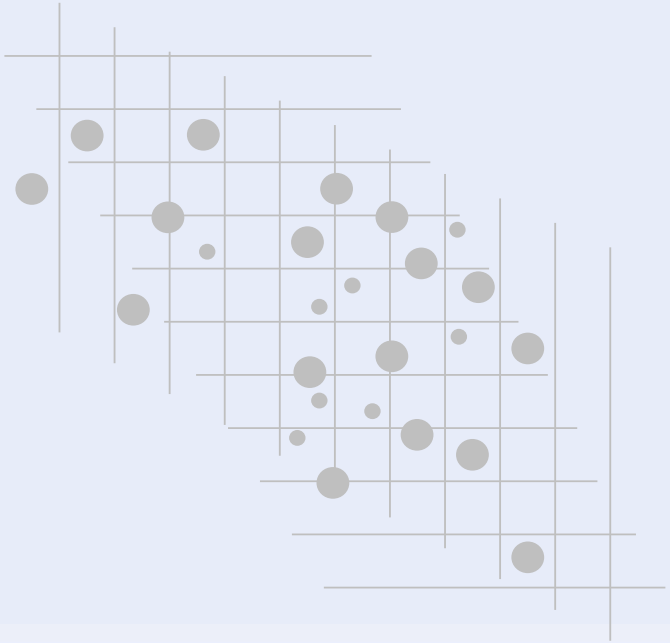
Bauhaus Universität Weimar
Faculty of Media
Degree Programme: M.Sc. Digital Engineering

COHERENCE BASED TEXT QUALITY IN SEARCH-SUPPORTED WRITING

MASTER'S THESIS DEFENCE



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

Supervised By:

Dr. Michael Völske **Dr. Magdalena Wolska**

Presenter:
Bibek khadayat

Date: 14.01.2022

Motivation

1. Numbers of texts, essays, journals, articles, books are written.
2. Every one would like to assess their written text .
3. Numbers of measures for judgement of text quality
4. Coherence is one of the most important factor to be checked in text.

OUTLINE

INTRODUCTION

- Areas of Impact
- Search-Supported writing
- Research Questions

CONCLUSION

- Conclusions
- Future Work

DATA AND DATASETS

- Data Acquisition Setup
- Data Collected
- Prior Analysis

RESULTS AND ANALYSIS

- Different Editing Type
- Effects of Editing Type
- Effects of Writing Styles.

DATA PREPROCESSING

- Data Extraction
- Data Cleaning
- Identification of Major Changes
- Data Analysis

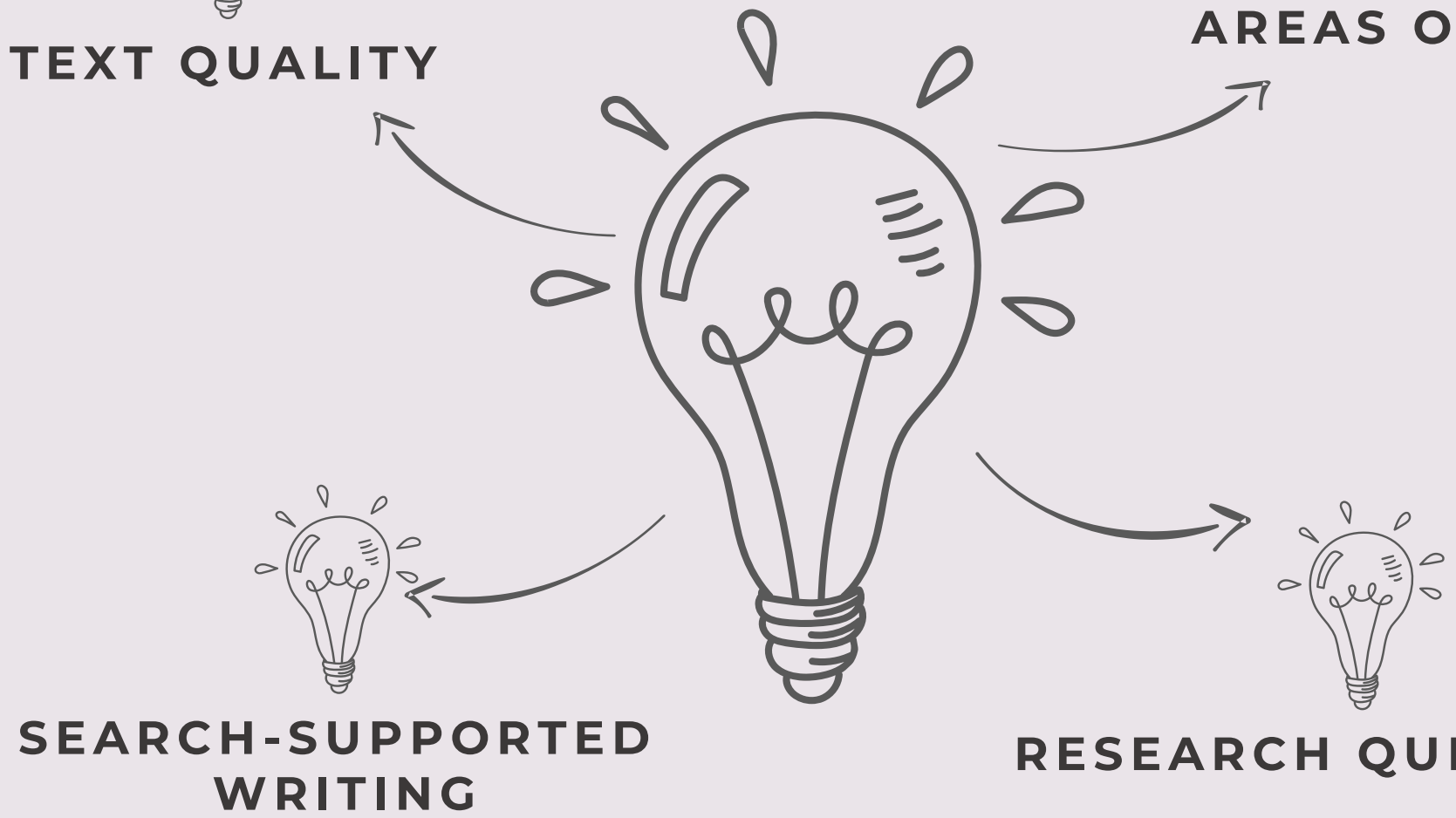
EDITING TYPES

- Examples of Editing Types.
- Automated Detection of Editing Type

TEXT QUALITY MEASURES

- Methodology for Coherence
- Methodology for Type-Token Ratio
- Methodology for Readability

INTRODUCTION



INTRODUCTION

TEXT QUALITY

FACTORS THAT DEFINE TEXT QUALITY

- Spelling.
- Vocabulary.
- Grammar.
- Structure.
- Organization.
- Readability, and so on.

INTRODUCTION

TEXT QUALITY

- COHERENCE
- COHESION
- READABILITY

INTRODUCTION

TEXT QUALITY

Coherent text	Cohesive Text
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.

INTRODUCTION

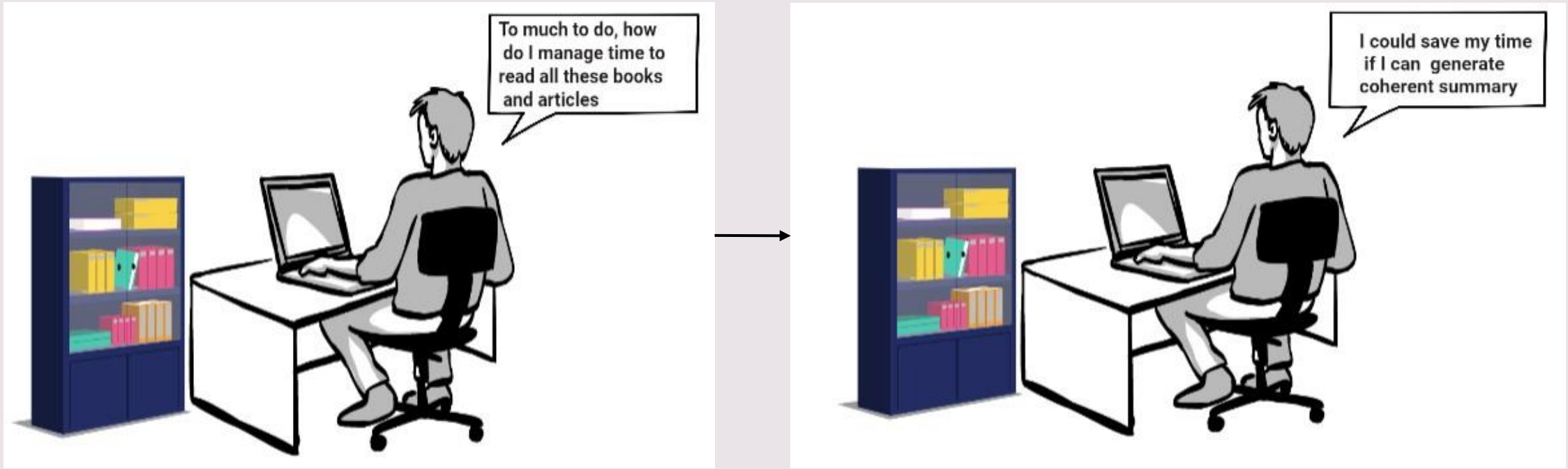
AREAS OF IMPACT

Web search recommendation

INTRODUCTION

AREAS OF IMPACT

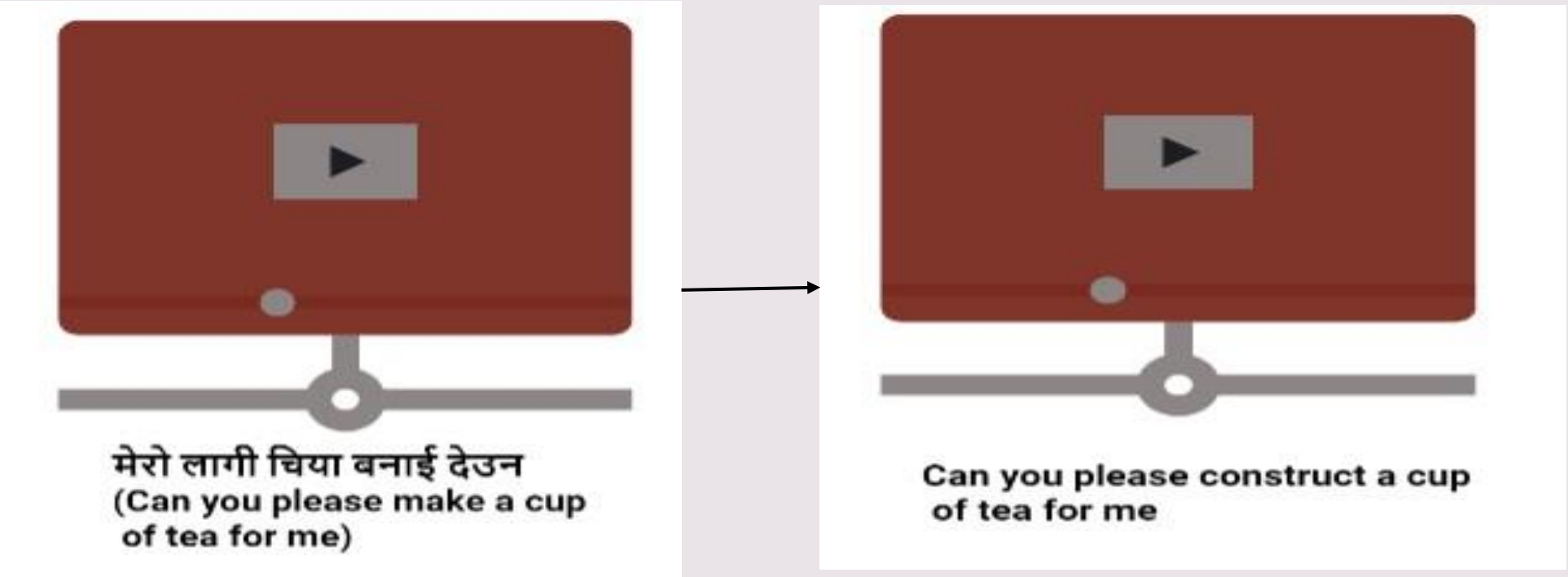
Automatic Summarization



INTRODUCTION

AREAS OF IMPACT

Machine Translation



INTRODUCTION

AREAS OF IMPACT

Writing Assessment

INTRODUCTION

SEARCH-SUPPORTED WRITING

01 Use of Internet for research purpose and gather information.

02 Use this information in writing.

INTRODUCTION

RESEARCH QUESTIONS

RQ1 What are the different types of editing techniques in search-supported writing ?

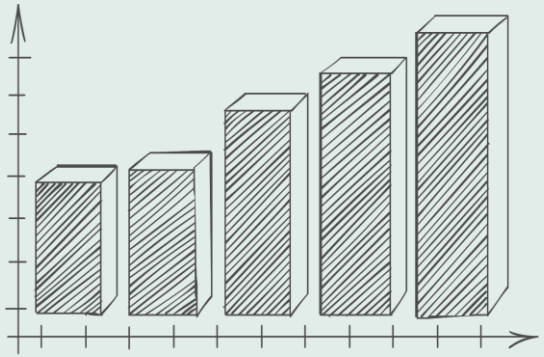
RQ2 How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3 How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

DATA AND DATASETS

DESCRIPTION OF DATASETS

DATA ACQUISITION SETUP



PRIOR ANALYSIS

COLLECTED DATA

DATA AND DATASETS

DESCRIPTION OF DATASETS

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

Text Reuse Corpus 2012 (Webis-TRC-12)

150 long essays written by professional writers.

Information from ClueWeb09 corpus.

Detailed interaction logs and revision history.

Available on webis website

DATA AND DATASETS

DESCRIPTION OF DATASETS

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

DATA AND DATASETS

DATA ACQUISITION SETUP

M. Potthast, M. Hagen, M. Völske, and B. Stein, “CrowdsourcingInteraction Logs to Understand Text Reuse from the Web,” in 51stAnnual Meeting of the Association for Computational Linguistics (ACL2013), P. Fung and M. Poesio, Eds., Association for Computational Linguistics, Aug. 2013, pp. 1212–1221. [Online]. Available: http://www.aclweb.org/anthology/P13-1119.

Editor, Search Engine, and Datasets were provided to writer.

Dataset used: a set of topics, and a set of web pages to search.

ChatNoir is used as search engine

DATA AND DATASETS

COLLECTED DATA

M. Potthast, M. Hagen, M. Völske, and B. Stein, “CrowdsourcingInteraction Logs to Understand Text Reuse from the Web,” in 51stAnnual Meeting of the Association for Computational Linguistics (ACL2013), P. Fung and M. Poesio, Eds., Association for Computational Linguistics, Aug. 2013, pp. 1212–1221. [Online]. Available: http://www.aclweb.org/anthology/P13-1119.

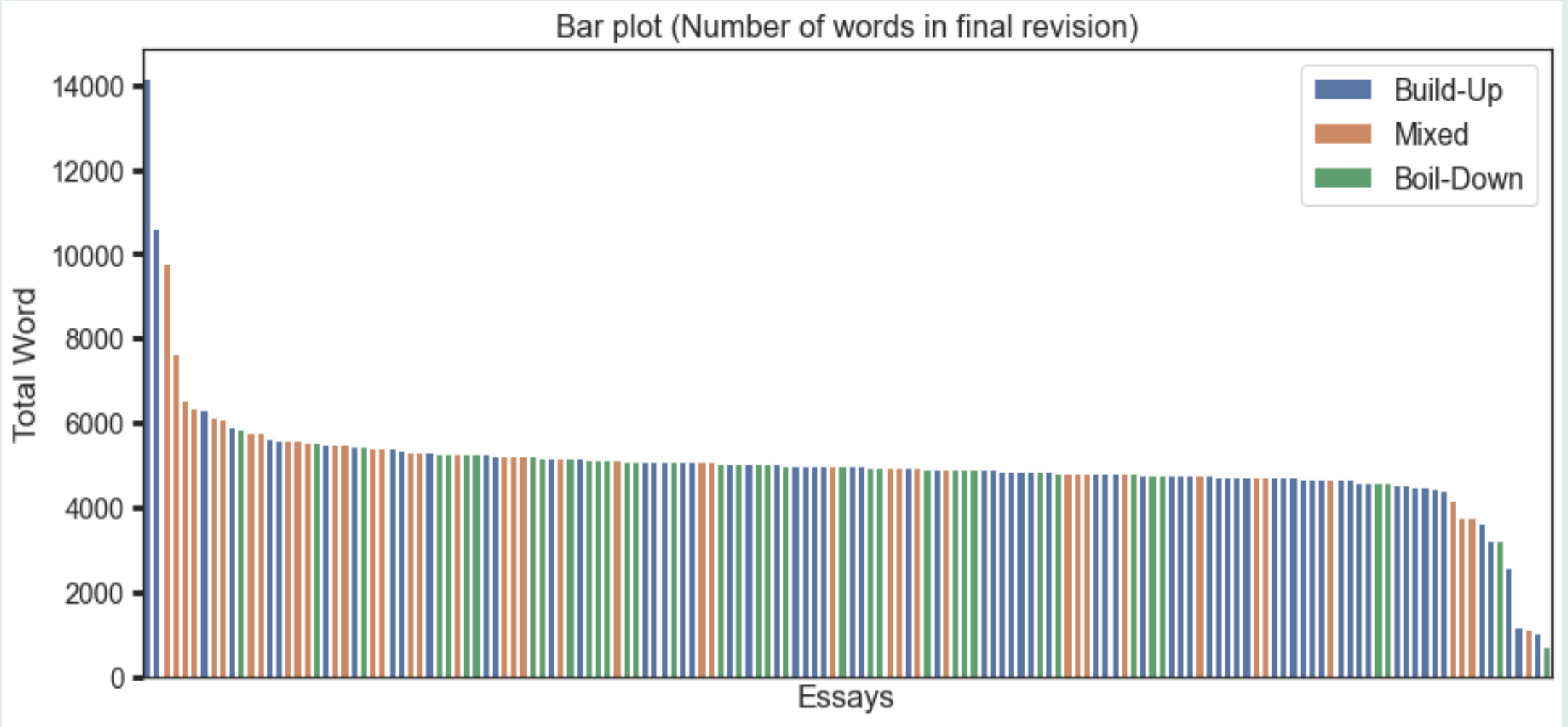
Different revisions created in each essays or topics.

Final revisions of most of the essays are around 5000 words.

Essays: 150
Writer: 12
Revision/Essay: 2827
Words/Final revision: 5040

DATA AND DATASETS

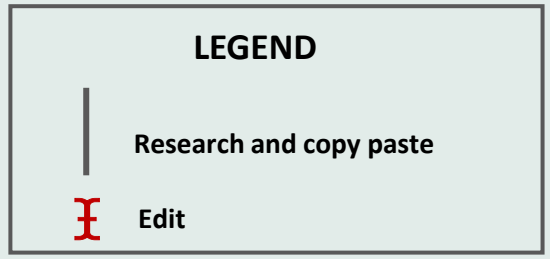
COLLECTED DATA



DATA AND DATASETS

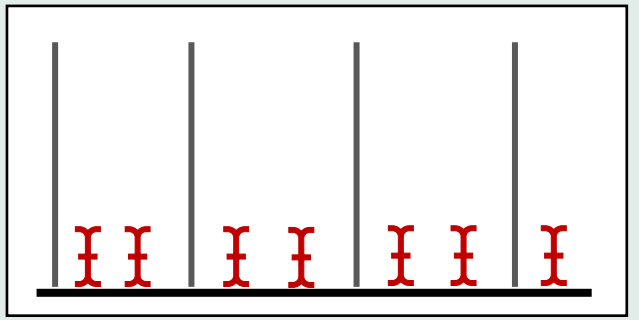
PRIOR ANALYSIS

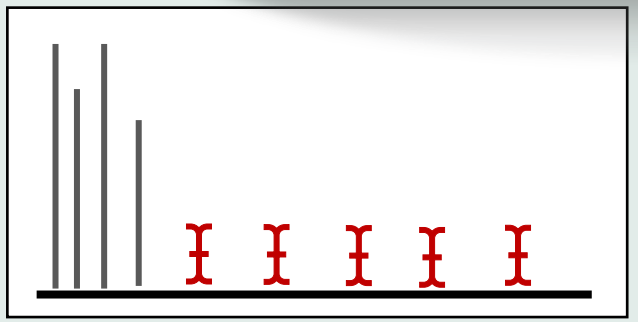
M. Potthast, M. Hagen, M. Völske, and B. Stein, "CrowdsourcingInteraction Logs to Understand Text Reuse from the Web," in 51stAnnual Meeting of the Association for Computational Linguistics (ACL2013), P. Fung and M. Poesio, Eds., Association for Computational Linguistics, Aug. 2013, pp. 1212–1221. [Online]. Available: http://www.aclweb.org/anthology/P13-1119.



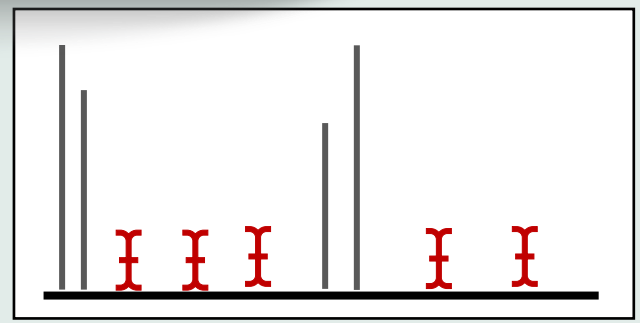
Writing Style

Build-up: continuous lengthening of the essay over the whole writing.



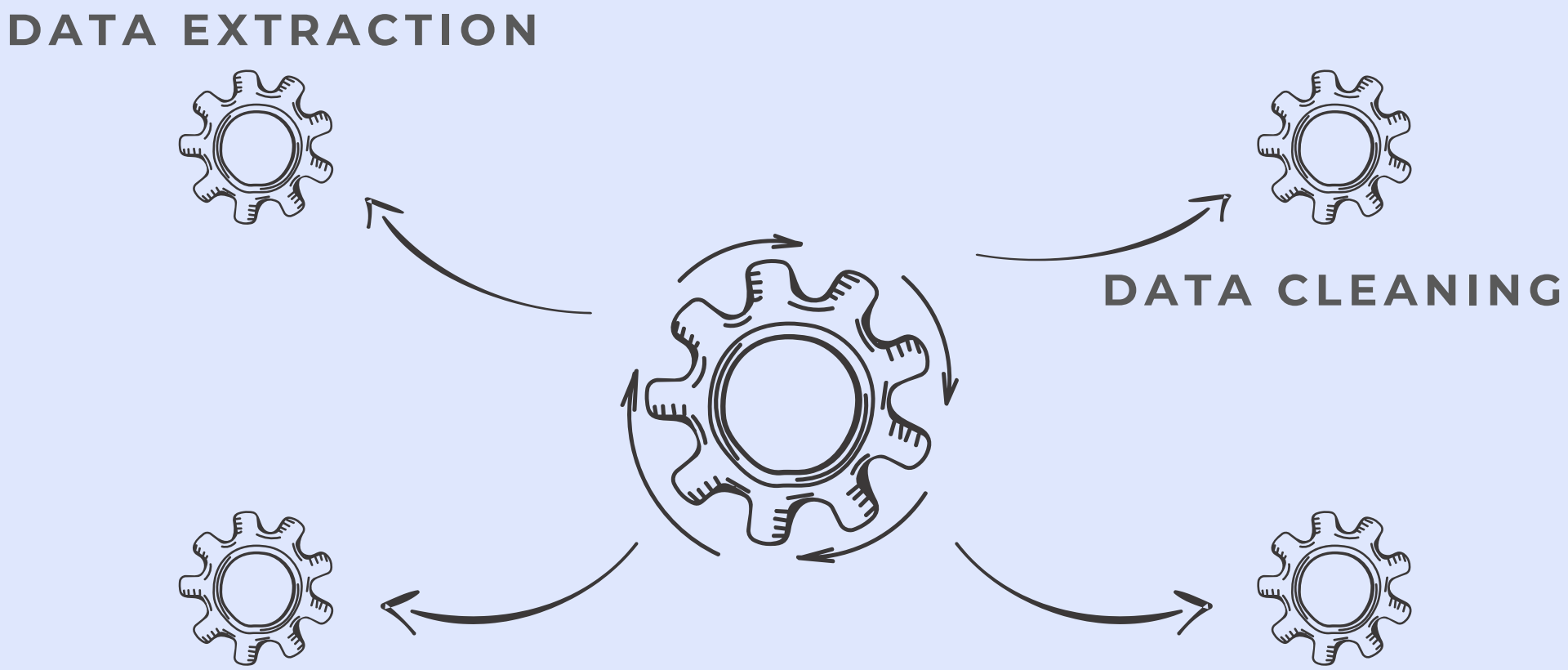


Boil-down: quick early length growth and then shorting



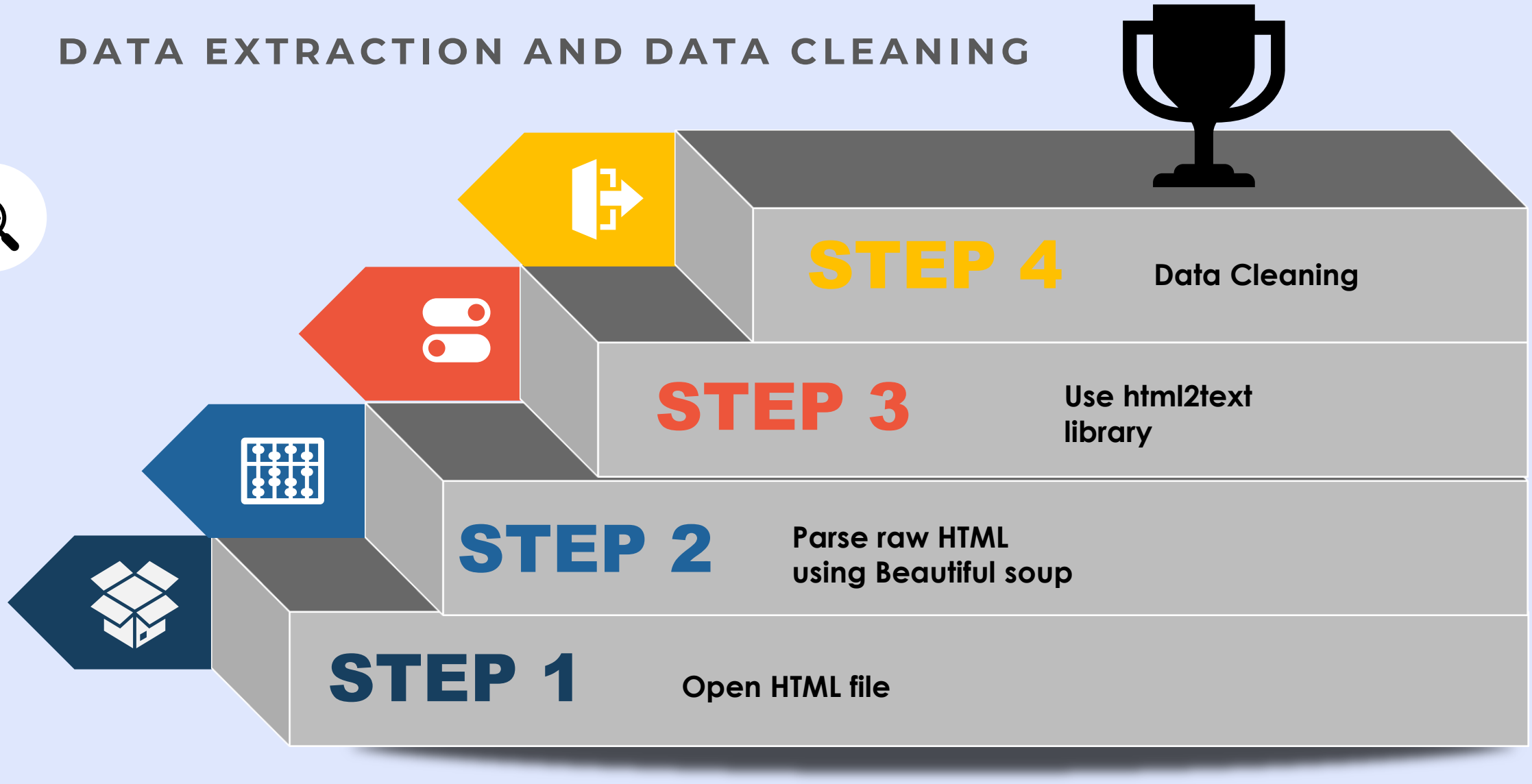
Mixed writing style have both Build-Up and Boil-down aspects.

DATA PREPROCESSING



DATA PREPROCESSING

DATA EXTRACTION AND DATA CLEANING



DATA PREPROCESSING

IDENTIFICATION OF MAJOR CHANGES

01 Too much of data to visualize

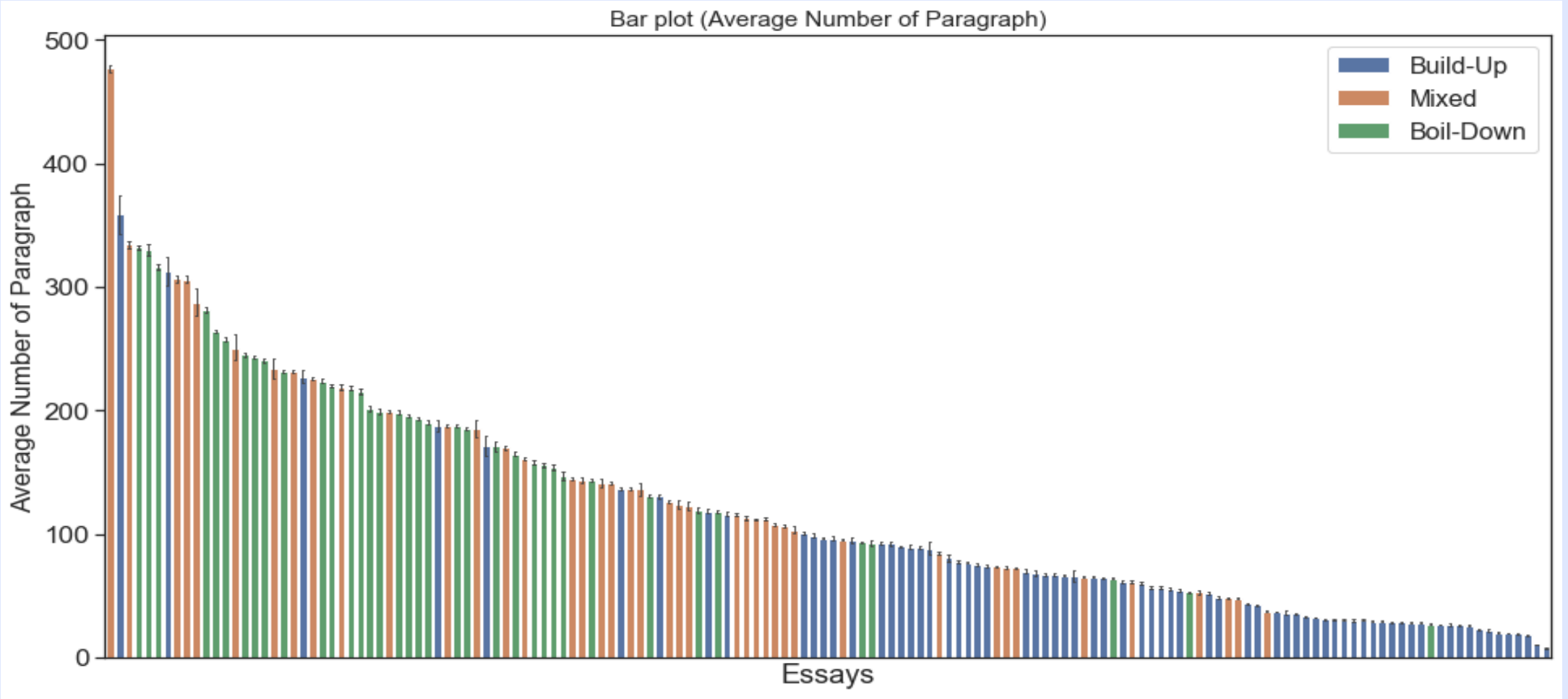
02 Check coherence score across adjacent revision

03 Select only major changes

DATA PREPROCESSING

DATA ANALYSIS

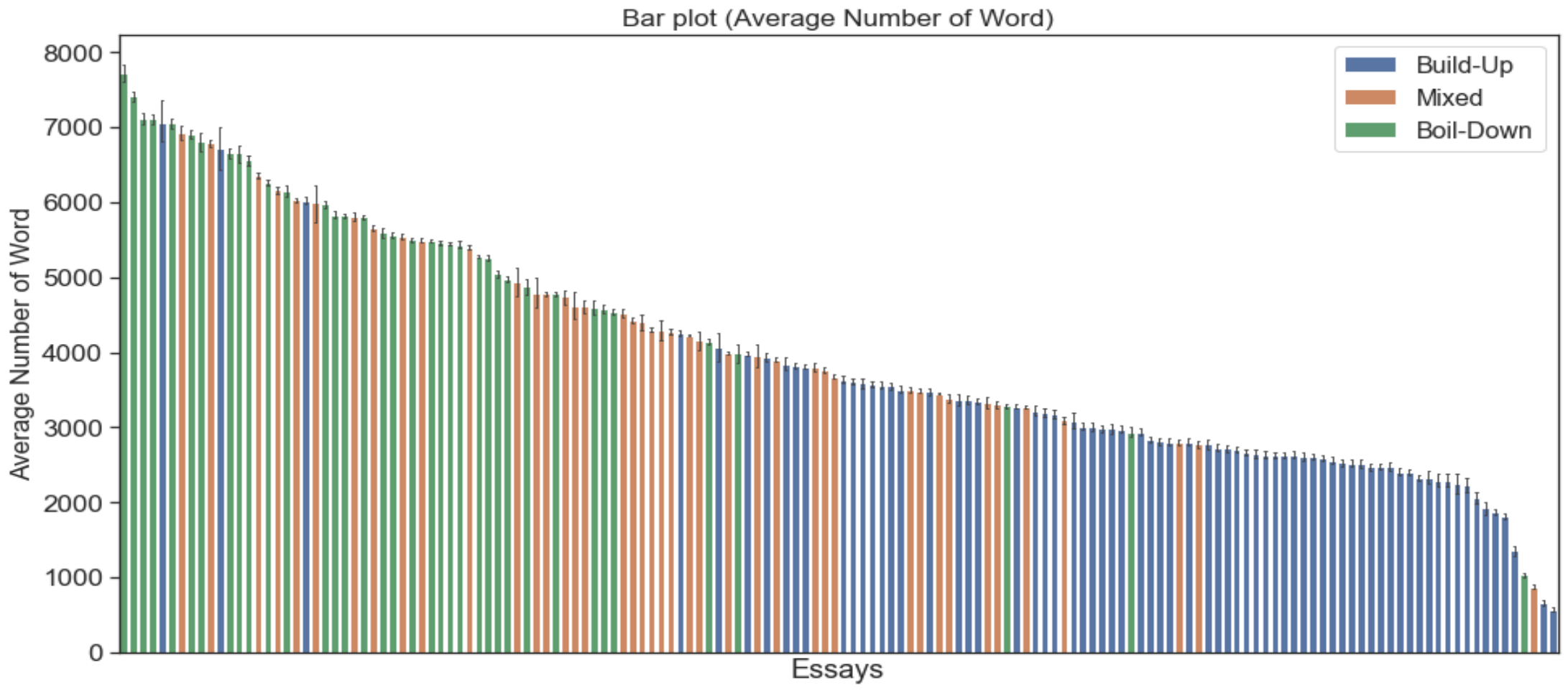
Average number of paragraphs.



DATA PREPROCESSING

DATA ANALYSIS

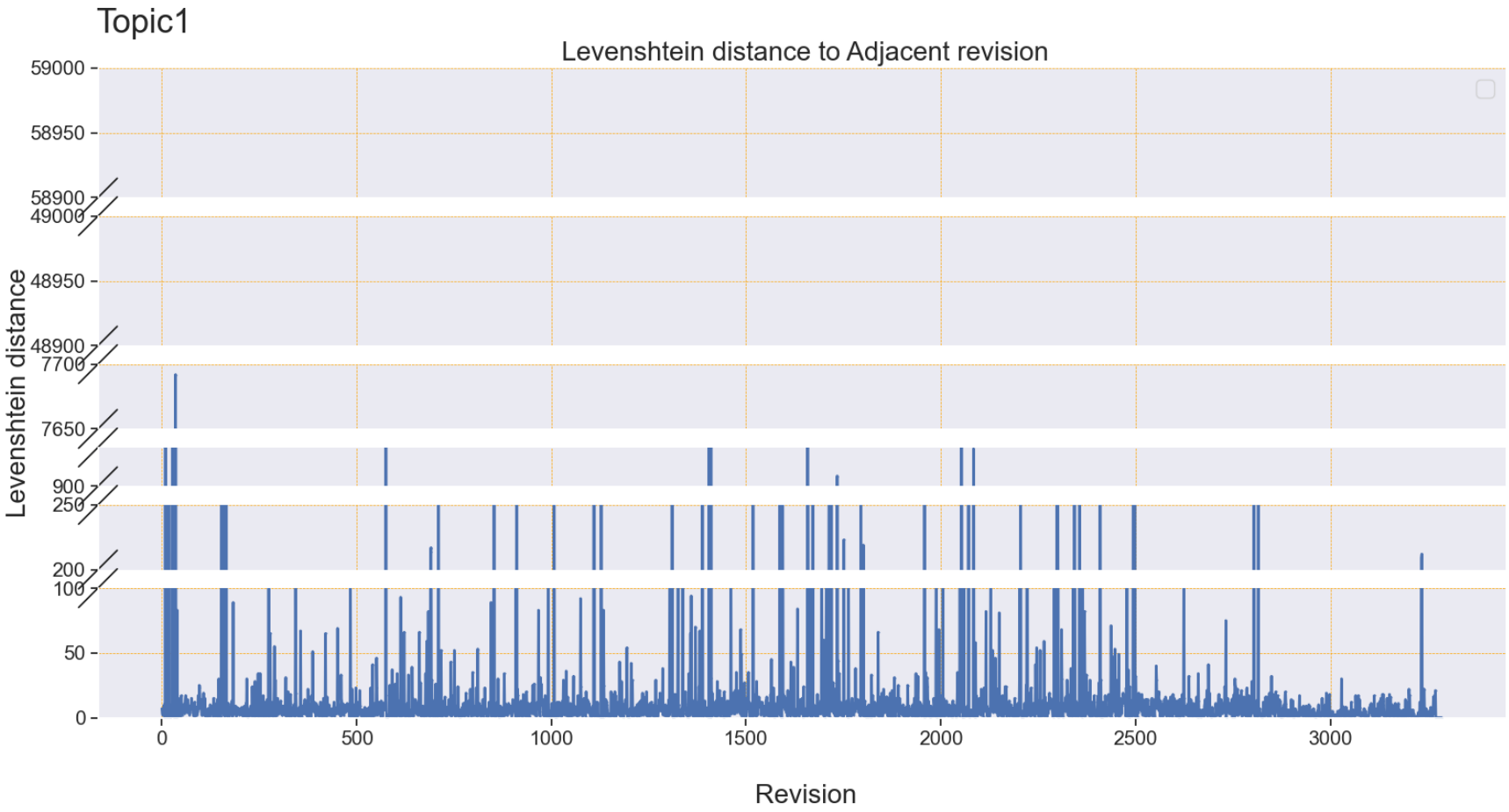
Average number of words.



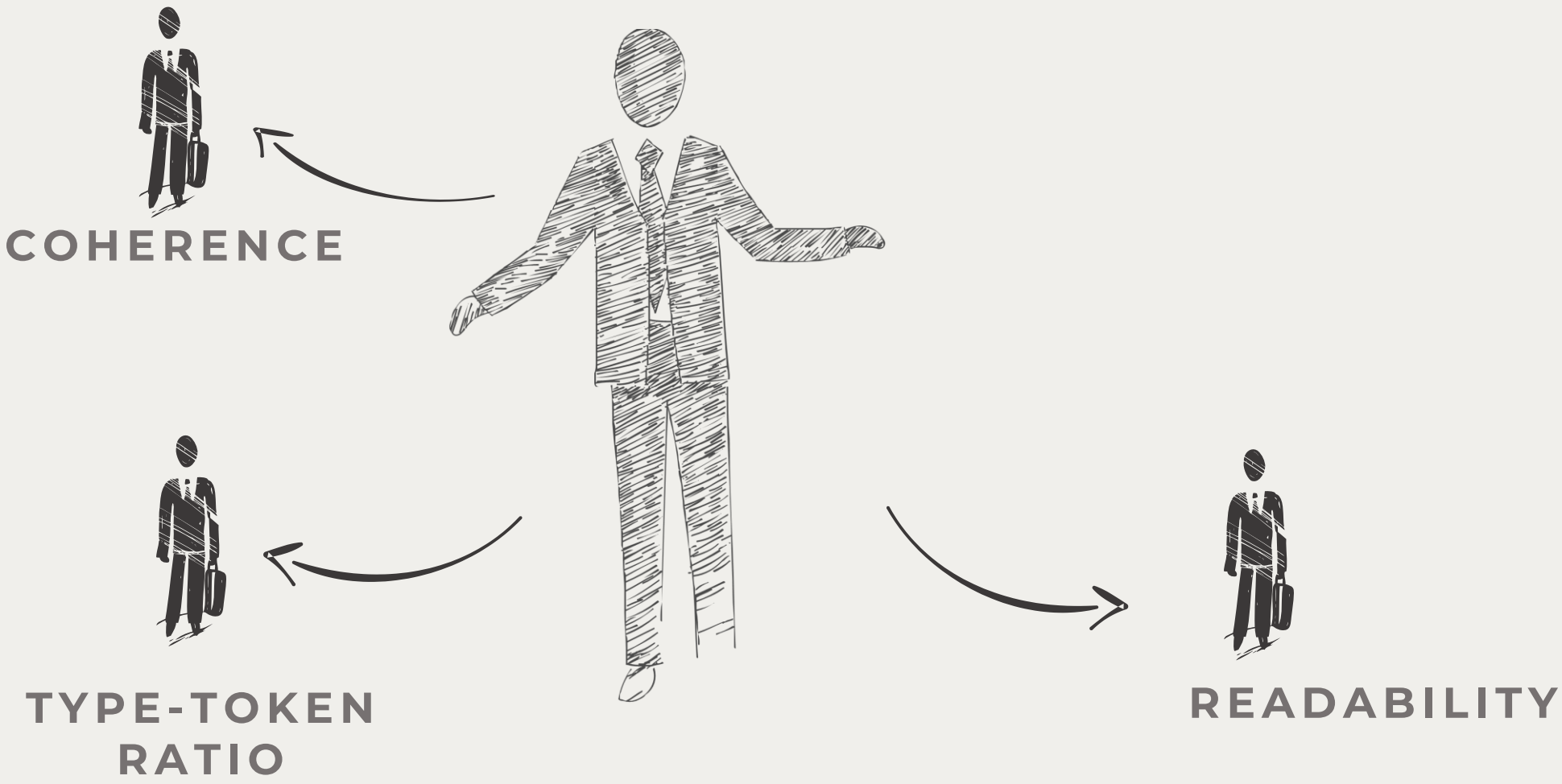
DATA PREPROCESSING

DATA ANALYSIS

The Levenshtein distance.



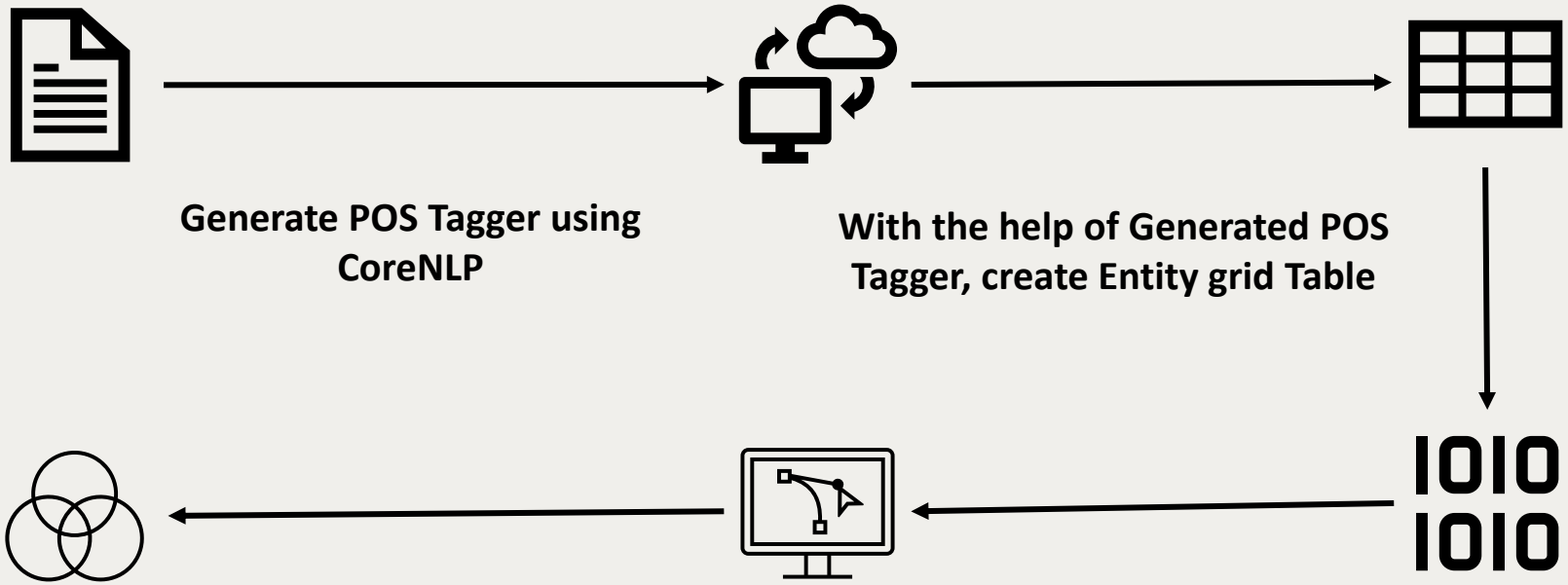
TEXT QUALITY MEASURES



TEXT QUALITY MEASURES

COHERENCE USING ENTITY GRID MODEL

(Modeling Local Coherence: An Entity-based Approach; Regina Barzilay, Mirella Lapata)



TEXT QUALITY MEASURES

READABILITY

- Flesch reading Ease Formula.

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

- Textstat python library.

TEXT QUALITY MEASURES

TYPE-TOKEN RATIO

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


Word        Frequency
at          3
the         3
time        1
of          1
obamas      1
landslide   1
victory     1
poll        1
america     1
was         1
crossroad   1


Type(Unique words)=11--- Total Tokens=15


TTR score       0.7333
```

RESEARCH QUESTION RQ1 SETUP

RQ1 What are the different types of editing techniques in search-supported writing ?

RQ2 How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3 How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

EDITING TYPES

EDITING TYPES

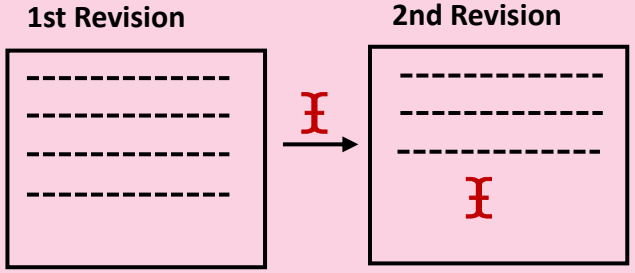
EXAMPLES OF EDITING TYPES

AUTOMATIC IDENTIFICATION OF EDITING TYPES

EDITING TYPES

EDITING TYPES

01 How is the second revision edited compared to the first revision.



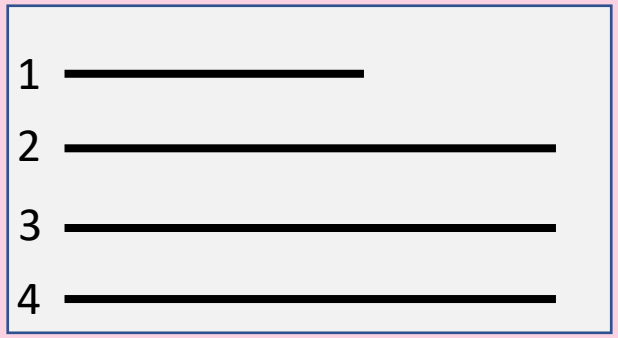
02 Analyse across pairs of adjacent revisions.

03 Identified five different editing types.

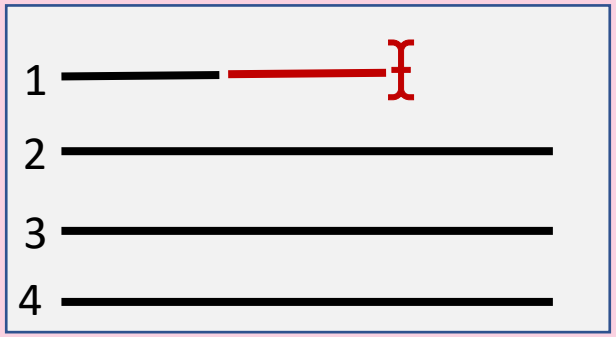
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block Edit



Before



After

French Lick Resort and Casino

Oh Wow! Did you catch the latest news

French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas, we are proud to announce the addition of the French Lick Casino, returning gaming to the Springs Valley for the first time since 1949. And, with the reopening of the luxurious West Baden Springs Hotel, French Lick has truly created the Midwest's premiere resort and casino destination.

French Lick Resort and Casino

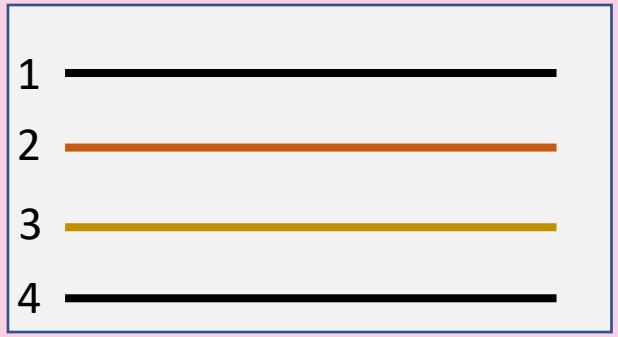
Oh Wow! Did you catch the latest news from Indiana? The State with the motto "The Crossroads of America" is not just a great place to watch motor races like the Indy 500, it's also a great place

French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas, we are proud to announce the addition of the French Lick Casino, returning gaming to the Springs Valley for the first time since 1949. And, with the reopening of the luxurious West Baden Springs Hotel, French Lick has truly created the Midwest's premiere resort and casino destination.

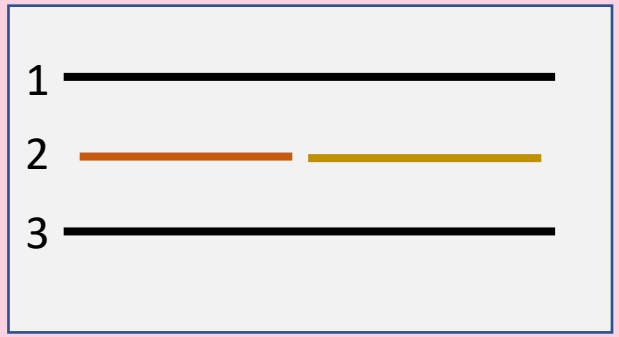
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block Merged



Before



After

Right now (drum roll please) ... Places like Circle City Escorts have every variety of independent escort you could dream of in Indiana. Indiana Escort Referrals recommends Naughtynightlife.com - your free guide to independent escorts, escort agencies and erotic madame and monsieur masseurs. Fancy a blond escort for the night? Escort Service in Circle City can provide the companion of your dreams. Feel like a taste of your own favorite fetish? Heaven 'n Heels everywhere in Indiana has a directory of the most elegant, beautiful and erotic Indiana independent escorts that belong in paradise.

What a fantastic place to locate a Casino. No wonder the designers chose to add a touch of wonderland to the magic state when they built French Lick resort and Casino on

8670 W. State Road 56
French Lick, IN 47432 (Map it)

French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas,

Right now (drum roll please) ... Places like Circle City Escorts have every variety of independent escort you could dream of in Indiana. Indiana Escort Referrals recommends Naughtynightlife.com - your free guide to independent escorts, escort agencies and erotic madame and monsieur masseurs. Fancy a blond escort for the night? Escort Service in Circle City can provide the companion of your dreams. Feel like a taste of your own favorite fetish? Heaven 'n Heels everywhere in Indiana has a directory of the most elegant, beautiful and erotic Indiana independent escorts that belong in paradise.

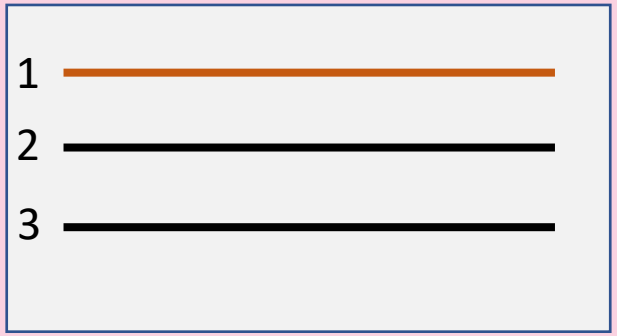
What a fantastic place to locate a Casino. No wonder the designers chose to add a touch of wonderland to the magic state when they built French Lick resort and Casino on 8670 W. State Road 56
French Lick, IN 47432 (Map it)

French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas,

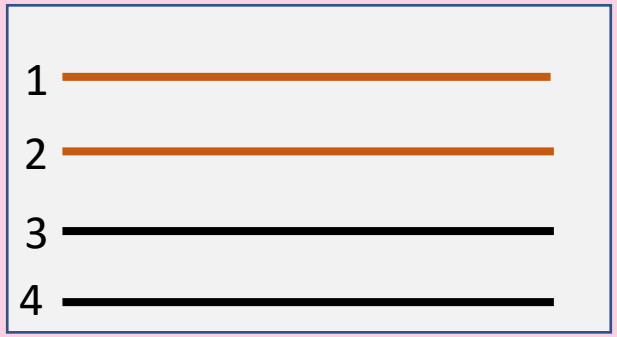
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block Split



Before



After

French Lick Resort and Casino

Oh Wow! Did you catch the latest news French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas, we are proud to announce the addition of the French Lick Casino, returning gaming to the Springs Valley for the first time since 1949. And, with the reopening of the luxurious West Baden Springs Hotel, French Lick has truly created the Midwest's premiere resort and casino destination.

For generations, our beautiful retreat has offered a scenic environment in which to relax and enjoy nature. Guests can stroll shaded walkways and visit the famous gazebo housing the Pluto mineral springs, nestled amidst lush gardens of colorful flowers and carefully trimmed greenery. The shaded walkways provide quiet solitude at mid-day or for an evening stroll. Our manicured grounds also provide an impeccable backdrop for all kinds of events, from weddings and corporate picnics, to family cookouts.

French Lick Resort and Casino

Oh Wow! Did you catch the latest news

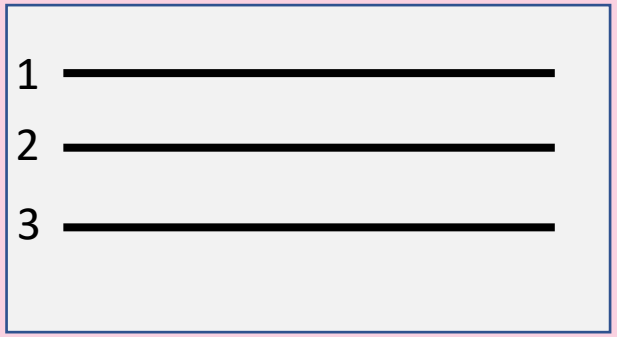
French Lick Resort has embarked on an incredible $500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas, we are proud to announce the addition of the French Lick Casino, returning gaming to the Springs Valley for the first time since 1949. And, with the reopening of the luxurious West Baden Springs Hotel, French Lick has truly created the Midwest's premiere resort and casino destination.

For generations, our beautiful retreat has offered a scenic environment in which to relax and enjoy nature. Guests can stroll shaded walkways and visit the famous gazebo housing the Pluto mineral springs, nestled amidst lush gardens of colorful flowers and carefully trimmed greenery. The shaded walkways provide quiet solitude at mid-day or for an evening stroll. Our manicured grounds also provide an impeccable backdrop for all kinds of events, from weddings and corporate picnics, to family cookouts.

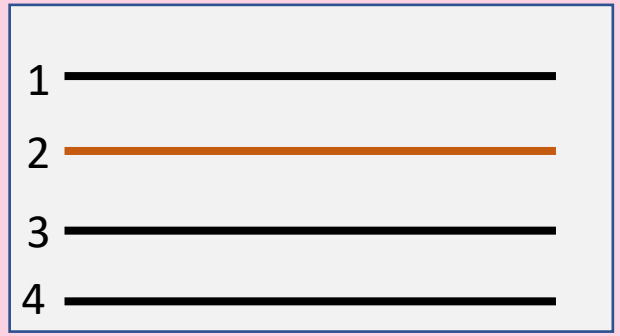
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block Insertion



Before



After

French Lick Resort and Casino.Write an advertising brochure for the French Lick Resort and Casino in Indiana. Interesting things could be the history of the casino, discounted packages for staying at the resort, are there close by other casinos, what could be job opportunities, etc.

French Lick Indiana got its name from early French settlers and the ineral licksin the area. French traders came to the area and discovered the mineral springs bubbling from the ground in the vicinity of what is now French Lick. Wildlife came to lick the mineral deposits left on the ground and rocks.In the early 1800's settlers began to bottle and sell the "Pluto Water" from the springs. In the early 1800's Doc Bowles built the first hotel, a three story frame building. The community thrived and there was an influx of tourist traffic coming to drink and soak in the mineral waters. In the 1850's French Lick was a key station in the "underground railway".The French Lick Springs Resort and Spa was built in the late 1800's. Tom Taggert purchased the property in 1901 and, with the help of the Monon Railroad, the former Indianapolis mayor turned the sleepy little resort into an international attraction. Many hoosiers traveled to French Lick by train. The old train depot remains in downtown French Lick.

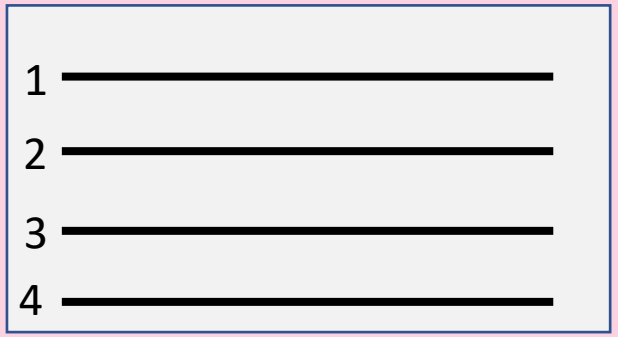
Today French Lick and West Baden remain a favorite hoosier vacation destination along with Brown County Indiana.

French Lick Resort and Casino.Write an advertising brochure for the French Lick Resort and Casino in Indiana. Interesting things could be the history of the casino, discounted packages for staying at the resort, are there close by other casinos, what could be job opportunities, etc.

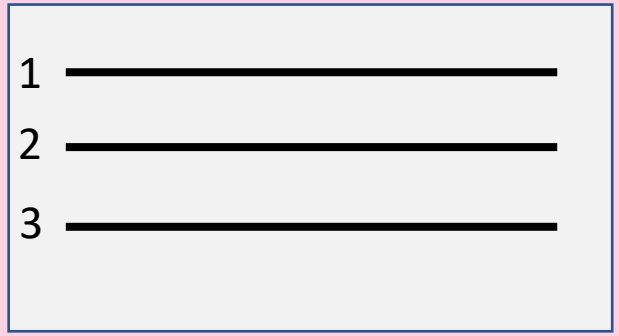
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block Deletion



Before



After

On the other hand (if you prefer)

Historical Sites	Lincoln Boyhood National Memorial, George Rogers Clark National Historical Park, Amish Acres, Conner Prairie Pioneer Settlement, Historic Fort Wayne
Points of Interest	Wyandotte Cave, Indianapolis Motor Speedway, Indiana Dunes, Holiday World, Brown County craft shops

What a fantastic place to locate a Casino. No wonder the designers chose to add a touch of wonderland to the magic state when they added French Lick Resort and Casino at 8670 West State Road, 56 French Lick, Indianapolis, 47432.

On the other hand (if you prefer)

What a fantastic place to locate a Casino. No wonder the designers chose to add a touch of wonderland to the magic state when they added French Lick Resort and Casino at 8670 West State Road, 56 French Lick, Indianapolis, 47432.

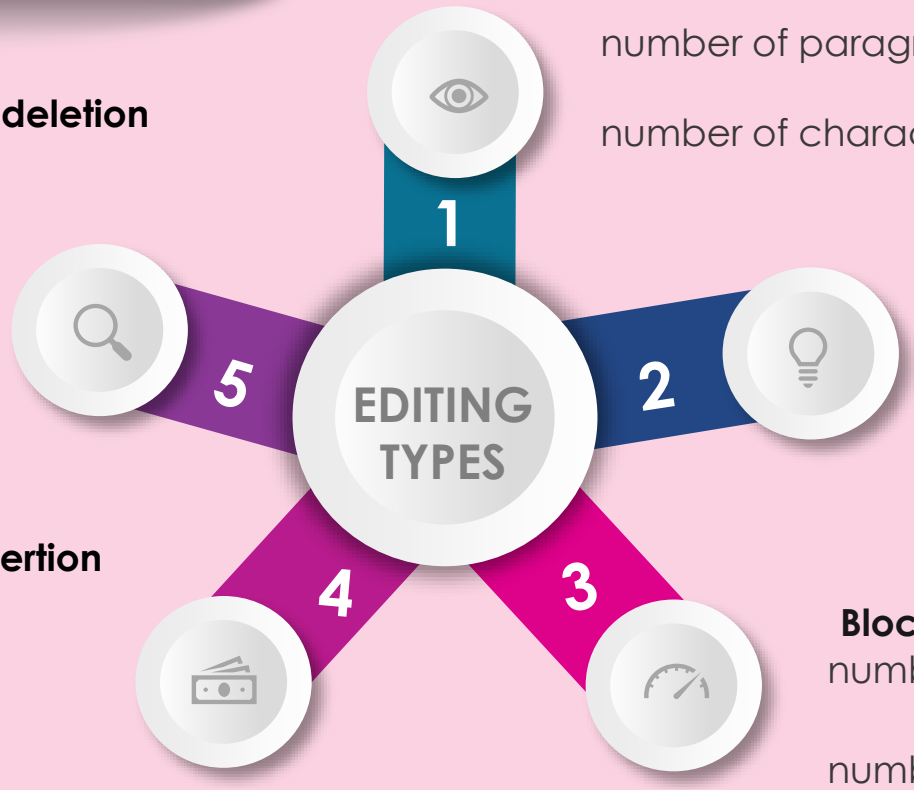
Incredibly, French Lick Resort has now embarked on an absolutely amazing $500 million

EDITING TYPES

AUTOMATIC IDENTIFICATION OF EDITING TYPES

LEGENDS	
=	No Changes
↑↓	Changes
↓	Decrease
↑	Increase

01 Number of characters and paragraphs play a vital role.



Block edited

number of paragraphs =

number of characters ↓ ↑

Block deletion

number of paragraphs ↓

number of characters ↓

Block merged

number of paragraphs ↓

number of characters =

Block insertion

number of paragraphs ↑

number of characters ↑

Block split

number of paragraphs ↑

number of characters =

RESULT AND ANALYSIS

EDITING TYPES ?

EFFECTS OF EDITING TYPES?

EFFECTS OF WRITING STYLES?

RESULT AND ANALYSIS

RESEARCH QUESTION RQ1

RQ1 What are the different types of editing techniques in search-supported writing ?

RQ2 How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3 How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

RESULT AND ANALYSIS

RQ1: DIFFERENT 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

RESULT AND ANALYSIS

RESEARCH QUESTION RQ2

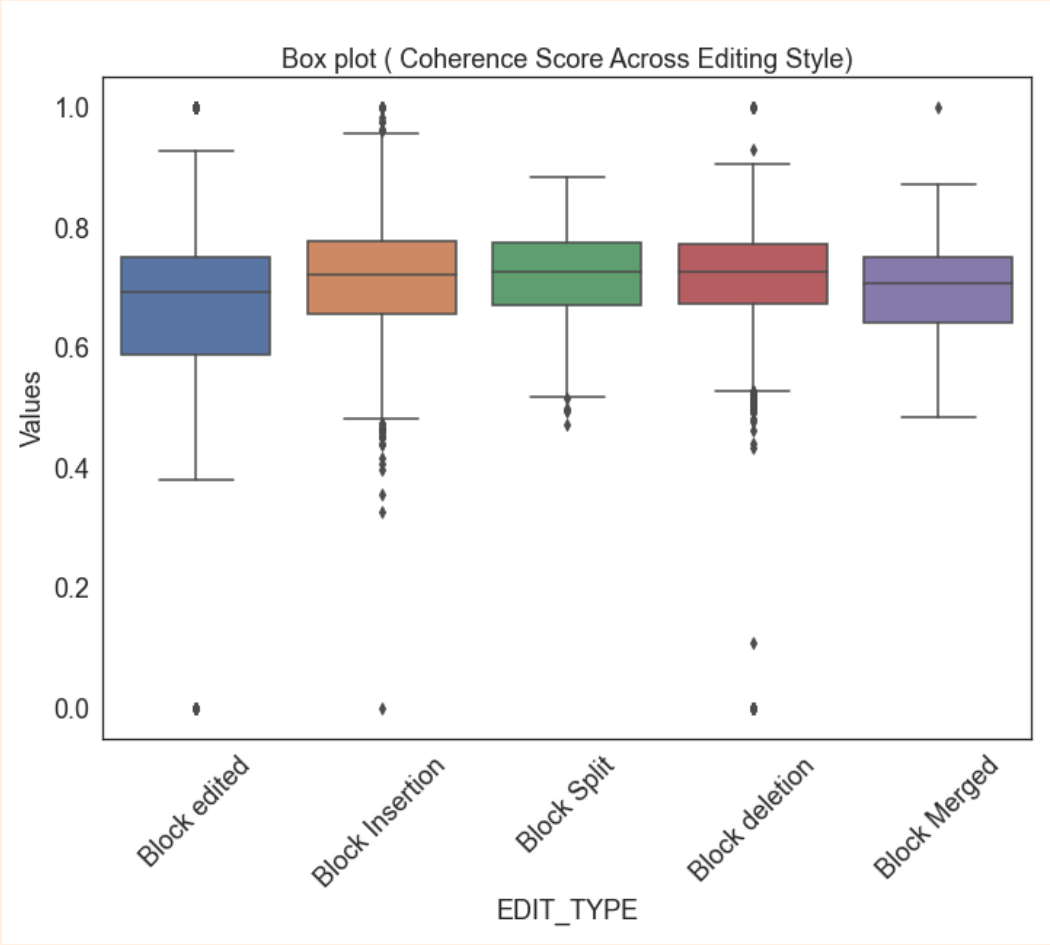
RQ1 What are the different types of editing techniques in search-supported writing ?

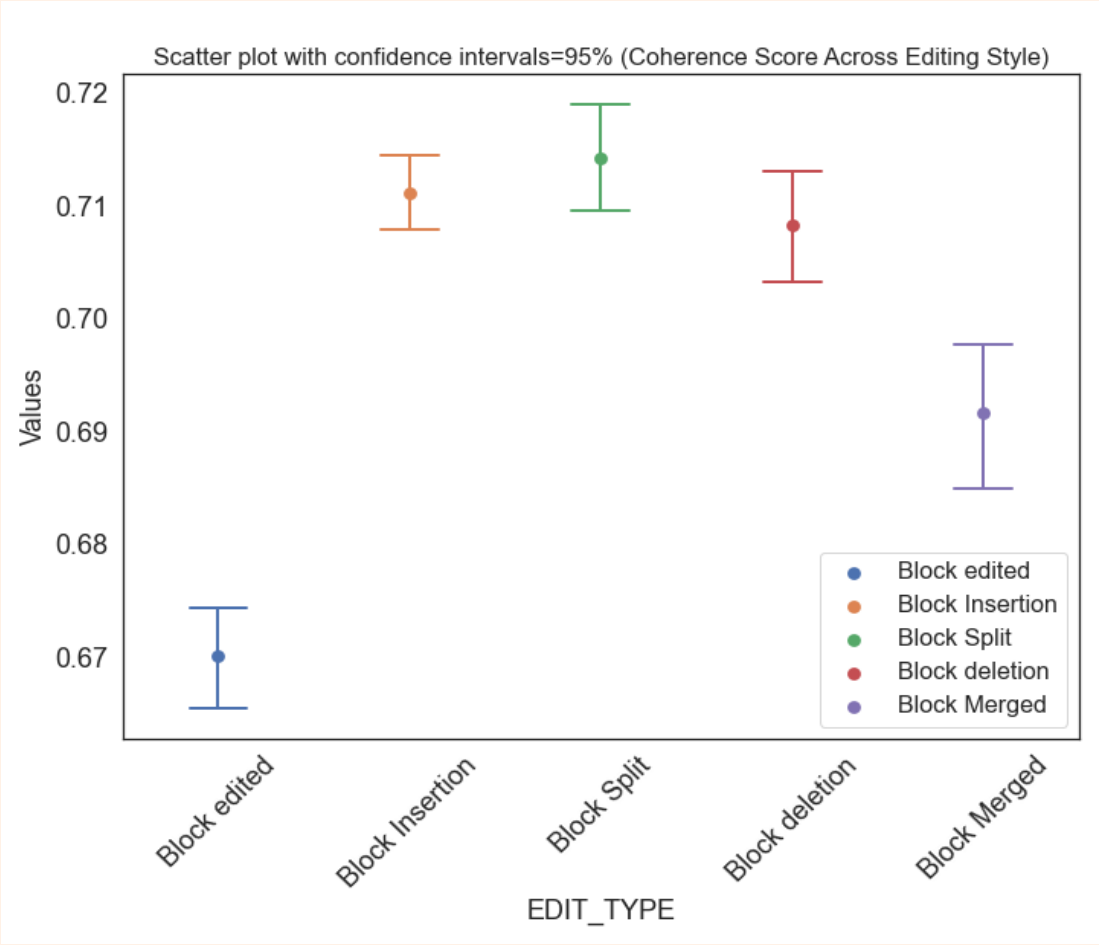
RQ2 How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3 How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

RESULT AND ANALYSIS

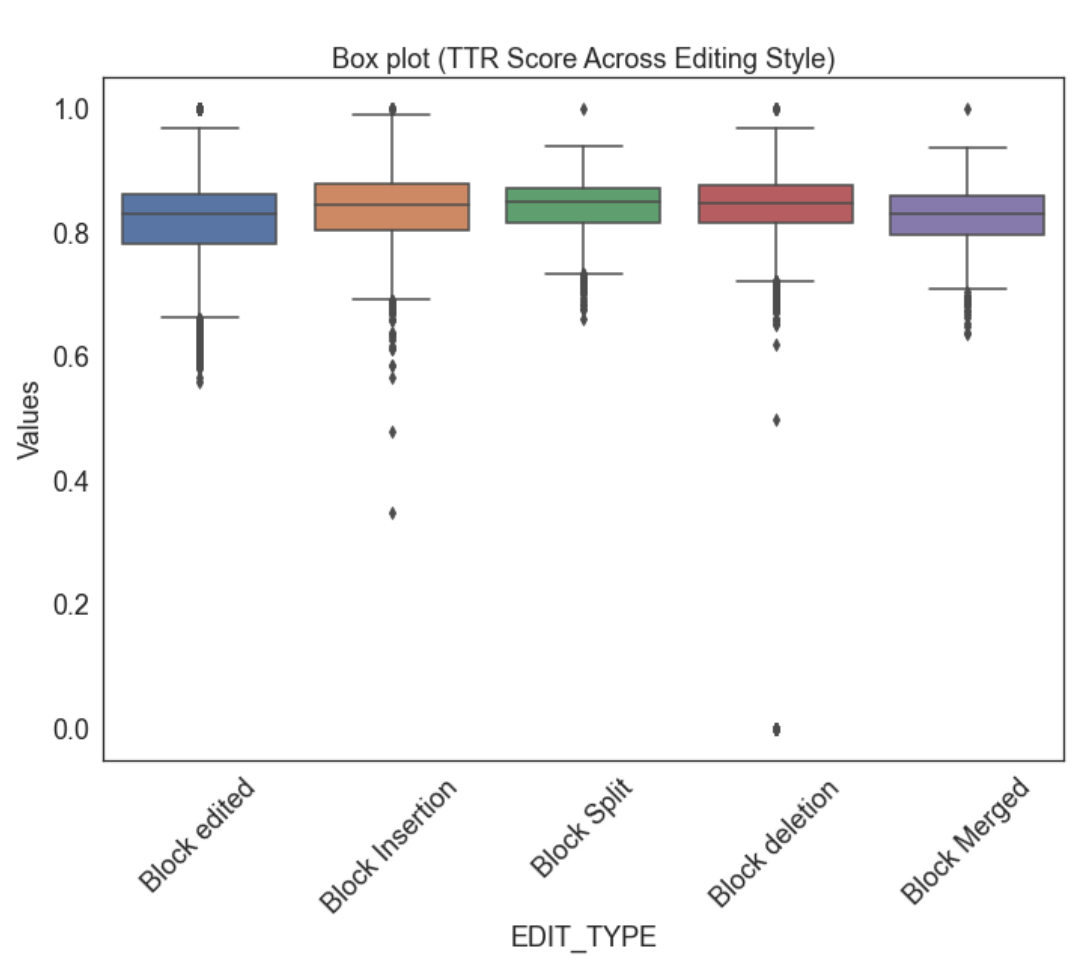
RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY

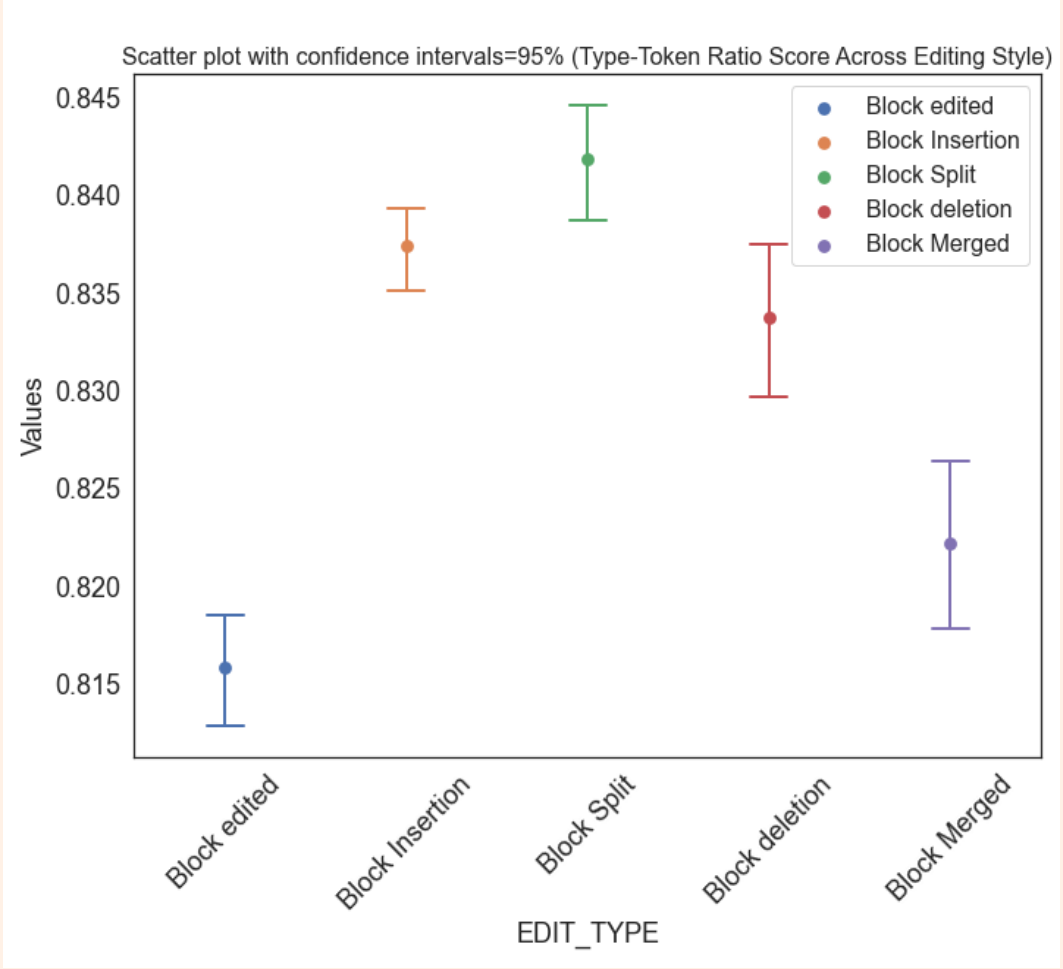




RESULT AND ANALYSIS

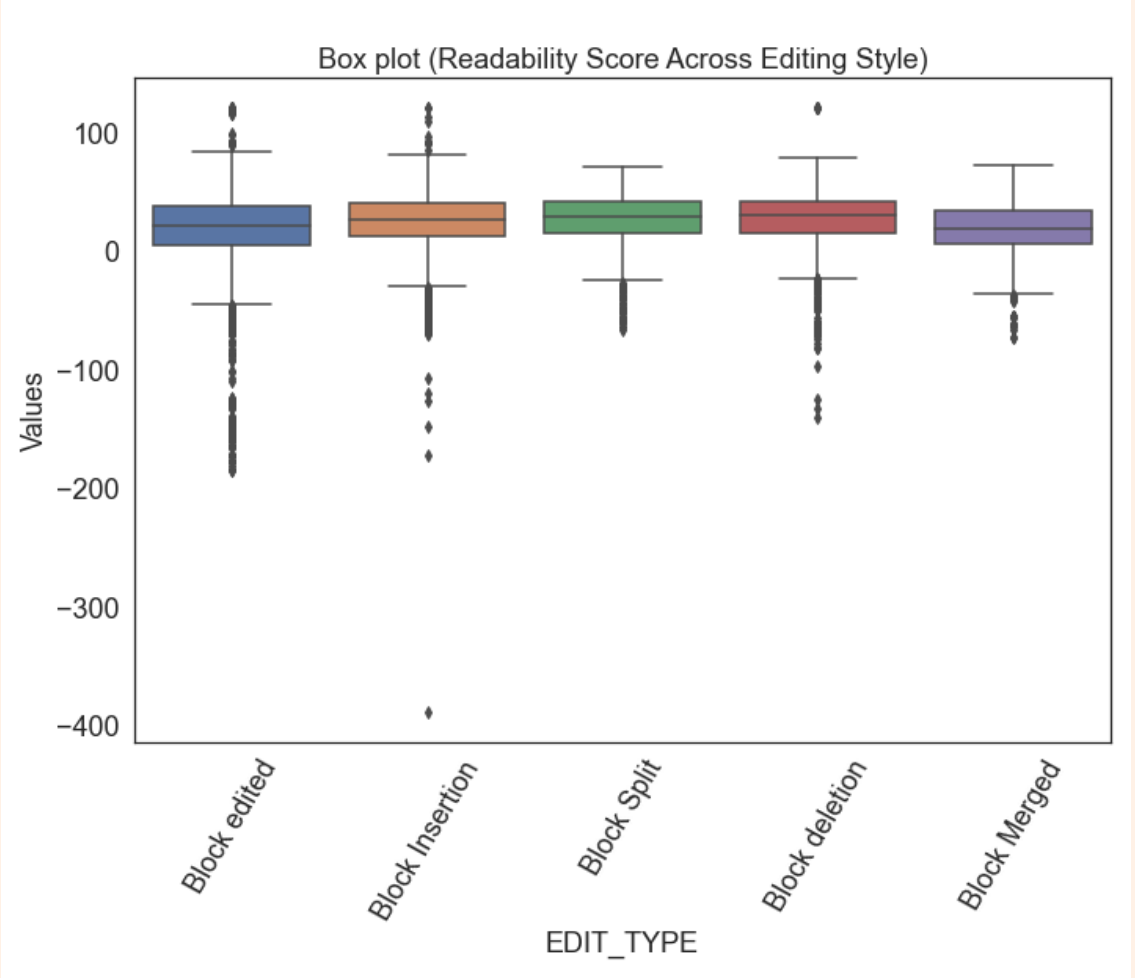
RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY

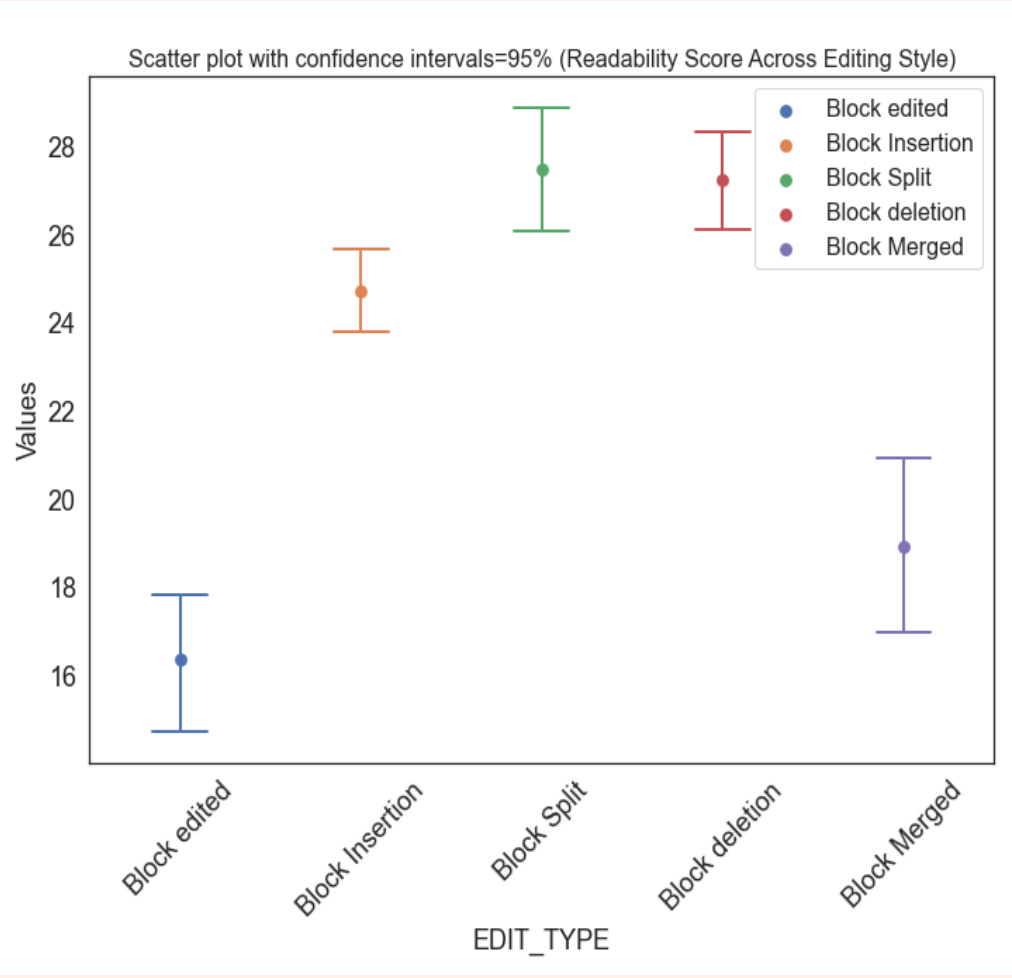




RESULT AND ANALYSIS

RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY





RESULT AND ANALYSIS

RESEARCH QUESTION RQ3

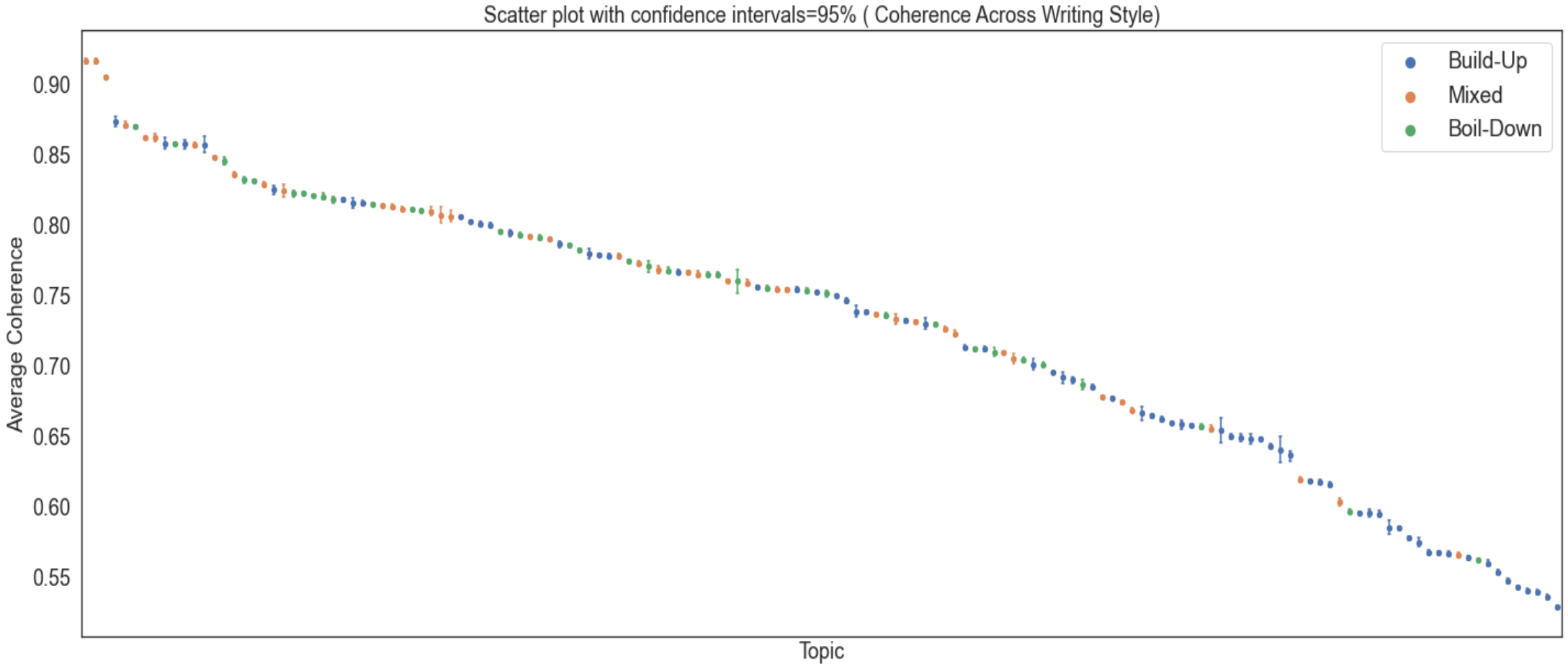
RQ1 What are the different types of editing techniques in search-supported writing ?

RQ2 How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3 How do different essay writing strategies affect Coherence, Type-Token ratio, and Readability?

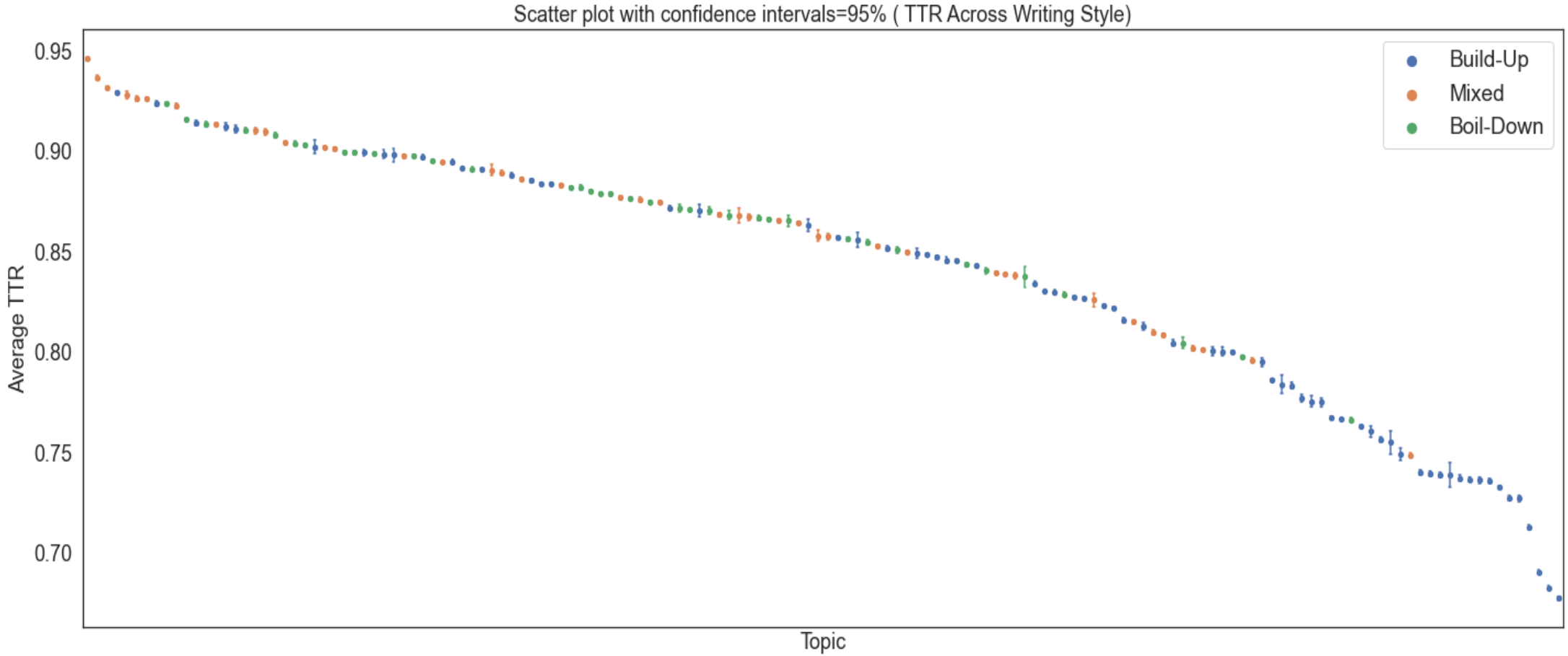
RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



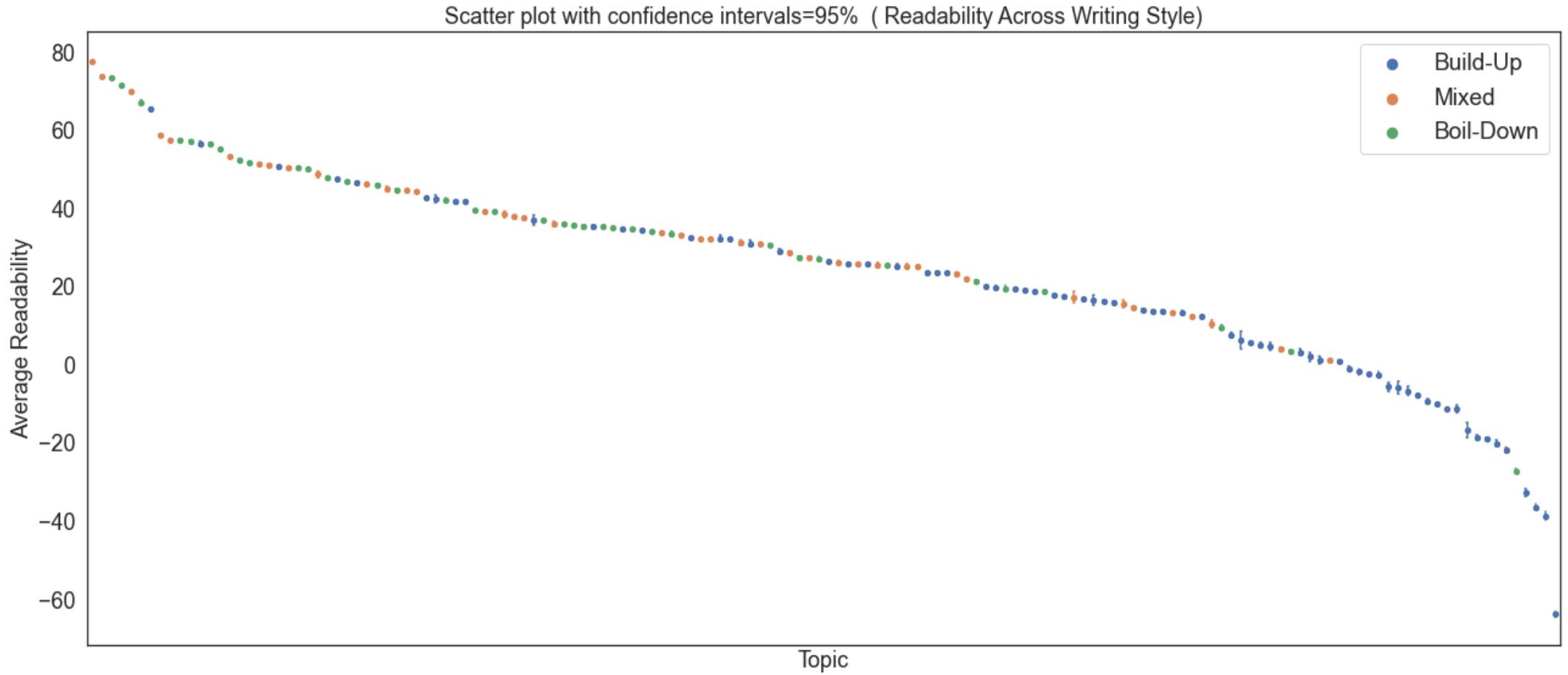
RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



CONCLUSION

MAIN FINDINGS

FUTURE WORK

CONCLUSION

MAIN FINDINGS

RQ1

Different editing types: Block Edited, Block Merged, Block Split, Block Insertion, and Block Deletion.

Editing types affect readability, but Coherence and Type token ratio have same pattern in every editing types.

RQ2

RQ3

Most of the essays with build-up writing are having lower coherence, TTR, and readability scores

Coherence, TTR, and readability are higher in mixed writing

CONCLUSION

FUTURE WORK

01 Investigate the lower coherence, TTR, and Readability in Build-up writing.

02 Investigate effects of introducing new entities in text.

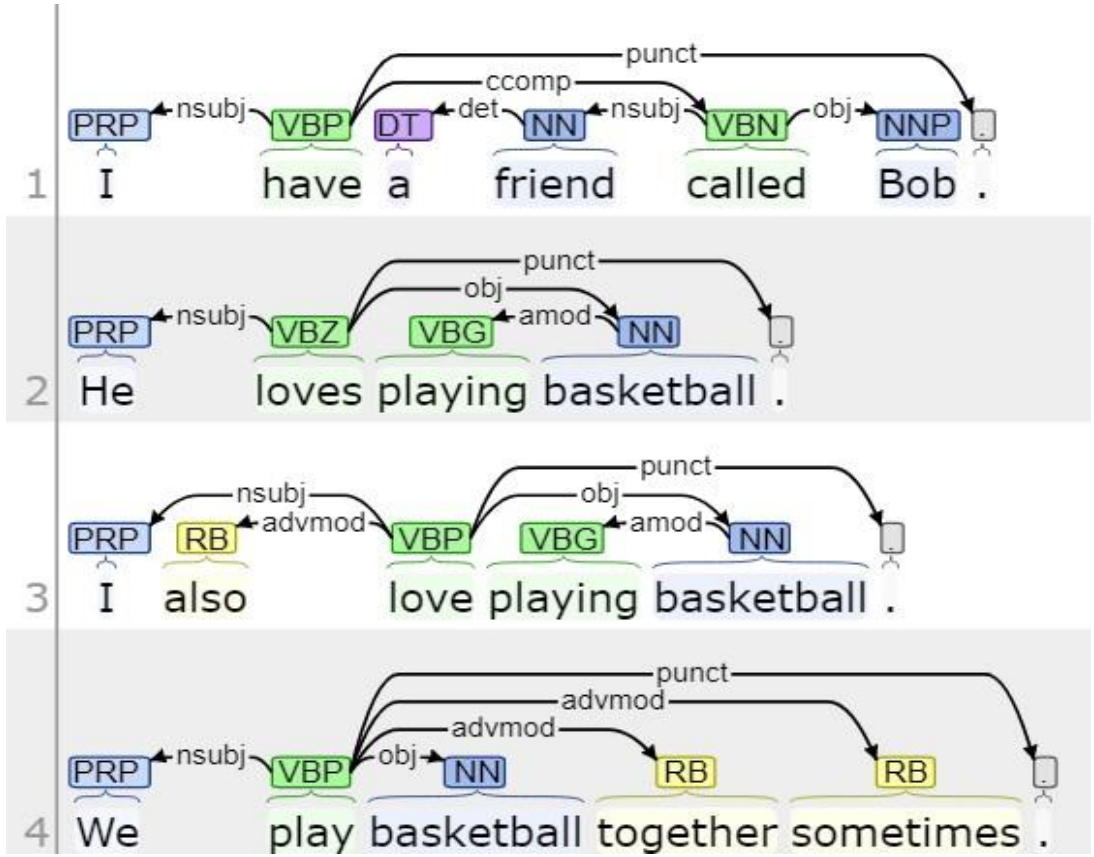
03 How coherence is affected by cohesion.

Thank You for Your Attention!

Appendix

COHERENCE (Modeling Local Coherence: An Entity-based Approach; Regina Barzilay, Mirella Lapata)

For example we take "I have a friend called Bob. He loves playing basketball. I also love playing basketball. We play basketball together sometimes."



LEGEND

- PRP - Personal Pronoun
- NN - Noun, Singular
- NNP - Proper Noun, Singular

Index	I	friend	Bob	basketball	we
1	S	S	O	-	-
2	-	S	-	O	-
3	S	-	-	O	-
4	-	-	-	O	S

Appendix

COHERENCE (Modeling Local Coherence: An Entity-based Approach; Regina Barzilay, Mirella Lapata)

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

SS	SO	S -	OS	OO	O -	- S	- O	- -
0.066	0	0.2	0	0.133	0.066	0.133	0.066	0.33

Appendix

Edit type	Qual. measure		mean	std	min	25%	50%	75%	max
Block Insertion	Coherence	2	0.711	0.094	0.000	0.657	0.723	0.777	1.000
	Readability	3	24.734	25.985	-387.983	12.729	27.422	40.990	121.220
	TTR	2	0.837	0.060	0.348	0.803	0.845	0.878	1.000
Block Merged	Coherence	4	0.692	0.082	0.484	0.641	0.707	0.752	1.000
	Readability	4	18.937	23.838	-72.090	6.700	19.656	35.455	72.826
	TTR	4	0.822	0.054	0.637	0.797	0.829	0.859	1.000
Block Split	Coherence	1	0.714	0.077	0.473	0.672	0.727	0.775	0.884
	Readability	1	27.492	22.934	-65.581	15.986	30.111	42.862	72.100
	TTR	1	0.842	0.047	0.662	0.816	0.850	0.871	1.000
Block deletion	Coherence	3	0.708	0.105	0.000	0.674	0.727	0.772	1.000
	Readability	2	27.244	24.743	-139.391	16.154	31.750	42.026	121.220
	TTR	3	0.834	0.090	0.000	0.815	0.847	0.877	1.000
Block edited	Coherence	5	0.670	0.124	0.000	0.588	0.693	0.751	1.000
	Readability	2	16.378	39.646	-184.310	5.126	21.838	38.517	121.220
	TTR	5	0.816	0.075	0.559	0.783	0.831	0.862	1.000