Bauhaus Universität Weimar
Faculty of Media
Degree Programme: M.Sc. Digital Engineering

COHERENCE BASED TEXT QUALITY IN SEARCH-SUPPORTED WRITING

Master's Thesis

Bibek Khadayat
Born: 28 Feb. 1993. In Dadeldhura, Nepal

Matriculation Number:
119505

1 Referee: Prof. Dr. Benno Stein
2 Referee: Prof. Dr. Ing. Volker Rodehorst

Supervised By:
Dr. Michael Völske

Dr. Magdalena Wolska

Submission Date: .....

Bauhaus Universität Weimar
Faculty of Media
Degree Programme: Msc. Digital Engineering

COHERENCE BASED TEXT QUALITY IN SEARCH - SUPPORTED WRITING

Master's Thesis

Bibek Khadayat
Born: 28 Feb. 1993, In Dadeldhura, Nepal

Matriculation Number
119505

1 Referee: Prof. Dr. Benno Stein
2 Referee: Prof. Dr. Ing. Volker Rodehorst

Supervised By:
Dr. Michael Völske
Dr. Magdalena Wolska

Submission date: .....

Declaration

I, Bibek Khadayat, hereby declare that the work presented here is genuine work done by myself and the work has not been submitted or published in the same or in a similar version elsewhere for the requirement of a degree program to any another Examination Board.

Any literature, data or work done by others are cited in this dissertation and have been given due acknowledgement and listed in the reference section.

.....	.....
Place, Date	**Bibek Khadayat**
	Signature

Abstract

The goal of this thesis is to investigate automatic measures for the identification of text quality patterns in search-supported essay writing. Text quality in terms of organization, coherence, cohesion, readability, or vocabulary typically varies throughout a text, and automatically finding lower-quality paragraphs can help streamline the workflow of editors. Coherence measures like Type-Token ratio, cosine similarity, or entity grid investigate whether the ideas, concepts, or themes in the text link together smoothly in the text flow. In this work, we compare a selection of established text coherence measures from the computational linguistics literature and apply them to a dataset of 150 long essays with detailed revision history. These essays were written in an experimental environment which recorded every revision authors made while developing their texts. This allows us to investigate not only how text quality develops across the paragraphs of the finished text, but also over time as the text is written.

" A paragraph is Coherent when the reader can move easily from one sentence to the next and read the paragraph as an integrated whole, rather than a series of separate sentences. "

James McNab McCrimmon,
Writing With a Purpose

Acknowledgements

It is a matter of immense pleasure that I got an opportunity to do this Master's thesis under the Webis group Weimar. I would like to express my sincere gratitude to Prof. Dr. Benno Stein for allowing me to write the thesis with the Webis group. I am immensely grateful to Prof. Dr. Volker Rodehorst for accepting my request for a second referee.

I would like to acknowledge my indebtedness to my supervisors Dr. Michael Völske and Dr. Magdalena Wolska for their constant support, guidance, inspiration and valuable suggestions and feedback that helped me to the completion of this thesis.

I am thankful to Baijanath Baba and Ghatal Baba for continuously showering blessings on me over time.

Last but not the least, I must express my sincere gratitude to my Mom, Dad, Uncle and all my family members along with my friends for their support, continuous encouragement, and blessing that led to the successful completion of this research work.

Bibek Khadayat
November, 2021

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

INTRODUCTION

On a regular basis, we have to deal with different kinds of texts like newspaper articles, essays, novels, and books. We encounter both poor-quality and high-quality texts or articles. Many of these texts are written with the support of a search engine. In search supported writing, writers use the search engine to research their topic and collect information and data.And, they use the collected information and data in their writing or articles. The quality of the resulting text is influenced by many factors like spelling, vocabulary, grammar, organization, readability, coherence, cohesion, etc. How text quality is perceived depends on the language skills of the reader. Normally, we judge or measure text quality spontaneously on the basis of spelling, vocabulary, grammar, readability, and cohesion. This thesis explores how text quality can be computed automatically based on coherence.

There are several fields or situations where we would like to measure text quality automatically. Few of them are:

Academic fields (Crossley, DeFore, Kyle, Dai, and D. S. McNamara 2013): Writing assessment is one area which can immensely benefit from this kind of system or measurement. Teachers have to grade students' essays and give feedback. Currently, there are many web tools or such systems which are in use or being researched. Some of them are commercially developed for this purpose.

Web search recommendation: In today's world, a huge number of articles and blogs appear in search engine indexes. People

query about different articles or blogs in the search engine, and it returns numbers or relevant articles which may not be coherent with the desired query or topics. Therefore, it would be very helpful for users if top results which are recommended are cohesive and coherent with the query.

Automatic summarization (Parveen and Mohsen Mesgar 2016): In the 21st century, everyone is so busy that people do not have enough time to read e.g., entire news articles. Automatic coherence measures can help generate coherent summaries.

"Coherence exists in a sequence of words, sentences, and paragraphs in which the reader can perceive connections and understand the structure and therefore the meaning as he reads." (Brostoff 1981)

Anita Brostoff

Coherence is a measure to evaluate whether the sentences are logically consistent throughout a text (Cui, Y. Li, Zhanga, and Z. Zhang 2017). Text coherence defines the logical and conceptual connection that a reader or listener distinguishes in a text (See section 2.1 for more details on Coherence). It is believed that Coherence is one of the essential qualities for informal writing as well as academic writing. Lack of coherence hampers the readability of text (Enago Academy 2020). By contrast, Cohesion means the presence of explicit words or hints in the text that allow or help the reader to make connections or to relate between the ideas in the text (Crossley and D. S. McNamara 2011). Cohesion is the grammatical and lexical linking within a text or sentences that holds a text together to give it meaning (Indiafreenotes 2020).

Table 1.1: Examples of Coherence and Cohesion

Coherence	**Cohesion**
Nepal is a small and beautiful country. Many tourists visit Nepal because of its beauty and nature. White mountains, green forests, wildlife reserves etc., lure people from different countries.	**Nepal** is a small country. **Nepal** is located in between India and China. **Its** area is 147516 sq. kilometers.

Table 1.1 illustrates an example contrasting coherence and cohesion. Sentences in the cohesion example are connected by specific words like “Nepal,” and “Its.” Whereas, sentences in the coherence example are linked by some idea or information like the nature and beauty of Nepal. The categories which show the cohesion relation are: Repeated words, reference, substitution, ellipsis, conjunction and lexical cohesion (Halliday and Hasan 1976a). Cohesion sticks the text with repeated words, reference, substitution, ellipsis, and conjunction. Cohesion is a lexical and grammatical connection within the text, it is not conceptual connection within the text. In this work, we try to evaluate text quality on the basis of the connection and linkage to the same idea or concept throughout the text. (The measures used are explained in detail in section 2.1)

This thesis is a step towards presenting an automatic method to calculate text quality based on coherence.

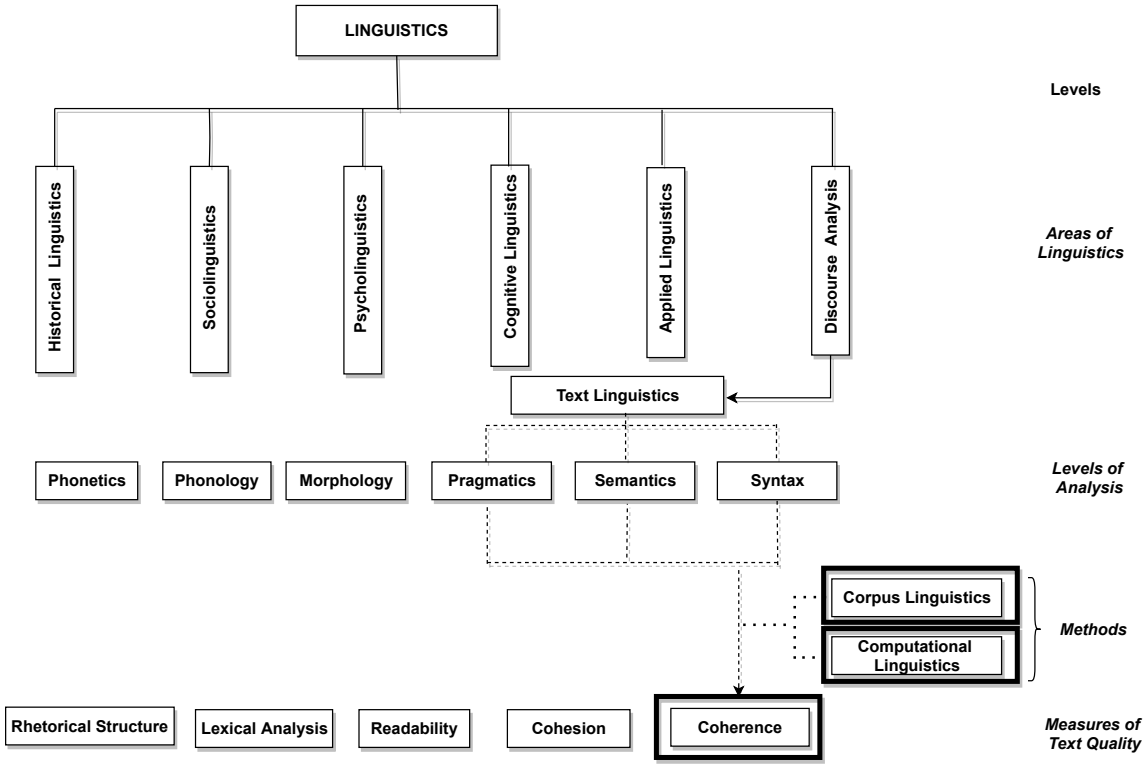


Figure 1.1: Different Levels: Areas of linguistics, levels of analysis, methods, and measures of text quality used in this thesis

Ultimately, all our initiatives and methods are based in Linguistics. Linguistics is a broad field and has different areas

and different levels of analysis in each area. All levels of analysis and areas of linguistics used in this thesis are illustrated in Figure 1.1. It shows Linguistics flow through the thesis, which uses computational linguistics and corpus linguistics as methods to calculate text quality, and measures coherence mainly using three different levels of analysis.

Every measure used in this thesis is based in Text Linguistics, which is hence the main focus of Figure 1.1.

This thesis focuses on calculating text quality using coherence, which is a measure of text linguistics, and a hot topic of research. It is difficult to automatically capture coherence. The problem of automatic coherence assessment was first proposed in the 1980s (Cui, Y. Li, Zhanga, and Z. Zhang 2017). Since then, different analysis methods and techniques have been developed. In this thesis, we have implemented an entity based approach to compute coherence (see chapter 3).

This thesis focuses, firstly, on local coherence, which evaluates logical connections at the level of paragraphs. We use the entity grid approach to evaluate text coherence at the paragraph level (Barzilay and Lapata 2008). We generate the entity grid of a text with the help of a parser, and calculate cosine similarity across entity transitions (see section 3.1).

Secondly, we apply our approach to the Webis Text Reuse Corpus 2012 (Potthast, Hagen, Völske, Gomoll, and Stein 2012) which is composed of 150 long essays with detailed revisions, written by 12 different writers (see section 2.3).

Thirdly, we investigate patterns in text quality not only across the paragraphs of the final revision, but also over time as the text is written by a writer.

The following three research questions frame this thesis:

- What are the different types of editing techniques in search-supported writing?
- How do different types of editing affect Coherence, Type-Token ratio, and Readability?

- How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

The writing strategies studied in the context of the final research question refer to the build-up and boil-down writing styles identified in previous work (Potthast, Hagen, Völske, and Stein 2013a). They are explained in detail in section 2.3.

The remainder of this thesis is structured as follows: Chapter 2 provides more details about text quality and the different ways it can be measured, and elaborates on related work and datasets. In Chapter 3, we explain the approaches we used to measure text quality in this work, and explain how we have implemented them. Chapter 4 analyzes the datasets we used, and details the setup and results of our experiments. Last but not the least, Chapter 5 concludes the thesis and discusses possible future work.

Chapter 2

BACKGROUND

Linguists, psychologists, and philosophers have long attempted to address the issues and scientifically study about language with new and modern techniques. Linguistics is the science which focuses exclusively on languages, their properties, and the evolution of languages over time, whereas text linguistic is one of the independent and young sub-disciplines of linguistics which has been in progress since 1960 and which studies text: its formation and perception, text structure, different methods of text analysis, etc.

The beginning of text linguistics is based on the idea of focusing on discourse alone proposed by *Zellig S. Harris* (Ashurova and Galiyeva 2012):

"The Analysis of the occurrence of an element in the text is applied only in respect to the text alone that is, in respect to the other elements in the same text, and not in respect to anything else in the languages." (Harris 1952)

Discourse is a unit of single or more sentences that is used to exchange ideas and thoughts. Discourse analysis is concerned with what a text is saying, but also with how it is saying it, and what the patterns are (Harris 1952). It investigates real data both at the sentence-level and beyond sentence boundaries and develops models of semantic and pragmatic phenomena. Text linguistics deals with the description and analysis of written text – its syntactic, semantic, and discourse properties – in the contexts of communication (Zienkowski, Östman, and verschueren 2011). Text is construed as meaningful communication which is used to express information, emotions,

and feelings in a written format. Quality of a text can be measured based on various criteria such as cohesion, coherence, readability, among others. In the following sections we briefly introduce various aspects of measuring text quality in more detail.

2.1 Text Quality

Before directly diving into text quality, let's start with what actually is a text? The word 'text' comes from Latin word '*texere*' which means 'to weave'. From the Latin meaning we can define text as structure "woven" out of words, or, more generally, signs. It is the arrangement of words, phrases, sentences with certain intentions in a meaningful way. For Teun A. van Dijk a text is "the abstract theoretical construct underlying discourse" (Dijk 1976) or "a sentence sequence with macrostructure". According to Halliday and Hasan "the word 'text' is used in linguistics to refer to any passage, spoken, or written, of whatever length, that does form a unified whole" (Halliday and Hasan 1976b).

"A text is a passage of discourse which is coherent in these two regards: it is coherent with respect to the context or situation, and therefore consistent in register; and it is coherent with respect to itself, and therefore cohesive."

M. A. K Halliday and Ruqaiya Hasan
Cohesion In English

Text quality is a vast term to define. Many factors affect the definition of text quality. Text quality does not involve a single measure. It is a combination of diverse measures including spelling, grammar, organization, coherence, cohesion, readability, rhetorical structure, lexical analysis, and many more. Profession and audience also affect rating or definition of text quality. Audience can vary with respect to age, educational levels, technical expertise, and those without cognitive disabilities (Louis 2013). Depending on the audience, same text may be rated with different text quality level. While surface-level text quality aspects include grammar, spelling, vocabulary, and even length of text, teachers assess student's

essay writing skills on the basis of a larger set of properties such as writing technique, organization, vocabulary, grammar, idea development, sentence fluency, etc. Some of those properties reflect discourse phenomena of coherence, cohesion, and readability which we discuss in more detail.

Factors That Affect Text Quality

Cohesion: A text is cohesive if it is well-connected (Mathis, Kanojia, Patel, Agrawal, and Bhattacharyya 2018). Cohesive devices establish links between sentences on the basis of grammar, words, or phrasing which give meaning to the text. Cohesion ties together the words, clauses, and sentences in text at a semantic level (D. S. McNamara, Graesser, McCarthy, and Z. Cai 2014).

Table 2.1: Example: Cohesion where **bold** and *italics* words are used to link sentences

	Cohesion
1	My favourite color is *blue*. *Blue* sports cars looks so beautiful. It goes **Very fast**. Driving in **this way** is dangerous and can cause many *car crash*. I had a *car accident* once and **broke my leg**. I was very sad because I had to miss holidays in Europe because of **the injury**.

Cohesion idea can be summed up into few distinct categories: Repeated words, reference, substitution, ellipsis, conjunction (Halliday and Hasan 1976a).
Repeated words: Repeated word or synonyms word is used to calculate cohesion of a text. In another word, we can say, a text is called cohesion if there are repeated words and synonym words in the text. In the above example, the word **blue** is repeated, and **car crashes** and **car accident** are synonyms which helps to make a text cohesive.
Reference: Reference words are the words which refer or assign to the another word which is already mentioned in the text. Most commonly pronouns comes in uses in this cases. In the above example, **It** refer to the blue cars.
Substitution: Substitution means to use one or more words to refer or substitute for one or more words which is already

mentioned in the text. Grammatically, it is similar to reference word. Reference is a relation on the semantic, whereas substitution is a relation on the Lexico grammatical, the level of grammar and vocabulary (Halliday and Hasan 1976c). In the above example, the clause **this way** substitute the clause **very fast**.

Ellipsis: Ellipsis is very similar to substitution, ellipsis is 'substitution by zero' (M. A. K Halliday and Ruqaiya Hasan 1976). In Oxford Dictionary ellipsis is "The omission (deletion) of one or more words in order to avoid repetition."

Conjunction: Conjunction elements are cohesive not in themselves but indirectly, by virtue of their specific meanings; they are not primarily devices for reaching out into the preceding text, but they express certain meaning which presuppose the presence of other components in the discourse (Halliday and Hasan 1976d). The three different types of conjunction are:

- Adverbs:
  Simple adverbs, eg: for, and, then, next
  Compound adverbs in -ly, eg: accordingly, subsequently, actually
  Compound adverbs in there- and where-, eg: therefore, thereupon, whereat

- Other compound adverbs, eg: furthermore, nevertheless, anyway, instead, beside
  Prepositional phrase, eg: on the contrary, as a result, in addition

- Prepositional expressions with that or other reference item, the latter being
  Optional, eg: as a result of that, instead of that, in addition to that
  Obligatory, eg: in spite of that, because of that (Halliday and Hasan 1976d)

Coherence: It is believed that coherence began in nineteenth century by the Alexander Bain's first rule of paragraph: " The bearing of each sentence upon what precedes shall be explicit and unmistakable " (Bamberg 1965). The first systematic

formulation of paragraph theory appeared in march 1866 in Alexander Bain's English Composition and Rhetoric, a manual (Paul C. Rodgers 1965).
" A paragraph is Coherent when the reader can move easily from one sentence to the next and read the paragraph as an integrated whole, rather than a series of separate sentences."

James McNab McCrimmon,
Writing With a Purpose

Coherence is the measure of connection between the text by the sense or idea which makes reader easy to understand, it is a semantic property of text. It is measured across the sentences within the text, by themselves as well as with other sentences in the text (Van Dijk 1980). Coherence refer to how easy a text is to understand or to get theme or idea of a text by a reader. Paragraphs and sentences need to be logically or conceptually linked to each other to be a coherent text (Enquist, Oates, and Francis 2017). This means that, the sentences and part of paragraph need to be linked with each other in a logical way so that the reader easily understands the development of ideas or concept and argument. Coherence is generally accepted as a, *sine qua non* in written discourse. Text which lacks coherence may certainly fail to communicate its intended message to an audience (Bamberg 1965).

Table 2.2: Example: Coherence[01].

Coherence
My favourite colour is blue. I'm calm and relaxed. In the summer I lie on the grass and look up.

A text can be called coherent, if all parts of the text is related to each other and ensure the bigger picture clear. Coherence also mean "clarity of expression". Without coherence, texts are difficult to read and understand (Enago Academy 2020).

[11] Example is taken from the given site: `http://gordonscruton.blogspot.com/2011/08/what-is-cohesion-coherence-cambridge.html` and paper (Mathis, Kanojia, Patel, Agrawal, and Bhattacharyya 2018)

The three separate operations writers must follow to build a coherent discourse, and failure at any one or more of these operation leads towards incoherent text or writing are: *(1)* writers must make or sustain logical relationships, *(2)* they must put together a series of relationship in a consistent way or *(3)* they must reveal relationship adequately to the reader (Brostoff 1981). Text coherence helps to explain what role each clause plays with regard to the whole (J. Li and Hovy 2014). Coherence is a sequence of sentences having *textual* structure, where textual behaviour of sequence of sentences expressed by a discourse is the semantic property of coherence (Van Dijk 1980). There are two kinds of coherence, viz. *Local* and *Global* coherence. Local coherence is the relation between two adjacent sentences of a textual sequence. Whereas, Global coherence is a relationship in whole set of sentences. Global coherence is also called in more intuitive terms as "theme", "idea", "upshot" or "gist" of a discourse (Van Dijk 1980). Two sentences are connected if the idea or concept are related in some possible world. Often, these relations are conditional in nature (Van Dijk 1980).

Several researches on coherence are done till date inspired on: (Grosz and Sidner 1986), which defines three separate components of discourse structure: the structure of sequence of utterances, i.e., linguistic structure; a structure of purpose, i.e., Intention structure; and the state of focus of attention, i.e., Attention state; (Mann and Thompson 1988), which define different relations that lead clause interdependence, ordering and evolve to text tree structure and also offer different features which are used in discourse studies; (Grosz, Loshi, and Weinstein 1995), which is an initial attempt to represent a relationship among focus of attention, choice of referring expression, and recognized coherence of utterances within discourse segment; and (Van Dijk 1980), which present functional relations between sentences and between speech of acts in coherent discourse. It discusses the relation of semantic and pragmatic. This thesis is influenced by some of above mentioned research and (Barzilay and Lapata 2008).

Readability: Readability is an ease with which we audience read and understand the written text. Readability is defined as how easily written materials can be read and understood, that depends on several factors including the average length of sentence, the number of new words contained, and the grammatical complexity of the language used in a passage (Richards, J. Platt, and H. Platt 1992). Harry McLaghlin, creator of SMOG Readability formula, define readability as the degree to which a given categories of audience find certain reading matter compelling and comprehensible (Zamanian and Heydari 2012). What I personally believe is, human judgement of readability is totally dependent upon the skills set of particular audience. There are many research done on readability and different formulas and technique are derived to calculate readability of a text. Readability formulas or indices are the mean which is used to estimate how easy a text is to read. There are over 200 formulas derived so far to estimate readability of text (Štajner, Evans, Orasan, and Mitkov 2012). Some of the popular formulas are:

Flesch Reading Ease Readability Formula, which rates texts on a 100 point scale. Higher the scale, easier it is to understand the document. This is one of the most widely recognized and reliable measure of text readability. It calculate readability on the basis of average sentence length and average word length. Formula is

$$206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Where, ASL is the Average Sentence Length which is number of word divided by number of sentence, and ASW is the average of syllables per word which is number of syllables divided by the number of word.

Dale-Chall Formula, which is derived to solve certain shortcomings of Flesch reading Ease formula. This use word length to determine how difficult a word is for audience to understand. The percentage of difficult word is calculated and used, where difficult word are the word which are not in the list

[11] Given Score is taken from the given site:https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

Table 2.3: Flesch Reading Ease score[01]

Reading Ease Score	School Level (US)	Description
100-90	5th grade	very easy to read
90-80	6th grade	easy to read
80–70	7th grade	Fairly easy to read.
70–60	8th,9th grade	Plain English.
60–50	10th-12th grade	Fairly difficult to read.
50–30	College	Difficult to read.
30–10	College graduate	Very difficult to read.
10–0	Professional	Extremely difficult to read.

of 3000 familiar words list. This formula is best suited for rating text readability from four to ten. It depend on the percentage of difficult words and average sentence length. The formula for calculating score of Dale-Chall Readability score is :

$$0.1579\left(\frac{Difficult\ Words}{Words}\times 100\right)+0.0496\left(\frac{Words}{Sentences}\right)$$

Table 2.4: Dale-Chall score[02]

Score	Description
4.9 or lower	4th-grade student or lower
5.0–5.9	5th or 6th-grade student
6.0–6.9	7th or 8th-grade student
7.0–7.9	9th or 10th-grade student
8.0–8.9	11th or 12th-grade student
9.0–9.9	13th to 15th-grade (college) student

SMOG Index Readability Score SMOG stand for Simple measure of Gobbledygook, which is best for text of 30 sentences or more. G. Harry McLaughlin suggest to use this with formula for 30 sentences text or more doing following: count 10 sentences near beginning of text, 10 in the middle and 10 near the end of text. Count every word with three or more syllables. Square root the number and add three. This technique or

[12] Given Score is taken from the given site:`https://en.wikipedia.org/wiki/Dale-Chall_readability_formula`

formula is widely used in health sector. The formula for SMOG Index is

$$SMOG\ Grade = 3 + \sqrt{polysyllable\ count}$$

Computational Linguistics

Computational linguistics is the scientific discipline that concerned with the understanding of languages either in the written form or oral form in the computational perspective, and create artifacts that evidence the usefulness of process and creation of language. To an extend, language is the reflection of mind which provides insights into thinking and intelligence. Computational Linguistic is the medium that facilitates the interaction of our mind with machines (Schubert 2020). Computational linguistics is a discipline that automates the engineering and scientific study on human languages with the help of computers. Computational linguistics compute the mathematical properties of language with the help of computer and design and analyse natural language processing system. This branch is concern with the development and analysis of language which focuses on modelling the meaning of words, recognising the grammatical structure of sentence, assessment of semantic of a text, and so on. The main goal of computational linguistics is the formulation of grammatical and semantic structure of language and exploit statistical properties of language. In mid Nineteen, computational linguistics was evolved for solving issue of lexical formation and content, translation of one language to another, characterizing the question patterns, etc. Different fields of scientific research with the use of computational linguistics in a language are: checking syntax and parsing, semantic representation and interpretation, making connection between the sentences, extracting knowledge from the text, statistical analysis of the languages. Some of the application of computational linguistics are: Machine translation, document retrieval, knowledge extraction or summarizing, sentiments analysis, chat bots, etc. Main engineering part of computational

linguistics is Natural Language Processing which is used for machine translation, summarization, even our thesis computing text quality, question answering, and so forth. Natural language processing can be used for other scientific research of computational linguistics.

2.2 Related Work

A number of researches are done on text quality and the measure coherence. In this part, we try to relate our task and research with the prior work. We categorize some of the related papers which is illustrated in figure 2.1.

A research was proposed by (Grosz and Sidner 1986), which explains that utterances chain provides a clue for resolution of coherence and discourse structure. This theory says that, if a chain of utterances in a discourse can be determined, there is a tendency for related statement in a discourse. They proposed a theory that structure of any discourse is composite of three components: Linguistic structure, which is a sequence of utterances that make discourse. It has a relationship which holds discourse segment together; Intentional structure, which captures discourse segment and discourse segment purpose and their relationship expressed in each linguistic segment; Attentional state, which record the objects, properties, and relation of the discourse. Change in attentional state is model by set of focus spaces, i.e., transition rule that explains the condition for adding and deleting spaces. Each discourse segment express both local coherence, i.e., coherence among utterances and global coherence, i.e., coherence with other segments in discourse which are the components of attentional state.

(Mann and Thompson 1988) proposed descriptive theory of the organization of the text, which presents three component of Rhetorical structure theory: the predominance of nucleus structure patterns, hierarchic structure, and the communicative behaviour of text. RST has been used as a general way to figure out relations among clauses in a text, used as analytical tool, useful in analysing narrative discourse, and also provides a framework for investigating Relational Propositions. RST is useful in the study of text coherence because coherence depends on Relational Propositions. Implicit propositions which emerge from the combination of the clause which hang together is Relational Propositions (Mann and Thompson 1986).

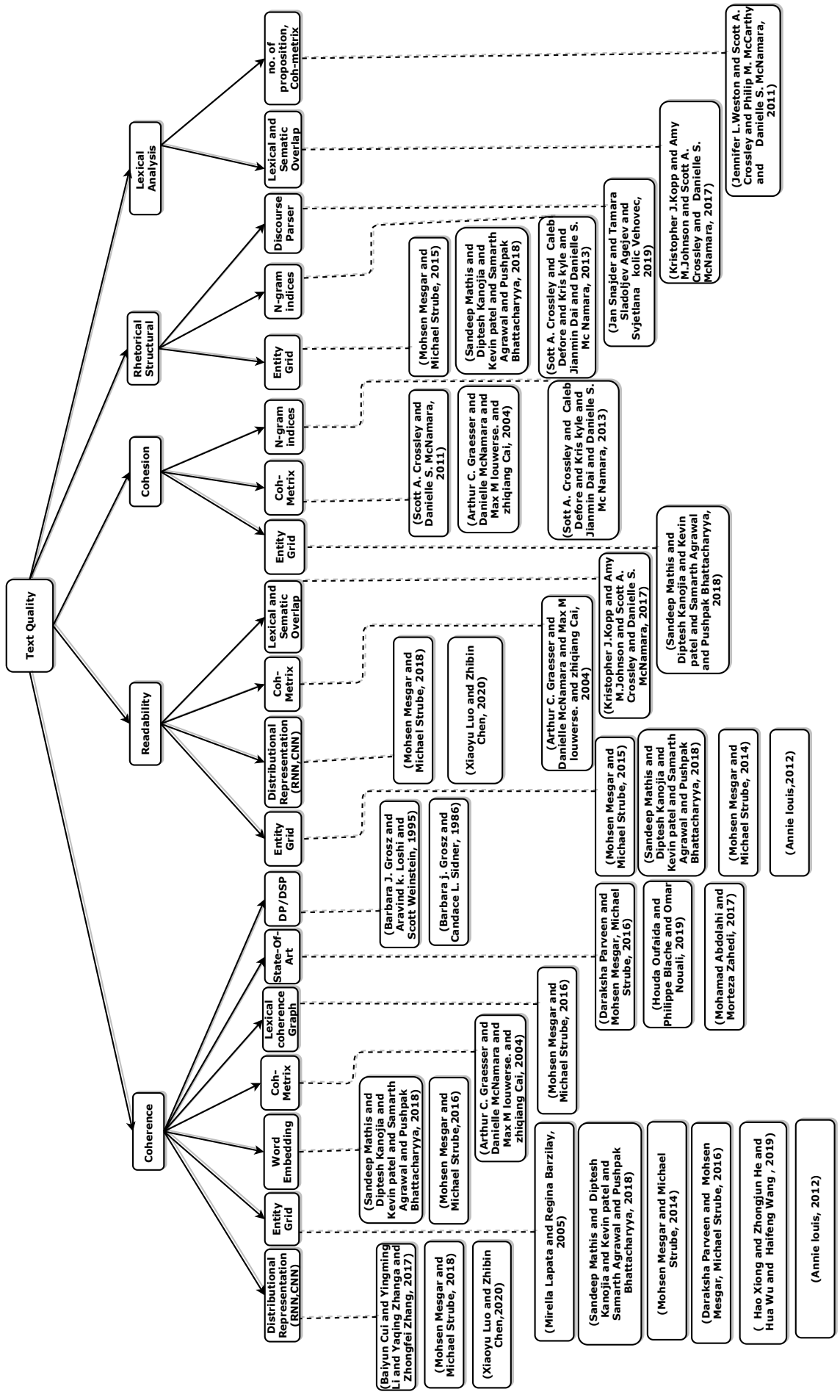


Figure 2.1: Categorization of prior work

(Grosz, Loshi, and Weinstein 1995) researched on the relationship among local coherence of utterances within a discourse segment and choice of referring expression, which believes that difference in coherence is made by compatibility between centering properties of an utterance and different types of referring expressions. Immediate focusing in discourse is related to identifying the entity that utterance is concerned. Sidner proposed a algorithm to track the immediate focus in discourse and rule how it is used to identify referring expression, i.e., demonstrative noun phrase and pronouns. This paper is based on (Grosz and Sidner 1986) and focuses on local coherence and its relation to the attentional state.

(Van Dijk 1980) presented a problem of linguistic theory, the relationship between sentences and speech acts in coherent discourse. This paper investigated semantic nature and pragmatic relation of a text and the delimitation of semantics and pragmatics against each other, where they tried to distinguish semantic and pragmatic functions. He also presented an issue that some proposition and speech act chain cannot only be accounted for the coherence of discourse, but also categorized functionally. Functional property of a speech act is the role of a speech that acts with respect to another.

(Barzilay and Lapata 2008) team presented an entity-based approach to discourse which computes coherence automatically from raw data. They focused on local coherence, which is important to generate global coherence. They abstracted a text into a set of entity transition sequences which gives distributional, syntactic, and referential information about discourse entities, which helps to learn the properties of local coherent. Entity Grid is used to obtain the local entity transition. Newspaper article and accident reports were used as data sets for the experiment. And the result of the model is evaluated against human coherent judgment and text ordering task algorithm.

(Lapata and Barzilay 2005) investigated to evaluate coherence in machine-generated text. They created a fully automated, linguistically rich model to calculate local coherence which

compares the result with human judgments. This research is focused on text organization at the level of sentences with sentence transitions, i.e., local coherence. The attention of this work is quantitative model of local coherence. Their model is based on two classes, i.e, *syntactic*, which characterizes how a same entity in different syntactic position are spread across sentences; and *semantic*, which calculates local coherence as a degree of connectivity across sentences. Experiment is done with different similarity measures: word-based, distributional, and taxonomy based which proves that both syntactic and semantic are best.

(Graesser, D. McNamara, louwerse., and Cai 2004) team developed Coh-Metrix which helps to analyze text over 200 measures of text, discourse which helps to investigate coherence, cohesion, readability and many more lexical features. This tool uses different components of computational linguistics. They state that explicit features, words, phrases, or sentences guide the audience to catch ideas in the text, connect ideas with other ideas in the text.

(Wachsmuth, AL-khatib, and Stein 2016) worked on argument mining to measure the argumentative structure of text. They developed a model which scores an essay's organization and its argument strength. This research is held of premises and conclusion, which are types of argumentative discourse units. The major application of this model is automatic grading. This research was done on International Corpus of learner English in which they first mine argumentative discourse units and then analyze their argumentative structure. Essay scoring is for Organization, Thesis Clarity, Prompt Adherence and Argument Strength.

(Mathis, Kanojia, Patel, Agrawal, and Bhattacharyya 2018) stated that predicting human ratings of text quality are better handled by gaze behaviour. For this research, the base measures for text quality are- organization, coherence, and cohesion. First, the text quality is calculated and then gaze behaviour is captured which helps to predict each of the properties. They believe that if a reader can understand the

text, they give better gaze behaviour. Gaze behaviour captures the effort needed by the reader to understand text.

(L.Weston, Crossley, McCarthy, and D. S. McNamara 2011) lighted on the relation between linguistic features of free writing and human assessment of free writing quality. This research was done on the previous model on linguistic free writing using the number of propositions instead of number of words. Because number of propositions leads toward number of ideas and which help to calculate the coherence. They believed that the number of words plays vital role in text quality of human judgement, because humans think that longer the text, higher the text quality. Therefore, here they proposed a model for a number of propositions. The number of proposition is measured by using part-of-speech tagger. This work is experimented in Prompt-based free writes which were collected from high school students in New York.

(Snajder, Agejev, and Vehovec 2019) developed a model that measure coherence automatically based on comparing rhetorical structure of summaries written by college student against expert summaries. The model is evaluated by comparing with coh-metrix index. And, they believed that RS scores correlate with both cohesion and coherence.

(Liang, feng, Liu, Y. Li, and X. Zhang 2018) experiment was done on short text using word embedding, which uses both local and global word embedding and helps to measure semantic relatedness information between a word that can be further used to strengthen the semantic coherence of topics. First, global word embedding is trained from a large corpus and obtains local embedding with negative sampling, which helps global word embedding to encode general semantic and syntactic information of words, whereas local word embedding contains word context information. Secondly, GPU model is employed, which changes the statistics of semantically related word and finally, a coherent topic is obtained by using maximum posterior estimation.

(Abdolahi and Zahedi 2017) presented a statistical model that uses the word2vec approach which calculates local and global

coherence automatically. This model relies on vector generated from word2vec combined with cohesive LD-n-gram perplexity to measure coherence.

(Crossley, DeFore, Kyle, Dai, and D. S. McNamara 2013) team research described an n-gram approach to automatically calculate text quality. They developed n-gram indices that examine rhetorical, syntactic, grammatical, and cohesion features of a paragraph and entire text. They believe that n-gram indices is the better approach to automatically calculate text writing quality. This study analysed the potential for n-gram indices related to different paragraph types. If n-gram is identified in quality text and human judgement, then those n-gram can be used to assign text quality.

(J.Kopp, M.Johnson, Crossley, and D. S. McNamara 2017) presented algorithm which measures question quality that gives feedback to the questions generated by students in iSTART (an intelligent tutoring system that teaches reading strategies). They worked on a corpus of 4575 questions using four-level taxonomy, i.e., 1-very shallow to 4-very deep. First, calculate NLP indices, i.e., lexical features for each question, and then machine learning is used to predict question quality.

(Parveen and Mohsen Mesgar 2016) presented a graph-based approach to summarize scientific papers. This model first generates a coherence pattern and then combines coherence important information and non-redundancy information to generate the summary. Coherence is measured by calculating out degree of sentence in graph representation. However, it has been a disadvantage because it computes the graph representation for one sentence at a time which is not sufficient to obtain coherent summaries and also it is obtained based on sentence connectivity in the document. Coherent is obtained by using discourse entities which relate sentences and extract whole sentences which are related and connected to these entities. And compare with human written summaries to evaluate.

(Mesgar and Strube 2016) presented lexical coherence graphs which represent lexical relations viz. repetition, synonymy,

hyperonymy, meronymy, etc., among sentences. Frequency in the graph observe connectivity style of sentences. Coherence of text is encoded by a vector of those frequencies. Kneser-Ney smoothing is used to smooth the frequencies which help to improve the performance. They used word embedding to represent text into vector and use the vector to check the occurrence of semantic relations between words. After finding relation lexical coherence is modeled using graph.

(Xiaohua and H. Luo 2015) experimented with the variance of Latent Semantic Analysis and Probability latent Semantic Analysis on judging text quality using an automated essay scoring tool. LSA and PLSA models are applied to the essay score generated by an automated essay scoring tool to evaluate correlation between them.

(X. Luo and Chen 2020) proposed an recurrent neural network model to analysis English text quality. They used neural network method to find the relationship between the context and the word. Word vector which is generated from neural network is used as input, and then uses recurrent neural network to classify the text. Result from this paper tries to prove that RNN improves the performance comparing to traditional methods.

(Mesgar and Strube 2015) performed an experiment to compare the readability of Wall Street Journal articles. They presented graph-based features for measuring readability. They first generate entity graph representation and graph representation of rhetorical relations between a sentence and then merge the entity graph and rhetorical graph together. With this model, they first measured the readability and then ranked the text according to readability.

(Xiong, He, Wu, and Wang 2019) researched on using discourse context to improve the translation quality from the discourse perspective. First they translate the sentences and then train the model to produce discourse coherent text. For entire text coherent translation, they first translate each sentence independently, and then make some modifications to make it entire text coherent and fluent. To do this, they have first

generated preliminary translation of each sentence and then modified each translation satisfying the discourse coherence rule.

(Crossley and D. S. McNamara 2011) studied the importance of human evaluation of coherence in predicting text quality. Computational indices related to text structure, semantic coherence, lexical sophistication, and grammatical complexity are used to calculate human judgement of text coherence. They model human judgments of coherence in order to understand which feature is important for a coherent text. First, they used text organization for analysing coherence and secondly, they used many linguistics sophistication indices from coh-metrix along with text structure, semantic coherence, lexical sophistication, and grammar. They believed that text organization plays a vital role in text coherence and coherence is an important essay quality.

(Mesgar and Strube 2018) presented a model which captures the flow of semantic connections within adjacent sentences. Semantic relations are represented by vectors and capture the semantic state of words, which helps to relate two adjacent sentences. RNN is used to extract information that relates two adjacent sentences, and then, CNN is used to encode the pattern of semantic information which helps to represent coherence. And the model is evaluated on two-task readability assessment and essay scoring.

(Oufaida, Blache, and Nouali 2019) used discourse analysis to identify coherence features. Model use that coherence feature to generate coherent summaries. Genetic algorithm is used to order the sentence in order to generate better summary. Models identify positive features, such as original thematic, and which share the entities of adjacent sentences; and negative feature, such as redundancy. Then the positive feature are maximized and negative features are removed to generate coherent summary.

(Cui, Y. Li, Zhanga, and Z. Zhang 2017) used Convolutional neural network to capture coherence of text. Each sentences are represented in a distributional representation using word

embedding. Then the relations between sentences are extracted by computing similarities of distributional representations using CNN and these similarities are concatenated with their corresponding sentences vector to estimate the coherence.

(Louis 2012) proposed a model in which text quality is predicted on the basis of content that is discussed, sentence level grammaticality, discourse coherence, and writing style in three different genres: academic publication, news articles about science, and machine generated text. The main property of this research is to consider a particular audience level for analysis text quality metrics, i.e., readability and coherence.

2.3 Data and Datasets

In this research, we are using a dataset, Webis Text Reuse Corpus 2012 (Webis-TRC-12) (Potthast, Hagen, Völske, Gomoll, and Stein 2012) of 150 long essays which are written by the professional writer hired at the crowd sourcing platform oDesk, who searched the ClueWeb09 corpus for source material, and reuse text what they found and all these activities were meticulously recorded. Those 150 essays which was the writing task for writer were the topics used at the TREC Web Tracks 2009-2011. This datasets also contain detailed interaction logs that covers the construction of corpus as well as the history of the search for source (Potthast, Hagen, Völske, and Stein 2013b). During the phase of construction of this corpus research on topic and writing effort about topic were done more or less simultaneously (Hagen, Potthast, Völske, Gomoll, and Stein 2016). It also captures the activities how average user perform exploratory search. Two datasets are used to form the basis for constructing a corpus: a set of topic to write about, and a set of web pages to research about the given topic. If a writer want to research about a topic he or she has to search with the helps of ChatNoir search engine in ClueWeb09. ClueWeb09 is a static web search environment that consist of more than one billion documents. ChatNoir employ BM25F as the retrieval model that use interface from commercial search engines. Tracking of the search record and different activities during

research about a topic is done by search engine and saved in log file. Search engine helps to save all search interaction alongside the revisions of the actual text (Hagen, Potthast, Völske, Gomoll, and Stein 2016). A new revision for a topic is created whenever the writer or author stops editing or typing for 300ms in the online editor provided for writing. Final revisions of most of the essays are around 5000 words which are illustrated in Figure 2.2. Some of the essays are shorter because of difficulties in finding related documents in ClueWeb.

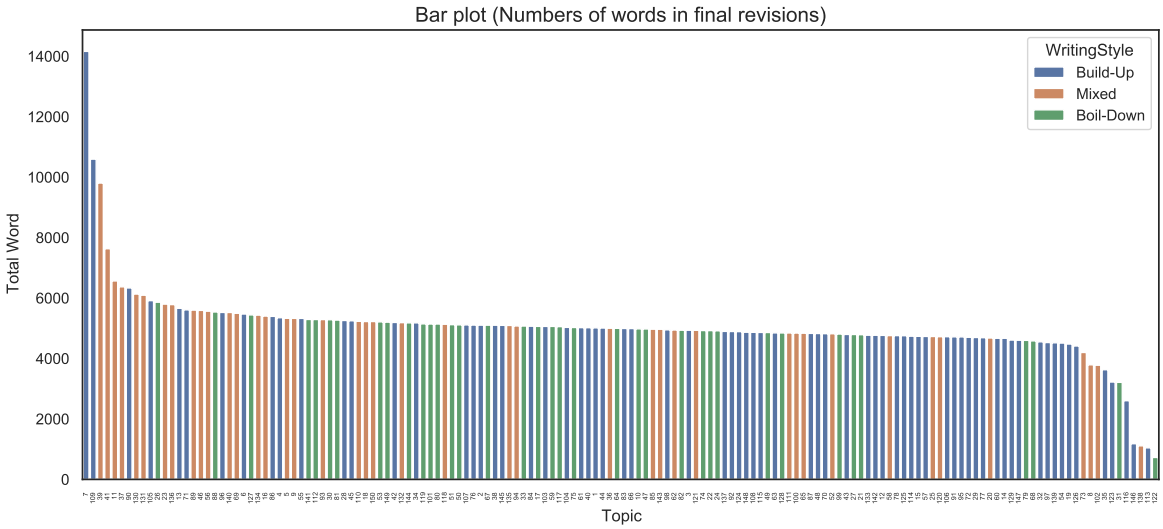


Figure 2.2: Numbers of word in final revisions of essays

Figure 2.2 also clearly illustrates that most of the final revision of 150 essays are around 5000 words. Figure 2.2 shows that writing style or writing strategy doesn't affect the length of final revision of any essays. All the writing strategies are equally distributed.

The dataset consists of 150 files for search log along with 150 essays. Search log file contains : queries along with the results, URL history and type, and revision numbers of text-writing interactions. Each interaction contains a timestamp and IP address which gives clue about different work station (Hagen, Potthast, Völske, Gomoll, and Stein 2016). Table in the Figure 2.3 clearly visualizes important statistics about the Writing session and the total duration to write an essay. Figure 2.3 shows that minimum number of days to write an essay is 1 and minimum working hours is 1.8, whereas maximum number

of days to write an essay is 56 days but working days for the respective essay is only 17 days and paused for 39 days. It also indicates that writer did not spend much time reading the viewed document. One reason could be that writer just copy paste text from viewed document and read while editing in the writing editor (Hagen, Potthast, Völske, Gomoll, and Stein 2016).

	Min	Q1	Mdn	Avg	Q3	Max	Sum
Writing sessions							
– per essay	11.0	28.0	42.0	46.3	59.5	178.0	6,943
– revisions (thousands)	0.2	1.8	2.9	2.9	3.8	6.8	–**
– words (thousands)	0.7	4.8	5.0	5.0	5.2	13.9	–**
– paste events	0.0	13.0	25.0	28.6	39.0	134.0	4,291
– references	3.0	11.0	16.0	18.4	21.0	69.0	2,761
Work time per essay							
– days passed	1.0	4.0	6.0	8.6	9.0	56.0	–**
– working days	1.0	4.0	5.0	5.5	7.0	17.0	–**
– working hours	1.8	5.2	7.5	7.9	9.8	23.0	1,191
– physical sessions	2.0	11.5	16.0	18.6	23.0	55.0	2,797
Minutes spent							
– reading per click	0.0	0.1	0.4	0.7	0.8	15.0	11,236
– writing per session	0.0	0.5	2.2	7.4	8.9	145.2	51,126

*Equal to some above value by definition.
**Sum not given to avoid misinterpretation.

Figure 2.3: Detail of duration required for the completion of Webis-TRC-12 dataset [01]

Writer Analysis: During the construction of this dataset a survey was held to gather the demographics detail of the writer. This data were collected based on the questionnaire and the resumes which were uploaded at oDesk platform by the respective writer. The detail statistics about the writer is shown in the table below. Most of the writer were from English speaking country, and almost all speak more than one language (Potthast, Hagen, Völske, and Stein 2013a). All demographics details are illustrated in the diagram below with different diagrams.

[11] This data is taken from the paper (Hagen, Potthast, Völske, Gomoll, and Stein 2016), we only used 2nd and 3rd portion from the table which is given in the paper (Hagen, Potthast, Völske, Gomoll, and Stein 2016)

Writer Demographics					
Age		*Gender*		*Native language(s)*	
Minimum	24	Female	67%	English	67%
Median	37	Male	33%	Filipino	25%
Maximum	65			Hindi	17%
Academic degree		*Country of origin*		*Second language(s)*	
Postgraduate	41%	UK	25%	English	33%
Undergraduate	25%	Philippines	25%	French	17%
None	17%	USA	17%	Afrikaans, Dutch,	
n/a	17%	India	17%	German, Spanish,	
		Australia	8%	Swedish each	8%
		South Africa	8%	None	8%
Years of writing		*Search engines used*		*Search frequency*	
Minimum	2	Google	92%	Daily	83%
Median	8	Bing	33%	Weekly	8%
Standard dev.	6	Yahoo	25%	n/a	8%
Maximum	20	Others	8%		

Figure 2.4: Demographics description of writer [01]

Writing Styles: Writing Styles are categorized into three types: build-up, boil-down, and mixed. Build-up is a fashion, in which continuous lengthening of the essay in a regular period of time over the whole period of writing is done. Whereas, in boil-down fashion, first quick length growth and then shorting happen. In simple way we can say, queries and click, i.e., research are done in regular interval and regularity of copy-paste happens over the duration of the writing process in Build-up. And, in Boil-down, all the research (queries and clicks) are done in the beginning of writing, copy-paste happens in the starting of the writing process, and the shorting of the essay is done. Our dataset consists of 150 essays some of them are written with build-up, some of them are with boil-down, and the rest are mixed.

[11] This data is taken from the paper (Potthast, Hagen, Völske, and Stein 2013a)

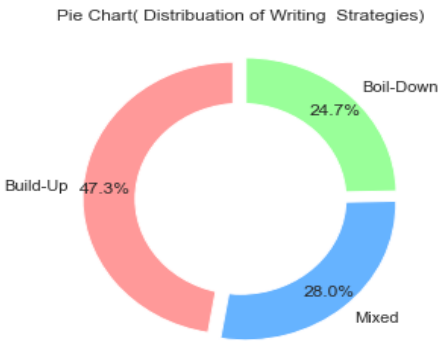


Figure 2.5: Distribution of writing style through all the topics

All the data used to create Figure 2.6 is taken from (Potthast, Hagen, Völske, and Stein 2013a). Figure 2.6 illustrates the details of the different authors who were responsible to create the dataset. This figure illustrates the statistics of different writing style implemented by different authors to create essays.

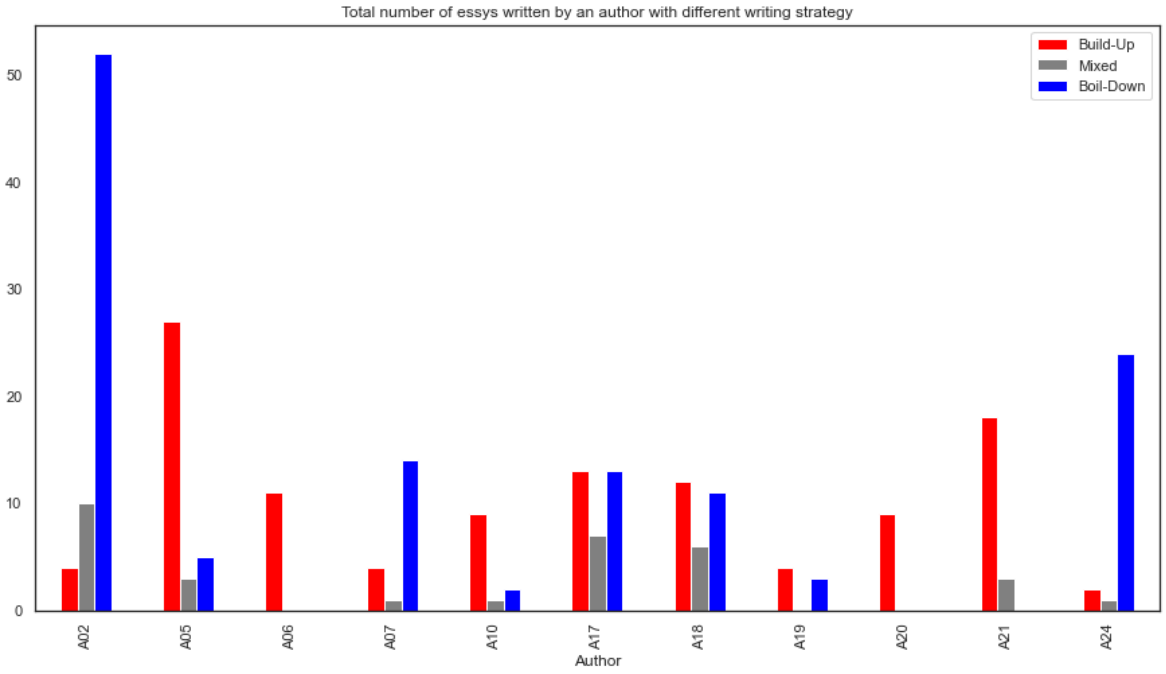


Figure 2.6: Visualization of statistics of essays written by an author with different writing strategy

Chapter 3

APPROACHES FOR TEXT QUALITY MEASURES

Coherence is an central measure to calculate text quality. Recently many researches are going on different measures of text quality. This section tries to explain the different approaches used to evaluate different measures of text quality that are implemented in our thesis. After passing our dataset through data preprocessing phase which is explained in section 4.1, we processed the clean text to different approaches for calculating the different measures of text quality which are explained below.

3.1 Coherence

Entity grid can be used as one of the approach to calculate coherence of text. Entity grid approach of discourse is inspired by Centering theory. This approach calculates coherence by capturing the distribution of entities across utterances. Text having same discourse entity are believed to be more coherent. From centering theory, text having certain kind of transitions are called more coherent than text having infrequent transitions (Grosz, Loshi, and Weinstein 1995). Entity grid is a two dimensional array which represents the distribution of discourse entities across adjacent sentences. Rows of grid represent sentences and columns represent discourse entities. Grid cell gives information about the presence or absence of entities in the corresponding sentence. Entity's absence from a sentence are marked by gaps (-) in the corresponding grid cell.

In addition, entity's presence in the sentence is marked by their syntactic role in their corresponding cell. Each grid cell thus correspond to the entity are reflecting whether the entity in the sentence is a subject (S), object (O), or neither (X).

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Figure 3.1: Example entity grid of discourse. [01]

Figure 3.1 illustrate an entity grid generated for Figure 3.2. In Figure 3.2 there are six sentences which are represented as six column in Figure 3.1. For example, in Figure 3.1 Evidence is recorded in column 1 and 3, because this entity is present in 1st and 3rd sentences in Figure 3.2 and is absent in rest of the sentences. Entity grid takes coreference resolution into account. For example, entity with the different linguistic forms such as Microsoft Corp., Microsoft, and the company are recorded in single row in the grid.

1 [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.
2 [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.
3 [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
4 [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
5 [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_O$.
6 [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

Figure 3.2: Sentences with syntactic annotations. [02]

[01] Figure is taken from the paper (Barzilay and Lapata 2008)
[02] Example in the figure is taken from the paper (Barzilay and Lapata 2008)

If a entity is noticed more than once with different grammatical role or linguistic form in a single sentence, then the grammatical role with highest ranking (subject > object > neither) is marked. For example, Microsoft comes twice in sentence 1 with role X (Microsoft Corp.) and S (Company), but represented by S in the cell (Barzilay and Lapata 2008). Grids of coherent text are supposed to have some dense columns. We can expect that entities corresponding to dense columns are subject or objects and these features will be less used in low coherent text. Coherence of text is defined with the probability that gives knowledge about how entities are distributed across text. Semantic relatedness between sentences can be measured by taking the mean of all the individual transitions.

$$Coherence(T)=\frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1}$$

where $sim(S_i, S_{i+1})$ is a measure of similarity between sentences s_i and s_{i+1}

Each sentence is represented by the mean of the vectors of its words, and similarity between two sentences is determined by the cosine similarity of their means.

$$sim(S_1, S_2)=cos(\mu(\vec{S_1}), \mu(\vec{S_2}))$$

$$=\frac{\sum_{j=1}^{n} \mu_j(\vec{S_1}), \mu_j(\vec{S_2})}{\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_1}))^2}\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_2}))^2}}$$

where $\mu(\vec{S_1}) = \frac{1}{|S_i|}\sum_{\vec{w}\epsilon S_i}\vec{w}$, $\vec{w}$ is a vector for word w (Lapata and Barzilay 2005) .

Implementation of entity grid approach for calculating coherence of text in this thesis is passed through different phase. First, we generate entity grid with the help of coreNLP.

CoreNLP

CoreNLP is a tool which facilitates to generate linguistic annotations for text, part of speech, named entities, numeric and time values, dependency and constituency parser, sentiment, etc. It takes raw text and generates a group of NLP annotation of text. As it is already mentioned that each occurrence of noun phrases, its role and presence are marked in respective cells corresponding to the respective sentence which is represented in row. So for finding Noun phrases corenlp use Parts Of Speech tagging.

Parts Of Speech is used to identify lexical role of the words in a sentence. For example, whether the word is verb, noun, adjective, etc. The sentence are preprocessed with Sanford Part-of-Speech tagger to extract information and role of words, sentence are tokenized into word and each word is assigned its POS tag shown in Figure 3.3(b) . A list of Part-of-speech tags is given in section A.1

There are two ways of using CoreNLP to generate entity grid of a text. If we have to generate entity grid of small text we can use CoreNLP online server [01] for NLP annotation and if we want to generate entity grid for large number of text then we have to run coreNLP as local server in our own device to generate annotation. After deriving POS annotation we generate entity grid. For example, *"I have a friend called Bob. He loves playing basketball. I also love playing basketball. We play basketball together sometimes."* In this example, we have 4 sentences so, in Figure 3.3(a) there is 4 row in the table. Each row is for individual sentence and each column has single noun phrases which are detected in the text with its role or presence written in each cell for individual sentence.

[01] Online coreNLP server for annotation `https://corenlp.run/`

Index	I	friend	Bob	basketball	we
0	s	s	o	-	-
1	-	s	-	o	-
2	s	-	-	o	-
3	-	-	-	o	s

(a) Example entity grid representation

(b) CoreNLP POS Annotation

Figure 3.3: Example entity grid representation and CoreNLP POS Tags

After deriving the entity grid, the next phase is to create a transition table in which a continuous sub-sequence of each column form an entity grid of length two is paired as shown in Figure 3.4. In Figure 3.4 we can notice that there is one row less than Figure 3.3(a) because all cells from each column are paired together with the size two.

Index	I	friend	Bob	basketball	we
0	('S', '-')	('S', 'S')	('O', '-')	('-', 'O')	('-', '-')
1	('-', 'S')	('S', '-')	('-', '-')	('O', 'O')	('-', '-')
2	('S', '-')	('-', '-')	('-', '-')	('O', 'O')	('-', 'S')

Figure 3.4: Example transition table.

After we create the transition table, we calculate the feature vector from the transition table. Figure 3.5 is the feature vector calculated from transition table. In Figure 3.4 There is a total 15 cells and there is 3 (s,-). Therefore, the feature vector for (s,-) can be calculated as $3 \div 15$ equals 0.2. Similarly, for (s,s) is $1 \div 15$ which equals 0.06666.

Index	SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	--
0	0.0666667	0	0	0.2	0	0.133333	0	0.0666667	0	0	0	0	0.133333	0.0666667	0	0.333333

Figure 3.5: Example feature vector.

And, at last, we compute the cosine similarity between the feature vectors which gives the coherence score. Cosine similarity is a comparison between the angle of two nonzero vector. The range of cosine similarity for information retrieval

is between [0,1]. It results 1 when two vectors are in the same orientation and if two vectors are far way, the similarity is 0. Cosine similarity of non-zero vector can be computed by calculating the dot product of two vectors divided by the product of their Euclidean norms. Dot product is normalized by the Euclidean norm so that each row vector has a length of 1. A Pseudocode for calculating Coherence is illustrated in algorithm 3.1

Algorithm 3.1: Pseudocode Coherence

```
Data: List of Paragraph
Result: Coherence_Score
Text ← [Paragraph];          /* [Paragraph] is lists of preprocessed
 paragraphs [p1,p2,p3,..] where p1,p2,p3 are paragraph */
for i = 0 to len(Text) do
    text = Text[i];
    /* Check element of list paragraph is not empty              */
    if len(text) > 0 then
        grid ← EntityGrid(text) ; /* Compute entity grid using CoreNLP
         server */
        Transition_table ← TransitionTable(grid);        /* make pair of
         transitions from a grid of length two */
        Feature_vector ← FeatureVector(Transition_table) ;    /* feature
         vector is ratio of pair of transition's frequency divided
         by total number of pair of transition in Transition_table,
         i.e., total number of cells in Transition_table */
        Coherence_Score ← CosineSimilarity(Feature_vector) ;
    else
        Coherence_Score ← 0;
    end
end
```

3.2 Type-Token Ratio

Type-Token ratio is a measurement specially used to study the lexical richness of a text. It helps to study the complexity of the text or sentences. Type-Token Ratio is defined as the ratio of unique words to the total words in a text. As vocabulary plays a vital role in the text quality, a well-used rich vocabulary in a text gives better quality of text. Hence, type-token ratio is used to measure the variety of vocabulary in the text. Till date,

many researcher have proposed different formulas to calculate type-token ratio, some of them are:

$$TTR = \frac{UniqueWord}{Totalwords}$$

$$RootTTR \text{(Guiraud 1950)} = \frac{UniqueWord}{\sqrt{Totalwords}}$$

$$CorrectedTTR \text{(Carroll 1964)} = \frac{UniqueWord}{\sqrt{2 \times Totalwords}}$$

$$LogTTR \text{(Herdan 1964)} = \frac{\log UniqueWord}{\log Totalwords}$$

We have implemented the normal TTR which is ratio of unique words, i.e., type to the total number of words, i.e., tokens in the text. An example how TTR is calculated is as shown in the Figure 3.6.

In the Figure 3.6, we have taken one single sentence as an example, but we can do the same to a bigger text too. We tokenize the sentences into word and count the unique word. Then the unique word is divided by total number of words. For example, in the figure there are 11 unique word and 15 total word, The type-token ration score is $\frac{11}{15}$, i.e., 0.733.

```
at the time of obamas landslide victory at the poll america was at the
crossroad


Counter({'at': 3, 'the': 3, 'time': 1, 'of': 1, 'obamas': 1, 'landslide':
1, 'victory': 1, 'poll': 1, 'america': 1, 'was': 1, 'crossroad': 1})


         Word  Frequency
0          at          3
1         the          3
2        time          1
3          of          1
4      obamas          1
5   landslide          1
6     victory          1
7        poll          1
8     america          1
9         was          1
10  crossroad          1


Type=11---Total Token=15


TTR Score=0.7333333333333333
```

Figure 3.6: Example Type-Token ratio

A simple Pseudocode for Type-Token Ratio is illustrated in algorithm 3.2. While calculating TTR, we take a list of preprocessed paragraphs as input. Paragraph tokenization is done by splitting text with newline. We take a single paragraph at one time and then we tokenize the paragraph into words using NLTK word tokenizer. After that, We count the unique words from the tokenized list which is type. Then, we check length of words to verify that the list of paragraphs does not have an empty element. Type token ratio is computed by dividing total number of unique words with the total number of words. This process is clearly illustrated in the algorithm 3.2.

Algorithm 3.2: Pseudocode TypeToken Ratio

Data: List of Clean text
Result: TTR_Score

```
Text ← [cleaned_text];          /* [cleaned_text] is lists of preprocessed
 paragraphs [p1,p2,p3,..] where P1,P2,p3 are paragraph */
for i = 0 to len(Text) do
    text = Text[i];
    /* use nltk standard; word tokenized and count the type and tokens
       */
    tokenized_word ← nltk.tokenized_word(text) ;
    types ← nltk.Counter(tokenized_word) ;
    if len(tokenized_word) > 0 then
        TTR_Score ← len(types) / len(tokenized_word) ;
    else
        TTR_Score ← 0;
    end
end
```

3.3 Readability

In our thesis, we calculate readability by using Flesch Reading Ease Formula. Flesch Reading Ease Formula is one of the oldest and popular one which is suitable for all kinds of texts. Whereas some of the other formulas have criteria for text. Flesch is one of the widely recognized readability formula,which is one of the reliable measures of the text readability. We used textstat Python library. In this library, numbers of formula are implemented to calculate readability and complexity of text. We use flesch reading ease formula from the textstat library to compute readability. Readability of the text is detailed explained in section 2.1.

While calculating readability, we take a list of preprocessed paragraphs as input as in Type-Token ratio. We take a single paragraph at one time and then check length of the paragraph to avoid zero exception error. Because the textstat library only accept texts which have at least few characters. Readability is computed by using $flesch\_reading\_ease$ function form textstat. The whole algorithm is clearly illustrated in the algorithm 3.3.

Algorithm 3.3: Pseudocode Readability

```
Data: List of clean Text
Result: Readability_Score
Text ← [cleaned_text];        /* [cleaned_text] is lists of
 preprocessed paragraphs [p1,p2,p3,..] where P1,P2,p3
 are paragraph */
for i = 0 to len(Text) do
    text = Text[i];
    /* Check empty element to avoid zero exception    */
    if len(text) > 0 then
        Readability_Score ← flesch_reading_ease(text) ;
    else
        Readability_Score ← 0 ;
    end
end
```

3.4 Different Approaches

There are many research done in the field of different measures of text quality, and different approaches are proposed to calculate text quality. Some of them are: Using Neural Network, Coh-Metrix, Human Judgements etc.

Chapter 4

EXPERIMENTS AND ANALYSIS

In this section, we conduct experiments on the Webis Text Reuse Corpus 2012 (Webis-TRC-12) (Potthast, Hagen, Völske, Gomoll, and Stein 2012). We discuss our results for different measures of text quality and the patterns we discover.

4.1 Data Preprocessing

The Webis-TRC-12 dataset which is used in this thesis is in HTML format. So, we need to extract data from HTML files. To do so, we read the HTML code from each file, and use the Beautifulsoup software library to parse the raw HTML data into a structured format. Beautiful Soup automatically encodes and converts the raw HTML code; we pass the result through html2text to get the desired output. We then pass this data through a data cleaning process, and on to the next phase for calculating coherence, type-token ratio, and readability. All data cleaning and preprocessing processes used in this thesis are illustrated in Figure 4.1.

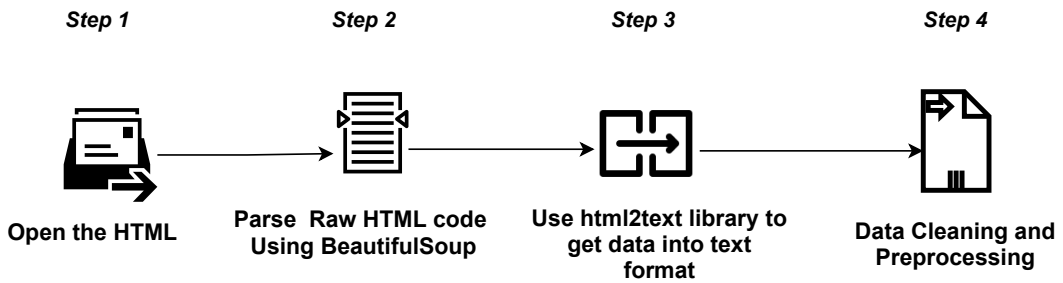


Figure 4.1: Data preprocessing work flow

Data Cleaning and Preprocessing

The extracted from the HTML files cannot be used as is, because of various unknown symbols or links. And, dirty data affects the result which we desire. Therefore, it is important that every data should pass through this phase. In data cleaning phase, we remove unwanted and unknown symbols and link them using regular expressions. For this, we tokenize all text into words, and check every word and letter through regular expressions which identify those unknown letters or symbols in English. In the first phase of data cleaning and preprocessing, we convert the whole text into lower case, tokenize the text into words, remove unknown symbols in English, lemmatize the words using Natural Language Toolkit (NLTK) and join them again after cleaning. In the second phase of data cleaning and preprocessing, we tokenize paragraphs. After paragraph tokenization, there are some empty list elements in the paragraph list which we remove.

Identify Major Changes in the Coherence Score Across Revisions

It is not feasible to visualize coherence changes across every revision of each of the 150 essays, some of which have thousands of revisions. Therefore, we decided to identify only the major changes across adjacent revisions. We checked the coherence scores of every pair of adjacent revisions; if they were the same, we skipped the result from the second revision and compared its results with the following revision. We only kept the revisions whose coherence scores differed from each other. We computed coherence across paragraphs in every revision of the 150 essays; this results in lists of coherence scores for each individual revision. To identify the major changes, we subtract the coherence scores of corresponding paragraphs across adjacent revisions. If all values after subtraction are equal to 0, that means there is no change in the coherence of paragraphs across adjacent revisions. Otherwise, we count the values not equal to 0; if this was the case for more than the half of the paragraphs in the respective revision, we consider it as a major change, otherwise we do not identify it as a major change.

Data Analysis

We computed the average number of words and the average number of paragraphs in all revisions of the 150 essays, and sort them by their average across revisions to analyze the writing style of all the essays. Figure 4.2 shows that most of the Build-up essays have fewer words, whereas Boil-Down essays have a higher number of words compared to others. The error bars in the figure show the 95 % confidence intervals across each essay's revisions. For example, topic 64 is clearly distinct as the topic with the highest number of words and topic 113 is clearly identified as the topic with the lowest number of words because the error bar are not overlapping for these two topics. For other pairs of essays, the error bars do overlap, meaning one is not consistently longer or shorter than the other over time.

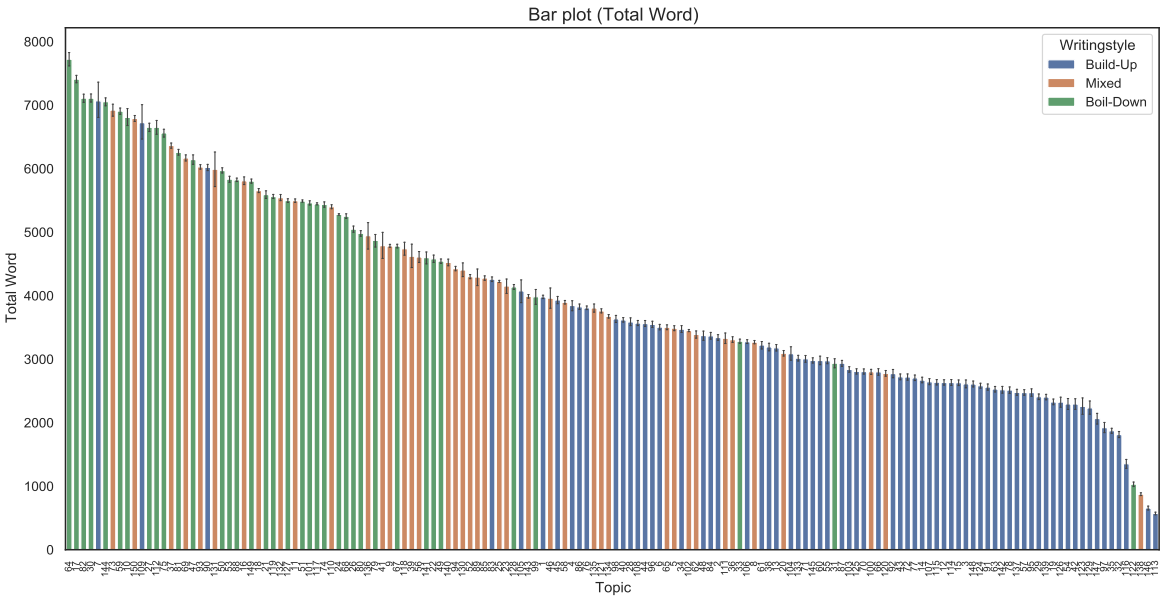


Figure 4.2: Average number of word in 150 essays

Figure 4.3 visualizes the average number of paragraphs in the 150 essays. In this figure, most of the Build-up essays have a smaller number of paragraphs, but writing styles group less clearly towards the higher end than they do for the number of words. Topic 37 is clearly distinct in that it has the highest number of paragraphs, and topic 113 has the lowest number of paragraphs. Table A.2 presents quantitative statistics of the 150 essays.

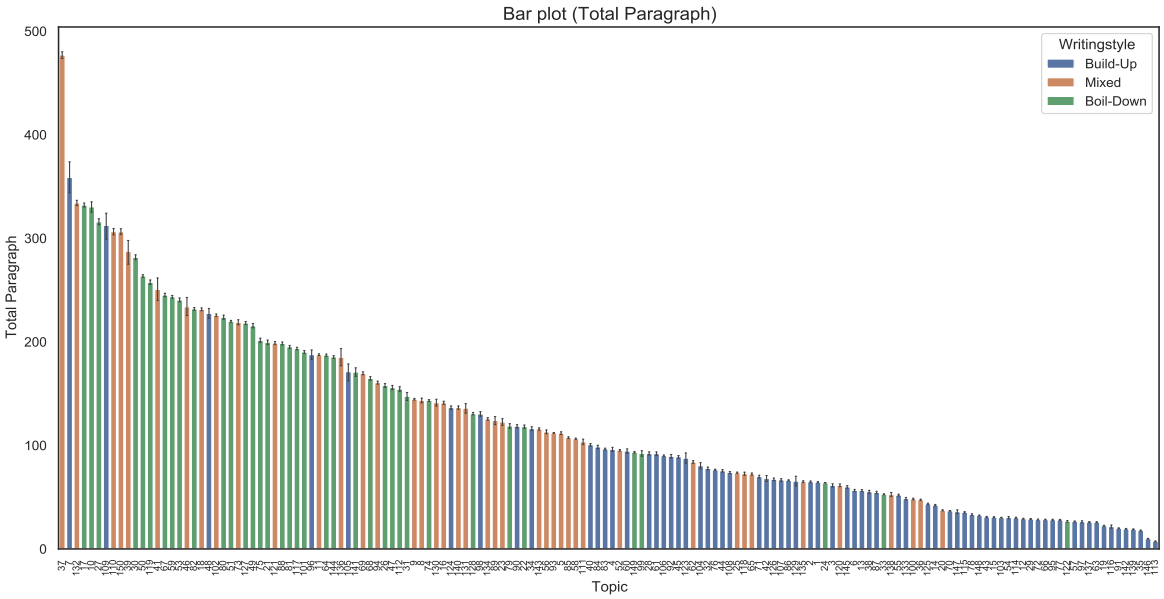


Figure 4.3: Average number of paragraph in 150 essays

Based on the previous work, we know that there were three writing styles: Build-up, Mixed, and Boil-Down (Potthast, Hagen, Völske, and Stein 2013a). Besides that, we attempt further analysis into how the essays were written. We calculate the Levenshtein distance between every pair of revisions of the essays. The Levenshtein distance gives the minimum number of insert, delete, and replace operations needed to convert the first revision into the second.

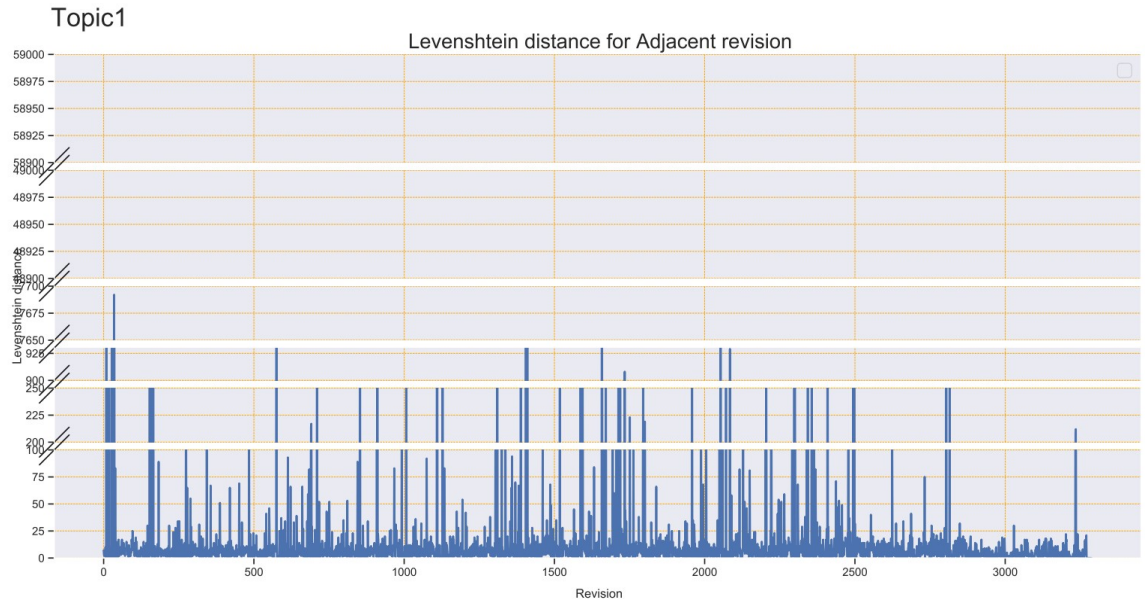


Figure 4.4: Levenshtein distance

Figure 4.4 visualizes the Levenshtein distance of adjacent revision of Topic001. A conclusion from the above figure is, every peak in the plot corresponds to times when the writer has copied and pasted content into the revision, or deleted large amounts of text, leading to a large edit distance between one revision and the next. Low points show that the writer has manually edited the text in the respective revision in a more fine-grained manner. To cross-check our conclusion from the figure, we have manually reviewed all the writing history of the topic in web visualization tool of the dataset, i.e., essay viewer (Potthast, Hagen, Völske, and Stein 2013b). While the plot shows only one essay, the same pattern holds for the others: peaks are copy-paste and low points are manual edits.

Besides this, we also analyze the distribution of descriptive statistics about the dataset. In this process we analyze for the total number of words and the total number of paragraphs in all revisions of the 150 topics.

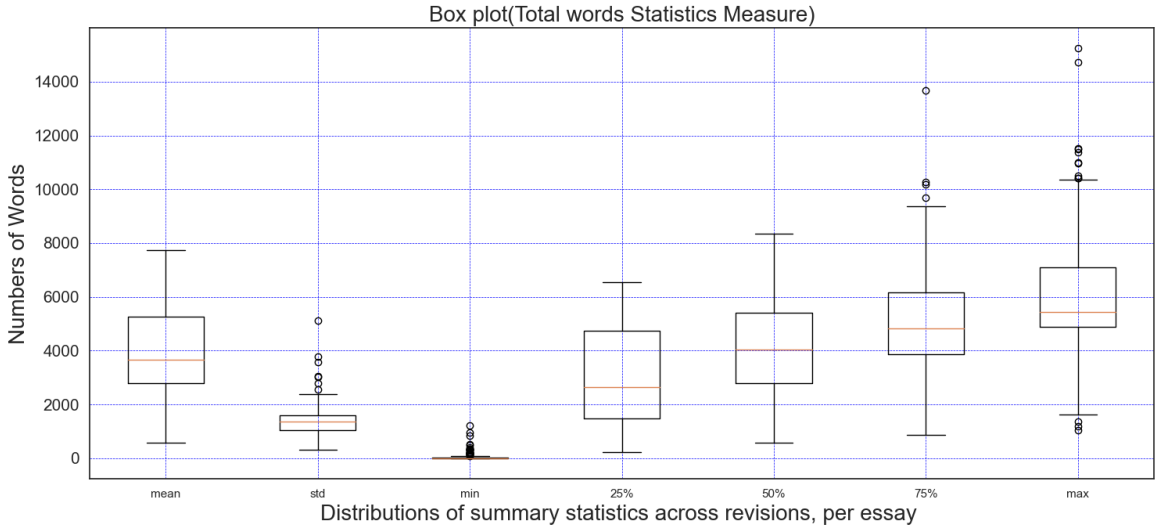


Figure 4.5: Box plots for descriptive statistics of the total numbers of words.

For each essay, we compute the mean, standard deviation, minimum, maximum, and quartiles of the word counts across all revisions. A box plot shows the distributions of these descriptive statistics across the 150 essays. Figure 4.5 shows that there are around 4000 words on average in most of the revisions across topics. It also shows that in some revisions there are more than 14000 words, and that the longest revision

contains nearly 6000 words on average across the 150 essays. All the descriptive statistics for revision word counts across all 150 essays are shown in Table A.3 in the appendix.

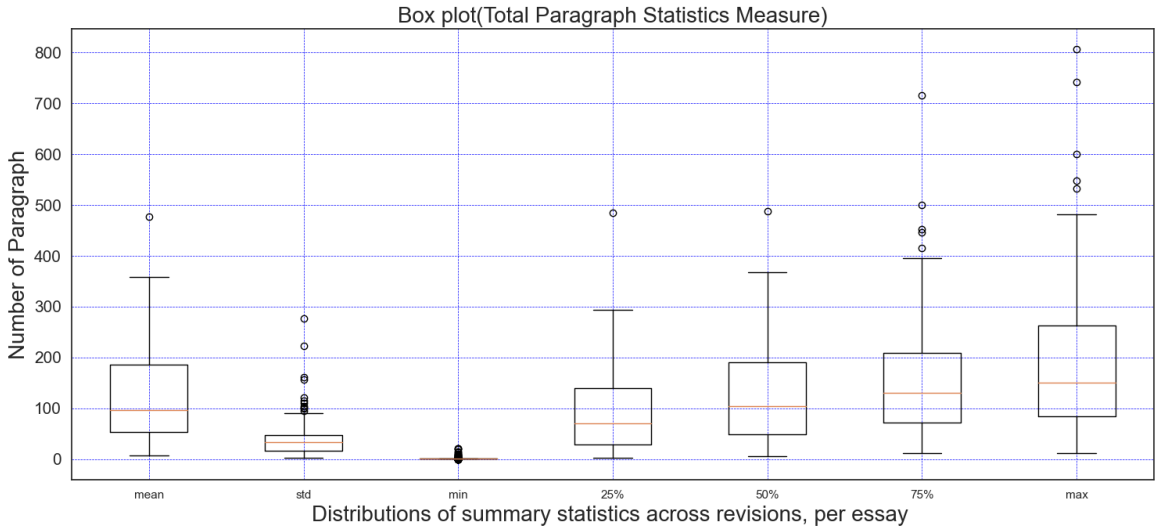


Figure 4.6: Box plots for descriptive statistics of total numbers of paragraphs.

For every essay, we also compute the mean, standard deviation, minimum, maximum, and quartiles of the paragraph count across all revisions. The box plots in Figure 4.6 are created from all descriptive statistics for 150 essays of the paragraph count which was computed across all revisions. Figure 4.6 shows that there are on average around 100 paragraphs in most of the revisions across topics. It shows that in some revisions there are more than 800 paragraphs, but the revision with the most paragraphs has about 150 paragraphs on average across the essays. All the descriptive statistics for the number of paragraphs of the revisions across the 150 essays are given in Table A.4.

4.2 Editing Types

We discovered five different editing types in this dataset. In this section, we discuss how we analyze the editing types and how we automatically identify all different types.

Analysis of editing operations

We analyze different editing types across pairs of consecutive revisions. Editing types mean how the second revision is edited compared to first revision. We have noticed that there are 5 different editing types across all revisions in the 150 essays. The five different editing techniques are: Block edited, Block merged, Block split, Block insertion, and Block deletion. In Block edited, writers manually type text into the respective block in the next revision. In Block merged, two different blocks are merged in the next revision compared to the previous revision. In Block split, one block is split into two different blocks in the next revision. Block insertion is the insertion of a new block in the next revision. And, Block deletion means deletion of a block in the next revision, where the block refers to a paragraph in our case. We first analyze the editing operations in selected revisions manually, and then distill our findings into an automated process with all the cases we recognized during editing operations in different revisions. The automation for identifying editing types is explained in the next section. Table 4.1 presents all different editing types with a brief description.

Table 4.1: Editing types

Editing Types	Description
Block edited	Blocks are manually edited
Block merged	Two different blocks are merged
Block split	One block split into two
Block insertion	New block is inserted
Block deletion	Block is deleted

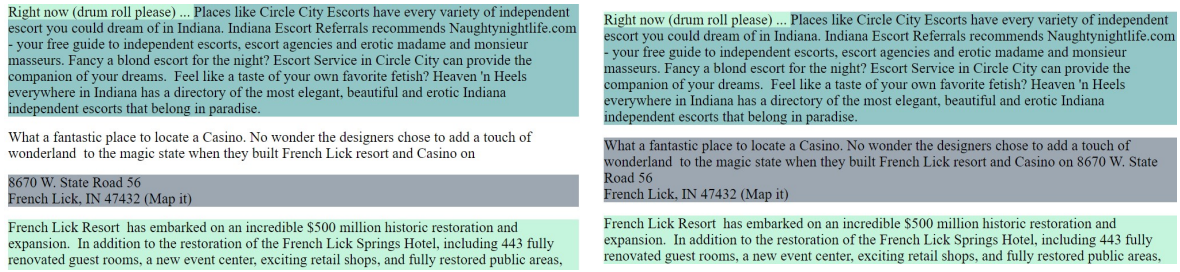
The figures below illustrate examples of the different editing types. Figure 4.7 shows an example of a Block edit in the first paragraph, with Figure 4.7(a) showing the paragraph before the edit, and Figure 4.7(b) showing the same paragraph after the edit.

(a) Before Block edited

(b) After Block edited

Figure 4.7: Example Block edited

In Figure 4.8(a), there are two different paragraphs in the middle, which are merged into one in Figure 4.8(b).



(a) Before Block merged

(b) After Block merged

Figure 4.8: Example Block merged

The first paragraph in Figure 4.9(a) has been split into two paragraphs in Figure 4.9(b) after the corresponding editing operation.

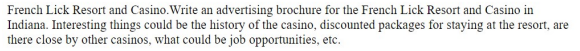
(a) Before Block split

(b) After Block split

Figure 4.9: Example Block split

In Figure 4.10(a), there is only a single paragraph (in this case, the essay's writing prompt), and in Figure 4.10(b), two additional paragraphs have been pasted into the essay in a Block insertion.

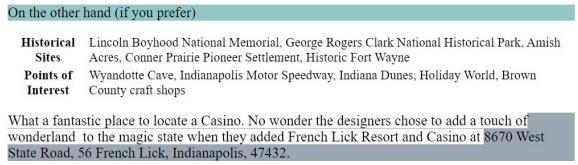
(a) Before Block insertion



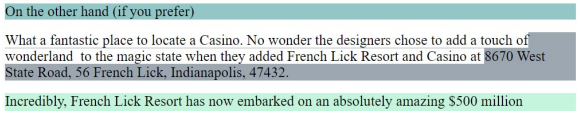
(b) After Block insertion

Figure 4.10: Example Block insertion

In Figure 4.11(a) there are three different paragraphs, whereas in Figure 4.11(b) the middle paragraph has been deleted from the previous revision.



(a) Before Block deletion



(b) After Block deletion

Figure 4.11: Example Block deletion

Automated Identification of Editing Types

To discover the different editing types, the number of characters and the number of paragraphs plays a vital role. We record the number of characters and number of paragraphs in adjacent revisions. When the number of paragraphs stays the same, but the number of characters changes, this is evidence of *Block edited*. Figure 4.12 shows an example output of the automation process, where the number of paragraphs is the same in both revisions, and the number of characters is different, that means it is Block edited.

```
number of paragraphs in 1st and 2nd revision 47--47
number of Characters in 1st and 2nd revision 12509--12415
Block edited
```

Figure 4.12: Example block edited identification

If the number of paragraphs in the first revision is higher than in the second revision, but the number of characters is the same in both revisions, then it is *Block merged*. Figure 4.13 shows an example output from the automation process.

```
number of paragraphs in 1st and 2nd revision 47--46
number of Characters in 1st and 2nd revision 12509--12509
Block Merged!!!!!!!
```

Figure 4.13: Example block merged identification

If the number of paragraphs in the first revision is smaller than the number of paragraphs in the second revision, but the number of characters is the same in both revisions, then it is *Block split*. Figure 4.14 presents an example from the automation process where the number of paragraphs in the first revision is less than in the second revision, and the number of characters is the same in both revisions, that means it is Block split.

```
number of paragraphs in 1st and 2nd revision 46--47
number of Characters in 1st and 2nd revision 12509--12509
Block Split!!!!!!!!
```

Figure 4.14: Example block split identification

Similarly, if the number of paragraphs in the first revision is smaller than the number of paragraphs in the second revision, and the number of characters in the first revision is less than the number of characters in the second revision, then it is *Block insertion*. Figure 4.15 shows an example output from the automation process.

```
number of paragraphs in 1st and 2nd revision 49--50
number of Characters in 1st and 2nd revision 12410--12413
Block Insertion!!!!!!!
```

Figure 4.15: Example block insertion identification

Finally, if the number of paragraphs in the first revision is higher than the number of paragraphs in the second revision, and the number of characters in the first revision is higher than

the number of characters in the second revision, then it is *Block deletion*. Figure 4.16 shows a corresponding example output from the automation process.

```
number of paragraphs in 1st and 2nd revision 52--51
number of Characters in 1st and 2nd revision 12802--12554
Block deletion!!!!!!!!
```

Figure 4.16: Example block deletion identification

Following the above, algorithm 4.1 shows the pseudocode to identify editing types, which forms the basis for our implementation.

Algorithm 4.1: Pseudocode for Editing Types

```
Data: List of Revision's paragraph counts and List of character lengths
Result: Editing Types
Topic ← [Revisions];                    /* lists of paragraphs inside list of
 revisions[[p1,p2,p3,..],[p1,p2,p3,..]] where P1,P2 are paragraph */
CharSize_Revision ← [CharSize];   /* lists of total number of characters
 in each paragraph inside list of
 revisions[[10,20,25,..],[10,20,30,..]] where [10,20,25,..] is total
 number of character in [p1,p2,p2,..]  */
for i = 0 to len(Topic) do
    Revision = Topic[i];
    CharSize_Paragraph = CharSize_Revision[i] ;    /* len(Revision[j]) is
     the total number of paragraphs in jth revision and
     sum(CharSize_Paragraph[j]) is the total number of character in jth
     revision  */
    for j = 0 to len(CharSize_Paragraph) − 1 do
        if len(Revision[j]) = len(Revision[j + 1]) and
         sum(CharSize_Paragraph[j]) ≠ sum(CharSize_Paragraph[j + 1]) then
            Block Edited;
        else if len(Revision[j]) > len(Revision[j + 1]) and
         sum(CharSize_Paragraph[j]) = sum(CharSize_Paragraph[j + 1]) then
            Block Merged;
        else if len(Revision[j]) < len(Revision[j + 1]) and
         sum(CharSize_Paragraph[j]) = sum(CharSize_Paragraph[j + 1]) then
            Block Split;
        else if len(Revision[j]) < len(Revision[j + 1]) and
         sum(CharSize_Paragraph[j]) < sum(CharSize_Paragraph[j + 1]) then
            Block Insertion;
        else if len(Revision[j]) > len(Revision[j + 1]) and
         sum(CharSize_Paragraph[j]) > sum(CharSize_Paragraph[j + 1]) then
            Block Deletion;
        else
            No Edit;
        end
    end
end
```

Result Discussion

After analyzing the editing techniques of essays over time, we spotted five different techniques: Block edit, Block merge, Block split, Block insertion, and Block deletion. To identify these editing techniques, we observed all revisions of 150 essays and cross-checked it with the essay's writing histories in the web visualization tool of the dataset, i.e., essay viewer (Potthast, Hagen, Völske, and Stein 2013b).

4.3 Effect of Editing Types on Text Quality Measures

After computing the text quality measures for all 150 essays, we analyze the different measure along with the editing styles. The statistics illustrated in Table 4.2 summarize the effect of coherence, readability, and TTR across all five editing styles. The result discussion section below elaborates further on these statistics.

Table 4.2: Descriptive statistics across editing type and text quality measures

Edit type	Qual. measure	mean	std	min	25%	50%	75%	max
Block Insertion	Coherence	0.711	0.094	0.000	0.657	0.723	0.777	1.000
	Readability	24.734	25.985	-387.983	12.729	27.422	40.990	121.220
	TTR	0.837	0.060	0.348	0.803	0.845	0.878	1.000
Block Merged	Coherence	0.692	0.082	0.484	0.641	0.707	0.752	1.000
	Readability	18.937	23.838	-72.090	6.700	19.656	35.455	72.826
	TTR	0.822	0.054	0.637	0.797	0.829	0.859	1.000
Block Split	Coherence	0.714	0.077	0.473	0.672	0.727	0.775	0.884
	Readability	27.492	22.934	-65.581	15.986	30.111	42.862	72.100
	TTR	0.842	0.047	0.662	0.816	0.850	0.871	1.000
Block deletion	Coherence	0.708	0.105	0.000	0.674	0.727	0.772	1.000
	Readability	27.244	24.743	-139.391	16.154	31.750	42.026	121.220
	TTR	0.834	0.090	0.000	0.815	0.847	0.877	1.000
Block edited	Coherence	0.670	0.124	0.000	0.588	0.693	0.751	1.000
	Readability	16.378	39.646	-184.310	5.126	21.838	38.517	121.220
	TTR	0.816	0.075	0.559	0.783	0.831	0.862	1.000

In Figure 4.17, the editing types are shown along the x-axis, and the coherence scores along the y-axis. Figure 4.17 shows that there are some revisions whose editing types are Block edited, Block insertion, and Block deletion where the coherence score is zero, and only Block split never reaches a maximum coherence score of 1 in any revision. Block split and Block merged have relatively more stable coherence scores, without many outliers. However, the median coherence scores for most of the editing types are nearly the same, except for Block edit.

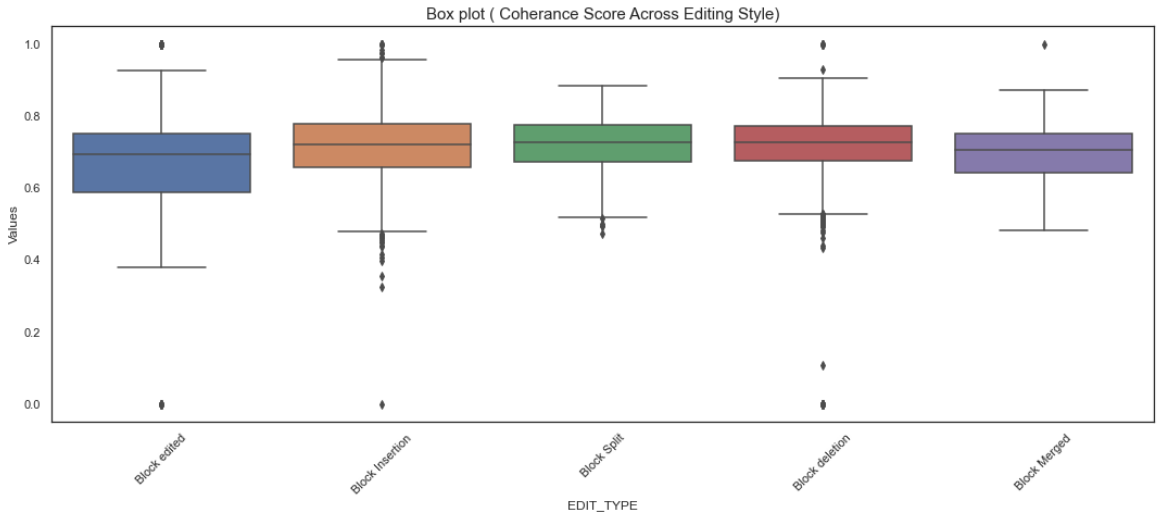


Figure 4.17: Box plot (Coherence score across different editing styles.)

Figure 4.18 shows that the TTR score of the Block edited type has high variance. For every editing type, there is some revision with a TTR score of 1, and all editing types have nearly the same median TTR score. Only a few revisions with the Block edited type have a TTR score of zero.

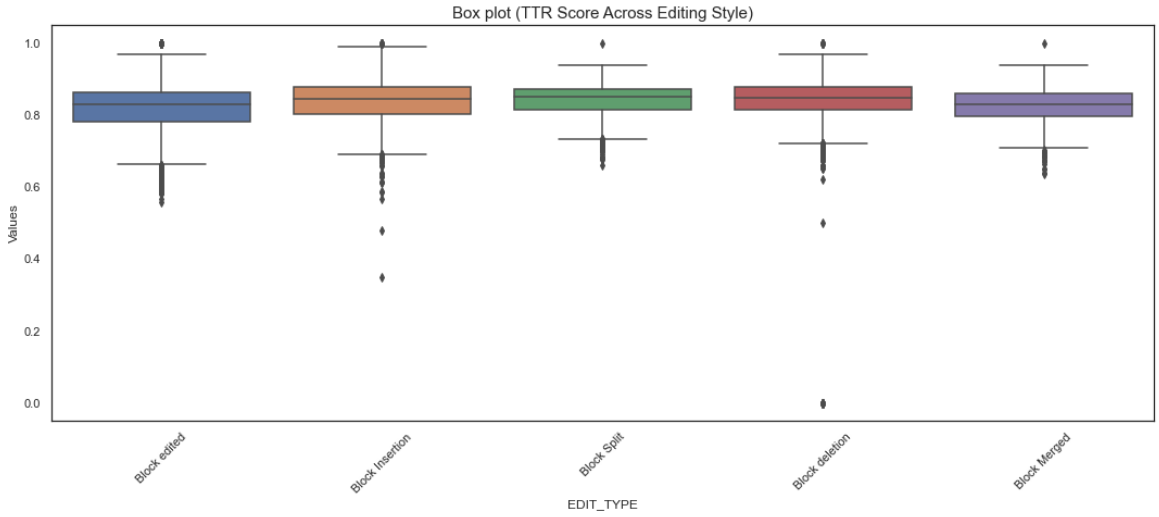


Figure 4.18: Box plot (Type-Token ratio score across different editing styles.)

Figure 4.19 shows that the readability score for most of the editing types has many outliers and high variance. A few revisions which are edited as Block insertion have a maximum value around 100 and minimum values around -400. The median readability score is in the range of 20–30 for all edit types.

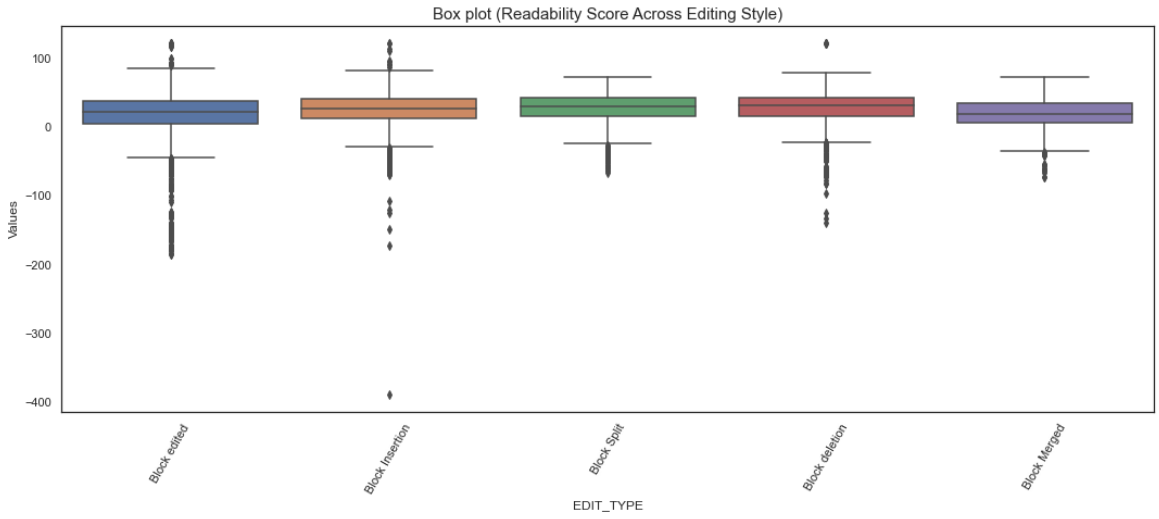


Figure 4.19: Box plot (Readability score across different editing styles.)

Result Discussion

After analyzing the edit types of essays across different text quality measures, we can conceptually order the different editing types by the quality dimensions. By way of average TTR and average coherence, the ordering is: Block split < Block insertion < Block deletion < Block merged < Block edited, whereas by average readability, the ordering is: Block split < Block deletion < Block insertion < Block merged < Block edited. From this, we can say that the average readability score is affected by the Block deletion and Block insertion editing types. The minimum and maximum values of coherence and TTR show the same pattern of sorting, whereas readability does not have the same pattern as coherence and TTR, which means the minimum and maximum values of readability are affected by different editing types. In Block merged and Block split, the scores of coherence, readability, and TTR vary less. The order of outliers variation in coherence is: Block edited < Block deletion < Block insertion < Block merged < Block split. The order of outliers variation in TTR is: Block deletion < Block edited < Block insertion < Block merged < Block split, whereas the order of outliers variation in readability is: Block edited < Block insertion < Block deletion < Block merged < Block split.

4.4 Effect of Writing Style on Text Quality Measures

After computing all text quality measures for the 150 essays, we tried to analyze the different measures along with different writing styles. The statistics illustrated in Table 4.3 summarize the effects of coherence, readability, and TTR across the three different writing styles.

Table 4.3: Descriptive statistics across writing style and text quality measures

Writing Style	Qual. measure	mean	std	min	25%	50%	75%	max
Boil-down	Coherence	0.727	0.081	0.000	0.690	0.731	0.776	1.00
	Readability	32.861	22.521	-132.590	22.042	34.794	46.165	121.22
	TTR	0.850	0.049	0.000	0.830	0.854	0.876	1.00
Build-up	Coherence	0.648	0.115	0.000	0.571	0.651	0.736	1.00
	Readability	6.902	36.554	-387.983	-7.382	14.731	26.526	121.22
	TTR	0.797	0.081	0.000	0.747	0.806	0.852	1.00
Mixed	Coherence	0.723	0.098	0.000	0.680	0.736	0.780	1.00
	Readability	29.634	19.262	-69.280	14.255	30.769	42.591	121.22
	TTR	0.846	0.069	0.000	0.810	0.846	0.885	1.00

Figure 4.20 is a scatter plot with error bars in each marker, where the y-axis shows the coherence score, and the x-axis shows the topic numbers sorted by coherence score. Each coherence score in the plot is the mean coherence score of the respective topic across revisions, and the error bar in each point is computed as a confidence interval around the mean. Points on the x-axis are sorted by the average coherence score, in descending order. The three different colors of markers indicate the three different writing styles: Build-up, Boil-down and Mixed. Figure 4.20 shows that, most of the essays written in the build-up style have lower coherence, whereas mixed writing styles have higher coherence, and boil-down essays are in between.

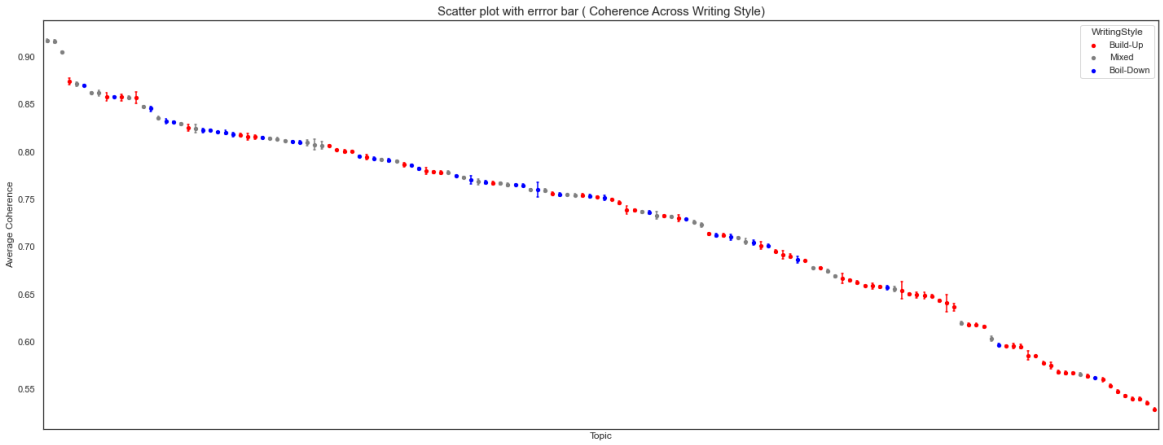


Figure 4.20: Scatter plot with error bars (Coherence across different writing styles).

Figure 4.21 is a scatter plot with error bars in each marker, where the x-axis shows the topic and the y-axis shows the TTR score. Otherwise, the plot follows the same style as Figure 4.20. But the Figure 4.20 is for coherence and this plot is for TTR. Figure 4.21 shows that most of the essays written in the build-up style have lower TTR, whereas essays with mixed writing style have higher TTR, and boil-down essays are in between.

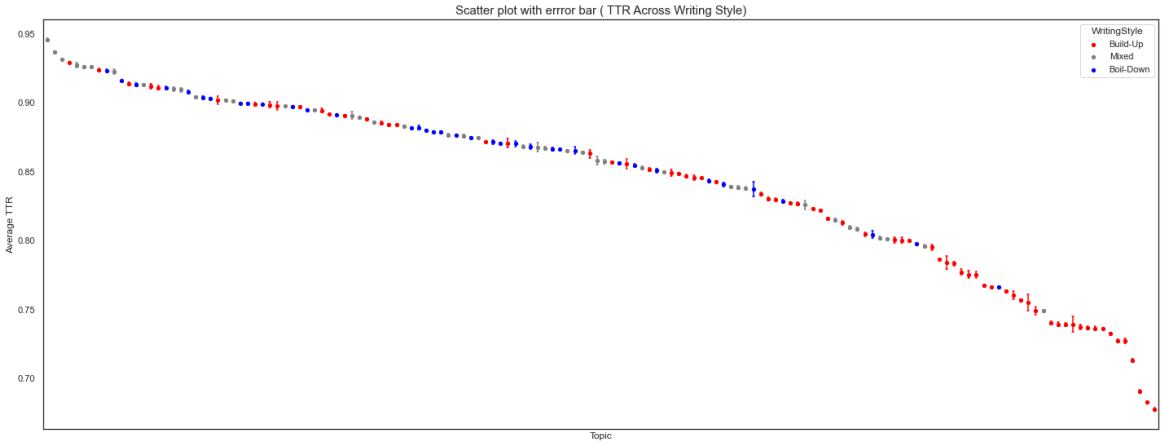


Figure 4.21: Scatter plot with error bars (TTR across different writing style).

Figure 4.22 is a scatter plot with error bars in each marker, where the x-axis shows the topic number and the y-axis shows the readability score. Otherwise, the plot is similar to Figure 4.20. But Figure 4.20 is for coherence and this plot is for readability. Figure 4.22 shows that most of the essays written

in the build-up style have lower readability, whereas essays with mixed writing style have higher readability, and boil-down essays are in between.

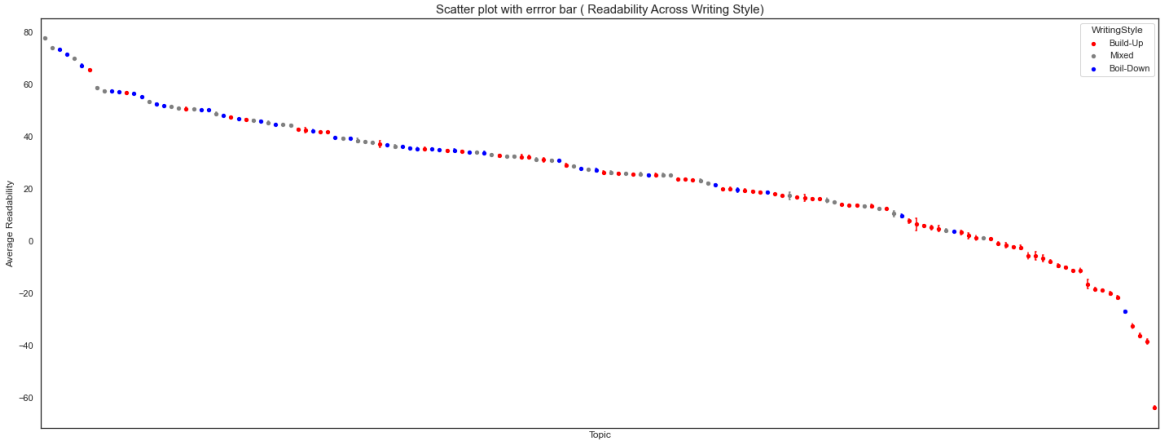


Figure 4.22: Scatter plot with error bar (Readability across different writing style.)

Result Discussion

After analyzing the writing techniques of essays across different text quality measures, we can order the writing strategies by their average text quality scores. For TTR, coherence, and readability the average ordering is always: Boil-down < Mixed < Build-up. The minimum value of coherence and TTR is 0 and maximum value of coherence and TTR is 1, whereas readability does not have same as coherence and TTR which means minimum and maximum value of readability is affected by the different writing types. From Table 4.3, Figure 4.20, Figure 4.21, and Figure 4.22, we observe that by amount of variation and outliers in coherence and TTR, the editing types can be ordered Build-up < Mixed < Boil-down. This means that the build-up writing style seems to result in higher quality variation, the mixed writing style has medium variation and the boil-down writing style has lower variation in coherence and TTR. By contrast, the build-up writing style has high variance, the boil-down writing style has medium variance, and the mixed writing style has lower variance and outliers in terms of readability.

Chapter 5

CONCLUSION

In this thesis, we used entity-grid based approach to calculate the coherence of essays written using search-supported writing. In addition, it also compares the readability and type-token ratio of the same text according to different writing styles and different editing types.

5.1 Main Findings

As one of our main contributions, we introduce an approach to identify different editing types. This thesis presents an automatic approach to distinguish the editing types "block edited," "block merged," "block insertion," "block split," and "block deletion." Our proposed approach identifies these editing types based on the number of characters and number of paragraphs modified in an edit.

After computing text quality scores using three different measures, we analyze their behavior across different writing and editing styles. Comparing the three measures across editing types, we notice that editing types affect readability. Coherence and type token ratio have same pattern of scores in every editing types.

To separate the results of different measures across writing styles, we create a scatter plot and sort the point with the mean values. We notice that most of the essays with build-up writing are having lower coherence, TTR, and readability scores. And, the maximum of coherence, TTR, and readability are in mixed writing.

5.2 Future Work

This thesis puts forward a detailed study on the approach for calculating different text quality measures: Coherence, Type-Token ratio, and readability, and analyzes the results according to different writing styles and different editing types. It also presents an in-depth analysis of Webis-TRC-12 essay-writing dataset. However, certain aspects and improvements in the approach are yet to be explored for improving the research and research results.

We notice that most of the essays written in the Build-up writing style have lower coherence scores. So, future work should investigate and explain the lower coherence in Build-up writing.

Similarly, we also noticed that most of the Build-up writing has lower TTR and Readability scores. So another task will be to investigate the reason and explain the lower TTR and Readability in Build-up writing.

Since entities play a vital role in the entity grid approach, another interesting task will be to investigate how edits to the essay that introduce new entities in particular affect the coherence score.

We can say that every coherent text is cohesive, but not every cohesive text is coherent. It will be interesting to research the relationship between the coherence and cohesion.

REFERENCES

[1] Enago Academy, *How coherence in writing facilitates manuscript acceptance*, 2020. [Online]. Available: https://web.archive.org/web/20210304154636/https://www.enago.com/academy/coherence-academic-writing-tips-strategies/, (accessed: 04.07.2021).

[2] Indiafreenotes, *Cohesion and coherence*, 2020. [Online]. Available: https://indiafreenotes.com/cohesion-and-coherence/, (accessed: 06.07.2021).

[3] X. Luo and Z. Chen, *English text quality analysis based on recurrent neural network and semantic segmentation*. Future Gener. Comput.Syst., 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167739X20302764?via%3Dihub.

[4] L. Schubert, "Computational Linguistics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Spring 2020, Metaphysics Research Lab, Stanford University, 2020.

[5] H. Oufaida, P. Blache, and O. Nouali, *A Coherence Model for Sentences ordering*. NLDB, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02299211/file/NLDB2019_Coherence.pdf.

[6] J. Snajder, T. S. Agejev, and S. kolic Vehovec, *Analysing Rhetorical Structure as a Key Feature of Summary Coherence*. BEA@ACl, 2019. [Online]. Available: https://www.aclweb.org/anthology/W19-4405.pdf.

[7] H. Xiong, Z. He, H. Wu, and H. Wang, *Modeling Coherence for Discourse Neural Machine translation*. AAAI/CoRR, 2019. [Online]. Available: https://arxiv.org/pdf/1811.05683.pdf.

[8] W. Liang, R. feng, X. Liu, Y. Li, and X. Zhang, *GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts*. IEEE Access, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8425711.

[9] S. Mathis, D. Kanojia, K. Patel, S. Agrawal, and P. Bhattacharyya, *Eyes are the windows to the soul:Predicting the rating of Text Quality Using Gaze Behaviour*. ACL/CoRR, 2018. [Online]. Available: https://arxiv.org/pdf/1810.04839.pdf.

[10] M. Mesgar and M. Strube, *A Neural Local Coherence Model for text Quality Assessment*. EMNLP, 2018. [Online]. Available: https://www.aclweb.org/anthology/D18-1464.pdf.

[11] M. Abdolahi and M. Zahedi, *Text coherence new method using word2vec sentence vectors and most likely n-grams*. ICSPIS, 2017. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8311598.

[12] B. Cui, Y. Li, Y. Zhanga, and Z. Zhang, *Text Coherence Analysis Based on Deep neural Network*. CIKM/CoRR, 2017. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3132847.3133047.

[13] A. Enquist, L. Oates, and jeremy Francis, *Just writing: Grammar, punctuation, and style for the legal writer*. Wolters Kluwer, 2017. [Online]. Available: https://books.google.de/books?hl=de&lr=&id=NvDfDgAAQBAJ&oi=fnd&pg=PT15&dq=Enquist,+A.+%26+Oates,+L.C.+2009.+Just+Writing:+Grammar,+Punctuation+and+Style+for+the+Legal+Writer+pdf&ots=4gXC3y2HK1&sig=YT7BF6yyihBKyCGT6q27DDQOcnU&redir_esc=y#v=onepage&q&f=false.

[14] K. J.Kopp, A. M.Johnson, S. A. Crossley, and D. S. McNamara, *Assessing Question Quality Using NLP*. AIED, 2017. [Online]. Available: https://files.eric.ed.gov/fulltext/ED577121.pdf.

[15] M. Hagen, M. Potthast, M. Völske, J. Gomoll, and B. Stein, "How Writers Search: Analyzing the Search and Writing Logs of Non-fictional Essays," in *1st ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2016)*, D. Kelly, R. Capra, N. Belkin, J. Teevan, and P. Vakkari, Eds., ACM, Mar. 2016, pp. 193–202. DOI: 10.1145/2854946.2854969.

[16] M. Mesgar and M. Strube, *Lexical Coherence Graph Modeling Using Word Embeddings*. HLT/NAACL, 2016. [Online]. Available: https://www.aclweb.org/anthology/N16-1167.pdf.

[17] D. Parveen and M. S. Mohsen Mesgar, *Generating Coherent Summaries of Scientific Articles Using Coherence Patterns*. EMNLP, 2016. [Online]. Available: https://www.aclweb.org/anthology/D16-1074.pdf.

[18] H. Wachsmuth, K. AL-khatib, and B. Stein, *Using Argument Mining to Assess the Argumentation Quality of Essays*. COLING, 2016. [Online]. Available: https://webis.de/downloads/publications/papers/stein_2016r.pdf.

[19] M. Mesgar and M. Strube, *Graph-based Coherence Modeling for Assessing Readability*. SEM@NAACl-HLT, 2015. [Online]. Available: https://www.aclweb.org/anthology/S15-1036.pdf.

[20] K. Xiaohua and H. Luo, *Using LSA and PLSA for text quality analysis*. esac, 2015. [Online]. Available: https://www.atlantis-press.com/proceedings/esac-15/25836910.

[21] J. Li and E. Hovy, *A Model of Coherence Based on Distributed Sentence Representation*. EMNLP, 2014. [Online]. Available: https://aclanthology.org/D14-1218.pdf.

[22] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, 2014. [Online]. Available: https://books.google.de/books?hl=en&lr=&id=xSPeAgAAQBAJ&oi=fnd&pg=PR13&dq=Automated+Evaluation+of+Text+and+Discourse+with+Coh-Metrix&ots=R62hImPv0P&sig=gVAvnQ4bfmjLuogvajlXLG6MN5g&redir_esc=y#v=onepage&q=Automated%20Evaluation%20of%20Text%20and%20Discourse%20with%20Coh-Metrix&f=false.

[23] S. A. Crossley, C. DeFore, K. Kyle, J. Dai, and D. S. McNamara, *Paragraph Specific N-Gram Approaches to Automatically Assessing Essay Quality*. EDM, 2013. [Online]. Available: https://www.educationaldatamining.org/EDM2013/papers/rn_paper_31.pdf.

[24] A. Louis, *Predicting text Quality: Metrics for content, organization and reader Interst*. Pennsylvabia university, 2013. [Online]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1823&context=edissertations.

[25] M. Potthast, M. Hagen, M. Völske, and B. Stein, "Crowdsourcing Interaction Logs to Understand Text Reuse from the Web," in *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, P. Fung and M. Poesio, Eds., Association for Computational Linguistics, Aug. 2013, pp. 1212–1221. [Online]. Available: http://www.aclweb.org/anthology/P13-1119.

[26] ——, "Exploratory Search Missions for TREC Topics," in *3rd European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR 2013)*, M. Wilson, T. Russell-Rose, B. Larsen, P. Hansen, and K. Norling, Eds., ser. Lecture Notes in Computer Science, vol. 1033, Aug. 2013, pp. 11–14. [Online]. Available: http://ceur-ws.org/Vol-1033/.

[27] D. Ashurova and M. Galiyeva, *Text linguistics*. Uzbek state world Langugae University, 2012. [Online]. Available: https://hozir.org/pars_docs/refs/38/37709/37709.pdf.

[28] A. Louis, *Automatic Metrics for Genre-specific Text Quality*. HLT/NAACL, 2012. [Online]. Available: https://www.aclweb.org/anthology/N12-2010.pdf.

[29] M. Potthast, M. Hagen, M. Völske, J. Gomoll, and B. Stein, *Webis text reuse corpus 2012*, Zenodo, Sep. 2012. DOI: 10.5281/zenodo.1341602. [Online]. Available: https://doi.org/10.5281/zenodo.1341602.

[30] S. Štajner, R. Evans, C. Orasan, and R. Mitkov, "What can readability measures really tell us about text complexity," in *Proceedings of workshop on natural language processing for improving textual accessibility*, Citeseer, 2012, pp. 14–22. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?url=10.1.1.353.9628&rep=rep1&type=pdf#page=19.

[31] M. Zamanian and P. Heydari, "Readability of texts: State of the art.," *Theory & Practice in Language Studies*, vol. 2, no. 1, 2012. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.4528&rep=rep1&type=pdf#page=45.

[32] S. A. Crossley and D. S. McNamara, *Text Coherence and Judgement of Essay Quality:Models of Quality and Coherence*. CogSci, 2011. [Online]. Available: https://cogsci.mindmodeling.org/2011/papers/0290/paper0290.pdf.

[33] J. L.Weston, S. A. Crossley, P. M. McCarthy, and D. S. McNamara, *Number of Words versus Number of Ideas: Finding a Better Predictor of Writing Quality*. FLAIRS Conference, 2011. [Online]. Available: https://alsl.gsu.edu/files/2014/03/Number-of-Words-versus-Number-of-Ideas-Finding-a-Better-Predictor-of-Writing-Quality.pdf.

[34] J. Zienkowski, J.-O. Östman, and J. verschueren, *Discursive Pragmatics*. John Nenjamins, 2011. [Online]. Available: https://books.google.de/books?hl=de&lr=&id=SRydcvRVBAsC&oi=fnd&pg=PA286&dq=text+linguistics&ots=V1uaGydKJZ&sig=TlrMGFJEQ6bg2gG8p_vl_BF1yno&redir_esc=y.

[35] R. Barzilay and M. Lapata, *Modeling Local Coherence: An Entity-Based Approach*. ACL/CoRR, 2008. [Online]. Available: https://aclanthology.org/J08-1001.pdf.

[36] M. Lapata and R. Barzilay, *Automatic Evaluation of Text Coherence :Models and representations*. IJCAI, 2005. [Online]. Available: http://homepages.inf.ed.ac.uk/mlap/Papers/505.pdf.

[37] A. C. Graesser, D. McNamara, M. M. louwerse., and zhiqiang Cai, *Coh-Metrix: Analysis of text on cohesion and language*. Psychonomic Society, 2004. [Online]. Available: https://www.researchgate.net/publication/8358727_Coh-Metrix_Analysis_of_text_on_cohesion_and_language.

[38] B. J. Grosz, A. K. Loshi, and S. Weinstein, "Centering: A framework for modeling the local coherence of discourse," *Comput. Linguistics*, vol. 21, pp. 203–225, 1995. [Online]. Available: https://www.aclweb.org/anthology/J95-2003.pdf.

[39] J. C. Richards, J. Platt, and H. Platt, "Longman dictionary of language teaching," *Applied Linguistics*, vol. 288, 1992.

[40] W. Mann and S. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text & Talk*, vol. 8, pp. 243–281, 1988. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/text.1.1988.8.3.243/html.

[41] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Comput. Linguistics*, vol. 12, pp. 175–204, 1986. [Online]. Available: http://www1.cs.columbia.edu/~julia/papers/grosz-sidner86.pdf.

[42] W. Mann and S. Thompson, “Relational propositions in discourse,” *Discourse Processes*, vol. 9, pp. 57–90, 1986. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/01638538609544632?needAccess=true.

[43] A. Brostoff, “Coherence: "next to" is not "connected to",” *College Composition and Communication*, vol. 32, no. 3, pp. 278–294, 1981, ISSN: 0010096X. [Online]. Available: http://www.jstor.org/stable/356191.

[44] T. A. Van Dijk, “The semantics and pragmatics of functional coherence in discourse,” *Speech act theory: Ten years later*, pp. 49–65, 1980. [Online]. Available: http://www.discourses.org/OldArticles/The%20semantics%20and%20pragmatics%20of%20functional%20coherence%20in%20discourse.pdf.

[45] T. A. V. Dijk, *PHILOSOPHY OF ACTION AND THEORY OF NARRATIVE*. Elsevier, 1976. [Online]. Available: http://www.discourses.org/OldArticles/Philosophy%20of%20action%20and%20theory%20of%20narrative.pdf.

[46] M. A. K. Halliday and R. Hasan, *Cohesion in English*. 1976, vol. I, p. 13. [Online]. Available: https://kupdf.net/download/cohesion-in-english-halliday-amp-hasan-1976_58cb70e0dc0d60db13c34635_pdf.

[47] ——, *Cohesion in English*. LONGMAN GROUP LIMITED LONDON, 1976, p. 23. [Online]. Available: https://kupdf.net/download/cohesion-in-english-halliday-amp-hasan-1976_58cb70e0dc0d60db13c34635_pdf.

[48] ——, “Cohesion in English,” *LONGMAN GROUP LIMITED LONDON*, vol. I, p. 89, 1976. [Online]. Available: https://kupdf.net/download/cohesion-in-english-halliday-amp-hasan-1976_58cb70e0dc0d60db13c34635_pdf.

[49] ——, *Cohesion in English*. 1976, vol. I, p. 226. [Online]. Available: https://kupdf.net/download/cohesion-in-english-halliday-amp-hasan-1976_58cb70e0dc0d60db13c34635_pdf.

[50] M. A. K Halliday and Ruqaiya Hasan, “Cohesion in English,” vol. I, 1976, p. 149. [Online]. Available: https://kupdf.net/download/cohesion-in-english-halliday-amp-hasan-1976_58cb70e0dc0d60db13c34635_pdf.

[51] B. Bamberg, “What makes a text coherent?” *College Composition and Communication*, vol. 34, no. 4, pp. 417–429, 1965. [Online]. Available: https://www.jstor.org/stable/357898?seq=1#metadata_info_tab_contents.

[52] J. Paul C. Rodgers, “Alexander bain and the rise of the organic paragraph,” *Quarterly Journal old speech*, vol. 51, no. 4, pp. 399–408, 1965. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00335636509382739.

[53] J. B. Carroll, *Language and thought. englewood cliffs, new jerseyl prentice-hall*, 1964.

[54] G. Herdan, “Quantitative linguistics,” 1964.

[55] Z. S. Harris, “Discourse Analysis,” *Linguistic society of America*, vol. 28, no. 1, pp. 1–30, 1952. [Online]. Available: https://www.jstor.org/stable/409987?seq=1#metadata_info_tab_contents.

[56] P. Guiraud, *Problèmes et méthodes de la statistique linguistique*. D. Reidel, 1950, vol. 2.

Appendix A

APPENDIX

A.1 Part-Of-Speech Tags

Table A.1: Part-of-speech Tags

Type	Description
CC	coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Conjunction, subordinating or preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Verb, modal auxillary
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Noun, proper singular
NNPS	Noun, proper plural
PDT	Predeterminer
PRP	Pronoun, personal
PRP$	possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Adverb, particle
SYM	Symbol
TO	Infinitival to
UH	Interjection
VB	Verb, base form
VBZ	Verb, 3rd person singular present
VBP	Verb, non-3rd person singular present
VBD	Verb, past tense
VBN	Verb, past participle
VBG	Verb, gerund or present participle
WDT	wh-determiner
WP	wh-pronoun, personal
WRB	wh-adverb
#	Pound sign
$	Dollar sign
.	Sentence final punctuation
,	Punctuation mark, comma
:	Punctuation mark, colon
(	Left bracket character
)	Right bracket character
"	Straight double quote

A.2 Quantitative statistics of 150 Essays

Table A.2: Quantitative statistics of 150 Essays

WritingStyle	Topic	Total Revision	Average Para Number	Average Word Number	Average Char Number
Build-Up	1	3288	64.459854	3980.312956	18945.524939
Build-Up	2	3151	64.888289	3340.993335	15827.692479
Build-Up	3	2240	61.466964	2612.176339	11162.655357
Build-Up	4	2507	96.121659	3845.359793	18372.326286
Mixed	5	3679	111.891818	3488.295461	17976.748573
Build-Up	6	3565	56.553717	3505.015147	16315.212342
Build-Up	7	1167	358.755784	7067.926307	33786.158526
Mixed	8	2049	143.496828	3268.847243	16704.658858
Mixed	9	2408	144.695598	4783.414037	23083.881229
Boil-Down	10	1015	330.391133	6807.582266	33856.951724
Mixed	11	6530	187.626953	5496.403216	27014.488668
Build-Up	12	3224	29.194169	2633.085298	11862.651055
Build-Up	13	4693	56.392073	3178.849563	15580.056041
Build-Up	14	3190	42.292476	2670.613480	13337.442006
Build-Up	15	3835	30.707692	2629.966102	12656.519426
Mixed	16	1615	141.138700	5809.457585	29771.739319
Boil-Down	17	4506	332.114736	7409.941855	30991.278074
Mixed	18	2254	231.464064	5654.752440	26625.661934
Build-Up	19	3657	22.435876	2325.938201	11388.305989
Mixed	20	2843	37.301442	3093.786141	14751.097784
Boil-Down	21	1443	199.609148	5591.913375	27086.891892
Boil-Down	22	1298	118.175655	4580.713405	22089.291988
Mixed	23	1319	122.654284	4152.551933	20893.633813
Boil-Down	24	6948	63.911917	5279.766983	28186.925158
Mixed	25	2940	73.621088	4225.502381	20776.434014
Boil-Down	26	3132	157.934547	5047.779055	26944.246807
Boil-Down	27	4146	315.954414	6652.140135	32868.174144
Build-Up	28	2903	92.190493	3587.033414	16057.838787
Build-Up	29	3586	28.894311	2409.068879	11339.052984
Boil-Down	30	2065	281.693462	7109.492494	35564.876998
Boil-Down	31	1032	147.475775	2937.030039	14247.728682
Build-Up	32	2188	77.809415	1812.910878	9087.983547
Boil-Down	33	4771	52.828967	3282.654580	15904.024733
Build-Up	34	2316	116.252591	3472.565199	16556.476252
Build-Up	35	2568	17.850078	1873.475078	8762.521417
Mixed	36	3884	47.480947	3307.651390	16173.251287
Mixed	37	2196	477.057377	6365.140255	29642.979053
Build-Up	38	2036	55.293222	3193.545678	14079.090864
Mixed	39	1434	287.423291	4620.775453	23186.201534
Build-Up	40	2806	100.631148	3618.750891	17925.908054
Mixed	41	792	250.688131	4788.179293	27035.631313
Build-Up	42	1510	68.058278	2294.204636	10874.825828
Build-Up	43	3666	30.875341	2722.974086	12466.180851
Build-Up	44	3006	75.464072	3561.335329	17490.740852
Build-Up	45	5084	88.987805	3932.519276	19867.920732
Mixed	46	735	233.846259	3961.180952	21561.851701
Boil-Down	47	1466	155.877899	6143.227831	29501.347885
Build-Up	48	1732	227.458430	3372.053695	17769.345266
Boil-Down	49	1816	215.615088	4544.713656	19505.147577
Boil-Down	50	4017	263.657954	5970.509833	27957.192432
Boil-Down	51	5720	219.876224	5490.045455	25574.942657
Mixed	52	2174	95.218491	4300.319687	20232.003680
Boil-Down	53	2553	240.452017	5835.206816	29020.091657
Build-Up	54	907	30.245865	2294.939361	11160.952591
Build-Up	55	3449	52.033923	2975.750072	14649.712090
Mixed	56	1908	113.046646	4609.264675	22457.829665
Build-Up	57	3866	26.524573	2474.826436	12727.872478
Mixed	58	3061	106.396929	3896.183927	19337.943809
Boil-Down	59	5480	243.556752	6906.233942	31164.645985
Build-Up	60	1842	94.614007	2979.185125	15509.275244
Build-Up	61	3076	92.073797	3222.958062	15601.316970
Mixed	62	3197	84.040663	3389.920238	17388.859556
Build-Up	63	3367	25.844966	2530.098010	12005.770419
Boil-Down	64	1554	187.213642	7723.525740	38736.350708
Mixed	65	2538	72.323877	3502.178093	17295.394405

Continued on next page

Table A.2 – *Continued from previous page*

WritingStyle	Topic	Total Revision	Average Para Number	Average Word Number	Average Char Number
Build-Up	66	3767	28.359437	2801.166976	13202.910273
Boil-Down	67	3795	245.251647	4781.238472	23644.894335
Boil-Down	68	4542	164.782034	5250.971378	27177.708278
Mixed	69	4430	169.686907	6168.401354	28128.225959
Build-Up	70	5067	36.816065	2805.462996	13334.660746
Build-Up	71	3654	69.875205	3006.959496	14539.865900
Build-Up	72	3230	28.363158	2720.023529	12746.220743
Mixed	73	3537	218.850438	6922.486288	31491.847894
Boil-Down	74	3644	143.476948	5436.829857	25737.669868
Boil-Down	75	1897	201.496574	6563.021086	32233.529784
Build-Up	76	5040	76.440476	3807.449206	17110.614484
Build-Up	77	3358	28.107207	2705.784693	12680.567004
Build-Up	78	2643	33.232312	2514.470299	11019.228907
Boil-Down	79	1610	118.819255	4871.876398	23914.017391
Boil-Down	80	1808	223.773230	4981.872235	23248.614491
Boil-Down	81	2020	195.259406	6257.996535	32033.004455
Boil-Down	82	5133	231.827002	7110.260471	37203.787064
Build-Up	83	2494	96.549719	4259.173617	20943.262229
Build-Up	84	2443	98.528449	3367.503479	16691.164552
Mixed	85	2750	107.430545	4278.440364	21873.916000
Build-Up	86	4879	66.297602	3824.903464	18158.848739
Build-Up	87	4088	54.633317	2936.503425	14747.318493
Boil-Down	88	5650	198.458230	5824.181947	28590.781947
Mixed	89	572	124.090909	4293.159091	21870.246503
Build-Up	90	2709	118.567368	6020.437431	28356.461794
Build-Up	91	3511	20.004842	2560.977784	11672.018513
Build-Up	92	1737	89.500288	2773.435233	12728.367300
Mixed	93	2667	112.005999	6027.521185	29512.626172
Mixed	94	2071	160.639305	4426.070014	20064.555770
Build-Up	95	2476	28.152666	2474.158320	10961.300485
Build-Up	96	2586	187.680588	3547.612916	17654.051044
Build-Up	97	830	26.254217	1921.779518	9882.120482
Build-Up	98	4667	130.367045	3635.501821	17585.990144
Boil-Down	99	1018	92.327112	3983.044204	21524.068762
Mixed	100	3777	48.261319	2802.897802	12848.735240
Boil-Down	101	3500	190.180000	5462.075429	26960.653714
Mixed	102	4101	225.726164	3450.775908	17256.270666
Build-Up	103	4542	30.294364	2838.617129	13547.012550
Build-Up	104	538	80.412639	3086.804833	15104.918216
Build-Up	105	573	171.198953	4075.483421	21465.169284
Build-Up	106	3911	90.111480	3275.472002	16492.523907
Build-Up	107	3836	66.740355	2644.474192	12215.969239
Build-Up	108	2971	73.886234	3568.539886	17291.602154
Build-Up	109	704	312.502841	6724.599432	38575.507102
Mixed	110	4505	306.451498	5400.376915	27635.254606
Mixed	111	2317	103.504100	3329.798446	16456.803194
Boil-Down	112	1778	154.450506	6651.953318	32979.345332
Build-Up	113	942	7.354565	570.581741	2592.125265
Build-Up	114	3338	30.142001	2632.428700	12343.064709
Build-Up	115	4045	35.389122	2637.431397	14622.408158
Build-Up	116	489	21.586912	1353.815951	6901.327198
Boil-Down	117	5486	193.566533	5446.263215	28775.951695
Mixed	118	941	72.724761	4739.578108	23036.376196
Boil-Down	119	3755	257.520107	5561.114514	26265.184820
Mixed	120	4069	61.459818	2776.625461	12624.013025
Mixed	121	2359	198.905468	3760.593048	19652.317507
Boil-Down	122	480	26.629167	1032.216667	5262.008333
Build-Up	123	249	87.682731	2257.650602	11577.714859
Build-Up	124	1822	136.568606	2582.522503	12299.600439
Build-Up	125	4211	43.542626	2809.039183	13276.532415
Build-Up	126	901	67.320755	2324.406215	11587.870144
Boil-Down	127	4212	218.199430	5498.300095	27655.615147
Boil-Down	128	2473	130.672058	4138.234937	20599.630004
Build-Up	129	713	65.472651	2233.302945	11260.945302
Mixed	130	1513	141.335757	4404.443490	20516.089227
Mixed	131	485	135.989691	5990.352577	32299.002062
Mixed	132	2959	334.178439	5543.043934	24792.167286
Build-Up	133	2527	48.625643	3015.432133	15309.410764
Mixed	134	3270	125.588685	3675.474924	17140.854128
Mixed	135	3842	65.091619	3806.126497	18109.758459
Mixed	136	601	185.076539	4945.108153	26953.856905

Continued on next page

Table A.2 – *Continued from previous page*

WritingStyle	Topic	Total Revision	Average Para Number	Average Word Number	Average Char Number
Build-Up	137	3767	25.967613	2477.571808	11683.745686
Mixed	138	820	52.592683	874.535366	4398.251220
Build-Up	139	3621	19.034797	2404.063518	11482.509528
Mixed	140	2834	136.509880	4524.151376	21519.321807
Boil-Down	141	1293	170.826759	4599.163186	22977.403712
Build-Up	142	3096	19.393411	2519.415698	12452.391796
Mixed	143	2827	115.902724	3990.845065	18934.172975
Boil-Down	144	3976	185.421781	7054.964537	33875.688380
Build-Up	145	3559	59.902782	2980.251194	14882.763979
Build-Up	146	396	9.906566	656.265152	3406.154040
Build-Up	147	934	35.980728	2065.746253	10184.962527
Build-Up	148	3803	32.304759	2609.488825	12398.836182
Boil-Down	149	2716	93.455817	5804.815906	29304.638071
Mixed	150	6374	306.336994	6790.545968	36332.721839

A.3 Different Statistics Measure for Total Word of 150 Essays

Table A.3: Different Statistics Measure for Total Word of 150 Essays

Topic	count	mean	std	min	25%	50%	75%	max
1	3288.0	3980.312956	773.338020	1.0	3599.00	4000.5	4665.00	5014.0
2	3151.0	3340.993335	1175.515195	45.0	2671.50	3346.0	4637.00	5111.0
3	2240.0	2612.176339	1506.480161	1.0	1697.75	2595.0	3401.00	4982.0
4	2507.0	3845.359793	2090.228423	1.0	1482.00	4673.0	5485.00	5604.0
5	3679.0	3488.295461	1197.592242	0.0	2580.50	3013.0	4859.00	5426.0
6	3565.0	3505.015147	1389.565255	1203.0	2379.00	3496.0	4792.00	5594.0
7	1167.0	7067.926307	5123.509450	1.0	3033.00	4739.0	13685.50	14713.0
8	2049.0	3268.847243	610.325023	1.0	3045.00	3466.0	3707.00	3991.0
9	2408.0	4783.414037	598.880698	1.0	4533.00	4928.0	5054.00	5328.0
10	1015.0	6807.582266	2198.516607	1.0	5016.00	6493.0	8620.00	9842.0
11	6530.0	5496.403216	1066.792533	1.0	4916.25	5822.0	6206.00	6884.0
12	3224.0	2633.085298	1169.056533	139.0	1496.75	2583.0	3720.00	4835.0
13	4693.0	3178.849563	1727.842293	0.0	1447.00	4016.0	4820.00	5684.0
14	3190.0	2670.613480	1247.370171	218.0	1529.00	2424.5	3850.00	4671.0
15	3835.0	2629.966102	1436.885253	1.0	1324.00	2648.0	3722.50	4736.0
16	1615.0	5809.457585	1270.002064	309.0	5165.50	6079.0	6634.00	7889.0
17	4506.0	7409.941855	2036.430956	1.0	5190.00	7997.0	9376.00	10496.0
18	2254.0	5654.752440	796.804303	0.0	5278.00	5592.5	6136.00	8122.0
19	3657.0	2325.938201	1429.025278	2.0	1046.00	2304.0	3589.00	4478.0
20	2843.0	3093.786141	1204.875975	1.0	2346.00	3360.0	4007.50	4681.0
21	1443.0	5591.913375	1152.147248	1.0	5717.00	6094.0	6182.00	6197.0
22	1298.0	4580.713405	1093.442835	2.0	4364.25	4627.0	5020.00	6263.0
23	1319.0	4152.551933	2191.770643	1.0	1396.00	5815.0	5979.00	6080.0
24	6948.0	5279.766983	476.964936	58.0	5161.00	5280.0	5540.00	5816.0
25	2940.0	4225.502381	464.814357	1.0	3882.00	4292.0	4541.00	5171.0
26	3132.0	5047.779055	1451.923701	64.0	5423.00	5547.5	5633.00	6040.0
27	4146.0	6652.140135	2144.412053	1.0	5157.00	6387.0	8345.50	10420.0
28	2903.0	3587.033414	1694.370874	2.0	2729.00	4334.0	4970.00	11366.0
29	3586.0	2409.068879	1422.057443	196.0	1050.00	2274.0	3519.00	4715.0
30	2065.0	7109.492494	1675.189697	1.0	5956.00	7896.0	8355.00	9322.0
31	1032.0	2937.030039	1132.082034	1.0	2791.00	3377.0	3533.00	4782.0
32	2188.0	1812.910878	1112.406466	158.0	962.00	1492.0	2579.25	4568.0
33	4771.0	3282.654580	1238.072574	2.0	1856.00	3355.0	4053.00	5130.0
34	2316.0	3472.565199	1353.731712	0.0	3139.00	3220.0	4388.00	5179.0
35	2568.0	1873.475078	1035.935838	70.0	1110.00	1800.0	2807.00	3959.0
36	3884.0	3307.651390	1406.628269	1.0	2046.00	3391.0	5019.00	5354.0
37	2196.0	6365.140255	961.216956	2.0	6538.00	6665.0	6685.00	6715.0
38	2036.0	3193.545678	1369.958487	852.0	2161.75	3511.0	4617.00	5103.0
39	1434.0	4620.775453	3580.066724	2.0	1481.25	3854.0	8168.50	11505.0
40	2806.0	3618.750891	953.591693	41.0	2853.00	4063.5	4301.00	5021.0
41	792.0	4788.179293	3042.778011	4.0	1347.00	5696.5	7920.00	8725.0
42	1510.0	2294.204636	1627.638477	0.0	759.00	1433.0	3542.25	5539.0
43	3666.0	2722.974086	1432.177697	3.0	1688.50	2724.5	3975.00	4796.0

Continued on next page

Table A.3 – *Continued from previous page*

Topic	count	mean	std	min	25%	50%	75%	max
44	3006.0	3561.335329	1297.043734	313.0	2039.00	3568.5	4652.00	5026.0
45	5084.0	3932.519276	2088.701674	1.0	2738.00	3657.5	6217.00	6795.0
46	735.0	3961.180952	2129.364305	35.0	1474.00	4637.0	5963.00	6161.0
47	1466.0	6143.227831	1435.353127	1.0	4980.00	6254.0	7290.00	8475.0
48	1732.0	3372.053695	1565.332628	0.0	2346.00	3806.5	4851.00	5035.0
49	1816.0	4544.713656	772.251822	1.0	4506.75	4769.0	4878.00	5506.0
50	4017.0	5970.509833	1386.624066	2.0	5058.00	5822.0	7058.00	8705.0
51	5720.0	5490.045455	704.818932	1.0	5128.00	5462.0	6145.00	7455.0
52	2174.0	4300.319687	732.747684	489.0	4376.00	4476.0	4624.75	15238.0
53	2553.0	5835.206816	1256.908921	3.0	5972.00	6116.0	6494.00	7098.0
54	907.0	2294.939361	1361.189855	15.0	1006.00	2220.0	3650.50	4510.0
55	3449.0	2975.750072	1382.693508	1.0	1623.00	2921.0	4048.00	5366.0
56	1908.0	4609.264675	1896.123116	193.0	3497.75	5198.0	6262.00	6390.0
57	3866.0	2474.826436	1430.395288	3.0	1120.00	2529.5	3802.00	4759.0
58	3061.0	3896.183927	815.286052	12.0	3605.00	3728.0	4540.00	5894.0
59	5480.0	6906.233942	1884.227532	1.0	6128.00	7076.0	8556.00	9199.0
60	1842.0	2979.185125	1469.537404	1.0	1739.00	2962.0	4599.75	4741.0
61	3076.0	3222.958062	1723.936627	156.0	1528.00	4227.0	4938.00	5118.0
62	3197.0	3389.920238	1584.743886	0.0	1758.00	3889.0	4870.00	5075.0
63	3367.0	2530.098010	1383.457540	13.0	1307.00	2441.0	3654.50	4871.0
64	1554.0	7723.525740	2152.623897	1.0	6227.75	7336.0	10258.50	10998.0
65	2538.0	3502.178093	1013.434228	32.0	3155.00	3510.0	4033.00	4837.0
66	3767.0	2801.166976	1542.616251	1.0	1420.50	2877.0	4194.50	5044.0
67	3795.0	4781.238472	952.289926	2.0	4481.00	4824.0	5325.00	7065.0
68	4542.0	5250.971378	1231.040352	1.0	4583.00	4794.5	5766.75	7854.0
69	4430.0	6168.401354	1664.850072	1.0	6153.00	6873.0	6997.00	7176.0
70	5067.0	2805.462996	1411.073821	1.0	1687.50	3158.0	3970.50	4821.0
71	3654.0	3006.959496	1672.700297	356.0	1550.00	2862.0	4434.00	5612.0
72	3230.0	2720.023529	1353.409984	54.0	1421.00	3081.0	3788.75	4704.0
73	3537.0	6922.486288	2776.750328	2.0	3859.00	8156.0	9237.00	10045.0
74	3644.0	5436.829857	1267.130582	1.0	4909.00	5139.0	6173.00	9234.0
75	1897.0	6563.021086	1433.897900	0.0	5351.00	6899.0	7414.00	9904.0
76	5040.0	3807.449206	1056.927388	2.0	2662.75	4041.5	4964.00	5155.0
77	3358.0	2705.784693	1293.185261	526.0	1665.00	2516.5	3859.50	5088.0
78	2643.0	2514.470299	1366.579784	1.0	1391.00	2444.0	3614.00	4757.0
79	1610.0	4871.876398	2025.530615	1.0	4195.00	4452.0	6938.00	7641.0
80	1808.0	4981.872235	933.955832	1.0	5088.00	5186.0	5335.00	5532.0
81	2020.0	6257.996535	1037.874295	0.0	5271.00	6088.0	6789.00	8946.0
82	5133.0	7110.260471	2380.211127	0.0	5144.00	6660.0	9681.00	10360.0
83	2494.0	4259.173617	947.972644	335.0	3398.00	4614.0	5035.50	5331.0
84	2443.0	3367.503479	1382.985224	1.0	2043.00	4116.0	4432.00	5069.0
85	2750.0	4278.440364	904.896251	1.0	4048.00	4421.5	4843.75	5010.0
86	4879.0	3824.903464	1532.248103	8.0	2611.00	3772.0	4966.50	5991.0
87	4088.0	2936.503425	1510.364713	0.0	1227.00	3528.5	4228.00	5348.0
88	5650.0	5824.181947	1027.341924	0.0	5302.00	5792.0	6658.00	7490.0
89	572.0	4293.159091	1655.690600	4.0	4131.00	5080.0	5278.00	5601.0
90	2709.0	6020.437431	1301.742315	1.0	5661.00	6053.0	6398.00	7585.0
91	3511.0	2560.977784	1454.238305	1.0	1182.00	2567.0	3810.00	4717.0
92	1737.0	2773.435233	1356.589031	105.0	1773.00	2022.0	4795.00	5808.0
93	2667.0	6027.521185	915.171944	1.0	5108.00	6277.0	6921.00	7268.0
94	2071.0	4426.070014	782.343645	0.0	4204.50	4653.0	4904.00	5172.0
95	2476.0	2474.158320	1371.333963	2.0	1306.75	2763.0	3604.25	4715.0
96	2586.0	3547.612916	1285.963062	959.0	2346.00	4185.0	4550.00	5530.0
97	830.0	1921.779518	1155.790023	22.0	1271.00	1622.5	2442.75	4528.0
98	4667.0	3635.501821	1842.211811	3.0	2688.50	4385.0	5335.00	5623.0
99	1018.0	3983.044204	1893.990968	0.0	2781.50	4943.0	5185.75	5571.0
100	3777.0	2802.897802	1297.869644	23.0	2137.00	2819.0	3796.00	4868.0
101	3500.0	5462.075429	1004.287020	1.0	5090.00	5417.0	5789.25	7620.0
102	4101.0	3450.775908	523.819859	1.0	3441.00	3538.0	3738.00	3854.0
103	4542.0	2838.617129	1468.247988	8.0	1593.25	2989.0	3972.75	5062.0
104	538.0	3086.804833	1247.032995	3.0	1892.00	3054.0	4264.00	5030.0
105	573.0	4075.483421	2218.747080	4.0	2325.00	4487.0	5914.00	6939.0
106	3911.0	3275.472002	908.970931	4.0	2506.50	2805.0	3875.50	4966.0
107	3836.0	2644.474192	1578.056766	22.0	1441.00	2537.0	3893.00	5118.0
108	2971.0	3568.539886	1112.135590	3.0	2333.00	3572.0	4730.00	4918.0
109	704.0	6724.599432	3779.005358	6.0	3092.00	8351.0	10196.00	10980.0
110	4505.0	5400.376915	1064.960587	2.0	5211.00	5521.0	6126.00	6696.0
111	2317.0	3329.798446	2041.488200	1.0	694.00	4617.0	4836.00	6287.0
112	1778.0	6651.953318	2373.120165	332.0	5299.00	8170.0	8409.00	8648.0
113	942.0	570.581741	327.902482	50.0	217.25	568.0	881.75	1043.0
114	3338.0	2632.428700	1320.677127	212.0	1675.50	2639.0	3770.75	4738.0

Continued on next page

Table A.3 – *Continued from previous page*

Topic	count	mean	std	min	25%	50%	75%	max
115	4045.0	2637.431397	1479.397999	124.0	1231.00	2610.0	4227.00	4954.0
116	489.0	1353.815951	803.283601	10.0	639.00	1403.0	2086.00	2603.0
117	5486.0	5446.263215	617.461267	1.0	5089.00	5397.0	5842.00	6438.0
118	941.0	4739.578108	1564.436126	35.0	4841.00	5366.0	5580.00	5982.0
119	3755.0	5561.114514	955.069507	2.0	4985.00	5157.0	6651.50	7191.0
120	4069.0	2776.625461	1480.697604	3.0	1161.00	2694.0	4140.00	4942.0
121	2359.0	3760.593048	814.710893	185.0	3461.00	3874.0	4319.50	4933.0
122	480.0	1032.216667	387.695222	3.0	728.00	1198.0	1335.00	1616.0
123	249.0	2257.650602	1047.849385	168.0	1562.00	2899.0	3073.00	3222.0
124	1822.0	2582.522503	852.225093	389.0	1560.25	2975.0	3108.00	4908.0
125	4211.0	2809.039183	1390.869322	3.0	1683.00	2802.0	3986.00	5136.0
126	901.0	2324.406215	1243.283748	5.0	1216.00	2312.0	3517.00	4414.0
127	4212.0	5498.300095	940.127544	0.0	5281.75	5608.0	6255.00	7012.0
128	2473.0	4138.234937	1004.626199	328.0	3666.00	4065.0	4504.00	10174.0
129	713.0	2233.302945	1401.402532	119.0	1088.00	2131.0	3512.00	4625.0
130	1513.0	4404.443490	2134.922703	1.0	1772.00	5890.0	6174.00	6298.0
131	485.0	5990.352577	3025.137085	3.0	2754.00	6714.0	8838.00	9595.0
132	2959.0	5543.043934	1272.447543	2.0	5091.00	5948.0	6367.00	6649.0
133	2527.0	3015.432133	1114.956233	1.0	1992.00	2556.0	3902.00	4772.0
134	3270.0	3675.474924	803.506851	272.0	3229.00	3685.0	3936.00	5435.0
135	3842.0	3806.126497	1924.321755	3.0	2026.00	5347.0	5490.00	5598.0
136	601.0	4945.108153	2544.661231	3.0	2928.00	5787.0	7311.00	8260.0
137	3767.0	2477.571808	1436.777745	3.0	1159.00	2377.0	3598.50	5162.0
138	820.0	874.535366	319.499242	1.0	537.00	1040.0	1085.00	1367.0
139	3621.0	2404.063518	1292.777626	6.0	1191.00	2418.0	3536.00	4533.0
140	2834.0	4524.151376	1485.018211	1.0	4778.25	5023.0	5134.00	5606.0
141	1293.0	4599.163186	1735.784953	1.0	3479.00	5354.0	5651.00	7069.0
142	3096.0	2519.415698	1456.854173	92.0	1183.25	2512.5	3841.25	4769.0
143	2827.0	3990.845065	725.135417	0.0	3636.00	4070.0	4494.00	4967.0
144	3976.0	7054.964537	2086.498608	1.0	5570.00	7068.0	9049.25	10412.0
145	3559.0	2980.251194	1300.141902	0.0	1853.00	2464.0	4346.00	5099.0
146	396.0	656.265152	316.271015	50.0	386.25	645.5	909.25	1180.0
147	934.0	2065.746253	1287.042848	238.0	1097.25	1687.0	3292.25	4604.0
148	3803.0	2609.488825	1483.669260	0.0	1368.50	2484.0	3875.50	4869.0
149	2716.0	5804.815906	869.538161	1.0	5340.00	5786.5	6013.25	9026.0
150	6374.0	6790.545968	1889.033717	1.0	5395.00	6007.0	7797.75	11513.0

A.4 Different Statistics Measure for Total Paragraph of 150 Essays

Table A.4: Different Statistics Measure for Total Paragraph of 150 Essays

Topic	count	mean	std	min	25%	50%	75%	max
1	3288.0	64.459854	14.632967	1.0	57.00	65.0	78.00	82.0
2	3151.0	64.888289	21.329097	1.0	50.00	70.0	84.00	95.0
3	2240.0	61.466964	34.214816	1.0	38.00	58.0	82.00	117.0
4	2507.0	96.121659	46.818168	1.0	47.00	108.0	135.00	137.0
5	3679.0	111.891818	37.075797	0.0	85.00	93.0	159.00	171.0
6	3565.0	56.553717	28.494841	21.0	33.00	50.0	70.00	109.0
7	1167.0	358.755784	276.663469	1.0	123.00	266.0	716.00	806.0
8	2049.0	143.496828	49.700029	1.0	147.00	168.0	175.00	198.0
9	2408.0	144.695598	17.298569	1.0	140.00	143.0	149.00	177.0
10	1015.0	330.391133	80.176136	1.0	294.00	353.0	379.00	437.0
11	6530.0	187.626953	31.770098	1.0	182.00	188.0	201.00	236.0
12	3224.0	29.194169	12.488138	2.0	19.00	28.0	42.00	54.0
13	4693.0	56.392073	28.794894	0.0	30.00	73.0	81.00	103.0
14	3190.0	42.292476	18.212261	3.0	29.00	39.0	61.00	68.0
15	3835.0	30.707692	11.675218	1.0	21.00	32.0	38.00	49.0
16	1615.0	141.138700	29.587320	3.0	134.00	142.0	147.00	192.0
17	4506.0	332.114736	65.137249	1.0	288.00	348.5	392.00	423.0
18	2254.0	231.464064	32.053956	0.0	219.00	230.0	251.00	326.0
19	3657.0	22.435876	12.016460	1.0	12.00	23.0	32.00	40.0
20	2843.0	37.301442	12.516599	1.0	31.00	41.0	46.00	52.0
21	1443.0	199.609148	42.534228	1.0	213.00	216.0	218.00	221.0

Continued on next page

Table A.4 – *Continued from previous page*

Topic	count	mean	std	min	25%	50%	75%	max
22	1298.0	118.175655	27.182895	1.0	109.00	127.0	136.00	160.0
23	1319.0	122.654284	59.594013	1.0	56.00	157.0	174.00	180.0
24	6948.0	63.911917	10.367732	1.0	59.00	65.0	73.00	75.0
25	2940.0	73.621088	13.872382	1.0	61.00	79.0	84.00	94.0
26	3132.0	157.934547	52.802270	1.0	178.00	180.0	182.00	188.0
27	4146.0	315.954414	89.913238	1.0	256.00	291.0	396.00	460.0
28	2903.0	92.190493	38.745194	1.0	70.00	102.0	127.00	149.0
29	3586.0	28.894311	15.150873	1.0	19.00	29.0	43.00	54.0
30	2065.0	281.693462	54.584579	1.0	258.00	303.0	317.00	385.0
31	1032.0	147.475775	60.838415	1.0	139.00	157.0	169.25	234.0
32	2188.0	77.809415	30.057752	20.0	55.00	75.0	110.00	139.0
33	4771.0	52.828967	14.736415	1.0	41.00	57.0	68.00	70.0
34	2316.0	116.252591	45.736154	0.0	106.00	111.0	152.00	167.0
35	2568.0	17.850078	7.629655	2.0	14.00	17.0	24.00	34.0
36	3884.0	47.480947	16.629467	1.0	33.00	49.0	68.00	72.0
37	2196.0	477.057377	75.839925	1.0	485.00	488.0	500.00	548.0
38	2036.0	55.293222	31.670648	7.0	34.00	55.0	93.00	113.0
39	1434.0	287.423291	223.137633	1.0	95.25	210.0	446.00	742.0
40	2806.0	100.631148	32.552654	1.0	72.00	104.0	123.00	158.0
41	792.0	250.688131	156.442862	1.0	83.75	246.5	415.25	481.0
42	1510.0	68.058278	56.826214	0.0	13.00	24.5	126.00	178.0
43	3666.0	30.875341	16.274638	1.0	18.00	29.0	46.00	53.0
44	3006.0	75.464072	31.839331	8.0	32.00	83.0	97.00	111.0
45	5084.0	88.987805	45.631579	1.0	68.00	77.0	139.00	152.0
46	735.0	233.846259	114.928408	1.0	115.00	255.0	345.50	375.0
47	1466.0	155.877899	39.362975	1.0	125.00	156.0	191.00	223.0
48	1732.0	227.458430	103.884162	0.0	160.00	232.0	333.00	415.0
49	1816.0	215.615088	46.599377	1.0	188.00	216.0	253.00	261.0
50	4017.0	263.657954	36.827239	1.0	241.00	274.0	288.00	309.0
51	5720.0	219.876224	32.615031	1.0	202.00	211.0	246.00	266.0
52	2174.0	95.218491	16.182536	14.0	84.00	103.0	104.00	279.0
53	2553.0	240.452017	45.614493	1.0	228.00	254.0	265.00	287.0
54	907.0	30.245865	15.879241	1.0	14.50	32.0	45.00	52.0
55	3449.0	52.033923	25.261986	1.0	27.00	50.0	74.00	94.0
56	1908.0	113.046646	40.363872	4.0	114.75	122.0	143.00	145.0
57	3866.0	26.524573	13.688119	1.0	15.00	28.0	36.00	59.0
58	3061.0	106.396929	16.759786	1.0	97.00	104.0	120.00	141.0
59	5480.0	243.556752	50.488868	1.0	250.00	255.0	259.00	276.0
60	1842.0	94.614007	45.485160	1.0	59.00	87.0	145.00	155.0
61	3076.0	92.073797	45.422126	15.0	44.00	122.5	137.00	146.0
62	3197.0	84.040663	36.981986	0.0	52.00	102.0	118.00	130.0
63	3367.0	25.844966	12.084190	1.0	17.00	24.0	38.00	46.0
64	1554.0	187.213642	20.439080	1.0	184.00	191.0	196.00	209.0
65	2538.0	72.323877	23.909201	1.0	63.00	67.0	84.00	108.0
66	3767.0	28.359437	9.787320	1.0	19.00	32.0	37.00	42.0
67	3795.0	245.251647	51.441044	1.0	217.00	250.0	289.00	371.0
68	4542.0	164.782034	60.301303	1.0	121.00	135.0	209.00	294.0
69	4430.0	169.686907	45.721784	1.0	171.00	188.0	194.00	199.0
70	5067.0	36.816065	12.041009	1.0	31.00	43.0	45.00	51.0
71	3654.0	69.875205	38.862146	7.0	38.00	70.0	104.00	127.0
72	3230.0	28.363158	13.003692	1.0	20.00	32.0	37.00	48.0
73	3537.0	218.850438	73.756685	1.0	141.00	262.0	280.00	290.0
74	3644.0	143.476948	23.626341	1.0	131.00	150.0	154.00	188.0
75	1897.0	201.496574	46.365352	0.0	181.00	213.0	235.00	290.0
76	5040.0	76.440476	17.210003	1.0	59.00	78.5	92.00	101.0
77	3358.0	28.107207	13.758908	4.0	16.00	27.0	42.00	49.0
78	2643.0	33.232312	17.555270	1.0	19.00	30.0	50.00	61.0
79	1610.0	118.819255	45.203775	1.0	103.00	106.0	159.00	182.0
80	1808.0	223.773230	43.180879	1.0	224.00	232.0	241.00	263.0
81	2020.0	195.259406	32.754328	0.0	169.00	192.0	207.00	292.0
82	5133.0	231.827002	49.626588	0.0	204.00	257.0	263.00	282.0
83	2494.0	96.549719	19.561266	4.0	77.00	106.0	112.00	118.0
84	2443.0	98.528449	38.382908	1.0	59.00	126.0	129.00	136.0
85	2750.0	107.430545	21.266641	1.0	106.00	109.0	116.00	126.0
86	4879.0	66.297602	16.041599	1.0	55.00	67.0	72.00	104.0
87	4088.0	54.633317	32.987884	0.0	18.00	64.0	75.00	112.0
88	5650.0	198.458230	52.491247	0.0	173.00	185.0	236.00	324.0
89	572.0	124.090909	45.407578	1.0	133.00	138.0	150.00	166.0
90	2709.0	118.567368	37.317131	1.0	90.00	105.0	151.00	169.0
91	3511.0	20.004842	9.225799	1.0	11.00	20.0	28.00	33.0
92	1737.0	89.500288	34.093564	1.0	65.00	73.0	132.00	152.0

Continued on next page

Table A.4 – *Continued from previous page*

Topic	count	mean	std	min	25%	50%	75%	max
93	2667.0	112.005999	11.398298	1.0	106.00	110.0	114.00	136.0
94	2071.0	160.639305	31.519625	0.0	138.00	156.0	189.00	204.0
95	2476.0	28.152666	13.089873	1.0	18.00	31.0	40.00	46.0
96	2586.0	187.680588	121.786937	22.0	56.00	274.0	299.00	332.0
97	830.0	26.254217	14.103416	1.0	19.00	23.0	31.00	59.0
98	4667.0	130.367045	74.748689	1.0	81.00	174.0	200.00	209.0
99	1018.0	92.327112	44.570738	0.0	42.25	120.0	121.00	138.0
100	3777.0	48.261319	22.894037	1.0	37.00	48.0	69.00	83.0
101	3500.0	190.180000	41.758128	1.0	161.00	172.0	222.00	303.0
102	4101.0	225.726164	38.311042	1.0	217.00	222.0	237.00	280.0
103	4542.0	30.294364	8.151092	1.0	26.00	32.0	35.00	42.0
104	538.0	80.412639	34.977311	1.0	40.00	93.0	110.00	126.0
105	573.0	171.198953	101.413414	1.0	80.00	197.0	274.00	290.0
106	3911.0	90.111480	20.184921	1.0	64.00	96.0	103.00	117.0
107	3836.0	66.740355	36.831217	1.0	35.00	71.0	98.00	120.0
108	2971.0	73.886234	23.929833	1.0	50.00	73.0	95.00	107.0
109	704.0	312.502841	161.097419	1.0	172.00	351.0	452.00	532.0
110	4505.0	306.451498	99.396252	1.0	237.00	332.0	391.00	446.0
111	2317.0	103.504100	65.387252	1.0	10.00	144.0	156.00	173.0
112	1778.0	154.450506	49.159521	9.0	112.00	191.0	191.00	211.0
113	942.0	7.354565	4.040914	1.0	3.00	6.0	12.00	13.0
114	3338.0	30.142001	14.380304	3.0	20.00	31.0	41.00	53.0
115	4045.0	35.389122	18.700775	2.0	17.00	38.0	54.00	66.0
116	489.0	21.586912	16.269664	1.0	9.00	16.0	42.00	45.0
117	5486.0	193.566533	41.738451	1.0	159.00	206.0	221.00	263.0
118	941.0	72.724761	22.709678	1.0	73.00	83.0	85.00	92.0
119	3755.0	257.520107	68.406595	1.0	211.00	226.0	332.00	363.0
120	4069.0	61.459818	40.942800	1.0	12.00	55.0	107.00	116.0
121	2359.0	198.905468	33.300849	11.0	182.00	200.0	226.00	260.0
122	480.0	26.629167	8.010431	1.0	26.00	28.5	32.00	34.0
123	249.0	87.682731	40.867308	2.0	76.00	105.0	117.00	127.0
124	1822.0	136.568606	34.035587	3.0	90.00	156.0	158.00	197.0
125	4211.0	43.542626	20.293178	1.0	25.00	47.0	58.00	76.0
126	901.0	67.320755	18.717677	1.0	62.00	70.0	80.00	83.0
127	4212.0	218.199430	48.079449	0.0	182.00	226.0	249.00	312.0
128	2473.0	130.672058	29.353491	10.0	118.00	125.0	138.00	288.0
129	713.0	65.472651	64.812280	3.0	22.00	39.0	158.00	179.0
130	1513.0	141.335757	65.724943	1.0	79.00	185.0	190.00	224.0
131	485.0	135.989691	51.338348	1.0	99.00	160.0	170.00	192.0
132	2959.0	334.178439	74.680156	1.0	274.00	368.0	394.00	401.0
133	2527.0	48.625643	29.787568	1.0	20.00	42.0	74.00	92.0
134	3270.0	125.588685	33.901436	6.0	94.00	133.0	144.00	190.0
135	3842.0	65.091619	23.390061	1.0	54.00	74.0	86.00	90.0
136	601.0	185.076539	94.990425	1.0	103.00	236.0	261.00	287.0
137	3767.0	25.967613	14.150320	1.0	14.00	24.0	35.00	56.0
138	820.0	52.592683	28.040655	1.0	34.00	38.0	62.00	111.0
139	3621.0	19.034797	8.269822	1.0	12.00	20.0	26.00	34.0
140	2834.0	136.509880	42.938041	1.0	148.00	151.0	153.00	161.0
141	1293.0	170.826759	76.804017	1.0	105.00	227.0	234.00	269.0
142	3096.0	19.393411	11.627482	1.0	8.00	19.0	29.00	37.0
143	2827.0	115.902724	28.534035	0.0	93.00	115.0	139.00	162.0
144	3976.0	185.421781	38.589664	1.0	181.00	192.0	207.00	221.0
145	3559.0	59.902782	36.282782	0.0	27.00	58.0	93.00	119.0
146	396.0	9.906566	2.891079	1.0	9.00	11.0	12.00	12.0
147	934.0	35.980728	26.660496	5.0	16.00	18.0	68.00	82.0
148	3803.0	32.304759	12.059882	0.0	22.00	31.0	42.00	54.0
149	2716.0	93.455817	17.209142	1.0	80.00	85.0	110.00	130.0
150	6374.0	306.336994	110.186194	1.0	239.00	287.0	375.00	600.0