a minute or two she	20	6	a minute or two he	32
6	she paused and then said	20	6	shook his head he said	32
8	that she was going to	19	8	that he was going to	30
9	but i don't think she	17	9	for a moment then he	28
9	she was very fond of	17	9	he glanced at his watch	28
9	shook her head she said	17	11	he paused and then went	27
12	a matter of fact she	16	12	he leaned back in his	26
12	for the first time she	16	12	moment or two then he	26
12	i asked her if she	16	14	it was as though he	24
12	i do not think she	16	15	a minute then he said	20
12	she shook her head i	16	15	a moment then he said	20
12	she was going to be	16	15	he looked at his watch	20
18	it was as though she	15	15	he was a man of	20
18	moment or two then she	15	15	he went out of the	20
18	she paused a minute and	15	15	shook his head no he	20
18	she rose to her feet	15	21	a moment or two he	19

Both of the lists consist of the same pentagrams such as *she/he paused and then said* and *a minute or two she/he*, so the difference in the use of these pronouns is unclear.

Thus, we would like to see the collocates of *she* and *he* using the MI score. Table 5.2 shows the top collocates of *she* and *he* in Christie's texts whose MI scores are over 3.00.

Table 5.2: Top collocates of *she* (left) and *he* (right) in the Christie corpus (MI score >=3.00)

No	Word (she)	MI	Word (he)	MI	No	Word (she)	MI	Word (he)	MI
1	fainted (VBD)	4.11	snored	5.05	24	breathlessly	3.25	muttered	3.16
2	stricken	4.01	mumbled	3.76	25	hated	3.22	himself	3.14
3	wept	3.87	employed	3.59	26	burst	3.20	winced	3.14
4	bit	3.86	laughs	3.59	27	hates	3.18	paused	3.14
5	frightens	3.80	spluttered	3.59	28	realizes	3.18	kills	3.13
6	choked	3.69	hummed	3.51	29	accusingly	3.17	wheeled	3.12
7	fainted (VBN)	3.65	stressed	3.46	30	owns	3.16	added	3.11
8	minded	3.62	ushered	3.42	31	clung	3.13	encountered	3.11
9	sobbed	3.60	drinks	3.38	32	appealingly	3.11	sorted	3.11
10	lied	3.60	pretends	3.36	33	unreliable	3.11	strode	3.11
11	sensed	3.55	supposed	3.33	34	added	3.10	grunted	3.08
12	unwrapped	3.55	paced	3.32	35	says	3.06	wrong	3.07
13	fragile	3.53	puffed	3.31	36	collided	3.05	faint	3.06
14	shrank	3.51	admits	3.28	37	walks	3.03	hopes	3.05
15	tartar	3.47	cripple	3.28	38	exhausted	3.03	wagged	3.05
16	riding	3.46	unfolded	3.26	39	wistfully	3.03	straightened	3.04
17	talks	3.40	moody	3.25	40	bustled	3.02	concerned	3.03
18	herself	3.33	painting	3.24	41	clenched	3.02	dislikes	3.03
19	supposed	3.31	delivered	3.23	42	screamed	3.01	smoked	3.03
20	appealed	3.31	doubted	3.20	43	scared	3.01	consulted	3.01
21	insists	3.31	drunk	3.20	44	likes	3.00	stated	3.01
22	boasted	3.31	adored	3.17	45			chuckled	3.00
23	shivered	3.26	barked	3.17					

The collocates of *she* in Christie's works consist of words related to crying such as *wept*, *sobbed* and *screamed*. There are also many words related to actions indicative of fear or sadness such as *fainted*, *shrank*, *shivered*, *clung* and *scared*. Moreover, there is a word related to weakness like *fragile*. Here are the examples of the use of these collocates in Christie's novels.

(23) I had great trouble with Suzanne. She argued, she pleaded, she even *wept* before she would let me carry out my plan. But in the end I got my own way. She promised to carry out my instructions to the letter and came down to the station to bid me a tearful farewell. (*The Man in the Brown Suit*)

(24) "I will answer no more questions," *screamed* the dancer. She tore herself away from Poirot's restraining hand, and flinging herself down on the floor in a frenzy, she *screamed* and *sobbed*. A frightened maid came rushing in. (*The Mystery of the Blue Train*)

(25) Flora raised her hand to her throat, gave a little cry, and I hurried to catch her as she fell. She had *fainted*, and Blunt and I carried her upstairs and laid her on her bed. Then I got him to wake Mrs Ackroyd and tell her the news. (*The Murder of Roger Ackroyd*)

(26) A dream-like feeling of contentment came over Victoria. If only she could drive on like this for ever. If only she wasn't such a miserable liar. She *shrank* like a child at the thought of the unpleasant denouement ahead of her. (*They Came to Baghdad*)

(27) She *shivered* and as she did so Mosgorovsky rose. His voice, smooth, silky, persuasive, seemed curiously far away. (*The Seven Dials Mystery*)

(28) "No, not that... Never leave me... never leave me."
She *clung* to him she was crying, the tears coursing down her cheeks. He felt her shudder. (*Sparkling Cyanide*)

(29) A parlourmaid opened the door of the opposite wing to us. She looked *scared* but slightly contemptuous when she saw Taverner. (*Crooked House*)

(30) "No no there is a certain rather startling callousness... but no, I really cannot envisage the hatchet. She is a *fragile* looking creature." (*After the Funeral*)

It seems that Christie often uses these collocates with *she* to describe the weakness of the female gender. We can see stereotypical femininity from these collocates.

On the other hand, the collocates of *he* in Christie's works make a great contrast to those of *she*, such as *laughs*, *hummed* and *chuckled*. Here are the examples of these collocates in Christie's novels.

(31) Stephen Farr has a high-bridged nose, a habit of throwing his head back when he *laughs*, and a trick of stroking his jaw with his forefinger. (*Hercule Poirot's Christmas*)

(32) Our eyes met. Poirot's were pleasantly vague. He got up and *hummed* a little tune. I watched him suspiciously. (*The Mysterious Affair at Styles*)

(33) "He gave me the reply he would give to a reporter yes." Poirot *chuckled*. (*Lord Edware Dies*)

Of course, these three words also occur with *she*, but they are used more frequently with *he* than with *she*.

Moreover, the significant collocates of *she* include *choked*. This is in contrast to the collocates of *he* which include *kills*. There is also *strangled*, having an MI score a bit less than 3 (2.95). The examples of the concordance lines of these words are shown below (Figures 5.5, 5.6 and 5.7).

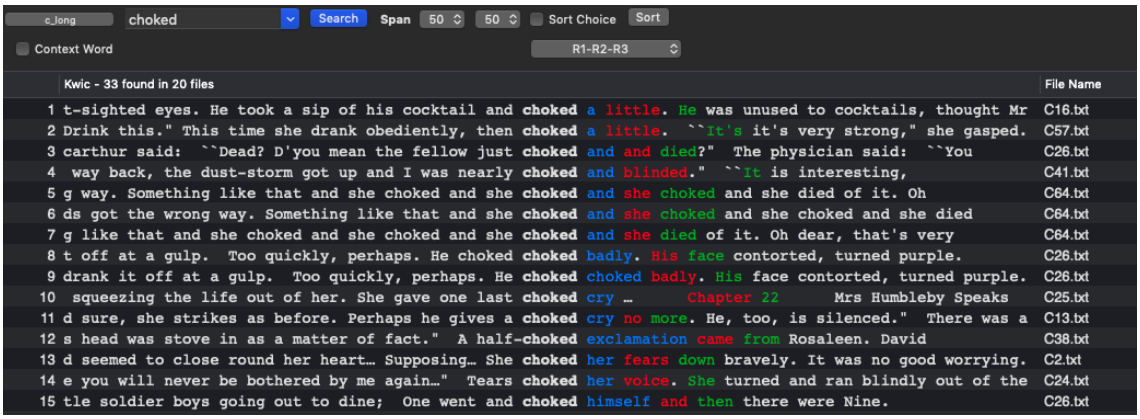

Figure 5.5: An example of the concordance lines including *choked*

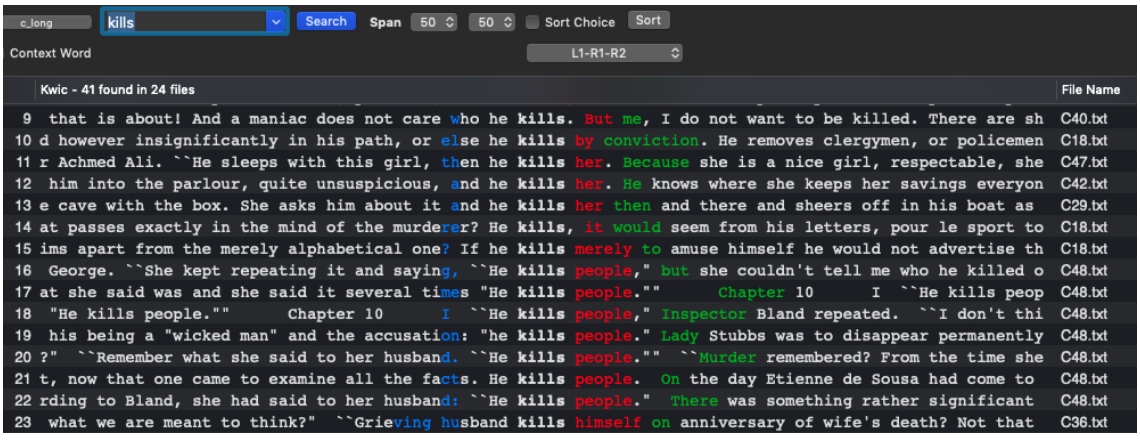

Figure 5.6: An example of the concordance lines including *killed*

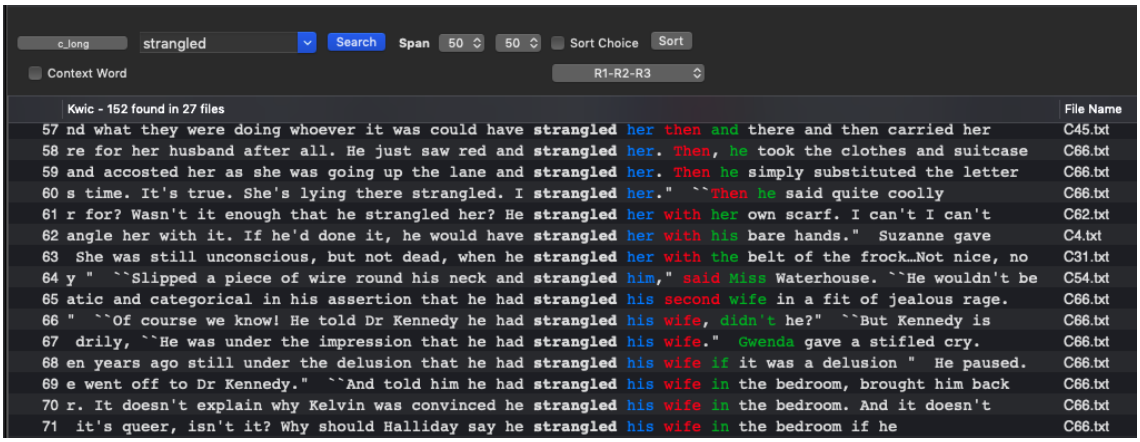

Figure 5.7: An example of the concordance lines including *strangled*

Choked is often used in phrases such as *she choked*. *Kills* tends to be used with the subject *he* and the object *her*. *Strangled* tends to occur as the sequences such as *he strangled her* or *he strangled his wife*.

We can see from these collocates that it is *she* who is *choked* or who *choke(s)* or who is *strangled* while it is *he* who *kills* someone.

From these uses of the pronouns, we can see that Christie seems to describe women as much weaker than men.

5.4 Result 2: Supervised method

In this section, we employ RF to classify Christie's works into two groups and extract the keywords from each group. Table 5.3 and Figure 5.8 show an example of the result of the RF classification. According to the multi-dimensional scaling plot in Figure 5.8, although only C50 is classified into the earlier group, the accuracy of the classification is still high. The best accuracy of classification is 97.14%.

Table 5.3: A result of Random Forests classification experiment on the three groups using the most frequent 600 words

Call:	
randomForest(formula = author ~.,	
data=csa.all, proximity=T, importance=T, ntree=10000)	
Type of random forest: classification	
Number of trees:	10000
No. of variables tried at each split:	26
OOB estimate of error rate:	2.86%

Confusion matrix:

	Earlier	Later	class.error
Earlier	20	0	0.00000000
Later	1	14	0.06666667

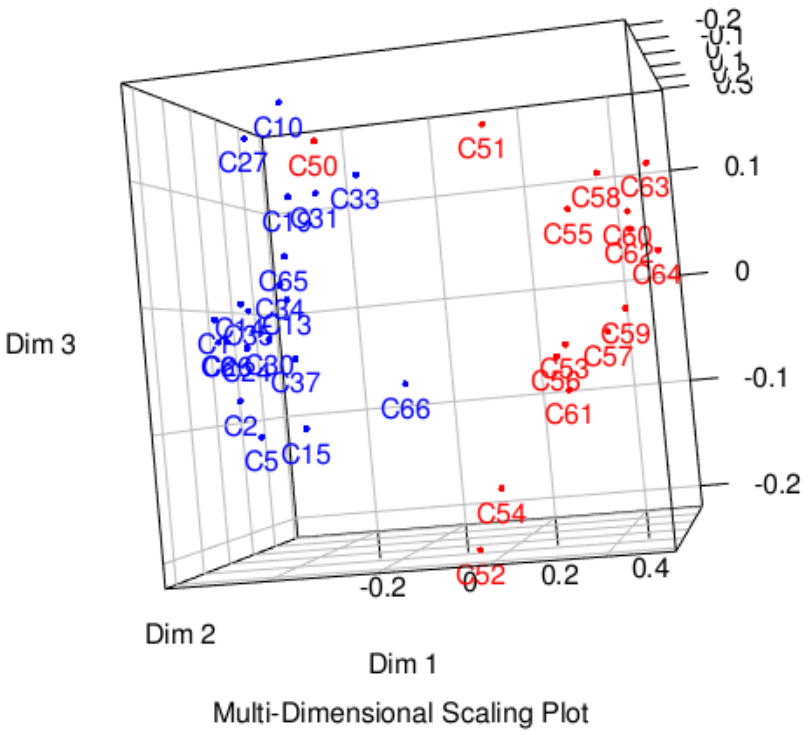

Figure 5.8: MDS plot of the result of Random Forests

Figure 5.9: Wordcloud of the characteristic words that contributed to the Random Forest classification the most

Figure 5.9 is a word cloud of the keywords that contributed to RF classification the most, and Table 5.4 shows the top 100 keywords out of the keywords in Figure 5.9.

One of the features of the keywords classified into the earlier group is that out of the top 100 keywords showed in Table 5.4, there is no present tense or base form verb that classified into the earlier group. Past tense verbs such as *murmured*, *cried*, *nodded*, *stared*, *broke*, *heard* and *shook* are classified into this group. We would like to investigate how Christie used these verbs in her novels.

Table 5.4: The top 100 key words of each group extracted by Random Forests

Earlier
minute_nn, fellow_nn, murmured_vbd, face_nn, eyes_nns, voice_nn, little_jj, upon_in, cried_vbd, curious_jj, gentleman_nn, into_in, absolutely_rb, nodded_vbd, clear_jj, stared_vbd, quietly_rb, the_dt, broke_vbd, opinion_nn, heard_vbd, night_nn, last_jj, dear_jj, shook_vbd, no_dt, open_jj, smile_nn, his_pp$, minutes_nns, against_in, window_nn, case_nn, suddenly_rb, truth_nn, after_in, head_nn
Later
or_cc, nowadays_rb, somehing_nn, particular_jj, really_rb, things_nns, someone_nn, want_vb, various_jj, who_wp, people_nns, know_vb, think_vbp, lot_nn, expect_vbp, mean_vbp, somebody_nn, might_md, seem_vbp, know_vbp, wanted_vbd, some_dt, usually_rb, many_jj, added_vbd, one_pp, much_jj, perhaps_rb, there_rb, about_in, nobody_nn, them_pp, age_nn, yes_uh, interesting_jj, well_rb, what_wp, can_md, somewhere_rb, get_vbp, well_uh, anything_nn, because_in, different_jj, they_pp, large_jj, years_nns, sort_nn, about_rb, it_pp, trouble_nn, girls_nns, better_rbr, interested_jj, think_vb, used_vbd, all_pdt, sometimes_rb, london_np, want_vbp, kind_nn, could_md

First, take a look at *murmured*. Table 5.5 is the list of two-word clusters and Table 5.6 is the list of three-word clusters in the Christie corpus, and the examples of sentences using *murmured* are shown below.

Table 5.5: Top two-word clusters with *murmured* in the Christie corpus

No	Cluster	Freq
1	he_pp murmured_vbd	99
2	she_pp murmured_vbd	42
3	murmured_vbd i_pp	32
4	murmured_vbd poirot_np	31
5	poirot_np murmured_vbd	26
6	i_pp murmured_vbd	21
7	and_cc murmured_vbd	18
7	murmured_vbd to_to	18
9	murmured_vbd the_dt	14
10	murmured_vbd you_pp	12
11	murmured_vbd it_pp	11
11	murmured_vbd something_nn	11
13	murmured_vbd a_dt	9
14	murmured_vbd _	7
14	murmured_vbd and_cc	7
14	murmured_vbd anthony_np	7
14	murmured_vbd miss_np	7
14	tuppence_nn murmured_vbd	7

Table 5.6: Top three-word clusters with *murmured* in the Christie corpus

No	Cluster	Freq
1	murmured_vbd to_to himself_pp	13
2	he_pp murmured_vbd i_pp	11
2	he_pp murmured_vbd to_to	11
3	murmured_vbd miss_np marple_np	7
4	she_pp murmured_vbd i_pp	6
5	murmured_vbd it_pp is_vbz	5
6	hercule_np poirot_np murmured_vbd	4
7	murmured_vbd i_pp wonder_vbp	4
7	murmured_vbd it_pp 's_vbz	4
7	murmured_vbd poirot_np it_pp	4
7	murmured_vbd something_nn about_in	4
7	poirot_np murmured_vbd i_pp	4

(34) Miss Johnson was white to the lips. *She murmured*: "Then it was not fancy. It was a trick a wicked trick! But who played it?" (*Murder in Mesopotamia*)

(35) "Do you mean to say," *he murmured* feebly, "that women use all these things?" (*The Body in the Library*)

It seems that Christie uses *murmured* to describe the actions of male characters more often than female characters. The most frequent three-word cluster is *murmured to himself*. This expression is mostly used to describe utterances of Poirot (nine times out of 13 cases).

(36) Poirot surveyed them with benign approval. He *murmured to himself*:
"C'est bien imaginé, ça!" (*Hercule Poirot's Christmas*)

1	heered up a little. ``A delightful creature," he	murmured to himself.	``A most	C5.txt
2	suredly, there is one who does not mourn," Poirot	murmured to himself.	A soft sound	C24.txt
3	. ``They go to bed early, these politicians," he	murmured to himself.	And suddenly a	C5.txt
4	Poirot walked to the other end of the terrace. He	murmured to himself:	``As I have	C24.txt
5	s. Poirot surveyed them with benign approval. He	murmured to himself:	``C'est bien	C24.txt
6	by their invitation. ``Je suis un peu snob," he	murmured to himself.	He had	C37.txt
7	please, please… We must be quick." The Commander	murmured to himself:	``Hope to	C30.txt
8	t is right, dragons." He was silent a minute. He	murmured to himself:	``I cannot	C14.txt
9	owned to himself. ``It is interesting, that," he	murmured to himself.	Magdalene said	C24.txt
10	ess. Anthony started; then smiled. ``Nerves," he	murmured to himself.	``Never knew I	C5.txt
11	here I do not comprehend." And once or twice he	murmured to himself.	``Pince-nez.	C13.txt
12	I said firmly. His gaze relaxed. He frowned and	murmured to himself:	``Since the	C6.txt
13	able, examining the charred fragment of paper. He	murmured to himself.	``What I need	C14.txt

Figure 5.10: Concordance lines of *murmured to himself*

Moving on to *cried*, Table 5.7 is the list of two-word clusters and Table 5.8 is the list of three-word clusters in the Christie corpus. The examples of sentences using *cried* are shown below.

(37) But the girl interrupted. Springing to her feet, *she cried out* angrily:
"What do you mean? What are you trying to suggest? That Mr Brown is Julius? Julius my own cousin!" (*The Secret Adversary*)

(38) "What? After all these years?" *he cried* incredulously. "Impossible." (*The Secret of Chimneys*)

Table 5.7: Top two-word clusters with *cried* in the Christie corpus

No	Cluster	Freq
1	cried_vbd out_rp	44
2	i_pp cried_vbd	40
3	she_pp cried_vbd	34
4	he_pp cried_vbd	24
5	cried_vbd poirot_np	18
6	vera_np cried_vbd	17
7	cried_vbd the_dt	16
8	cried_vbd i_pp	14
8	cried_vbd virginia_np	14
10	cried_vbd tuppence_np	13
11	cried_vbd but_cc	12
12	cried_vbd what_wp	9
13	cried_vbd you_pp	8
14	cried_vbd it_pp	7
14	cried_vbd julius_np	7
14	cried_vbd miss_np	7
17	ah_uh cried_vbd	6
17	cried_vbd flora_np	6
17	cried_vbd frankie_np	6

Table 5.8: Top three-word clusters with *cried* in the Christie corpus

No	Cluster	Freq
1	she_pp cried_vbd out_rp	17
2	poirot_np i_pp cried_vbd	7
3	cried_vbd m _np	6
4	:_: i_pp cried_vbd	5
4	i_pp cried_vbd i_pp	5
4	i_pp cried_vbd you_pp	5
7	:_: he_pp cried_vbd	4
7	cried_vbd out_rp i_pp	4
7	good_jj heavens_nns cried_vbd	4
7	he_pp cried_vbd out_rp	4
7	i_pp cried_vbd what_wp	4

In Table 5.7, we can see more female proper nouns (*Vera*, *Virginia*, *Tuppence* and *Flora*) and words related to female (*she* and *miss*) than male proper nouns (*Poirot*, *Julius* and *Frankie*) or words related to male (*he*). It seems that Christie describes female and male characters distinctively: she describes female characters as more emotional than male characters.

Table 5.9: Top two-word clusters with *nodded* in the Christie corpus

No	Cluster	Freq
1	poirot_np nodded_vbd	106
2	he_pp nodded_vbd	58
3	nodded_vbd and_cc	51
4	nodded_vbd he_pp	46
5	she_pp nodded_vbd	44
6	nodded_vbd yes_np	43
7	nodded_vbd his_pp$	40
8	nodded_vbd her_pp$	39
9	i_pp nodded_vbd	36
10	nodded_vbd i_pp	34
11	tuppence_nn nodded_vbd	20
12	nodded_vbd thoughtfully_rb	19
13	nodded_vbd it_pp	17
13	nodded_vbd that_dt	17
13	tommy_np nodded_vbd	17
16	marple_np nodded_vbd	14

Table 5.10: Top three-word clusters with *nodded* in the Christie corpus

No	Cluster	Freq
1	nodded_vbd his_pp$ head_nn	37
2	nodded_vbd her_pp$ head_nn	30
3	miss_np marple_np nodded_vbd	14
3	she_pp nodded_vbd her_pp$	14
5	nodded_vbd he_pp said_vbd	13
5	nodded_vbd that_dt 's_vbz	13
7	poirot_np nodded_vbd he_pp	11
8	poirot_np nodded_vbd yes_np	10
9	he_pp nodded_vbd his_pp$	9
9	poirot_np nodded_vbd his_pp$	9
11	hercule_np poirot_np nodded_vbd	7
11	nodded_vbd yes_np i_pp	7
11	poirot_np nodded_vbd and_cc	7
14	and_cc he_pp nodded_vbd	6
14	he_pp nodded_vbd and_cc	6
14	i_pp nodded_vbd and_cc	6
14	i_pp nodded_vbd i_pp	6

Next, we focus on *nodded*. Table 5.9 is the list of two-word clusters and Table 5.10 is the list of three-word clusters in the Christie corpus.

As we can see in the example of the sentence below, the cluster *Poirot nodded* is much more frequent than the other clusters, so it seems that Christie used this fixed phrase frequently to describe Poirot.

(39) *Poirot nodded*. "Well, Hastings, I will tell you this, when you have seen the rest of the household, my statement will seem to you just as improbable as it is now." (*Curtain*)

Moreover, we can see that Christie uses *nodded his/her head* frequently. To see these clusters more closely, we expand the span of these clusters. Table 5.11 is the list of *nodded his head* and collocates after them, and Table 5.12 is the list of *nodded her head* and collocates following them.

Table 5.11: Top four-word clusters containing *nodded his head* in the Christie corpus (span: right only)

No	Cluster	Freq
1	nodded_vbd his_pp$ head_nn gently_rb	6
2	nodded_vbd his_pp$ head_nn and_cc	4
2	nodded_vbd his_pp$ head_nn in_in	4
2	nodded_vbd his_pp$ head_nn that_dt	4
2	nodded_vbd his_pp$ head_nn with_in	4
6	nodded_vbd his_pp$ head_nn he_pp	3
6	nodded_vbd his_pp$ head_nn slowly_rb	3
8	nodded_vbd his_pp$ head_nn approvingly_rb	2
8	nodded_vbd his_pp$ head_nn i_pp	2
8	nodded_vbd his_pp$ head_nn once_rb	2
8	nodded_vbd his_pp$ head_nn reassuringly_rb	2
8	nodded_vbd his_pp$ head_nn understandingly_rb	2
8	nodded_vbd his_pp$ head_nn vigorously_rb	2
8	nodded_vbd his_pp$ head_nn yes_np	2
15	nodded_vbd his_pp$ head_nn _nn	1
15	nodded_vbd his_pp$ head_nn a_dt	1
15	nodded_vbd his_pp$ head_nn affirmatively_rb	1
15	nodded_vbd his_pp$ head_nn after_in	1
15	nodded_vbd his_pp$ head_nn agreeably_rb	1
15	nodded_vbd his_pp$ head_nn also_rb	1
15	nodded_vbd his_pp$ head_nn appreciatively_rb	1
15	nodded_vbd his_pp$ head_nn as_in	1
15	nodded_vbd his_pp$ head_nn but_cc	1
15	nodded_vbd his_pp$ head_nn do_vbp	1
15	nodded_vbd his_pp$ head_nn gravely_rb	1
15	nodded_vbd his_pp$ head_nn had_vbd	1
15	nodded_vbd his_pp$ head_nn his_pp	1
15	nodded_vbd his_pp$ head_nn my_pp	1
15	nodded_vbd his_pp$ head_nn oh_uh	1
15	nodded_vbd his_pp$ head_nn thank_vb	1
15	nodded_vbd his_pp$ head_nn then_rb	1
15	nodded_vbd his_pp$ head_nn thoughtfully_rb	1
15	nodded_vbd his_pp$ head_nn to_to	1
15	nodded_vbd his_pp$ head_nn two_np	1
15	nodded_vbd his_pp$ head_nn well_uh	1
15	nodded_vbd his_pp$ head_nn wisely_rb	1
15	nodded_vbd his_pp$ head_nn yes	1
15	nodded_vbd his_pp$ head_nn you_pp	1

Table 5.12: Top four-word clusters containing *nodded her head* in the Christie corpus (span: right only)

No	Cluster	Freq
1	nodded_vbd her_pp$ head_nn and_cc	6
2	nodded_vbd her_pp$ head_nn slowly_rb	5
3	nodded_vbd her_pp$ head_nn in_in	4
3	nodded_vbd her_pp$ head_nn she_pp	4
5	nodded_vbd her_pp$ head_nn gently_rb	3
6	nodded_vbd her_pp$ head_nn at_in	2
6	nodded_vbd her_pp$ head_nn then_rb	2
6	nodded_vbd her_pp$ head_nn with_in	2
9	nodded_vbd her_pp$ head_nn a_dt	1
9	nodded_vbd her_pp$ head_nn across_in	1
9	nodded_vbd her_pp$ head_nn appreciatively_rb	1
9	nodded_vbd her_pp$ head_nn as_in	1
9	nodded_vbd her_pp$ head_nn comprehendingly_rb	1
9	nodded_vbd her_pp$ head_nn did_vbd	1
9	nodded_vbd her_pp$ head_nn dismissively_rb	1
9	nodded_vbd her_pp$ head_nn exactly_rb	1
9	nodded_vbd her_pp$ head_nn ii_np	1
9	nodded_vbd her_pp$ head_nn macwhirter_np	1
9	nodded_vbd her_pp$ head_nn nice_np	1
9	nodded_vbd her_pp$ head_nn oh_uh	1
9	nodded_vbd her_pp$ head_nn once_rb	1
9	nodded_vbd her_pp$ head_nn perhaps_rb	1
9	nodded_vbd her_pp$ head_nn regally_rb	1
9	nodded_vbd her_pp$ head_nn satisfied_vbn	1
9	nodded_vbd her_pp$ head_nn shrewdly_rb	1
9	nodded_vbd her_pp$ head_nn significantly_rb	1
9	nodded_vbd her_pp$ head_nn that_dt	1
9	nodded_vbd her_pp$ head_nn towards_in	1
9	nodded_vbd her_pp$ head_nn vigorously_rb	1
9	nodded_vbd her_pp$ head_nn we_pp	1
9	nodded_vbd her_pp$ head_nn while_in	1
9	nodded_vbd her_pp$ head_nn wisely_rb	1
9	nodded_vbd her_pp$ head_nn yes_np	1

In these tables, we can see various *-ly* adverbs. Christie tends to use fixed expressions like *nodded his/her head* frequently, but she uses various kinds of *-ly* adverbs to describe the

action of nodding one's head. She describes how one is nodding his/her head using these *-ly* adverbs properly.

(40) There was a minute's dead silence. Poirot *nodded his head gently*. (*Murder on the Orient Express*)

(41) She looked at us both, *nodded her head vigorously*, and withdrew satisfied with the effect she had produced. Griselda and I stared at each other. (*The Murder at the Vicarage*)

(42) He did so slowly and carefully. When he had finished, she *nodded her head comprehendingly*. (*Why Didn't They Ask Evans?*)

Next is *shook*. Table 5.13 is the list of two-word clusters and Table 5.14 is the list of three-word clusters.

In Christie's novels, *shook* is used mostly in two patterns: shaking one's head or shaking hands. From Table 5.13 and 5.14, it is clear that Christie uses *shook* to describe shaking one's head much more frequently than to describe shaking hands.

(43) He *shook his head*, puffed out his chest, and stood blinking at us. (*The Murder of Roger Ackroyd*)

(44) She *shook hands*, hoped I'd had a good journey, tossed off her hat, gave a cool nod to Mr Coleman and sat down. (*Murder in Mesopotamia*)

Table 5.13: Top two-word clusters with *shook* in the Christie corpus

No	Cluster	Freq
1	shook_vbd his_pp$	356
2	shook_vbd her_pp$	180
3	he_pp shook_vbd	73
4	poirot_np shook_vbd	62
5	she_pp shook_vbd	52
6	i_pp shook_vbd	49
7	shook_vbd my_pp$	47
8	and_cc shook_vbd	39
9	shook_vbd hands_nns	27
10	tommy_np shook_vbd	16
11	marple_np shook_vbd	13
12	tuppence_nn shook_vbd	11
13	merely_rb shook_vbd	9
13	then_rb shook_vbd	9
15	doctor_nn shook_vbd	7
15	julius_np shook_vbd	7
15	lord_np shook_vbd	7

Table 5.14: Top three-word clusters with *shook* in Christie' earlier works

No	Cluster	Freq
1	shook_vbd his_pp$ head_nn	350
2	shook_vbd her_pp$ head_nn	174
3	he_pp shook_vbd his_pp$	62
3	poirot_np shook_vbd his_pp$	62
5	shook_vbd my_pp$ head_nn	47
6	she_pp shook_vbd her_pp$	45
7	i_pp shook_vbd my_pp$	44
8	tommy_np shook_vbd his_pp$	16
9	and_cc shook_vbd his_pp$	15
10	marple_np shook_vbd her_pp$	13
10	miss_np marple_np shook_vbd	13
12	and_cc shook_vbd her_pp$	12
13	shook_vbd hands_nns with_in	11
14	tuppence_nn shook_vbd her_pp$	10
15	he_pp shook_vbd hands_nns	9
16	doctor_nn shook_vbd his_pp$	7
16	julius_np shook_vbd his_pp$	7
16	sighed_vbd and_cc shook_vbd	7

We next focus on the top two expressions, *shook his/her head*. Table 5.15 is the list of *shook his head* and collocates following them, and Table 5.16 is the list of *shook her head* and collocates following them.

Table 5.15: Top four-word clusters containing *shook his head* in Christie's earlier works (span: right only)

No	Cluster	Freq
1	shook_vbd his_pp$ head_nn i_pp	43
2	shook_vbd his_pp$ head_nn he_pp	27
3	shook_vbd his_pp$ head_nn no_uh	20
4	shook_vbd his_pp$ head_nn and_cc	15
5	shook_vbd his_pp$ head_nn sadly_rb	12
5	shook_vbd his_pp$ head_nn you_pp	12
7	shook_vbd his_pp$ head_nn not_rb	11
8	shook_vbd his_pp$ head_nn with_in	10
9	shook_vbd his_pp$ head_nn doubtfully_rb	8
9	shook_vbd his_pp$ head_nn gently_rb	8
9	shook_vbd his_pp$ head_nn it_pp	8
12	shook_vbd his_pp$ head_nn nothing_nn	6
12	shook_vbd his_pp$ head_nn oh_uh	6
12	shook_vbd his_pp$ head_nn that_dt	6
15	shook_vbd his_pp$ head_nn then_rb	5
15	shook_vbd his_pp$ head_nn we_pp	5
17	shook_vbd his_pp$ head_nn do_vbp	4
17	shook_vbd his_pp$ head_nn never_rb	4
17	shook_vbd his_pp$ head_nn well_uh	4

As we can see in these two tables, *shook his/her head* is also used with various *-ly* adverbs, the same as *nodded his/her head* (See also Figure 5.11 and 5.12).

(45) Inspector Graves *shook his head sadly* and said:

"That's where you're wrong, sir." (*The Moving Finger*)

(46) Renisenb *shook her head perplexedly*. She sat on there by the water after the others had gone in, trying vainly to understand the confusion in her mind. (*Death Comes as the End*)

Table 5.16: Top four-word clusters containing *shook her head* in Christie's earlier works (span: right only)

No	Cluster	Freq
1	shook_vbd her_pp$ head_nn i_pp	26
2	shook_vbd her_pp$ head_nn no_uh	20
3	shook_vbd her_pp$ head_nn she_pp	11
4	shook_vbd her_pp$ head_nn and_cc	10
5	shook_vbd her_pp$ head_nn it_pp	7
6	shook_vbd her_pp$ head_nn not_rb	6
6	shook_vbd her_pp$ head_nn oh_uh	6
8	shook_vbd her_pp$ head_nn he_pp	4
8	shook_vbd her_pp$ head_nn that_dt	4
8	shook_vbd her_pp$ head_nn there_ex	4
11	shook_vbd her_pp$ head_nn but_cc	3
11	shook_vbd her_pp$ head_nn no	3
11	shook_vbd her_pp$ head_nn sadly_rb	3
11	shook_vbd her_pp$ head_nn very_rb	3
11	shook_vbd her_pp$ head_nn with_in	3
16	shook_vbd her_pp$ head_nn energetically_rb	2
16	shook_vbd her_pp$ head_nn never_rb	2
16	shook_vbd her_pp$ head_nn no_dt	2
16	shook_vbd her_pp$ head_nn people_nns	2
16	shook_vbd her_pp$ head_nn perplexedly_rb	2
16	shook_vbd her_pp$ head_nn slowly_rb	2
16	shook_vbd her_pp$ head_nn the_dt	2
16	shook_vbd her_pp$ head_nn to_to	2

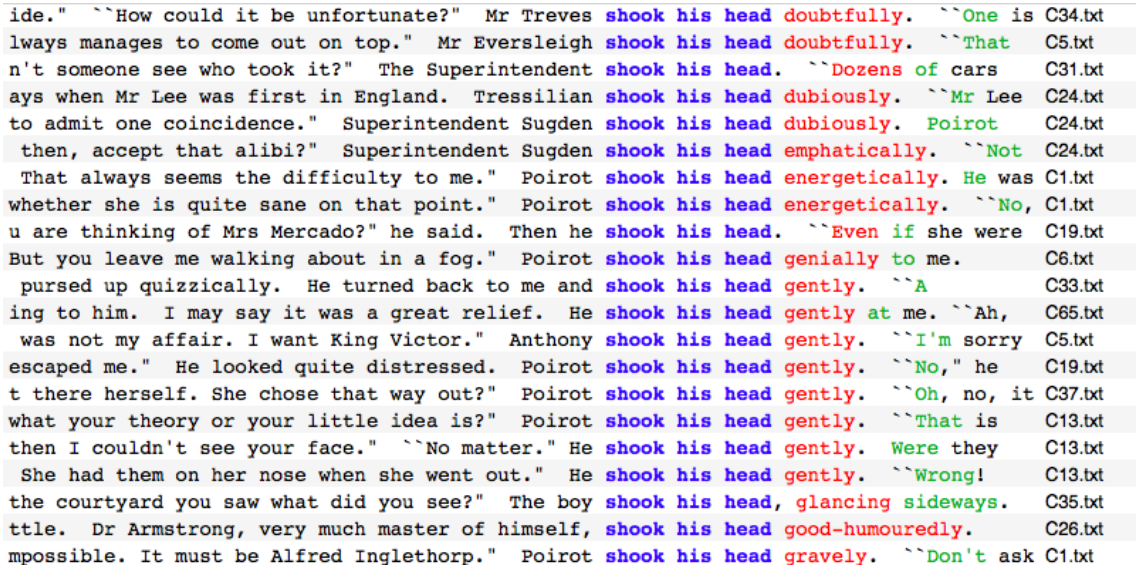

Figure 5.11: An example of concordance lines of *shook his head* (1)

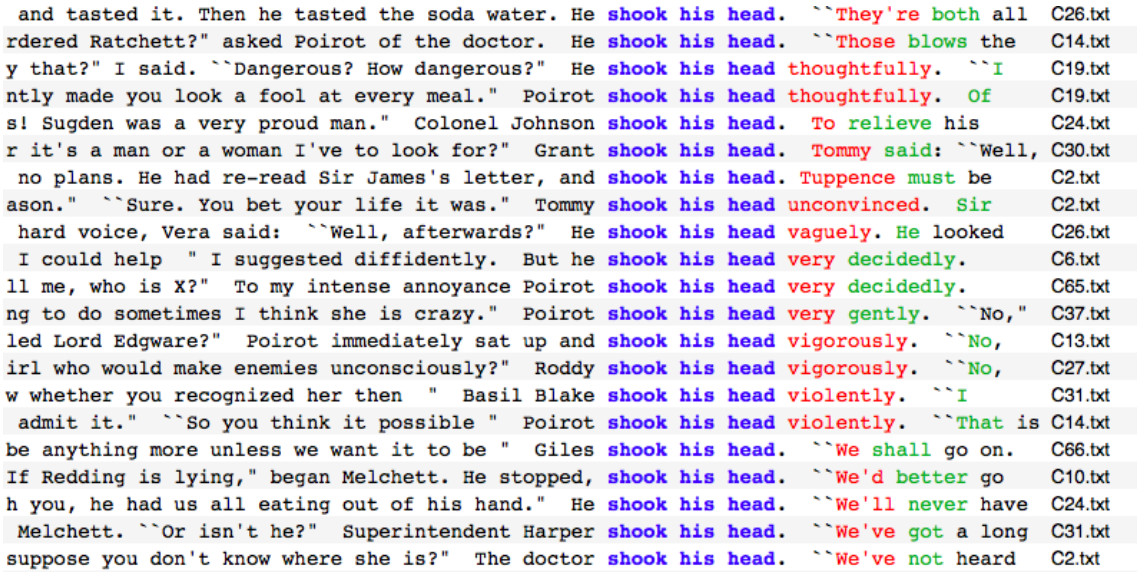

Figure 5.12: An example of concordance lines of *shook his head* (2)

We have focused on the keywords in the earlier group, and especially on past tense verbs. We can see that Christie uses these past tense verbs to describe characters' actions using various *-ly* adverbs.

When focused on verbs, the differentiating feature in the later group compared with the earlier one is that the keywords in the later group contain present tense (*think*, *mean* and *seem*)

or base form (*want* and *know*) of verbs. We investigate present tense verbs below.

Table 5.27 is the list of two-word clusters including *think* and the examples of sentences using *think* are shown below.

Table 5.17: Top two-word clusters with *think* in Christie's later works

No	Cluster	Freq
1	i_pp think_vbp	1822
2	you_pp think_vbp	557
3	think_vbp it_pp	341
4	think_vbp i_pp	299
5	think_vbp you_pp	210
6	think_vbp she_pp	209
7	think_vbp he_pp	138
8	think_vbp that_in	125
9	think_vbp so_rb	103
10	think_vbp of_in	75
11	think_vbp they_pp	74
12	think_vbp there_ex	69
13	think_vbp we_pp	61
14	think_vbp the_dt	56
15	they_pp think_vbp	54

Table 5.18: Top three-word clusters with *think* in Christie's later works

No	Cluster	Freq
1	i_pp think_vbp it_pp	238
2	do_vbp you_pp think_vbp	223
3	i_pp think_vbp i_pp	200
4	i_pp think_vbp you_pp	155
5	i_pp think_vbp she_pp	146
6	think_vbp it_pp 's_vbz	112
7	i_pp think_vbp he_pp	105
8	think_vbp it_pp was_vbd	103
9	but_cc i_pp think_vbp	75
10	and_cc i_pp think_vbp	72
10	you_pp think_vbp it_pp	72
12	n't_rb you_pp think_vbp	60
13	you_pp think_vbp i_pp	57
14	i_pp think_vbp so_rb	53
15	i_pp think_vbp there_ex	50

(47) "No," said Desmond, "that's where I don't agree with you. *I think it* would be quite possible." (*Elephants Can Remember*)

(48) "Good gracious, Mummy," said Jennifer, "what age *do you think* I am? Can I have some money, please? I haven't got any English money." (*Cat Among the Pigeons*)

We focus on *mean*. Table 5.19 is the list of two-word clusters and Table 5.20 is the list of three-word clusters including *mean*. The clusters *I mean* and *you mean* are much more frequent than others. It is likely that they are used in conversations.

(49) "Shouldn't think so. *I mean*, my grandmother said that was what people said. It wasn't in the last war. It was ages before that." (*Postern of Fate*)

(50) "*You mean*, Dr Calgary, that you agree with my attitude? You don't feel he was responsible for his actions?" (*Ordeal by Innocence*)

(51) "Excuse me, Mr Rudd, *do you mean* by that you can't or that you won't?" (*The Mirror Crack'd from Side to Side*)

Table 5.19: Top two-word clusters with *mean* in Christie's later works

No	Cluster	Freq
1	i_pp mean_vbp	907
2	you_pp mean_vbp	515
3	mean_vbp i_pp	117
4	mean_vbp it_pp	93
4	mean_vbp you_pp	93
6	mean_vbp the_dt	74
7	mean_vbp she_pp	67
8	mean_vbp that_in	53
9	mean_vbp he_pp	42
10	mean_vbp if_in	38
11	mean_vbp they_pp	37
12	mean_vbp there_ex	35
13	mean_vbp said_vbd	34
14	mean_vbp what_wp	29
15	mean_vbp to_to	27

Table 5.20: Top three-word clusters with *mean* in Christie's later works

No	Cluster	Freq
1	do_vbp you_pp mean_vbp	126
2	what_wp i_pp mean_vbp	116
3	what_wp you_pp mean_vbp	91
4	i_pp mean_vbp i_pp	88
5	i_pp mean_vbp it_pp	66
6	i_pp mean_vbp the_dt	49
6	i_pp mean_vbp you_pp	49
8	you_pp mean_vbp you_pp	44
9	i_pp mean_vbp she_pp	39
10	it_pp i_pp mean_vbp	38
11	you_pp mean_vbp that_in	36
12	i_pp mean_vbp if_in	34
12	mean_vbp it_pp 's_vbz	34
14	i_pp mean_vbp they_pp	32
15	well_rb i_pp mean_vbp	31

Moving on to *seem*, Table 5.21 is the list of two-word clusters and Table 5.22 is the list of three-word clusters including *seem*.

You seem is used frequently, but *I seem* is also used frequently compared with other clusters. *You seem* is often followed by *to have* or *to know*, and *I seem* is frequently used in the form of *I seem to remember*. The following verbs change according to the subject of *seem*. And as we can see in Figure 5.13 and 5.14, these clusters are also used in conversations.

Table 5.21: Top two-word clusters with *seem* in Christie's later works

No	Cluster	Freq
1	seem_vbp to_to	151
2	you_pp seem_vbp	54
3	i_pp seem_vbp	27
4	they_pp seem_vbp	23
5	seem_vbp very_rb	14
6	that_wdt seem_vbp	9
7	always_rb seem_vbp	7
7	things_nns seem_vbp	7
9	all_rb seem_vbp	6
10	people_nns seem_vbp	5
10	there_rb seem_vbp	5
12	it_pp seem_vbp	4
12	never_rb seem_vbp	4
12	seem_vbp in_in	4
12	seem_vbp so_rb	4
12	them_pp seem_vbp	4

Table 5.22: Top three-word clusters with *seem* in Christie's later works

No	Cluster	Freq
1	seem_vbp to_to have_vb	42
2	you_pp seem_vbp to_to	39
3	seem_vbp to_to be_vb	36
4	i_pp seem_vbp to_to	27
5	seem_vbp to_to know_vb	15
6	they_pp seem_vbp to_to	13
7	seem_vbp to_to remember_vb	12
8	always_rb seem_vbp to_to	7
9	seem_vbp to_to happen_vb	6
9	seem_vbp to_to think_vb	6
11	all_rb seem_vbp to_to	5
11	things_nns seem_vbp to_to	5
11	you_pp seem_vbp very_rb	5
14	never_rb seem_vbp to_to	4
14	of_in them_pp seem_vbp	4
14	people_nns seem_vbp to_to	4
14	that_wdt seem_vbp to_to	4
14	there_rb seem_vbp to_to	4

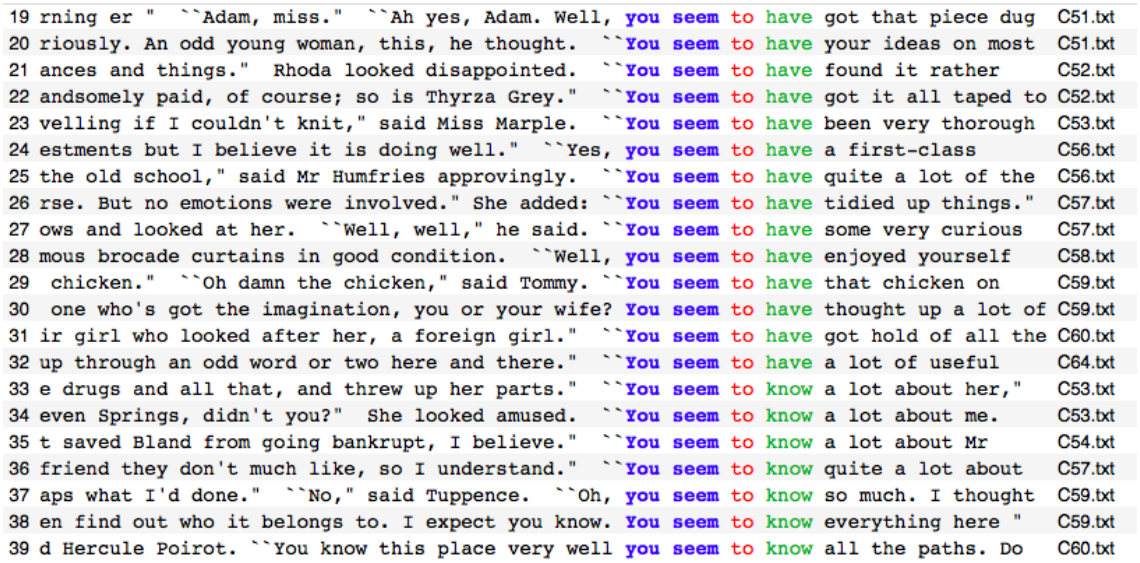

Figure 5.13: An example of concordance lines of *you seem*

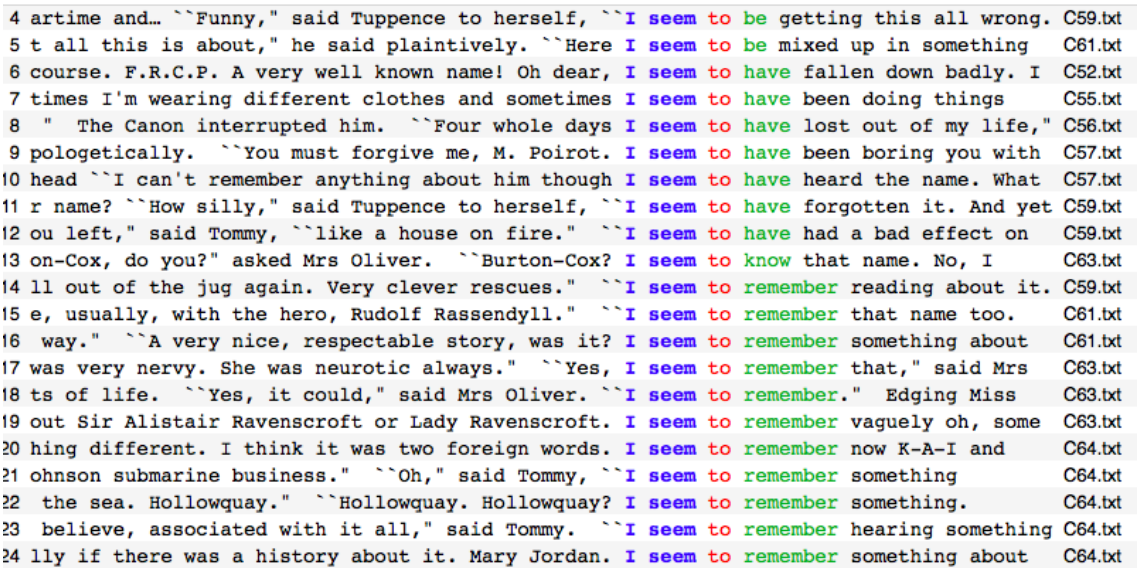

Figure 5.14: An example of concordance lines of *i seem*

The keywords in the later group include some verbs in their base form: *want*, *know* and *think*. The verbs in the base form are not classified into the earlier group in Table 5.4. We would like to investigate these base form verbs more closely.

Focusing on *want*, Table 5.23 is the list of two-word clusters and Table 5.24 is the list of three-word clusters including *want*.

Table 5.23: Top two-word clusters with *want* in Christie's later works

No	Cluster	Freq
1	n't_rb want_vb	360
2	want_vb to_to	350
3	not_rb want_vb	49
4	want_vb it_pp	23
5	to_to want_vb	21
5	want_vb you_pp	21
7	'd_md want_vb	19
7	want_vb me_pp	19
7	would_md want_vb	19
10	want_vb her_pp	17
11	'll_md want_vb	14
11	want_vb a_dt	14
13	might_md want_vb	12
14	want_vb him_pp	10

Table 5.24: Top three-word clusters with *want* in Christie's later works

No	Cluster	Freq
1	n't_rb want_vb to_to	224
2	do_vbp n't_rb want_vb	194
3	did_vbd n't_rb want_vb	128
4	want_vb to_to be_vb	38
5	not_rb want_vb to_to	29
6	want_vb to_to go_vb	23
7	does_vbz n't_rb want_vb	22
8	did_vbd not_rb want_vb	21
8	want_vb to_to know_vb	21
10	do_vbp not_rb want_vb	18
11	want_vb to_to do_vb	17
12	want_vb me_pp to_to	15
13	n't_rb want_vb you_pp	14
13	want_vb to_to kill_vb	14
13	want_vb to_to talk_vb	14

From these two tables, we can see that the base form of *want* is frequently used in negative

sentences. It is notable that negative sentences using contracted forms are much more frequent than sentences without contracted forms. This might be because Christie used a Dictaphone to write novels in her later life and thus reflects a feature of spoken language.

(52) "Mind you," he went on, "you *don't want to* rush at things. Take it steady, that's what I say. Steady is what does it." (*Cat Among the Pigeons*)

(53) Poirot picked up the receiver reluctantly. He did not want to talk to Mrs Oliver. He felt that she would urge upon him something which he did *not want to* do. (*Third Girl*)

Table 5.25: Top two-word clusters with *know* in Christie's later works

No	Cluster	Freq
1	n't_rb know_vb	1088
2	to_to know_vb	560
3	know_vb what_wp	313
4	not_rb know_vb	124
5	know_vb i_pp	120
6	know_vb who_wp	90
7	know_vb anything_nn	85
7	know_vb said_vbd	85
9	know_vb how_wrb	84
10	know_vb if_in	83
10	know_vb that_in	83
12	know_vb about_in	70
13	know_vb the_dt	68
14	know_vb why_wrb	66
15	really_rb know_vb	65

Table 5.26: Top three-word clusters with *know* in Christie's later works

No	Cluster	Freq
1	do_vbp n't_rb know_vb	799
2	did_vbd n't_rb know_vb	201
3	n't_rb know_vb what_wp	176
4	want_vbp to_to know_vb	110
5	n't_rb know_vb i_pp	68
6	to_to know_vb what_wp	61
7	like_vb to_to know_vb	60
8	n't_rb really_rb know_vb	58
9	do_vbp not_rb know_vb	56
9	n't_rb know_vb anything_nn	56
11	n't_rb know_vb how_wrb	53
11	n't_rb know_vb if_in	53
13	know_vb anything_nn about_in	51
14	n't_rb know_vb said_vbd	49
15	n't_rb know_vb who_wp	45

Table 5.27: Top two-word clusters with *think* in Christie's later works

No	Cluster	Freq
1	n't_rb think_vb	797
2	to_to think_vb	230
3	think_vb it_pp	169
3	think_vb of_in	169
5	think_vb so_rb	146
6	think_vb i_pp	143
7	think_vb she_pp	85
8	think_vb that_in	82
9	should_md think_vb	80
10	think_vb you_pp	72
11	think_vb he_pp	65
12	not_rb think_vb	62
13	think_vb there_ex	49
13	think_vb they_pp	49
15	you_pp think_vb	35

Table 5.28: Top three-word clusters with *think* in Christie's later works

No	Cluster	Freq
1	do_vbp n't_rb think_vb	614
2	n't_rb think_vb so_rb	111
3	n't_rb think_vb i_pp	110
4	n't_rb think_vb it_pp	97
5	i_pp should_md think_vb	76
6	did_vbd n't_rb think_vb	72
7	n't_rb think_vb she_pp	62
8	to_to think_vb of_in	61
9	n't_rb think_vb you_pp	49
10	think_vb it_pp was_vbd	47
11	n't_rb think_vb of_in	43
12	n't_rb think_vb he_pp	42
13	ca_md n't_rb think_vb	39
14	think_vb it_pp 's_vbz	37
15	think_vb of_in it_pp	36

Table 5.25 and 5.27 are the lists of two-word clusters and Table 5.26 and 5.28 are the lists of three-word clusters including *know* and *think*, respectively. As we observed in the clusters containing *want*, base forms of *know* and *think* are also used in negative sentences using contracted forms much more frequently than negative sentences without them. The examples of sentences using *know* and *think* in negative sentences are shown below.

(54) "Oh no," said Miss Marple. "I suppose you *don't know*. I've forgotten to tell you. Miss Bradbury-Scott very kindly asked me to come back and stay another night or two nights here." (*Nemesis*)

(55) "I *don't think* anything's too fantastic to be true," said Tommy. (*Postern of Fate*)

(56) Miss Gorringe looked almost shocked. It was to her impossible that anyone should *not know* Henry. (*At Bertram's Hotel*)

(57) "A young lady?" Calgary looked surprised. He could *not think* who was likely to visit him. (*Ordeal by Innocence*)

We have investigated the characteristic words of the later group focusing especially on the

verbs in present tense and base form. Throughout these investigations, the shared feature is that all of these verbs have a relationship with conversation and spoken language. This might be due to the change of Christie's writing style, which went from writing longhand and typing with a typewriter to using a Dictaphone.

Another feature of the characteristic words in the later group is that there are many indefinite terms such as *something*, *things* and *anything*. These words are considered to be a significant markers of one's decline in linguistic ability in previous studies such as Lancashire & Hirst (2009) and Le et al. (2011). They are indeed used in conversations frequently, but they are also used in narratives to some extent.

(58) "How interesting. I don't know that I've really thought about it. ⋯ It's so exciting when it comes. It doesn't very often, of course." Miss Bulstrode nodded in agreement. She had been right! This girl had *something*! (*Cat among the Pigeons*)

(59) Miss Marple gave her attention first to the main news on the front page. ⋯ Articles, comments, science, sport; then she pursued her usual plan, turned the paper over and had a quick run down the births, marriages and deaths, after which she proposed to turn to the page given to correspondence, where she nearly always found *something* to enjoy; (*Nemesis*)

If these indefinite words are used only in conversations, it is possible that Christie intended to use them as the speaker's habit. However, she also uses them in narratives, so it is natural to think that Christie could not use non-indefinite terms efficiently and her linguistic ability and vocabulary were declining when she wrote novels in her later life.

5.5 Summary of this chapter

In this chapter, we investigated the diachronic change in Christie's style through cluster analysis, PCA and RF and were able to see if her style changed after she had changed the way she wrote novels. We analysed this by making two groups: the earlier group and the later group of Christie's works.

In the cluster analysis and PCA, the classification was not so clear between Christie's earlier works and later works. However, works in the same group tended to be plotted near

each other. As we did in the previous chapter, in the analysis of PCA we obtained the top 20 words with the largest and the smallest value which contribute to the first principal component and the second one. Mirroring the result of the previous chapter, we could see a contrast between female and male pronouns and a contrast between first and third person pronouns. Initially, this result is influenced by the detectives of each novel and the narrative of the story, but it seems that they do not actually influence it a great deal. To see the difference between the use of the female and male pronouns, we saw the top pentagrams of *she* and *he* in Christie's texts. However, there was no significant difference between them. Thus, we investigated the collocates of *she* and *he* using the MI score. The collocates of *she* consist of words related to crying and weakness and words related to actions indicative of fear or sadness. On the other hand, the collocates of *he* make a great contrast to those of *she*. Moreover, the significant collocates of *she* include *choked*, though those of *he* consist of *kills* and *strangled*. Christie seems to describe women as much weaker than men.

In RF, almost all of the texts were classified into two groups correctly. We identified keywords in each group. In the keywords in the earlier group, there were words in the past tense, while there were words in their base form in the keywords in the later group. There were many verbs in the past tense in the keywords of the earlier group, and it seemed that Christie tended to employ fixed expressions using these words. She also seemed to describe stereotypical femininity using *cried*. In the keywords in the later group, there were more verbs in their base form than the earlier group. They were used with auxiliary verbs in their contracted forms such as *don't* and *doesn't*. The keywords of the later group have the characteristic of spoken language, and this might be because Christie changed her method of writing novels from typing with a typewriter to using a Dictaphone. Moreover, the keywords in the later group consisted of some indefinite words. This result was consistent with previous studies on the decline in her linguistic ability in her later life. Christie might not have been able to use the full extent of her vocabulary in her later life and might have used indefinite terms instead.

Chapter 6

Topic Modelling

We compared Christie's texts with the other three authors' in Chapter 4 and analysed Christie's novels diachronically in Chapter 5. In stylometry or digital humanities, researchers often focus on common words and identify characteristic words among them. This paper applied such methods in the previous chapters. It has been long considered that it is very difficult to deal with the semantic perspective. However, as the technology of machine learning has developed, we have become able to approach the style in the semantic level through the topic model. With the topic model, we can capture the feature of the style which cannot be seen through conventional methods. Here we investigate Christie's works compared with the other authors' works through the topic model. In this chapter, we will apply the topic model to see semantic characteristics compared with the results in the previous chapters, which focused on the common words. The research questions in this chapter are the following:

1. Can we distinguish Christie's texts from other authors' texts ones using the topic model?
2. What topics are characteristic in Christie's novels?

6.1 Topic model

Topic model is one of the machine learning methods. Blei (2012a) summarises Topic model as follows:

> Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time.

In this paper, we conduct the topic model using Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). According to Blei, LDA makes two assumptions. First, there are a fixed number of patterns of word usage, groups of words that tend to occur together in documents. Topic models find these sets of terms which tend to co-occur in the texts. They are called topics. Second, documents are regarded as a mixture of topics, and each document exhibits the topics to a different degree (Blei, 2012b). LDA reduces 'the dimensionality of a corpus to several hundred clusters or *topics*, represented as distributions over the full vocabulary' (Jockers & Mimno, 2013). It is possible for the same vocabulary to have different topics. For example, consider the word *play*. The topic of this word can be 'music' in the context such as *play the piano* on one hand, but on the other hand, it can be 'theatrical' if the word occurs together with *theatre*, *comedies* and so on.

The topic model has been applied in various areas. For example, DiMaggio et al. (2013) dealt with sociological data. They employed the topic model on the articles in five newspapers that contained any reference to government support for the arts in the United States. Grimmer (2010) dealt with political data. Jockers and Mimno (2013) employed the topic model on literary sources. Blei and Lafferty (2007) and Murakami et al. (2017) dealt with academic papers.

6.2 Data

We use all of the four authors' texts for the topic model. In this chapter, we use untagged data. All of the texts are over 30,000 words in length, and some texts are over 100,000 words in length. Since LDA looks at words' co-occurrence, the longer each of the texts, the more likely it is to find a set of words regarded as collocates. Therefore, topics made by LDA using the full length of novels become broader and more ambiguous. For this reason, we break all the novels into segments of 2000 words. Files that do not reach 2000 words are not used for the analysis. As a result of the segmentation, 130 novels are divided into 4,661 files.

In order to employ LDA and extract topics, we use MAchine Learning for LanguagE Toolkit[1] (MALLET), a Java-based toolkit for natural language processing (verion 2.0.7). When using MALLET, stopwords can be removed by setting a stopword list. In addition to

[1]http://mallet.cs.umass.edu/index.php

the default stopword list of MALLET, researchers can set their original stopword lists. When we extracted topics without setting stopwords of our own, some topics contained proper nouns in each author's novel such as Poirot, Wimsey and so on. Therefore, we created a stopword list containing proper nouns occurring in the four authors' novels, and used it as an extra stopword list.

The number of topics extracted by the topic model using LDA can be decided by each researcher depending on the data they use. There is no definite answer to what number of topics is the best for topic modeling using LDA. If the number of topics is too small, the topics extracted by LDA would become too ambiguous to interpret. On the other hand, if the number of topics is too many, extracted topics become too detailed and it is also difficult to interpret topics. In this study, we changed the number of topics to decide the number of topics capturing topics the best, and in the following section, the result of LDA with the 50 topics will be shown.

6.3 Surveying topics: The case of the number of topics set to 50

As a result of LDA, we obtained 50 topics. After the LDA using MALLET, we can obtain several files that are used for surveying obtained topics. We will show three files that we used in the following analysis:

1. topic-keys file: A file that shows keywords that have big weights in topics.
2. doc-topics file: A file that contains the topic composition of documents.
3. topic-word-weight-file: A file that shows each word's topic and its weight in the sentences in the documents.

Figure 6.1 is a network graph of the relationship between extracted topics and words that constitute topics. This graph was drawn using the topic-word-weight file. Tables 6.1 and 6.2 are the lists of topics and their keywords that can be obtained by the topic-keys file. Although proper nouns were removed from the analysis using the stopword list, some proper nouns remain in the keyword list. The tagging through Tree Tagger might be wrong when tagging

some proper nouns. Network graphs are described using Gephi, a visualisation software for graphs and networks[2].

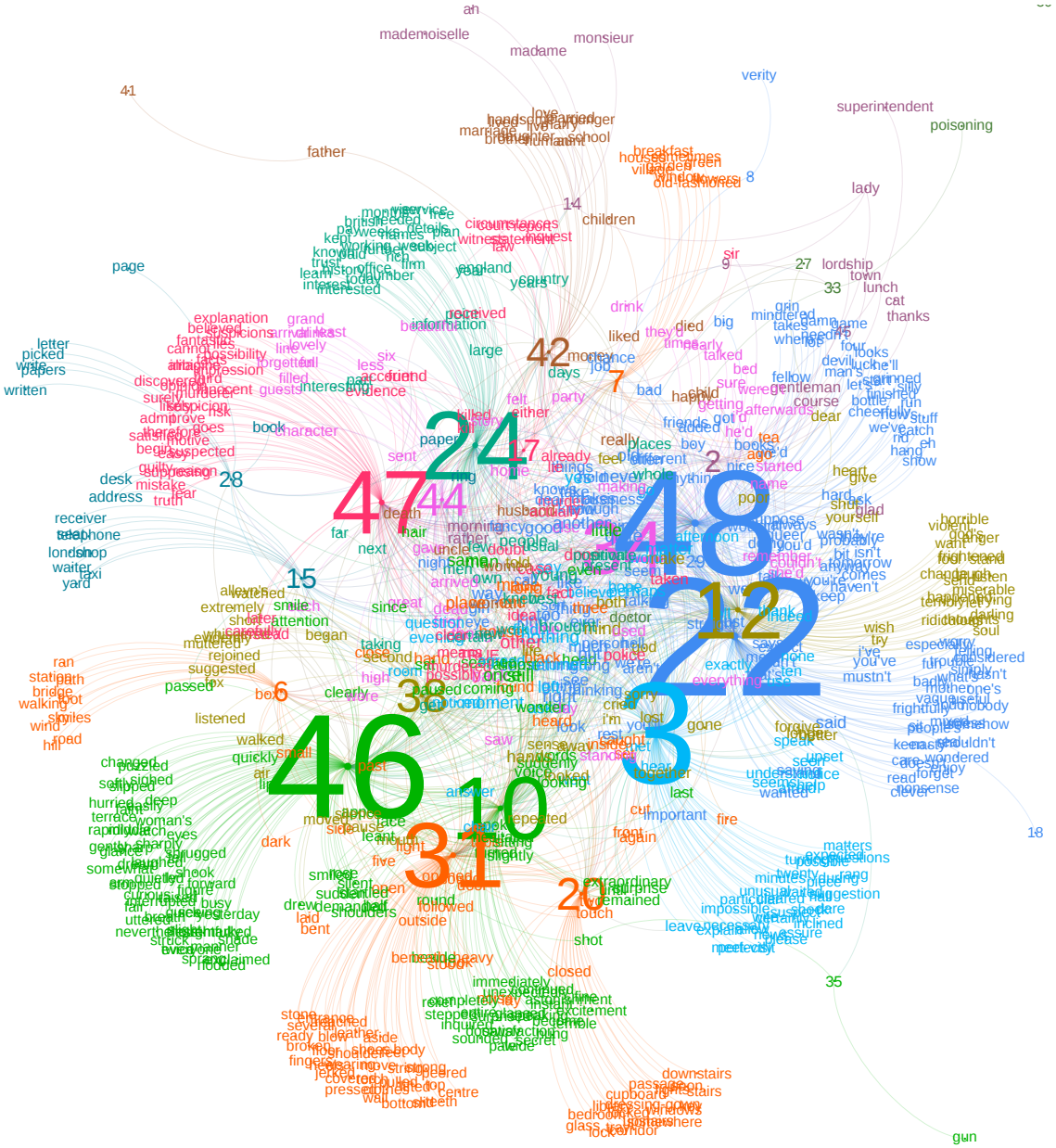

Figure 6.1: Network graph of topics and words

[2]https://gephi.org

Table 6.1: Key words in each topic (1)

Topic	Words
0	troy, dikon, dr, gaunt, questing, falls, miss, cost, panty, natouche, hewson, ancred, rickerby-carrick, bard, carmichael, lattienzo, bell, caley, millamant
1	troy, ferrant, syd, forrester, harkness, plank, gibson, moult, cuthbert, president, uncle, tree, window, sanskrit, husband, ricky's, thought, far, ambassador
2	lady, well, good, little, morning, say, dear, rather, old, poor, course, gentleman, young, make, way, much, time, day, never
3	said, yes, think, see, know, say, well, tell, miss, went, asked, quite, go, course, ask, looked, time, anything, told
4	lady, mandrake, hart, pastern, hersey, compline, dr, kettle, frid, nurse, phinn, breezy, licit, madame, lift, lisse, rose, manx, aunt
5	said, uncle, inspector, belle, room, cousin, potter, curley, oates, miss, godolphin, phillida, lucar, campion's, ritchie, went, lafcadio, great-aunt, studio
6	road, way, back, path, side, came, away, lane, go, far, time, see, bridge, past, walked, left, drive, river, round
7	house, old, tea, garden, place, nice, little, years, day, like, good, small, two, people, long, kitchen, always, much, used
8	verity, dr, nz, otterly, old, satipy, guiser, sobek, kait, dame, sons, mardian, westholme, carol, trixie, father, say, dance, lennox
9	sir, lady, miss, strange, yes, cabin, night, young, asked, boat, that's, treves, butler, fanthorp, deck, royde, bellefort, fellow, otterbourne
10	eyes, old, man, little, face, see, time, looked, still, came, like, back, first, spoke, seemed, turned, sat, room, last
11	troy, period, monsieur, moppett, dupont, desir, miss, bimbo, garbel, roqueville, truebody, period's, raikes, french, bantling, leiss, glande, callard, factory
12	i'm, like, go, know, dear, said, think, you're, let, away, own, thought, get, face, i've, never, mind, make, tell
13	legge, watchman, said, parish, cubitt, miss, bottle, harper, sir, decima, darragh, dart, island, glass, nark, bar, oates, brandy, fox
14	madame, monsieur, mademoiselle, ah, friend, little, yes, paris, see, lady, well, eh, ami, english, french, edgware, cried, know, last
15	street, man, telephone, taxi, london, flat, just, name, number, train, office, rang, door, back, shop, way, get, got, walked
16	garnette, miss, ogden, father, inspector, ravigne, upward, dr, spence, quayne, redding, book, fane, cup, pringle, candour, sunday, slack, wetherby
17	sir, inspector, body, found, evidence, police, case, man, time, made, chief, taken, coroner, inquest, death, dead, anything, witness, own
18	said, an', tae, man, bicycle, major, o', ay, duke, sweet, wi', gatehouse, sergeant, braceley, night, baron, questore, glasgow, train
19	inspector, miss, lady, school, girl, horse, dane, pavilion, girls, calthrop, rich, springer, sports, she's, mademoiselle, woman, things, sousa, vansittart
20	door, room, bed, went, came, back, window, house, hall, heard, open, opened, night, go, left, stairs, upstairs, two, light
21	miss, dean, said, college, whittaker, somebody, vine, new, till, better, quad, person, women, oxford, staircase, students, anything, cattermole, room
22	said, know, think, well, like, i'm, really, mean, just, yes, that's, say, things, rather, see, people, got, quite, she's
23	miss, fox, troy, sir, prentice, carrados, templett, wilde, withers, yes, bathgate, hall, lady, pilgrim, seacliff, friday, rector, malmsley, sort
24	man, time, name, people, own, work, two, money, nothing, young, other, new, quite, take, get, much, country, know, great

Table 6.2: Key words in each topic (2)

Topic	Words
25	train, calgary, compartment, argyle, bouc, micky, conductor, carriage, man, huish, platform, dr, arbuthnot, macqueen, passengers, corridor, american
26	inspector, razor, o'clock, umpelty, flat-iron, rock, bright, perkins, two, morecambe, darley, suicide, shore, glaisher, mare, body, yes, throat, murder
27	miss, dear, old, inspector, really, quite, warrender, major, anelida, cherry, hall, bellamy, gantry, pinky, she'd, hotel, templeton, kind, bunner
28	letter, paper, letters, read, written, book, write, desk, wrote, writing, envelope, note, found, took, put, papers, two, drawer, sheet
29	said, things, yes, know, think, well, something, miss, perhaps, people, years, name, girl, remember, other, kind, time, happened, spence
30	inspector, miss, captain, major, percival, narracott, hotel, sittaford, enderby, morning, three, man, house, rycroft, willett, webb, police, past, tea
31	back, hand, head, round, look, two, made, found, like, something, stood, again, right, light, hands, first, still, feet, end
32	peregrine, stage, martyn, theatre, play, macbeth, poole, yes, conducis, saint, dressing-room, go, scene, went, good, surbonadier, company, came, door
33	dr, doctor, death, bottle, poison, glass, nurse, drink, died, take, medical, patient, arsenic, took, stuff, put, suicide, patients, coffee
34	didn't, thought, went, came, got, time, know, wasn't, told, just, back, knew, see, never, he'd, couldn't, something, i'd, saw
35	said, shot, revolver, gun, pistol, fired, detective, nodded, gale, prenderby, spoke, shoot, meggie, house, man, girl, boy, bullet, crowther
36	nurse, dr, miss, o'callaghan, banks, yes, thoms, norton, sir, hyoscine, carrington, lansquenet, humbleby, father, patient, styles, theatre, inspector, luttrell
37	church, ain't, superintendent, tonker, says, rector, vicar, see, bells, puffett, lordship, wife, o', village, grave, bell, uncle, again, man
38	said, fox, looked, asked, i'm, turned, sir, hand, began, voice, hands, face, sort, look, moved, walked, air, opened, gave
39	sutane, cuddy, captain, merryman, father, dale, dillington-blick, cliff, konrad, deck, mcangus, miss, night, abbott, ship, sock, cabin, aubyn, mercer
40	world, plane, dr, baghdad, yes, olive, things, other, german, sprot, young, country, scheele, hotel, sir, commander, pauncefoot, clipp, blenkensop
41	father, canon, chief-inspector, davy, bassington-ffrench, bertram's, hotel, pennyfather, cust, lady, crome, inspector, sedgwick, gorringe, badger, photograph
42	father, woman, money, mother, old, always, young, life, family, married, years, wife, girl, child, never, husband, children, love, quite
43	superintendent, lady, sir, ruby, cade, thesiger, harper, kemp, victor, chimneys, seven, chief, king, bridge, room, prince, minute, table, farraday
44	thought, like, little, too, much, made, party, great, rather, each, other, perhaps, first, seemed, own, table, hair, really, wondered
45	oates, superintendent, yeo, got, that's, know, just, police, carados, boy, avril, street, campion's, mayo, appeared, sir, story, chief, they're
46	said, looked, head, man, face, just, voice, eyes, went, little, came, turned, back, shook, nodded, suddenly, slowly, again, two
47	murder, man, case, say, think, crime, person, time, other, murderer, point, thing, killed, two, death, fact, first, way, possible
48	said, got, that's, get, well, right, know, like, old, see, i've, good, i'm, say, look, bit, go, man, want
49	said, lugg, georgia, ere, guffy, yer, marlowe, penny, ave, campion's, went, know, biddy, old, professor, that's, lobbett, i'm, sir

The numbers in the network graph are topic numbers, and the thickness of the edges that connect the topic numbers and the words show the degree of the weight in the topics. The larger the size of the topic number, the larger the topic. A word does not necessarily have only one topic. For example, *said* appears in many topics, adding up to 13 topics. We can infer what the topic is about by the components of topics. For instance, we can infer that topic 20 is about the description of the inside of a house or room since it contains the words such as *door*, *bed*, *window* and *hall*, and topic 28 is about writing letters since it contains words such as *letter/letters*, *write/wrote/writing* and *envelope*. The interpretation of topics depends on each researcher, so there is no absolute and single answer for topic interpretation. We can identify the words that have large weights in topics by drawing a word cloud. Figures 6.2 and 6.3 are the word clouds of topic 20 and 28.The larger the size of the word in the word cloud, the larger the weight in the topic.

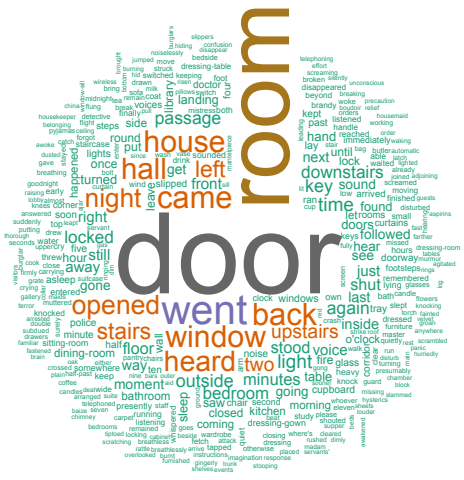

Figure 6.2: Word cloud of topic 20

Figure 6.3: Word cloud of topic 28

Figure 6.4 is a network graph that shows the relationship between topics and texts. We can see what topics are characteristic in each text. Figure 6.5 is a heatmap of the density of topics in each author's corpus.

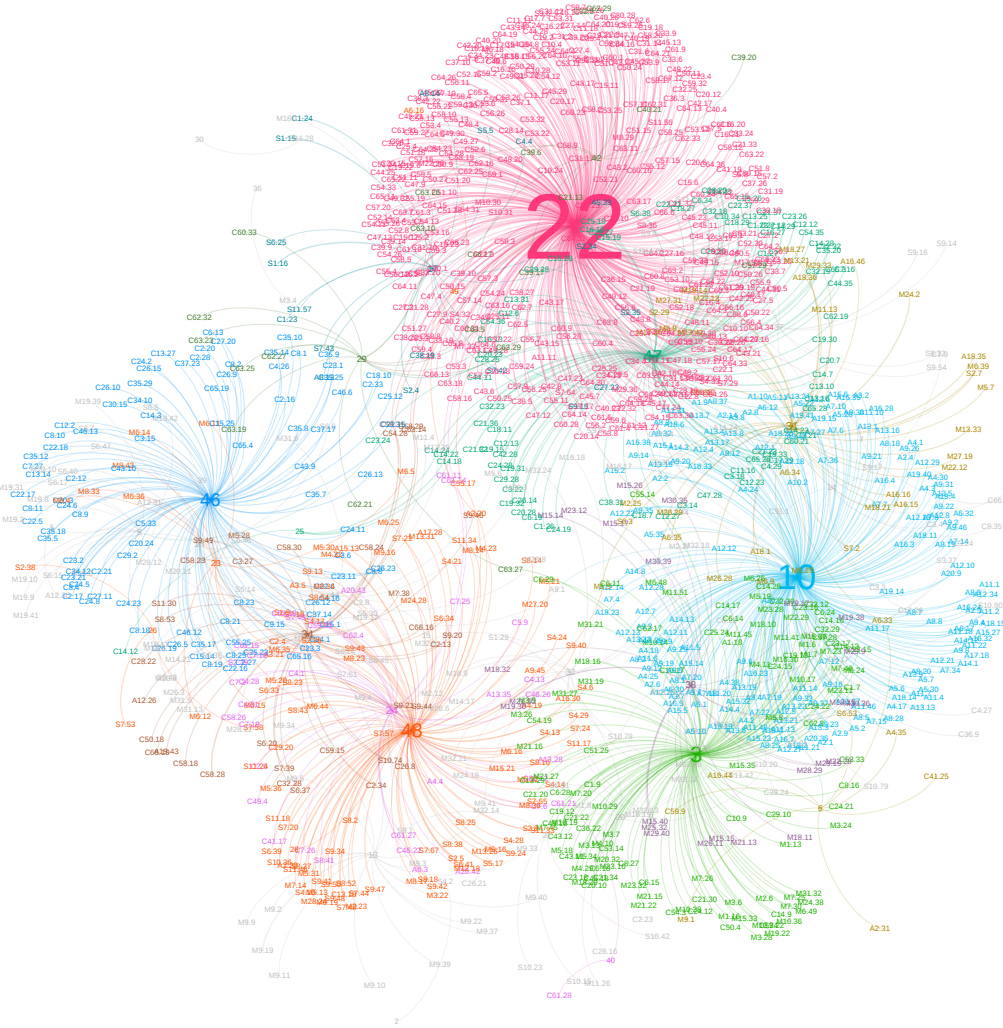

Figure 6.4: Network graph of topics and texts

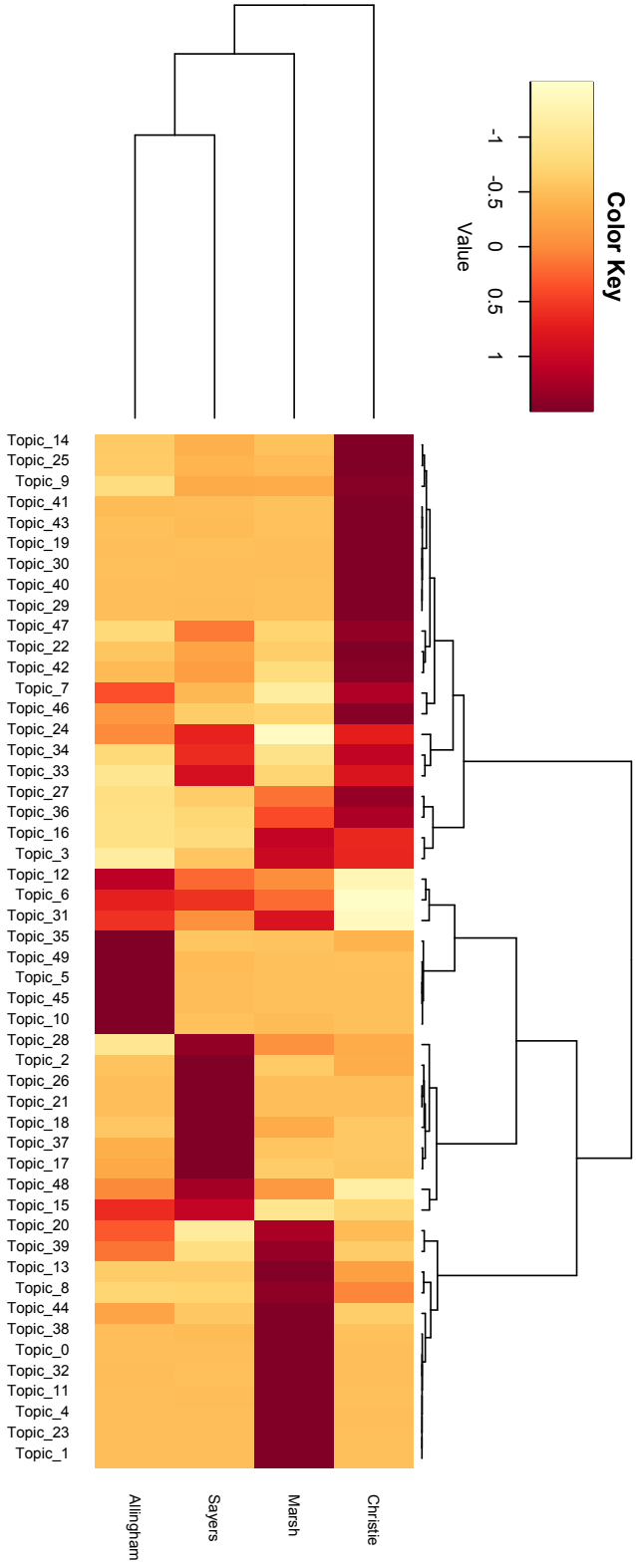

Figure 6.5: Heatmap of the topic density

From Figure 6.4, we can see that many of Christie's texts are connected with topics 22 and 46. We can also see from Figure 6.5 that the density of topics 14, 29, 40 and 42 is higher than that of other authors' texts. We will deal with each of these topics in the following section.

6.3.1 Topic 42: Family and life

Topic 42 includes the words related to family such as *father*, *mother*, *family*, and those related to life such as *life*, *married*, *died* and so on. Figure 6.6 is a word cloud of topic 42.

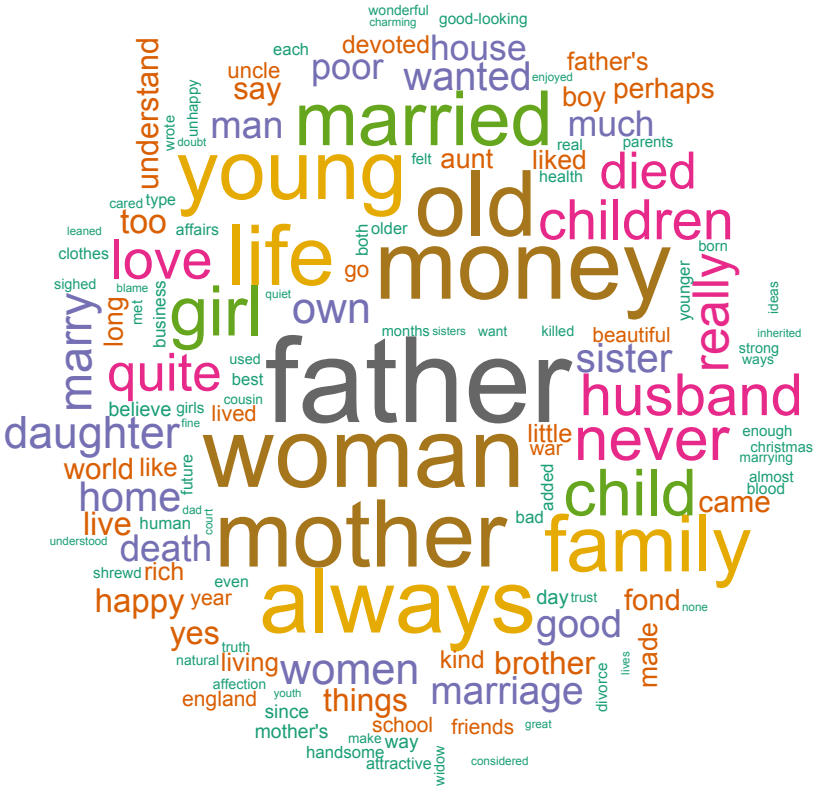

Figure 6.6: Word cloud of Topic 42

According to Figure 6.7 and the heatmap in Figure 6.5, this topic appears more frequently in Christie's works than in the other authors' works.

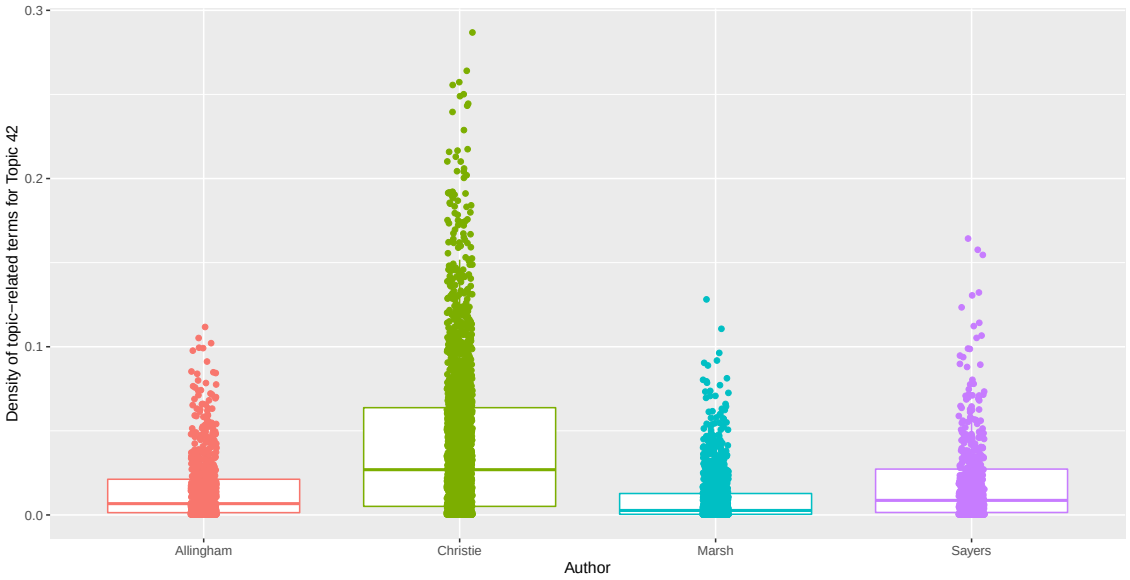

Figure 6.7: Density of topic-related terms for Topic 42

Figure 6.8 shows the density of topic-related terms for topic 42 in Christie's works. The texts are lined up from left to right in the chronological order of their publishing years. The density of this topic in her earlier works are low, but it becomes higher in her middle and later works.

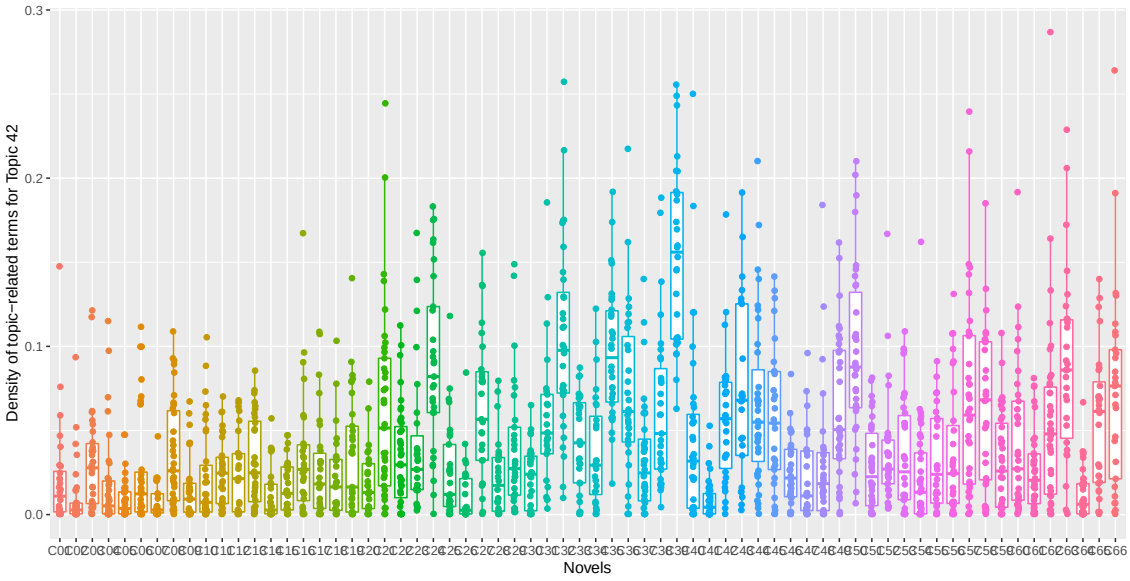

Figure 6.8: Density of topic-related terms for Topic 42 in Christie's novels

The density of this topic is the highest in C39 (*Crooked House*). In this novel, Aristide Leonides, the patriarch of the wealthy Leonides family, is killed in his house and members of the Leonides family are suspected to have killed him. Since the story focuses on them, it seems that the topic related to family is characteristic of this work.

6.3.2 Topic 40: Travelling around the world

Topic 40 contains many words related to travelling, such as *world*, *plane*, *Baghdad*, *hotel* and so on. Figure 6.9 is the wordcloud of this topic.

Figure 6.9: Wordcloud of Topic 40

Figure 6.10 shows the density of this topic in each author's corpus. We can see that there are many files in which the density of this topic is high.

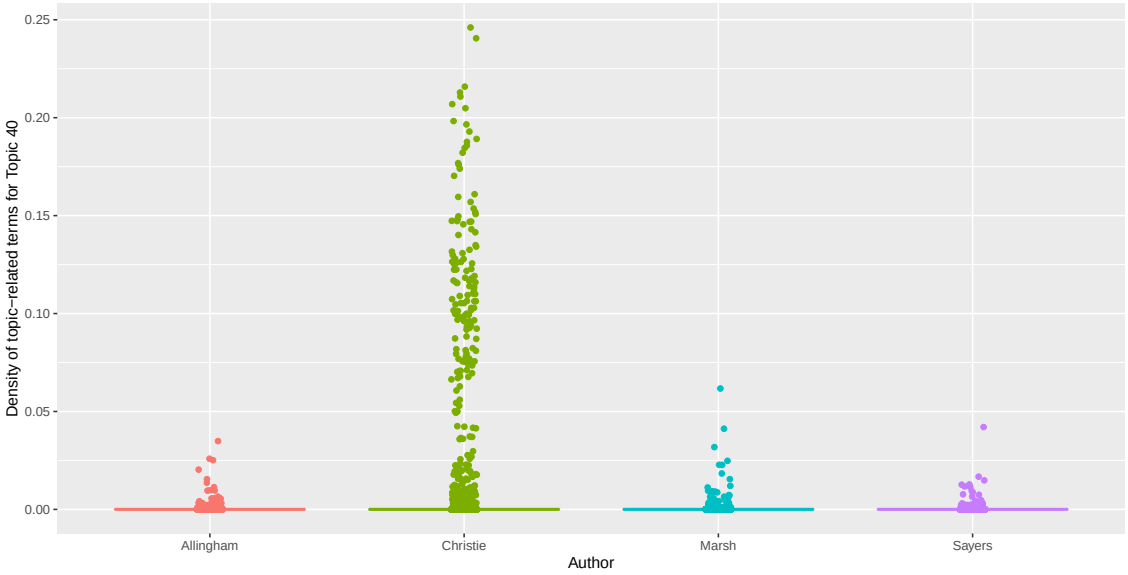

Figure 6.10: Density of topic-related terms for Topic 40

Figure 6.11 shows the density of topic-related terms for topic 40 in Christie's novels.

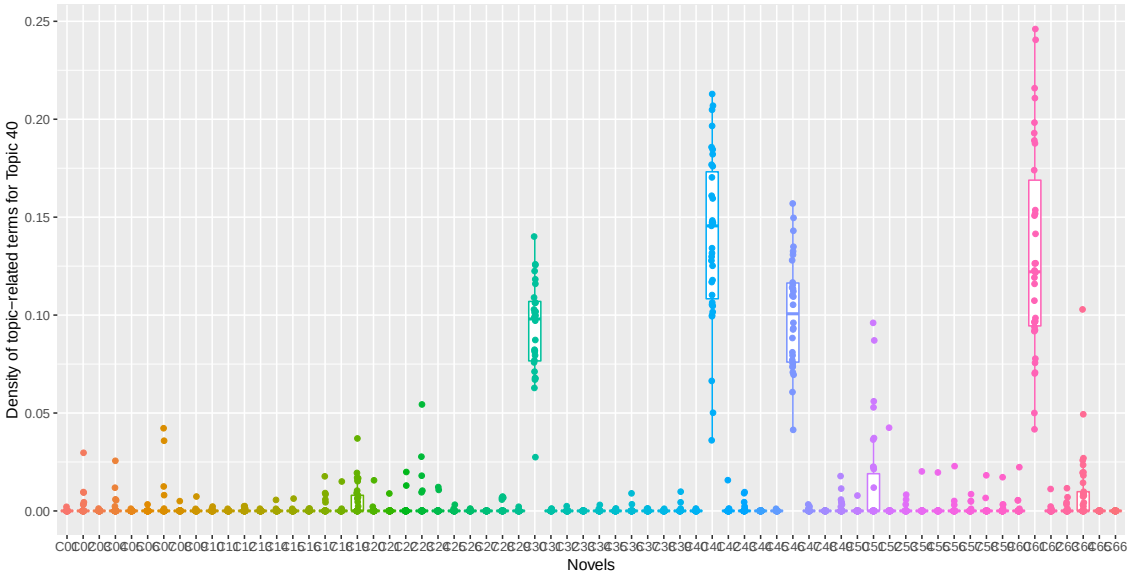

Figure 6.11: Density of topic-related terms for Topic 40 in Christie's novels

The density of this topic is high in some texts; C30 (*N or M?*), C41 (*They Came to Baghdad*), C46 (*Destination Unknown*) and C61 (*Passenger to Frankfurt*). All of these works were published after the 1940s. Christie divorced Archibald in the 1920s, and met her second

husband Max while she was travelling Mesopotamia in the 1930s. She loved the Middle East and often travelled there after she married Max. It is likely that her travel experiences affected her writing.

C30 (*N or M?*) is a spy thriller set during World War II. Feeling depressed, Tommy and Tuppence were visited by Grant, a British secret agent. Grant explained that another British agent had been killed, leaving the message: 'N or M. Song Susie'. Grant asked Tommy to go to Sans Souci, a hotel in Leahampton, to identify 'N' and 'M'. Only Tommy was sent to Sans Souci, but Tuppence arrived there before he did, and they began to investigate the spy together. This novel describes the situation in World War II and focuses on the hotel Sans Souci, thus the density of Topic 40 might be high.

The setting of C41 (*They Came to Baghdad*) is Baghdad. A young lady, Victoria Jones, who was recently fired, falls in love with a young handsome man named Edward, who is leaving for Baghdad the next day. Victoria could not give up him, so she decides to go to Baghdad for work and track him down. In Baghdad, she is involved in a murder case. In this novel, the story progresses from London and Baghdad, and the murder is carried out in a hotel room in Baghdad. This might explain why the density of Topic 40 is high in this work.

C46 (*Destination Unknown*) is a spy fiction set during the Cold War. A young American scientist named Thomas Betterton disappears. There seems to be a pattern of scientists disappearing, and Mr. Betterton is only the latest. Thomas' wife, who seems to have the key to this mystery, is dying from injuries she sustained in a plane crash. Meanwhile, a young woman named Hilary Craven, who resembles Thomas' wife, attempts to commit suicide but is stopped by a British secret agent named Jessop. Jessop asks Hilary Craven to impersonate Thomas' wife as he investigates the mystery of the disappearing scientists. In this novel, the location of Hilary Craven's attempted suicide is a hotel room, and Thomas' wife is injured in a plane crash. Therefore, the density of Topic 40 might be high in this work.

In C61 (*Passenger to Frankfurt*), the story begins in Frankfurt airport. Sir Stafford Nye, a middle-aged diplomat, is approached by a strange woman. She asks him to lend her his passport and mantle, claiming that she will be murdered if she continues to travel in her current manner. He lends her his passport and mantle, after which he is involved in a world-wide conspiracy. This novel treats the world-wide conspiracy and the story develops not only in

Germany but also in London. At the beginning of the story, the characters are at Frankfurt Airport. These might be the reason why the density of Topic 40 is high in this novel.

Other than the high density of Topic 40, these novels have another common point. All of these works are spy thrillers. Christie might have the tendency to write novels of a similar theme using similar topics.

6.3.3 Topic 22: Conversational expressions

As we can see in Figure 6.13, Topic 22 contains words such as *said*, *yes*, *know*, *think* and so on. This topic contains many words related to conversations. The density of this topic is high in Christie's texts, but not as high in the other authors' texts (Figure 6.14).

Figure 6.12: Partially enlarged figure of Figure 6.1 around Topic 22

Figure 6.13: Wordcloud of Topic 22

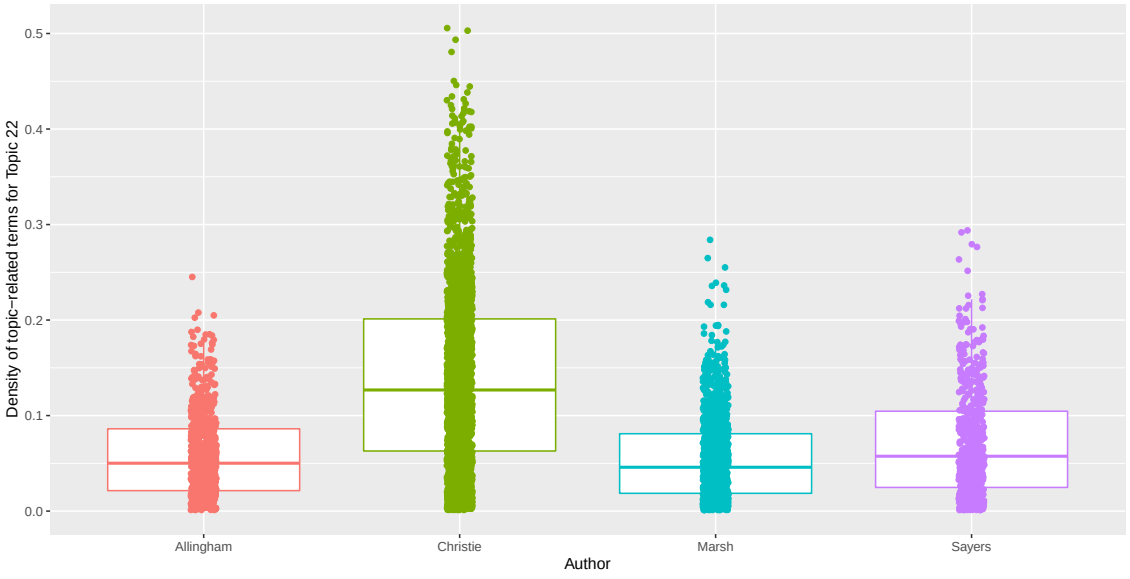

Figure 6.14: Density of topic-related terms for Topic 22

Christie's novels are propelled by conversations like the example below, so this topic seems to reflect the character of her novels.

(60) Poirot *said* quietly:

"You have remembered something, have you not?"

Roddy *said* doubtfully:

"*Yes* – but –"

"You wonder if you ought to tell me?"

"*Well*, *yes*..." (*Sad Cypress*)

In Figure 6.15, we can see that Christie tends to use the words related to conversational expression more frequently in her later works.

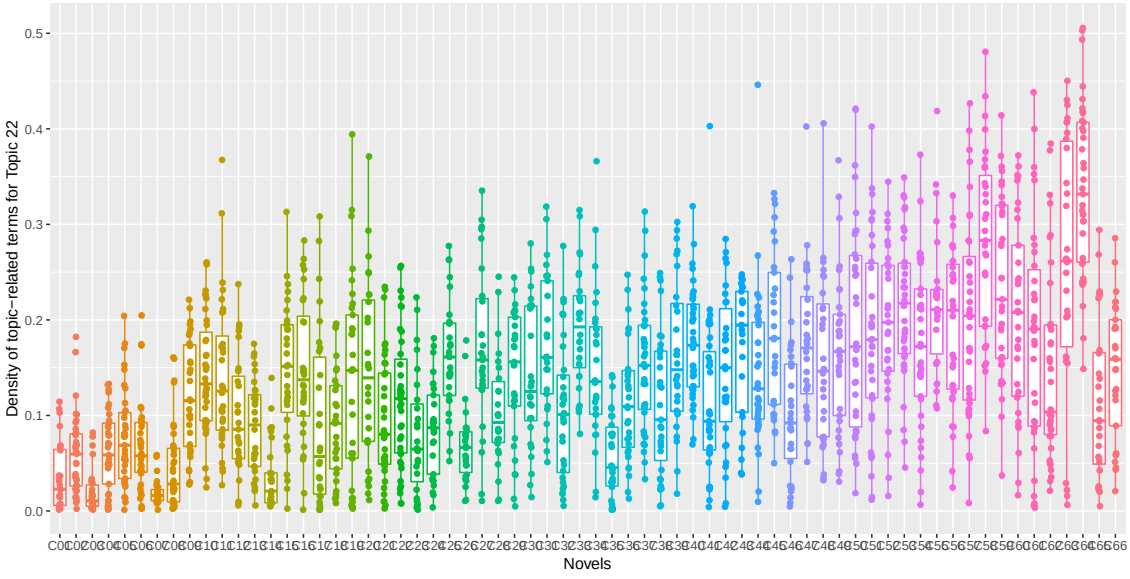

Figure 6.15: Density of topic-related terms for Topic 22 in Christie's novels

The verbs *know* and *think* are in the base form; it is highly possible that they are used in conversations in the present tense. These verbs indeed keep the base form when they are used with auxiliary verbs, but the clusters such as *I know*, *you know* and *I think* are used much more frequently than clusters containing auxiliary verbs (See Table 6.3). Thus these verbs are thought to be used frequently in conversations.

(61) “So that’s how *I know* Mr Curry didn’t go in by the front door. *I think* perhaps he got in somehow in the night and hid in an attic. Do *you think* that’s likely?” (*The Clocks*)

(62) “It mightn’t be that. It might be, *you know*, that they had a hold on her.” (*By the Pricking of My Thumbs*)

Table 6.3: Top two-word clusters with *know* (Left) and *think* (Right) in the Christie corpus

No	Cluster	Freq	No	Cluster	Freq
1	you know	5216	1	i think	5069
2	i know	2770	2	you think	2297
3	don’t know	2453	3	don’t think	1637
4	know what	1738	4	think it	983
5	to know	1638	5	think that	965
6	know that	1034	6	think i	959
7	know i	865	7	think of	952
8	know the	605	8	to think	899
9	not know	550	9	think so	831
10	didn’t know	547	10	think you	615

Moreover, this topic contains many contracted forms such as *I’m*, *that’s*, *didn’t*, *I’ve*, *wasn’t* and so on. Contracted forms are spoken rather than written, so it seems that these words also reflect the characteristic of her novels that the stories are mainly conversations.

(63) “Where were you at the time of the crime, *that’s* what he wants to know, Murgatroyd?” said Miss Hinchcliffe. She winked at Craddock. (*A Murder Is Announced*)

(64) “Supposing they *didn’t*?” said Virginia. “*I’m* pretty sure they *didn’t*, as a matter of fact." (*The Secret of Chimneys*)

6.3.4 Topic 46: Body parts and actions

Topic 46 contains words related to body parts such as *head*, *face*, *hand* and so on. There are also words related to actions such as *looked*, *shook*, *nodded* and so on (Figure 6.17). This topic seems to be related to the characters' actions accompanied by the physical expressions of their different body parts. This topic also contains *-ly* adverbs such as *slowly*, *suddenly*, *gently* and so on. Christie uses this topic much more frequently than the other authors (Figure 6.18).

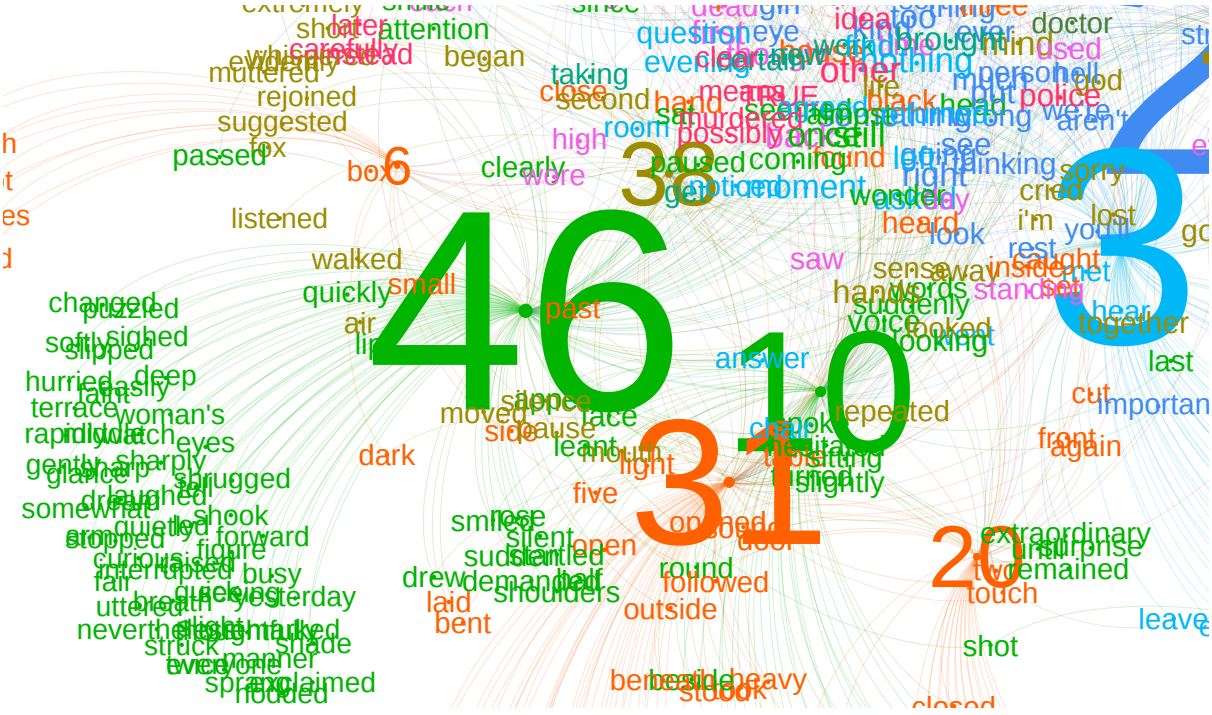

Figure 6.16: Partially enlarged figure of Figure 6.1 around Topic 46

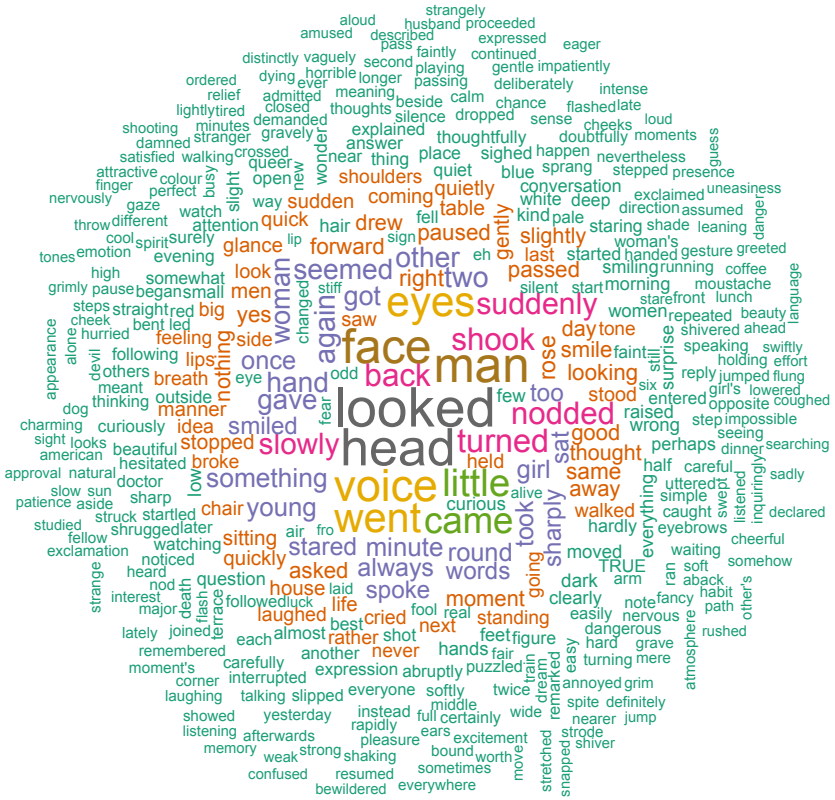

Figure 6.17: Wordcloud of Topic 46

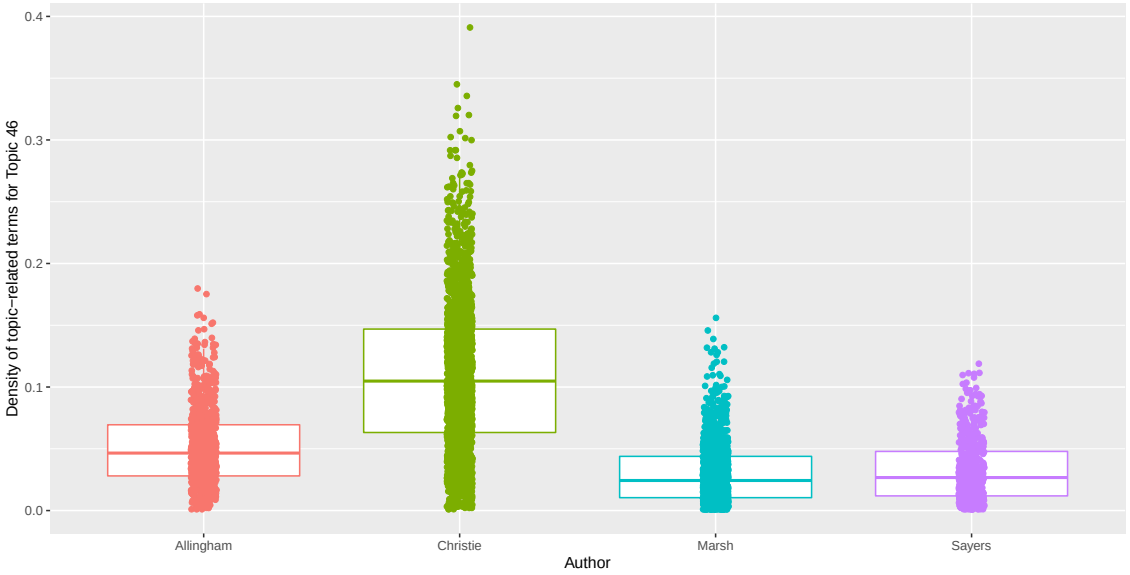

Figure 6.18: Density of topic-related terms for Topic 46

In Christie's novels, the density of this topic is larger in her earlier and middle works than in her later works (Figure 6.19). According to Figure 6.19, the densities in C63 and C64 are much smaller than those in other works. This result is the same as the result in the previous chapter, which focused on the diachronic change of Christie's style.

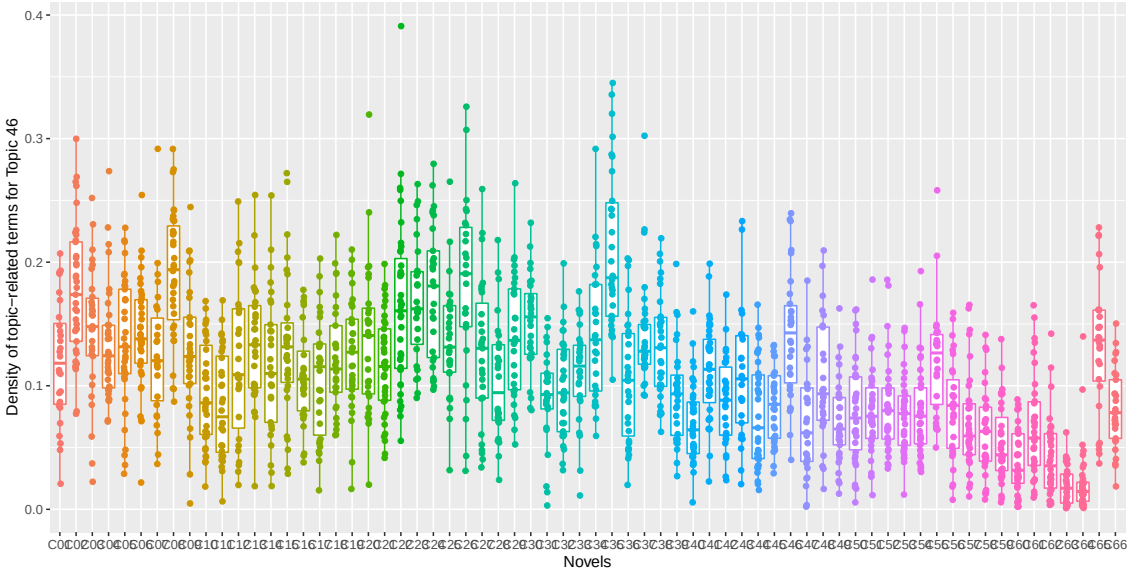

Figure 6.19: Density of topic-related terms for Topic 46 in Christie's novel

As we can see in the examples below, *-ly* adverbs seem to be used with the words related to body parts and/or actions.

(65) She said:

"Yes you understand you've known what it's like..."

She *turned suddenly*, a questioning look in her *eyes*:

"Aunt Laura..." (*Sad Cypress*)

(66) "It is as well you are in France," said Van Aldin. "We are made of sterner stuff in the States. Business comes before pleasure there."

Poirot did not take offence; indeed, he *smiled gently* and confidingly at the irate millionaire. (*The Mystery of the Blue Train*)

6.3.5 Topic 29: Signs of decline of linguistic ability

Topic 29 looks like Topic 22 at first sight. This topic contains the words related to conversations such as *said*, *yes*, *well* and so on. The major difference between Topic 29 and 22 is that the indefinite terms such as *things* and *something* have larger weights in Topic 29 than Topic 22 (Figure 6.21). These words are investigated in previous studies such as Lancashire & Hirst (2009) and Le et al. (2011) as the significant marker of the decline of linguistic ability.

Figure 6.20: Partially enlarged figure of Figure 6.1 around Topic 29

Figure 6.21: Wordcloud of Topic 29

In Christie's later works, the indefinite terms contained in this topic are used in a short span like the examples below.

(67) "I think probably that she was away at school in Switzerland at the time the tragedy happened and nobody told her much and, well, when you're a teenager or younger still you just accept *things* as *something* that happened, but that's not anything to do with you really." (*Elephants Can Remember*)

(68) Sooner or later, though, he had to ask questions. Certain *things* aroused a new and sinister apprehension in his mind. He was being persuaded, inveigled into *something*. But what? And whose doing was it? (*Passenger to Frankfurt*)

The densities of this topic in the other authors' novels are very small, but there are many files for Christie's novels which have a large density of this topic (Figure 6.22).

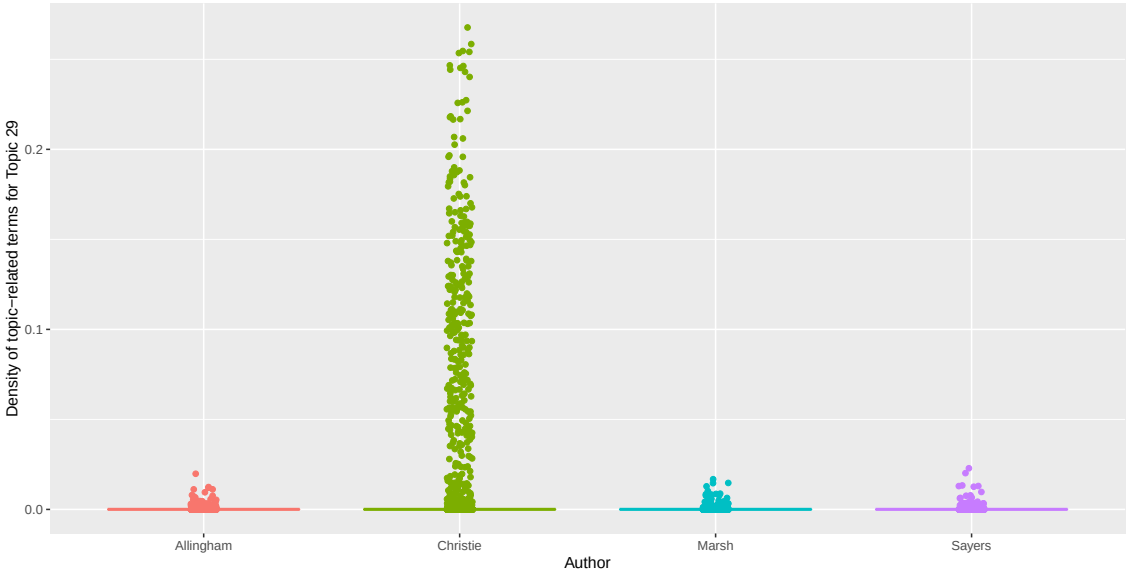

Figure 6.22: Density of topic-related terms for Topic 29

Figure 6.23 shows the density of Topic 29 in each text in the Christie corpus. We can see that while the densities in earlier works are small, the densities become large in her later works. The densities in C65 and C66 are small because they were written approximately 30 years earlier than their publishing years. The low connection of Topic 23 in C61 can be announced for by the genre the text belongs to; it is a spy thriller rather than a mystery in its classic sense. This also shows that this topic has some relationship with the decline of Christie's linguistic ability. The result that the topic of signs of decline in Christie's linguistic ability is significant in her later works is the same as the result of the previous chapter.

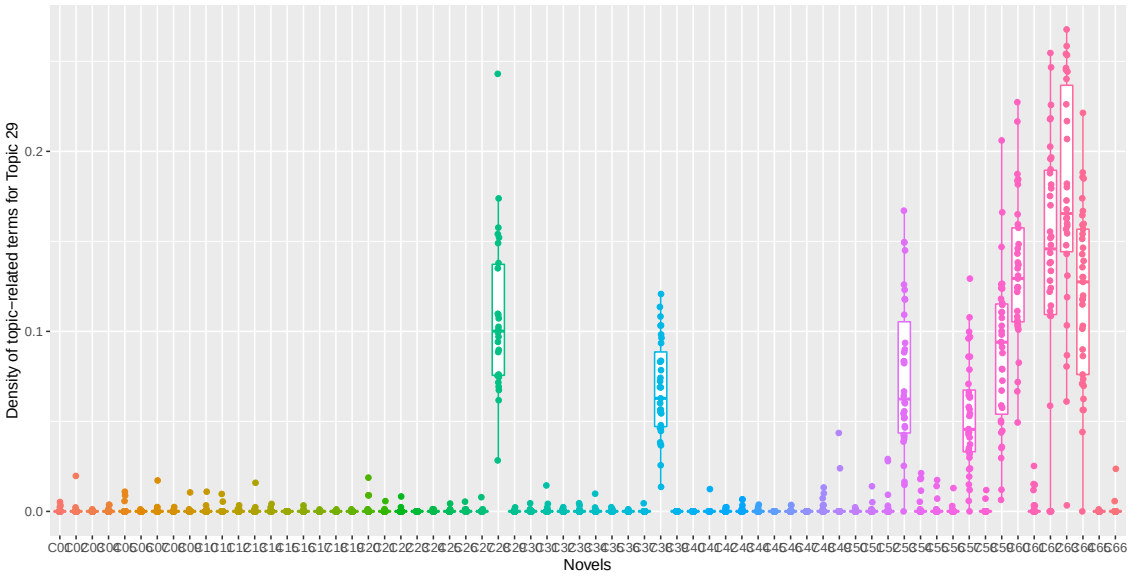

Figure 6.23: Density of topic-related terms for Topic 29 in Christie's texts

6.3.6 Topic 14: French and Poirot

Topic 14 contains words related to French such as *monsieur*, *madame*, *mademoiselle* and *Paris* (Figure 6.25). As we can see in Figure 6.26,the density of this topic in Christie's works is larger than that of the other authors' works. Hercule Poirot is a Belgian detective and often speaks French, and there are some characters speaking French in Christie's novels, so this topic reflects Poirot's and French speakers' utterances.

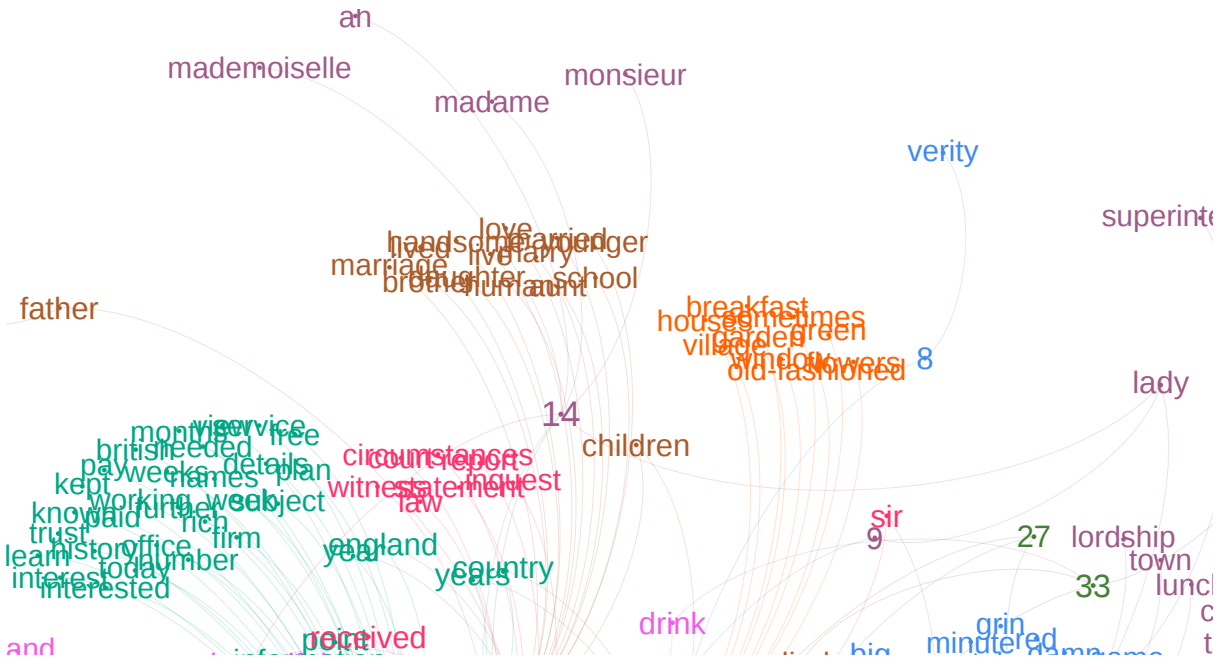

Figure 6.24: Partially enlarged figure of Figure 6.1 around Topic 14

Figure 6.25: Wordcloud of Topic 14

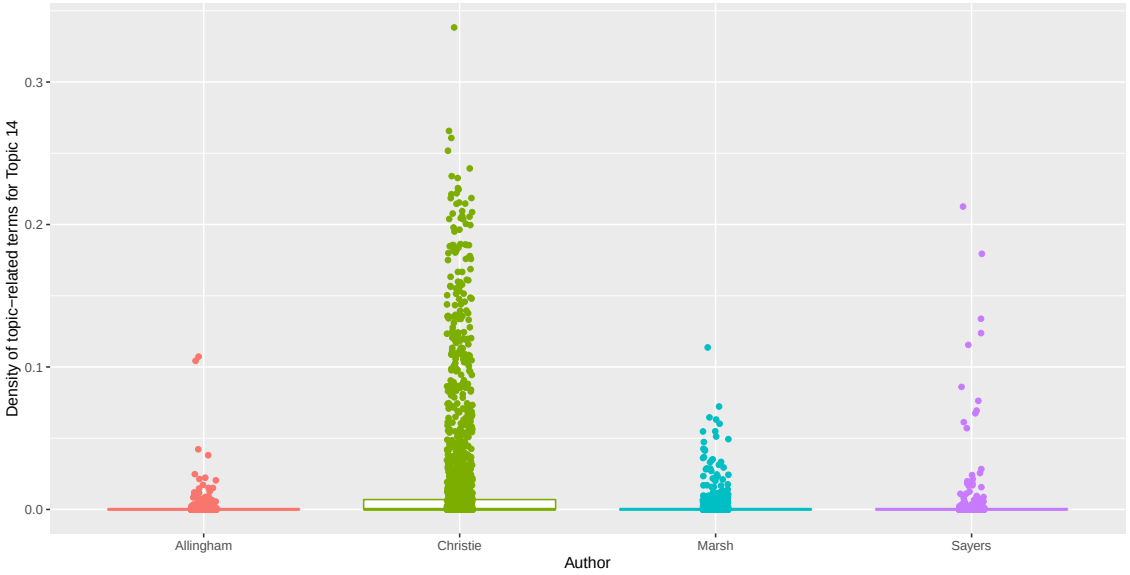

Figure 6.26: Density of topic-related terms for Topic 14

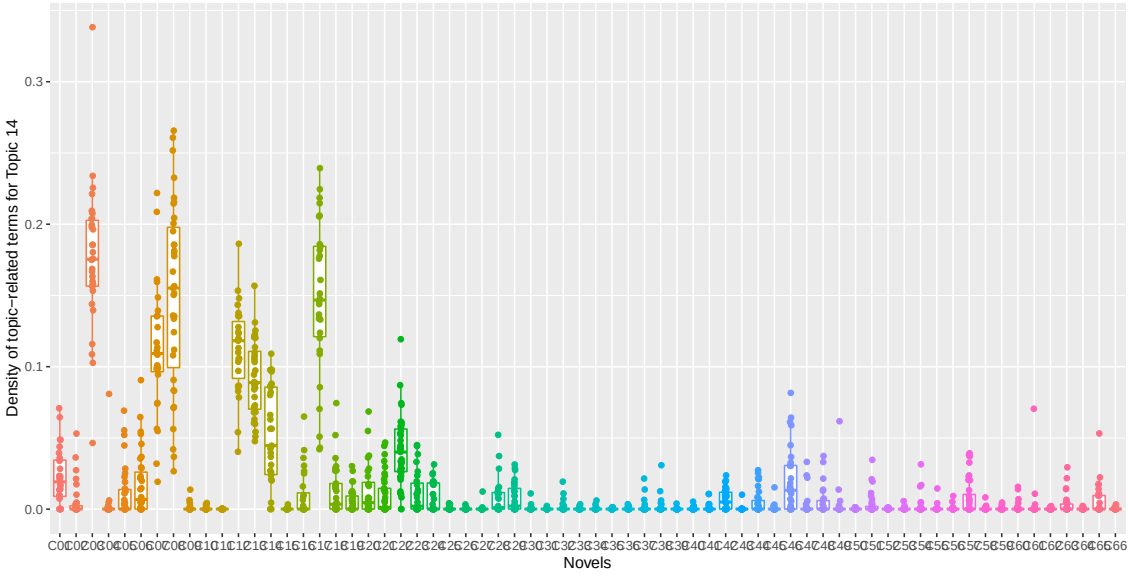

Figure 6.27: Density of topic-related terms for Topic 14 in Christie's novels

The words related to French are often used in utterances by Poirot like the examples below.

(69) "That is quite right, *Monsieur*. *Madame* trusted me. Ever since I entered her service I have carried out her instructions faithfully." (*Death in the Clouds*)

(70) "In what way can I be of service to you, *Mademoiselle*?" asked Poirot. (*Lord Edware Dies*)

This topic also contains words such as *murmured*, *nodded* and *head*. As we have seen in the previous chapter that these are words often used with *Poirot* to describe his behaviour, this topic reflects Poirot's behaviour. This topic also contains an interjection *ah*. As we can see in the examples below, this word is frequently used with *said Poirot*. Christie might use it as his favorite saying.

(71) Poirot *nodded* his *head* slowly in complete agreement. (*Evil Under the Sun*)

(72) Poirot *murmured*:
"Where would it have ended, I wonder?" (*Sad Cypress*)

(73) "*Ah*!" said Poirot. "You are very well informed very well informed. Someone has been very obliging." (*Lord Edware Dies*)

6.4 Surveying topics: The case of the number of topics set to 100

In this section, we will see another result of the topic model when the number of the topic is set to 100. When we increase the number of topics in LDA, we can obtain more detailed topics. However, there are more topics which occur in a certain text when increasing the number of topics than when the number of the topic is small. Therefore, we will exclude small topics which occur in a certain text and large topics which are pervasive in many texts by several authors because topics that are too small do not capture the characteristics of an author but the characteristics of a certain novel. Large pervasive topics are not very useful in identifying characteristic topics in Christie's works. In order to limit the topics which are used in the analysis, we use the alpha values. The alpha values show how large and pervasive the topics are. We can obtain these values in a topic-keys file. We exclude the topics with large alpha values which are pervasive in many texts and those with small alpha values which occur in the limited texts. We exclude the topics with the alpha values larger than 0.5 and smaller than 0.02. As a result, we obtain 36 topics. Table 6.4 is the list of the keywords in each topic, and Figure 6.28 is a heatmap which shows the density of each topic in each author's corpus.

According to Figure 6.28, we can see that the density of topics 1, 3, and 16 is higher in Christie's novels than in other authors' novels. These topics are almost the same as the topics obtained in the previous section. Topic 1 contains *world*, *plane*, *Baghdad* and so on, thus this topic is similar to the topic of traveling in the previous section. Topic 3 contains *monsieur*, *madame*, *mademoiselle* and so on; thus this topic is similar to the topic of French in the previous section. Topic 16 contains indefinite terms such as *things* and *something*, and this topic is similar to the topic of the signs of the decline of Christie's linguistic ability in the previous section.

Table 6.4: Key words in each topic

Topic	alpha	Words
0	0.30189	too, good, table, party, little, evening, dinner, three, people, glass, great, drink, like, lady, much, play, young, other, hand
1	0.02346	world, plane, baghdad, olive, young, man, scheele, pauncefoot, money, clipp, carmichael, rathbone, youth, country, hotel, england, passport, air
2	0.14545	well, man, rather, till, better, good, again, replied, gentleman, way, away, anything, course, lordship, somebody, found, morning, make, suppose
3	0.05438	monsieur, madame, mademoiselle, paris, ah, english, lady, french, comte, cannot, ami, husband, magistrate, hotel, daubreuil, villa, commissary, girl
4	0.03651	evidence, coroner, jury, deceased, court, inquest, witness, heard, duke, body, sir, prisoner, verdict, witnesses, case, impey, judge, lady, defence
7	0.20043	body, man, dead, shot, doctor, head, blood, found, death, police, revolver, killed, lying, accident, happened, men, gun, get, right
11	0.03049	says, man, ain't, see, old, say, o', that's, lordship, well, an', gentleman, ah, give, young, time, wot, bein', call
14	0.33563	old, like, little, great, two, long, first, large, place, white, new, big, years, way, round, small, much, never, side
16	0.38405	said, know, things, well, think, something, yes, people, name, remember, time, like, mean, old, years, say, ago, perhaps, other
18	0.35766	dear, lady, always, thought, quite, little, good, course, much, never, poor, like, nice, old, really, day, people, too, indeed
20	0.02793	train, compartment, man, bouc, platform, carriage, station, conductor, passengers, corridor, next, arbuthnot, door, just, american, monsieur, travelling
21	0.13587	sir, lady, man, young, gentleman, yes, chief, house, that's, butler, fellow, sure, secretary, gentlemen, strange, thank, home, master, years
25	0.09884	dr, doctor, nurse, patient, patients, father, quite, death, medical, doctors, nervous, doctor's, hospital, case, wife, bed, opinion, gave, sister
33	0.05243	nurse, bottle, poison, yes, arsenic, poisoning, taken, girl, poisoned, stuff, dose, died, suicide, left, morphine, tablets, case, medicine, death
34	0.04121	inspector, man, girl, murder, police, cust, webb, crescent, crome, house, martindale, rival, sheppard, clock, gudgeon, waterhouse, boys, inquest
35	0.30596	thought, well, sort, asked, say, course, perhaps, each, indeed, look, gave, wondered, added, darling, extremely, rather, air, almost, called
38	0.18889	letter, letters, paper, read, written, write, wrote, envelope, note, writing, desk, words, address, sheet, same, again, took, dear, message
42	0.03896	uncle, aunt, cousin, house, great-aunt, inspector, young, auntie, police, family, morning, stanislaus, featherstone, uncle's, oates, william's, close
43	0.06911	fox, sir, yes, asked, case, yard, sergeant, thompson, inspector, prints, he'd, sort, fancy, br'er, chief, reckon, job, what's
47	0.02870	lugg, ere, yer, ave, ain't, e's, that's, wot, knapp, doll, tiddy, bloke, lugg's, gaffer, ouse, oo, roly, feller, goin
48	0.15117	road, village, back, lane, drive, bank, river, drove, country, miles, little, left, bridge, way, hill, side, turned, great, green
51	0.23132	police, found, job, first, desk, way, set, appeared, report, returned, until, photograph, course, office, left, far, case, hope, find
52	0.16466	bed, night, sleep, morning, good, asleep, last, sat, still, slept, breakfast, drink, hour, tomorrow, awake, light, water, bedroom, wake
55	0.25729	street, man, telephone, london, name, taxi, shop, office, flat, number, road, receiver, address, yard, corner, walked, just, door, rang
57	0.22980	miss, quite, really, dear, course, young, old, little, woman, kind, just, anything, sure, get, perhaps, wasn't, ladies, afraid, girls
63	0.33191	father, mother, young, old, married, girl, years, family, child, wife, woman, husband, always, children, home, died, sister, money, daughter
74	0.11583	book, read, books, paper, write, page, reading, detective, work, written, writing, story, interesting, daily, newspaper, pages, copy, literary, stories
79	0.31537	hair, like, black, eyes, little, face, hat, clothes, dress, white, looked, wore, blue, girl, green, red, seen, woman, coat
82	0.17606	house, garden, path, window, place, gate, away, trees, side, drive, looking, terrace, turned, round, seen, came, village, led, small
84	0.03022	boat, hotel, yes, beach, rock, island, cliff, captain, rocks, morning, colgate, cove, bay, darnley, water, blatt, coast, tide, bathing
85	0.18338	money, pounds, business, hundred, thousand, much, good, paid, never, left, made, family, sum, died, pay, five, bank, year, three
88	0.09934	tea, coffee, cup, kitchen, table, day, house, tray, breakfast, bacon, family, food, eat, eggs, make, butter, murder, cook, sugar
89	0.39310	murder, crime, man, murderer, killed, case, evidence, police, first, motive, person, story, people, woman, truth, two, true, time, committed
92	0.02764	oates, man, superintendent, avril, just, police, old, street, canon, havoc, hutch, duds, sergeant, chief, anscombe, campion's, knew, men, we're
93	0.42015	door, room, hall, house, went, opened, window, open, stairs, heard, came, upstairs, left, floor, back, bed, go, front, bedroom
97	0.42736	found, two, back, took, look, small, table, left, looked, pocket, hand, box, used, find, put, see, other, nothing, paper

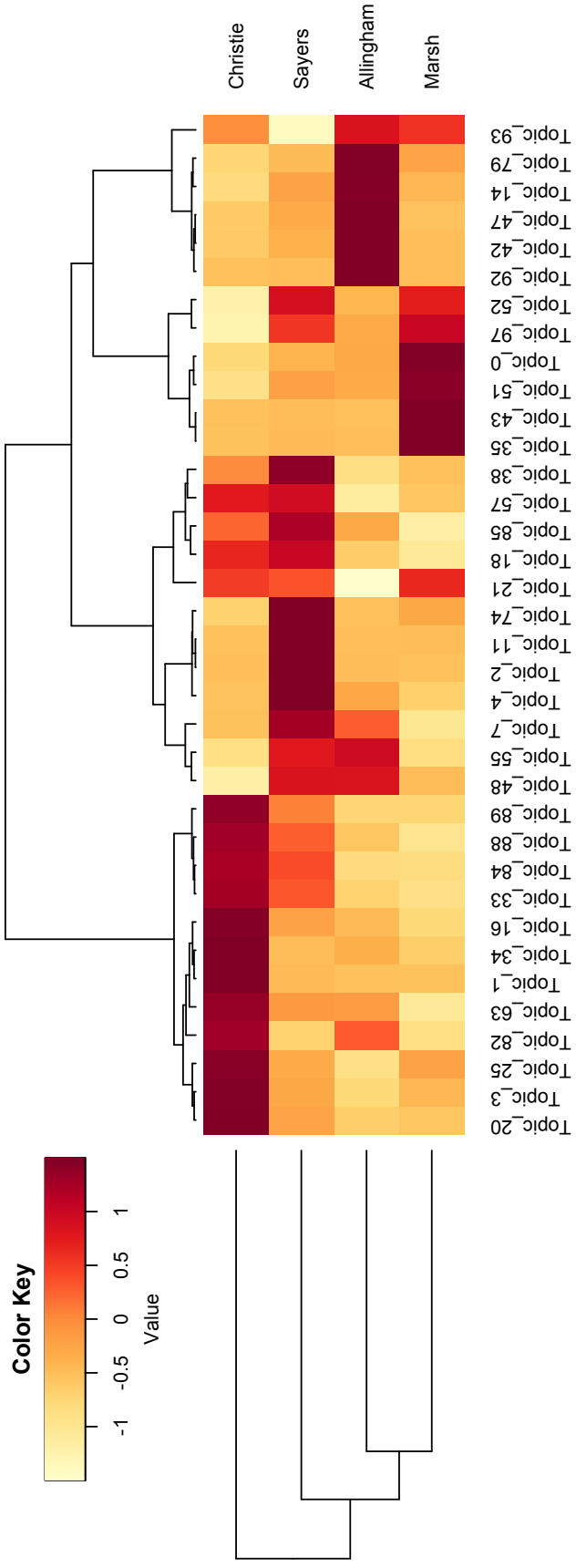

Figure 6.28: Heatmap of the topic density

According to Figure 6.29, the word cloud of Topic 20, Topic 20 contains *train*, *compartment*, *platform* and so on. This topic is related to travelling by train rather than plane. There is also a similar topic in the result when LDA is carried out setting the number of topics to 50, but the topic of travelling by train when the number of the topic is set to 100 seems to have more words with larger weights for this topic.

Figure 6.29: Word cloud of Topic 20

Figure 6.30 shows the density of Topic 20 in each author's corpus. The density of this topic is very small, although it is large in many of Christie's texts.

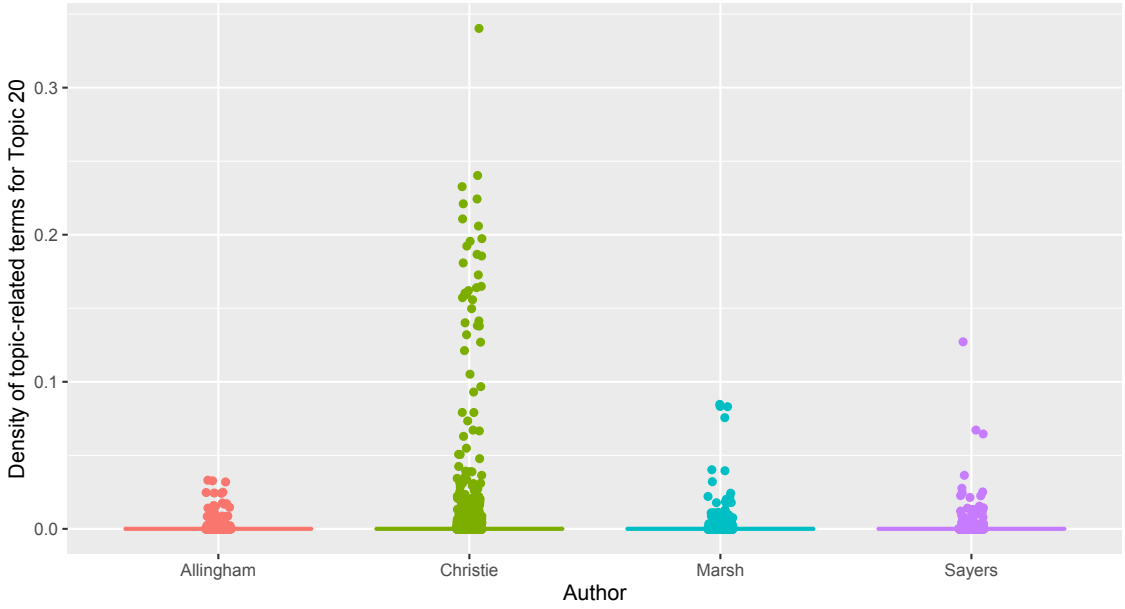

Figure 6.30: Density of topic-related terms for Topic 20

Figure 6.31 shows the density of this topic in Christie's novels. The density is the largest in C14 (*Murder on the Orient Express*). C8 (*The Mystery of the Blue Train*) is the second largest in the density of this topic, and some segments in C49 (*4.50 from Paddington*) have a large density of this topic.

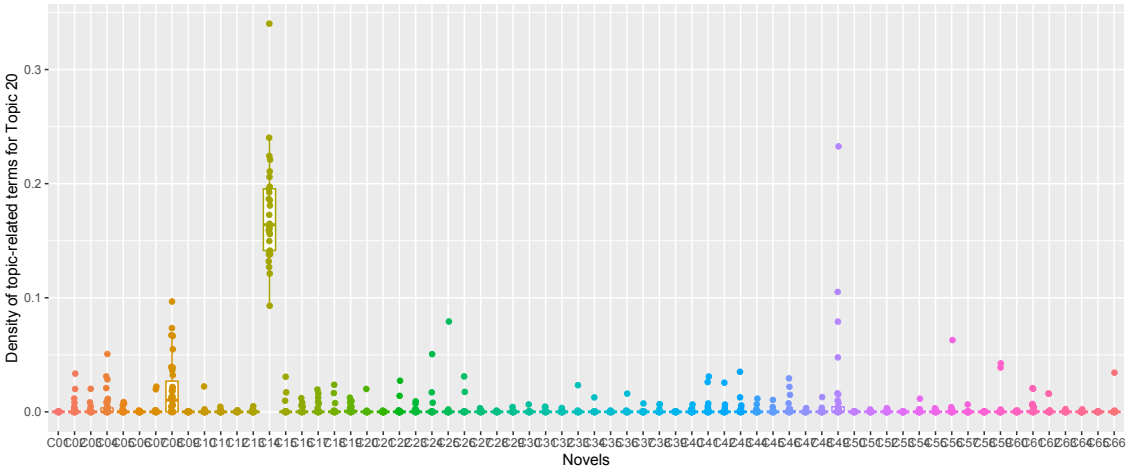

Figure 6.31: Density of topic-related terms for Topic 20 in Christie's novels

Needless to say, *Murder on the Orient Express* is a crime novel whose theme is a murder

in a sleeper train, and Poirot tries to settle the case. After settling a case in Syria, Poirot was planning to spend his holidays in Istanbul. However, he was told to return to London, and he managed to take the Orient Express. On this sleeper train, a wealthy man named Ratchett is killed, and Poirot tries to settle the case. It is natural that the density of this topic is by far larger in this work than others.

In *The Mystery of the Blue Train*, murder is also carried out on a train. Ruth Kettering, an American heiress leaving London to meet her former lover, is killed on Le Train Bleu. Poirot happened to board this train, so he is asked to settle this case. Therefore, the density of Topic 20 seems to be large in this novel.

In *4.50 from Paddington*, Mrs. Elspeth McGillicuddy, one of Miss Marple's friends, sees a woman killed by a man in a compartment of a train when she is boarding another train that leaves at 4:50. Murder is also carried out on a train in this novel, so the density of this topic seems to be large.

Figure 6.32: Word cloud of Topic 89

In addition, we would like to see Topic 89. Figure 6.32 is a word cloud of this topic. In this topic, there are words related to crime and murder, such as *murder*, *crime*, *killed* and so on. This seems to be the topic of crime and murder.

Figure 6.33 shows the density of Topic 89 in each author's corpus. Although the data used in this paper consist of crime fiction, the density of this topic is larger in Christie's corpus than the other authors'. A similar topic can be found in the result of the previous section. In the previous section, Sayers' corpus had the second largest density of crime and murder topic (Topic 47), but according to Figure 6.33, the density of Topic 89 in Sayers corpus is almost the same as Allingham's and Marsh's corpora.

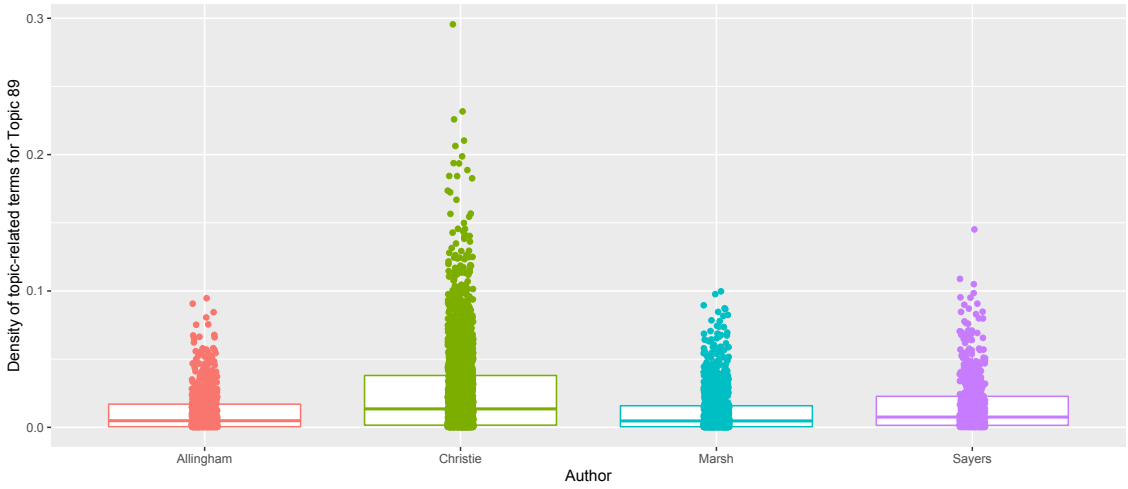

Figure 6.33: Density of topic-related terms for Topic 89

Figure 6.34 shows the density of this topic in Christie's works. We can see that this topic appears in almost all of her novels to some degree.

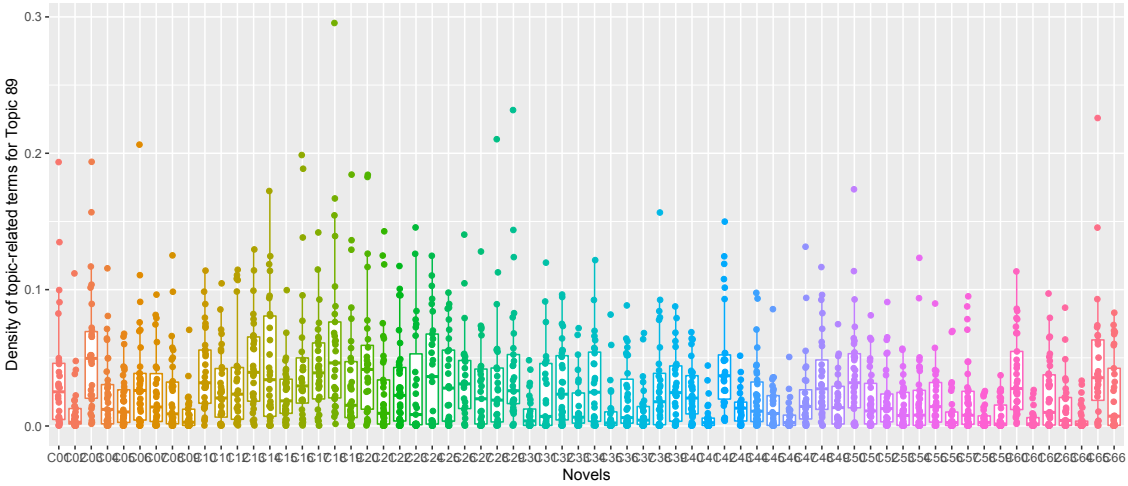

Figure 6.34: Density of topic-related terms for Topic 89 in Christie's novels

When we focus on the topic which is underused in Christie's texts but is used more frequently in other authors' novels, we can find Topic 48.

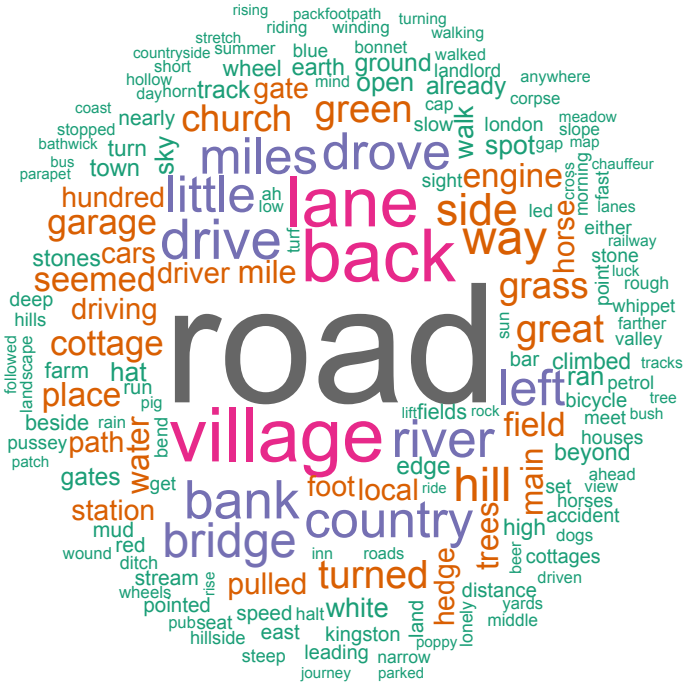

Figure 6.35: Wordcloud of Topic 48

According to the wordcloud in Figure 6.35, this topic contains words related to driving

such as *road*, *drive*, *miles* and so on. This topic seems to be about driving somewhere.

Figure 6.36 shows the density of this topic in each author's corpus. The density of this corpus is much smaller than other authors.

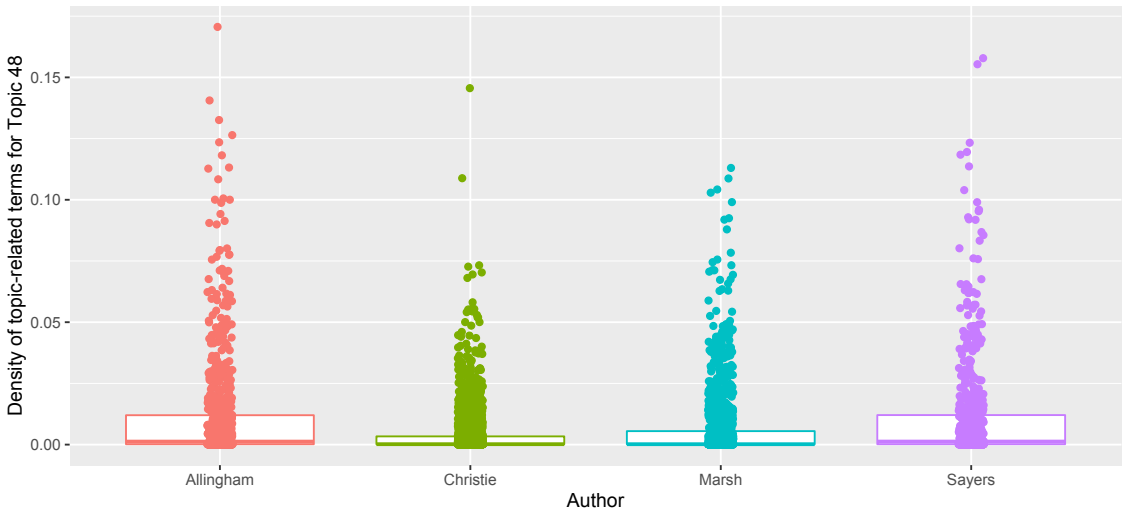

Figure 6.36: Density of topic-related terms for Topic 48

As we have seen, Christie tends to use topics of travelling using planes or trains. However, she rarely uses the topic of driving. As she loved to travel to The Middle East and often wrote about travelling by plane or train, she might not have been very interested in driving, thus only infrequently using the situation of driving.

When the number of topics is set to 100, and too-small topics and too large topics are excluded from the analysis, we could gain clearer topics to interpret than when the number of topics is set to 50.

6.5 Summary of this chapter

In this chapter, we employed topic modeling using LDA and investigated characteristic topics in Christie's works. As the result of LDA in which the number of the topic is set to 50, we obtained topics characteristic in Christie's novels compared with other authors, such as a topic of family and life, that of traveling, that of conversational expressions, that of body parts and actions, that of signs of Christie's decline of her linguistic ability and that of French and Poirot.

The topics of family and life and travelling around the world were the topics that were not found in the previous chapters. These topics were affected by the theme of the novels and a few works had a high density of these topics. The density of the topic of family and life was low in Christie's earlier works, but became higher in her middle and later works. The density of the topic of travelling was high in some novels, most of which were written in the 1930s and later, after she divorced Archibald Christie and married Max Mallowan. Since she first met Max while travelling in Mesopotamia and often traveled The Middle East with him, accompanying his excavations, it seems that she loved to describe stories related to The Middle East. Moreover, all of the novels in which the density of the topic of travelling was high were spy thrillers. It seems that Christie tended to use similar situations when writing novels in this genre.

As we have seen in the previous chapters, the topics of conversational expressions, that of body parts and actions and that of signs of the decline of Christie's linguistic ability were also characteristic in the Christie corpus. The topic of conversational expressions reflected the characteristic of her novels being propelled by conversations between characters. This topic contains some words in the base form such as *know* and *think*, and some contracted forms such as *I'm*, *that's* and so on. The words in the topic of body parts and actions seemed to be used frequently with *-ly* adverbs. The density of the topic of the decline of Christie's linguistic ability was much higher in her last several works. This topic contained some indefinite words such as *something* and *things*. This result was consistent with previous studies and the previous chapters. The topic of French was also obtained as a result of LDA. Since Poirot often spoke French, this topic had a strong relationship with him. This topic contained not only French words, but also the verbs describing Poirot's actions and interjection. Thus, this topic reflects his actions and utterances.

We also saw the result of LDA when we obtained 100 topics and excluded too-small topics and too-pervasive topics using alpha values. We could obtain a similar result when the number of topics was set to 50, but some topics became clearer to interpret. The topic of travelling by train might be among Christie's favorite themes, along with the topic of travelling around the world. However, the topic of driving is not used in Christie's works very frequently compared with the other authors. She might have loved to describe situations travelling by plane or train

more than driving.

Applying the topic model, we could obtain features in Christie's novels in the meaning level which could not be found through the conventional keywords analysis. Therefore, this method is useful in investigating a text's stylistic features.

Chapter 7

Conclusion

7.1 Christie's style in comparison with other authors

We have investigated stylistic features in Christie's novels compared with her three contemporaries, Sayers, Allingham, and Marsh, using stylometric methods. As a result of analysis through PCA, cluster analysis, RF and topic model, we were able to see important stylistic features of Christie's works.

First, the characteristic words in Christie's works are characterised by words related to females. When we survey the collocates of the female pronoun *she* in her works using MI score, we see stereotypically feminine manifest words related to crying, fear or sadness. Comparing the collocates of *she* with those of he in her works, the collocates of she make a great contrast to those of *he*; the collocates of *he* include the words which are used for the murderer–although those of *she* contain many words related to weakness of females. Christie describes women stereotypically; she describes them as much weaker than men.

Second, Christie's style is characterised by the great variety of *-ly* adverbs and the verbs they modify. Moreover, investigating the clusters including the *-ly* adverbs, we could find that Christie uses *-ly* adverbs in fixed expressions and uses them in her works repeatedly. She tends to use these fixed expressions to describe the manner of characters' utterances and behaviour. As we saw through the investigations in this study, Christie's novels contain many easy and plain vocabularies and many fixed expressions. Thus, anyone can enjoy reading her novels regardless of his or her age or nationality. This might be one of the reasons her novels are popular throughout the world.

7.2 Changes of Christie's style over years

We have investigated not only stylistic features in Christie's works compared with other authors but also a diachronic change of her style in her novels. Christie's career lasted approximately 50 years. We found diachronic changes of style in her novels. Previous studies by other researchers investigated diachronic changes of Christie's style in her novels, but this study analysed changes of her style using quantitative methods. The results of this study support the findings in previous studies by other authors.

First, it seems that the proportion of conversation becomes larger as Christie ages. This is illustrated by the results that the characteristic words in her later works include many verbs in their base form which are used in conversations, and that conversational expressions such as contracted forms are used frequently in her later works. One of the reasons for this might be that she changed her method of writing novels from typing to using a Dictaphone after she broke her wrist.

Second, in her last several works she tends to use many more indefinite terms such as *things* and *something* than throughout her career as a whole. The increased use of the indefinite terms can be seen clearly in the result of the topic model. This result is consistent with previous studies on the decline of her cognitive ability due to the speculation that she suffered from Alzheimer's disease in her later life. The findings in this research confirm the findings in previous studies.

7.3 Using topic modelling to investigate stylistic features

This paper applied the topic model in combination with conventional keyword analysis to investigate stylistic features of Christie's texts. Using topic modelling, we can find stylistic features in the meaning level which cannot be found through conventional keyword analysis such as PCA. In this research, we found that the topic of travelling around the world was characteristic in Christie's corpus, to which particular reference is made in spy thrillers. This method is useful in investigating stylistic features in one's texts and finding out features which cannot be obtained using conventional methods.

7.4 Limitations and future tasks

As this paper considers changes in style in Christie's novels over the years, all of her short stories are excluded from the data for the analysis. This is because short stories are often published as part of a short story collection, and they are not always published in the same time period in which they were written. In order to investigate stylistic features in greater detail, I would like to analyse characteristic style in Christie's works compared with other authors' works using short stories as well as long novels.

This research deals with only crime fiction written by four writers. It might be interesting to use data of works written by other contemporary authors.

This research used topic modeling and surveyed topics obtained through LDA. However, the survey of topics in this paper is still insufficient to capture stylistic features in Christie's texts more in detail. It is necessary to investigate the obtained topics more. I would also like to investigate changes of topics in one novel.

Bibliography

Akano, I. 赤野一郎, Hori, M. 堀正広 and Tono, Y. 投野由紀夫 (2014). *Eigo Kyōshi no tameno Kōpasu Katsuyou Gaido*『英語教師のためのコーパス活用ガイド』[*A Practical Guide to Using Corpora in English Language Teaching*]. Tokyo: Taishūkan Shoten 大修館書店.

Antonia, A., Craig, H. and Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, 29(2): pp.147–163.

Blei, D. (2012a). Probabilistic topic models: surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM*, 55(4): pp.77–84.

Blei, D. (2012b). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, 2(1): pp.8–11.

Blei, D. and Lafferty, J. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, 1(1): pp.17–35.

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: pp.993–1022.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45: pp.5–23

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. London: Clarendon Press.

Burrows, J. F. (1992). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing*, 7(2): pp.91–109.

Burrows, J. F. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3): pp.267–287.

Burrows, J. F. (2003). Questions of Authorship: Attribution and Beyond. *Computers and the Humanities*, 37: pp.5–32.

Burrows, J. F. and Hassall, A. J. (1988). Anna Boleyn and the Authenticity of Fielding's Feminine Narratives. *Eighteenth-Century Studies*, 21(4): pp.427–453.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

DiMaggio, P., Nag, M. and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of US government arts funding. *Poetics* 41(6): pp.570–606.

Ellegård, A. (1962). *Who was Junius?* Stockholm: Almqvist and Wiksell.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1): pp.1–35.

Hirst, G. and Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4): pp.405–417.

Hori, M. 堀正広 (2009). *Eigo Korokēshon Kenkyū Nyūmon* 『英語コロケーション研究入門』[*Introducing English Collocation Studies*]. Tokyo: Kenkyūsha 研究社.

Hori, M. 堀正広 (Ed.) (2012). *Korekara no Korokēshon Kenkyū*『これからのコロケーション研究』[*Collocation Studies in the Future*]. Tokyo: Hituzi Shobō ひつじ書房.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Inaki, A. 稲木昭子 (2013). *Nazotoki no Kotobagaku: Agasa Kurisuti no Eigo wo Tanoshimu* 『謎解きのことば学 : アガサ・クリスティの英語を楽しむ』[*Language Studies of Mystery: Appreciating Agatha Christie's English*]. Tokyo: EihōSha 英宝社.

Ishida, M. 石田基広 (2008). *R niyoru Tekisuto Mainingu Nyūmon*『R によるテキストマイニング入門』[*An Introduction to Text Mining Using R*]. Tokyo: Morikita Shuppan 森北出版株式会社.

Ishikawa, S. 石川慎一郎 (2008). *Eigo Kōpasu to Gengo Kyōiku: Dēta toshiteno Tekisuto* 『英語コーパスと言語教育 : データとしてのテクスト』[*English Corpora and Language Teaching: Texts as Data*]. Tokyo: Taishūkan 大修館書店.

Ishikawa, S. 石川慎一郎 (2012). *Bēsikku Kōpasu Gengogaku*『ベーシックコーパス言語学』[*Basic Corpus Linguistics*]. Tokyo: Hituzi Shobō ひつじ書房.

Ishikawa, S. 石川慎一郎, Maeda, T. 前田忠彦 and Yamazaki, M. 山崎誠 (Eds.) (2010). *Gengo Kenkyū no tameno Tōkei Nyūmon*『言語研究のための統計入門』[*Statistics for Language Studies: An Itroduction*]. Tokyo: Kuroshio Shuppan くろしお出版.

Jin, M. 金明哲 (2007). *R niyoru Dēta Saiensu: Dēta Kaiseki no Kiso kara Saishin Shuhou made*『R によるデータサイエンス : データ解析の基礎から最新手法まで』[*Data Science Using R: From Basics of Data Analysis to Latest Techniques*]. Tokyo: Morikita Shuppan 森北出版.

Jin, M. 金明哲 (2009). *Tekisuto Dēta no Tōkei Kagaku Nyūmon*『テキストデータの統計科学入門』[*Introducing Statistics for Textual Data*]. Tokyo: Iwanami Shoten 岩波書店.

Jin, M. 金明哲, Kabashima, T. 樺島忠夫 and Murakami, M. 村上征勝 (1993). Tōten to Kakite no Kosei「読点と書き手の個性」[The Positioning of Commas in Sentences and Writing Styles]. *Keiryō Kokugo Gaku*『計量国語学』[*Mathematical Linguistics*], 18: pp.382–391.

Jin, M. 金明哲 and Murakami, M. 村上征勝 (2007). RandomForests Hou niyoru Bunshō no Kakite no Dōtei「RandomForests 法による文章の書き手の同定」[Authorship

Identification Using Random Forests]. *Tōkei Sūri*『統計数理』[*Proceeding of the Instutute of Statistical Mathematics*], 55(2): pp.255–268.

Joannou, M. (Ed.) (2013). *The History of British Women's Writing, 1920–1945: Volume Eight*. London: Palgrave Macmillan.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature.* New York: Springer.

Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6): pp.750–769.

Johnson, K. (2008). *Quantitative Methods in Linguistics.* Malden, MA: Blackwell.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2): pp.263–275.

Kobayashi, Y. 小林雄一郎, Tanaka, S. 田中省作 and Tomiura, Y. 冨浦洋一 (2011). Randamu Foresuto wo Mochiita Eigo Kagaku Ronbun no Bunrui to Hyōka「ランダムフォレストを用いた英語科学論文の分類と評価」[Classification and Assessment of English Scientific Papers Using Random Forests]. *Jōhō Shori Gakkai Kenkyū Hōkoku*『情報処理学会研究報告 IPSJ SIG Technical Report』[*IPSJ SIG Technical Report*], 90(6): pp.53–68.

Lancashire, I. and Hirst, G. (2009). *Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study.* Paper presented at the 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, Toronto.

Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(4): pp.435–461.

Maeda, K. 前田啓朗 and Yamamori, K. 山森光陽 (Eds.) (2004). *Eigo Kyōshi no tameno Kyōiku Dēta Bunseki Nyūmon: Jugyō ga Kawaru Tesuto, Hyōka, Kenkyū*『英語教師のための教育データ分析入門：授業が変わるテスト・評価・研究』[*An English Teacher's Introduction to Quantitative Data Analysis*]. Tokyo: Taishūkan 大修館書店.

Malling, S. and Peters, B. (Eds.) (1997). *AZ Murder Goes: Classic*. Scottsdale: Poisoned Pen Press.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*, 9: pp.237–249.

Mori, H. 森英俊 (Ed.) (1998). *Sekai Misuteri Sakka Jiten: Honkaku-ha Hen*『世界ミステリ作家事典：本格派篇』[*Encyclopedia of Mystery Writers in the World*]. Tokyo: Tosho Kankō-kai 図書刊行会.

Mosteller, F. & Wallace, D. L. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, 58(303): pp.169–233.

Murakami, M. 村上征勝 (2004). *Shēkusupia ha Dare desuka? Keiryō Bunkengaku no Sekai*『シェークスピアは誰ですか?計量文献学の世界』[*Who is Shakespeare? An Introduction to Bibliometrics*]. Tokyo: Bunshun Shinsho 文春新書.

Murakami, M. 村上征勝, Jin, M. 金明哲, Tsuchiyama, G. 土山玄 and Uesaka, A. 上阪彩香 (2016). *Keiryō Bunkengaku no Shatei*『計量文献学の射程』[*The Scope of Research in Bibliometrics*]. Tokyo: Bensei Shuppan 勉誠出版株式会社.

Murakami, A, Thompson, P., Hunston, S. and Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora* Vol. 12(2): pp.243–277.

Nakamura, J. & Sinclair, J. (1995). The world of woman in the Bank of English: Internal criteria for the classification of corpora. *Literary and Linguistic Computing*, 10: pp.99–110.

Nakamura, T. 中村妙子 (1986). *Agasa Kurisuti no Shinjitsu*『アガサ・クリスティーの真実』[*Unveiling Agatha Christie*]. Tokyo: Shinkyō Shuppansha 新教出版社.

Özgür, A. (2004) *Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization.* Master thesis, Boğaziçi University: Istanbul Turkey.

Saito, T. 齋藤俊雄, Akano, I. 赤野一郎 and Nakamura, J. 中村純作 (2005). *Eigo Kōpasu Gengogaku: Kiso to Jissen Kaitei Shinpan*『英語コーパス言語学 : 基礎と実践 改訂新版』

[*English Corpus Linguistics: Basics and Practice, Revised Edition*]. Tokyo: Kenkyūsha Shuppan 研究社出版.

Sinclair, J. and Jones, S. (1974). English lexical collocations: A study in computational linguistics. In Teubert, W. and Krishnamurthy, R. (Eds.) (2007). *Corpus Linguistics: Critical Concepts in Linguistics*, Vol.1 (pp.223–269). London: Routledge.

Tabata, T. (1994). Dickens's Narrative Style: A Statistical Approach to Chronological Variation. *Revue Informatique et Statistique dans les Sciences Humaines*, 30: pp.165–182.

Tabata, T. (1995). Narrative Style and the Frequencies of Very Common Words: A Corpus-Based Approach to Dickens's First Person and Third Person Narratives. *English Corpus Studies*, 2: pp.91–109.

Tabata, T. (2014). Stylometry of Dickens's Language. In Arthur, P. and Bode, K. (Eds.) (2014) *Advancing Digital Humanities: Research, Methods, Theories* (pp.28–53). London: Palgrave Macmillan.

Takeuchi, O. 竹内理 and Mizumoto, A. 水本篤 (2014). *Gaikokugo Kyōiku Kenkyū Handobukku: Kenkyū Shuhō no Yori Yoi Rikai no tameni*『外国語教育研究ハンドブック：研究手法のより良い理解のために』[*The Handbook of Research in Foreign Language Learning and Teaching*]. Tokyo: Shōhakusha 松柏社.

Tsuchimura, N. (2017). How Did Agatha Christie Describe Women?: Behaviour of *She* in Christie's Novels. Poster presented at Digital Humanities 2017 (DH2017), McGill University, Montreal, Canada.

Tsuchimura, N. 土村成美 (2017). Topikku Moderu ni yoru Agasa Kurisuti Sakuhin no Keiryōteki Bunseki「トピックモデルによるアガサ・クリスティ作品の計量的分析」[Topic Modeling of Agatha Christie's Works]. *Jinmonkon 2017 Ronbunshū*『じんもんこん 2017 論文集』[*Proceedings of the Computers and the Humanities Symposium 2017*]: pp.177–182.

Tsuchimura, N. 土村成美 (2017). Agatha Christie Sakuhin no Keiryō Buntai Bunseki「Agatha Christie 作品の計量文体分析」[Stylistic Analysis of Works by Agatha Christie].

Tōkei Sūri Kenkyūsho Kyōdō Kenkyū Ripōto 386 Tekusuto Mainingu to Dejitaru Hyūmanitīzu『統計数理研究所共同研究リポート 386 テクストマイニングとデジタルヒューマニティーズ』[*The Institute of Statistical Mathematics Cooperative Research Report 386 Text Mining and Digital Humanities*]: pp.107–119.

Tsuchimura, N. 土村成美 (2017). Random Forests wo Siyō Shita Agatha Christie Sakuhin no Buntai Bunseki: Dorothy Sayers to Hikaku shite「RandomForests を使用した Agatha Christie 作品の文体分析：Dorothy Sayers と比較して」[A Stylometric Approach to Agatha Christie's Works Using Random Forests: In Comparison with Dorothy Sayers]. *Gengo Bunka Kyōdō Kenkyū Purojekuto Tekusuto Mainingu to Dejitaru Hyūmanitīzu 2016*『言語文化共同研究プロジェクト テクストマイニングとデジタルヒューマニティーズ 2016』[*Collaborative Research Projects in Language and Culture 2016 Text Mining and Digital Humanities*]: pp.43–54.

Tsuchimura, N. 土村成美 (2018). Igirisu ni okeru Misuterī Ougon Jidai no Josei Sakka Sakuhin no Hikaku Bunseki: Christie, Sayers to Allingham「イギリスにおけるミステリー黄金時代の女性作家作品の比較分析–Christie, Sayers と Allingham–」[An Analysis of Female Writers' Works during Britain's Golden Age of Crime Fiction: Christie, Sayers, and Allingham]. *Gengo Bunka Kyōdō Kenkyū Purojekuto Tekusuto Mainingu to Dejitaru Hyūmanitīzu 2017*『言語文化共同研究プロジェクト テクストマイニングとデジタルヒューマニティーズ 2017』[*Collaborative Research Projects in Language and Culture 2017 Text Mining and Digital Humanities*]: pp.19–29.

Uesaka, A. 上阪彩香 and Murakami, M. 村上征勝 (2014). Saikaku Ikōshū no Chosha ni kansuru Tōkei Bunseki: Hōjō Dansui no Ukiyozōshi tono Buntai Hikaku「西鶴遺稿集の著者に関する統計分析–北条団水の浮世草子との文体比較–」[A Statistical Analysis for the Authorship of Saikaku's Posthumous Works]. *Jinbun Kagaku to Konpyūta Shinpojiumu Ronbunshū*『人文科学とコンピュータシンポジウム論文集』[*Proceedings of the Computers and the Humanities Symposium 2014*], 2014(3): pp.113–118.

Uesaka, A. 上阪彩香 (2015). Janru no Kanten kara Saikaku Sakuhin no Bunshō no Sūryō Bunseki「ジャンルの観点から西鶴作品の文章の数量分析」[Quantitative Comparative Analysis for Saikaku's Works]. *Jinbun Kagaku to Konpyūta Shinpojiumu Ronbunshū*『人

文科学とコンピュータシンポジウム論文集』[*Proceedings of the Computers and the Humanities Symposium 2015*]: pp.15–22.