

COHERENCE BASED TEXT QUALITY IN SEARCH SUPPORTED WRITING

MASTER'S THESIS

1 Referee: Prof. Dr. Benno Stein

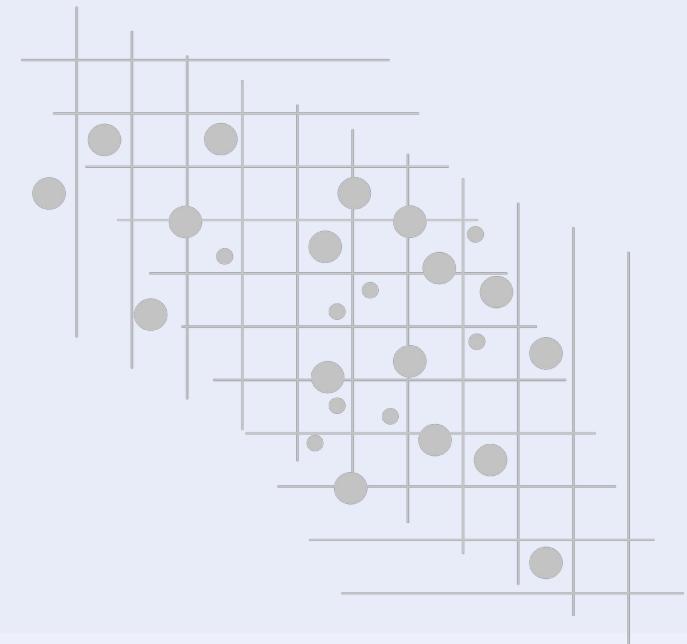
2 Referee: Prof. Dr. Ing. Volker

Rodehorst

Supervised By:

Dr. Michael Völske

Dr. Magdalena Wolska



Presenter:

Bibek khadayat

Date:

OUTLINE

INTRODUCTION

- ~~Definition and different factors~~
- ~~Area of Impacts~~
- ~~Search support writing~~
- Research Questions



DATA AND DATASETS

- ~~About Datasets~~
- Data Acquisition Setup
- Data Collected
- Prior Analysis



DATA PREPROCESSING

Data Extraction

- Data cleaning
- Identification of major changes
- Data Analysis



DIFFERENT TEXT QUALITY

MEASURES

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

CONCLUSION

- ~~Conclusion~~
- Future Works



RESULTS AND ANALYSIS

- ~~Results for research questions 1~~
- ~~Results for research questions 2~~
- ~~Results for research questions 3~~



be more precise: better

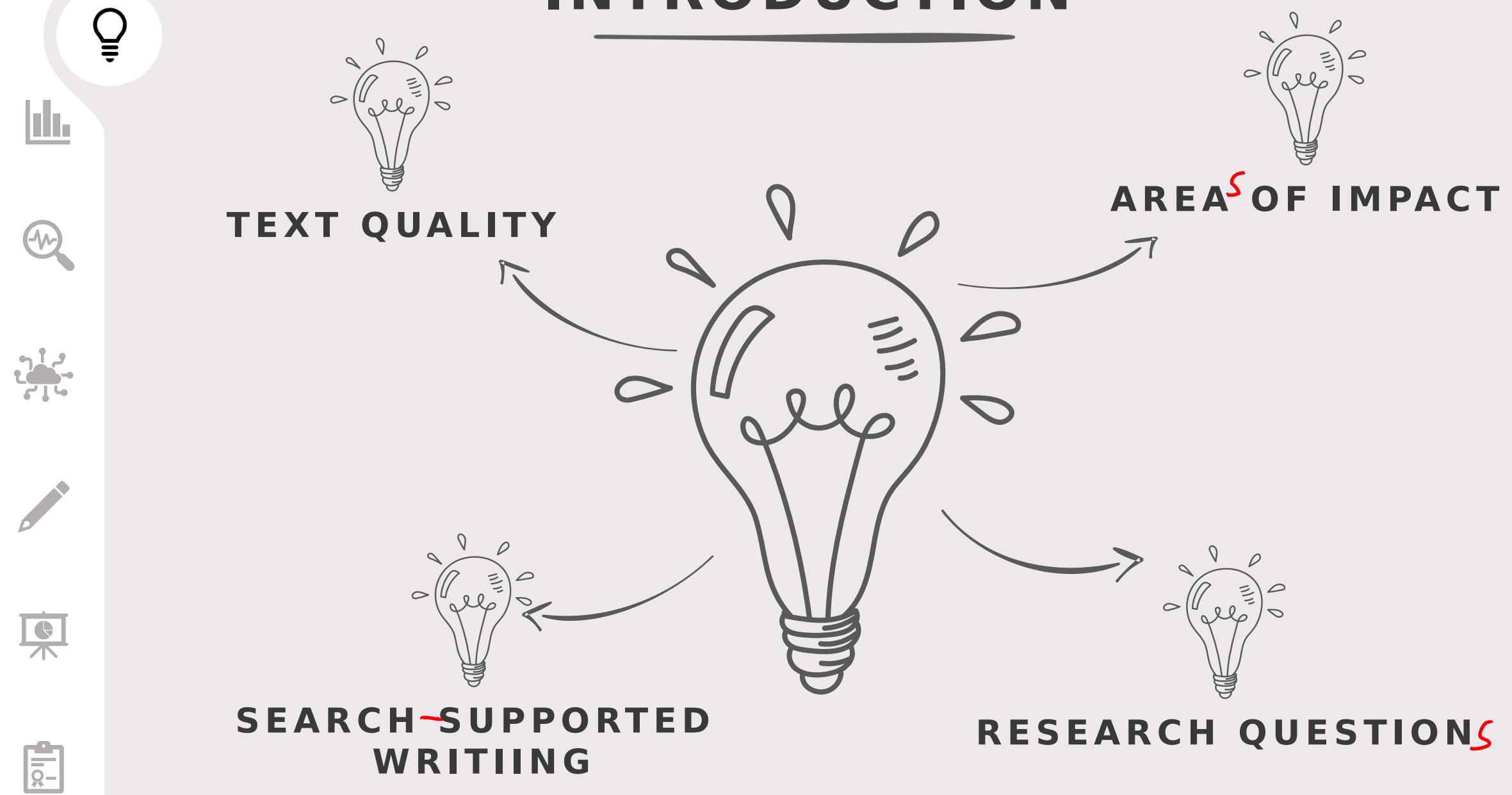


EDITING TYPES

- ~~About Editing Types~~
- ~~Example of editing types.~~
- ~~Automation of editing type~~

Better: automated detection

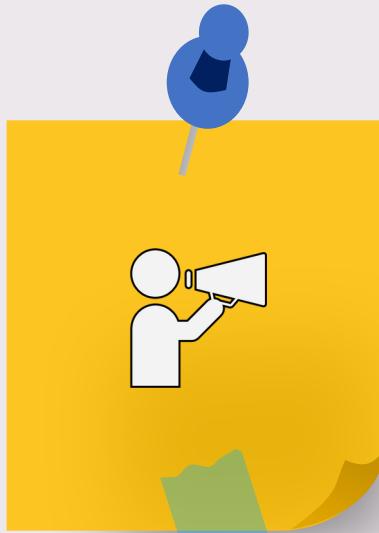
INTRODUCTION



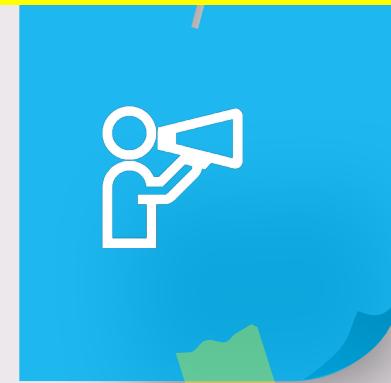
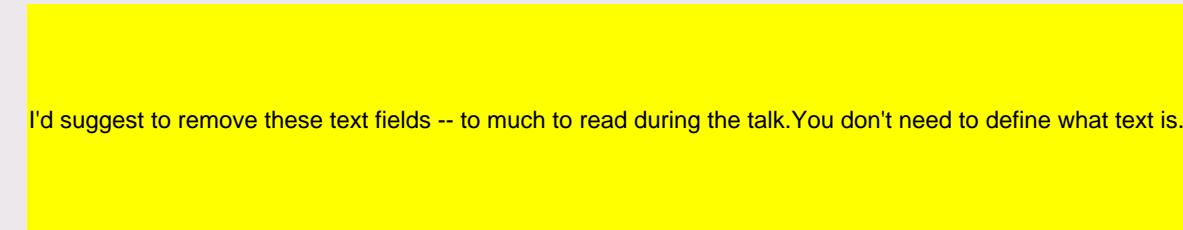
INTRODUCTION



TEXT QUALITY



Text is used in linguistics to refer to any passage, spoken, or written, of whatever length, that does form a unified whole.



Quality text is the text which is easily understandable by everyone.



Text quality is judge on the basis of spelling, vocabulary, grammar, structure, organization, readability, and so on.

INTRODUCTION



TEXT QUALITY

COHERENCE

COHESION

READABILITY



INTRODUCTION



TEXT QUALITY

Coherence

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.

Cohesion

Nepal is a small country. **Nepal** is located in between India and China. **Its** area is 147516 sq. kilometers.

INTRODUCTION



AREAS OF IMPACT

01

Web search
recommendation

02

Automatic
Summarization

03

Word
Prediction

04

Writing
Assessment

INTRODUCTION



SEARCH-SUPPORTED WRITING

01

Use of Internet for research purpose.



02

Gather information and data with the help of search engine.

03

Use these ^{is} 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?

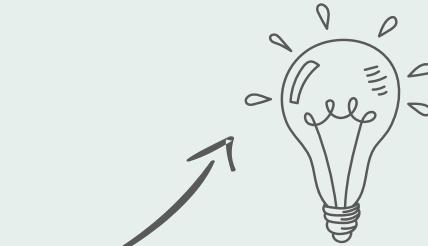
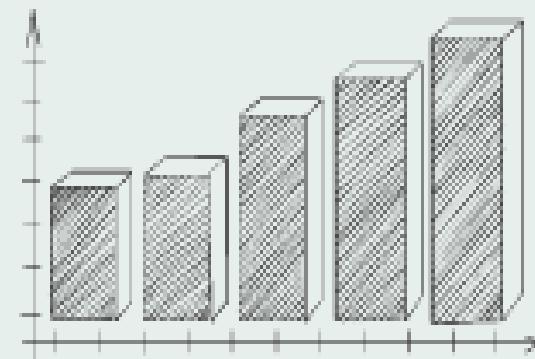
RQ3

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

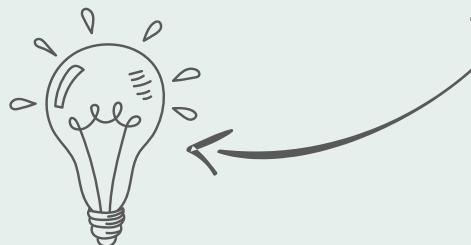
DATA AND DATASETS



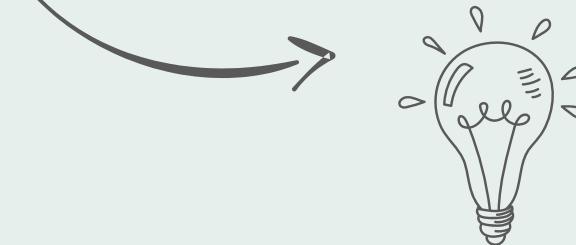
DESCRIPTION OF DATASETS



DATA ACQUISITION SETUP



PRIOR ANALYSIS



COLLECTED DATA



DATA AND DATASETS



DESCRIPTION OF DATASETS



Text Reuse Corpus 2012 (Webis-TRC-12)



150 long essays written by ~~the~~ professional writers



Information ~~are used~~ from ClueWeb09 corpus



~~Contain detailed interaction logs file + Revision history!~~



~~Dataset is available on webis website and can be explore how it is created in web visualization tool.~~



DATA AND DATASETS



DESCRIPTION OF DATASETS



Webis-TRC Essay Viewer 2%

1 54 of 3288

Whence Cometh Obama?

At the time of Obama's landslide victory at the polls America was at the crossroads.

1
Sung: Why I will vote for Barack Obama
By Eric Sung, Guest Columnist
Oct. 30, 2008

When this long journey first began last year on the steps of the old capitol in Springfield, Ill., I endorsed an African-American candidate who is running against the established bedrocks of the Democratic Party and against all odds.

This man proceeded to shock the world by capturing the Iowa caucus on Jan. 3 and went on to defeat Sen. Hillary Clinton in the toughest primary in recent history. Now we stand less than a week from Nov. 4, Election Day, and possible history in the making. I endorsed this man in the primaries and I will continue to stand by Barack Obama on the eve of the most important election in my life.

RELATED LINK

- Castellanes: Why I will vote for John McCain

was nine years old, her mother was diagnosed with Cancer. She was let go of her job and they lost health care. They filed for bankruptcy and that's when 9-year-old Ashley decided to do what she can to help her mom. As a child, there are very few things she could've done to support her mom. However, she knew that food was one of their highest expenses. She convinced her mom that what she really liked and really wanted to eat more than anything else was mustard and relish sandwiches, because that was the cheapest way to eat. She ate mustard and relish sandwiches for a year until her mom got better and their situation improved. That may not sound like a great sacrifice, but for a 9-year-old, it might be one of the few things that she could've done. She told everyone that the reason she stands with Barack was so she could help the millions of other children in the country who want and need to help their parents too. After she told her story, other people started sharing why they are supporting Barack. Everyone had a different reason ranging from Iraq to health care except for an elderly black man. He simply says, "I am here because of Ashley."

I stand here with Barack because of Ashley and others like her. We stand here together because we want to help the millions of Americans that need help in a time like this. Will you join me, Ashley and millions of Americans on Nov. 4 and vote for Barack Obama as the next president of the United States?

Deny cynicism and fear and vote for your hopes and dreams. Vote for the future and what it can be, and reject a continuation of the past and what it was. Vote Barack Obama for the next President of the United States of America.

Barack Obama was born at the Kapi'olani Medical Center for Women & Children in Honolulu, Hawaii.^{[3][4]} In 1961 Dunham, a White American from Wichita,

Dunham and grandfather Stanley Dunham, in Hawaii (early 1970s)

Of his early childhood, Obama has recalled, "That my father looked nothing like the people around me □ that he was black as pitch, my mother white as milk □ barely registered in my mind."^[15] In his 1995 memoir, he described his struggles as a young adult to reconcile social perceptions of his multiracial heritage.^[16] He wrote that he used alcohol, marijuana, and cocaine during his teenage years to "push questions of who I was out of my mind".^[17] At the 2008 Civil Forum on the Presidency, Obama identified his high-school drug use as his "greatest moral failure."^[18]

Some of his fellow students at Punahoa School later told the Honolulu Star-Bulletin that Obama was mature for his age, and that he sometimes attended college parties and other events in order to associate with African American students and military service people. Reflecting later on his formative years in Honolulu, Obama wrote: "The opportunity that Hawaii offered □ to experience a variety of cultures in a climate of mutual respect □ became an integral part of my world view, and a basis for the values that I hold most dear."^[19]

Following high school, Obama moved to Los Angeles,

DATA AND DATASETS

DATA ACQUISITION SETUP



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



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



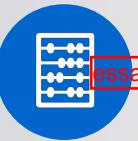
ChatNoir is used as search engine



DATA AND DATASETS



COLLECTED DATA



~~Writer or Author create different revisions during writing process in each essays or topics.~~



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



~~150 long essays were created with numbers of revisions in each essays by 12 professional writers.~~

you already said this on a previous slide

Instead, you could put here a recap/summary table like:



DATA AND DATASETS

COLLECTED DATA





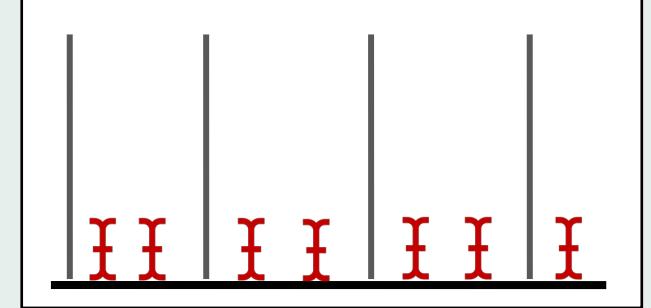
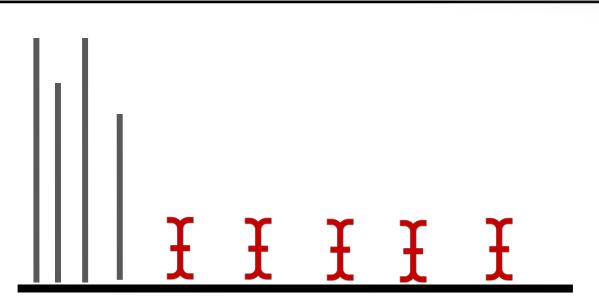
DATA AND DATASETS

PRIOR ANALYSIS

Writing Style



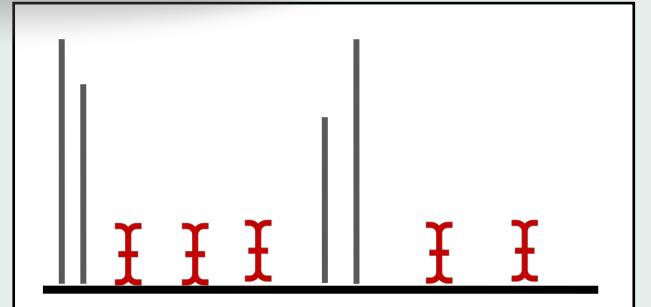
Build-up is ~~a fashion in which continuous lengthening of the essay in regular period of time over the whole period of writing is done. process~~ **process**



Boil-down is ~~a fashion in which first quick length growth and then shorting happens~~ **early** **process**



Mixed writing style ~~has included~~ **has** both Build-Up and Boil-down. **aspects**



LEGEND

Research and copy paste

Edit

DATA PREPROCESSING

DATA EXTRACTION



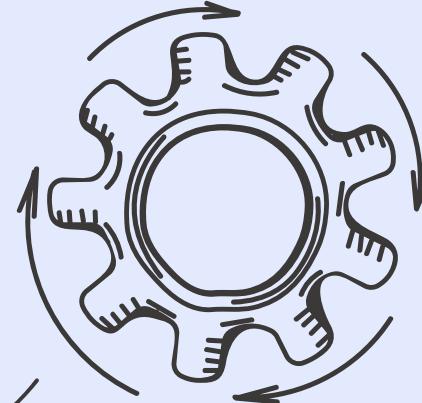
DATA CLEANING



DATA ANALYSIS

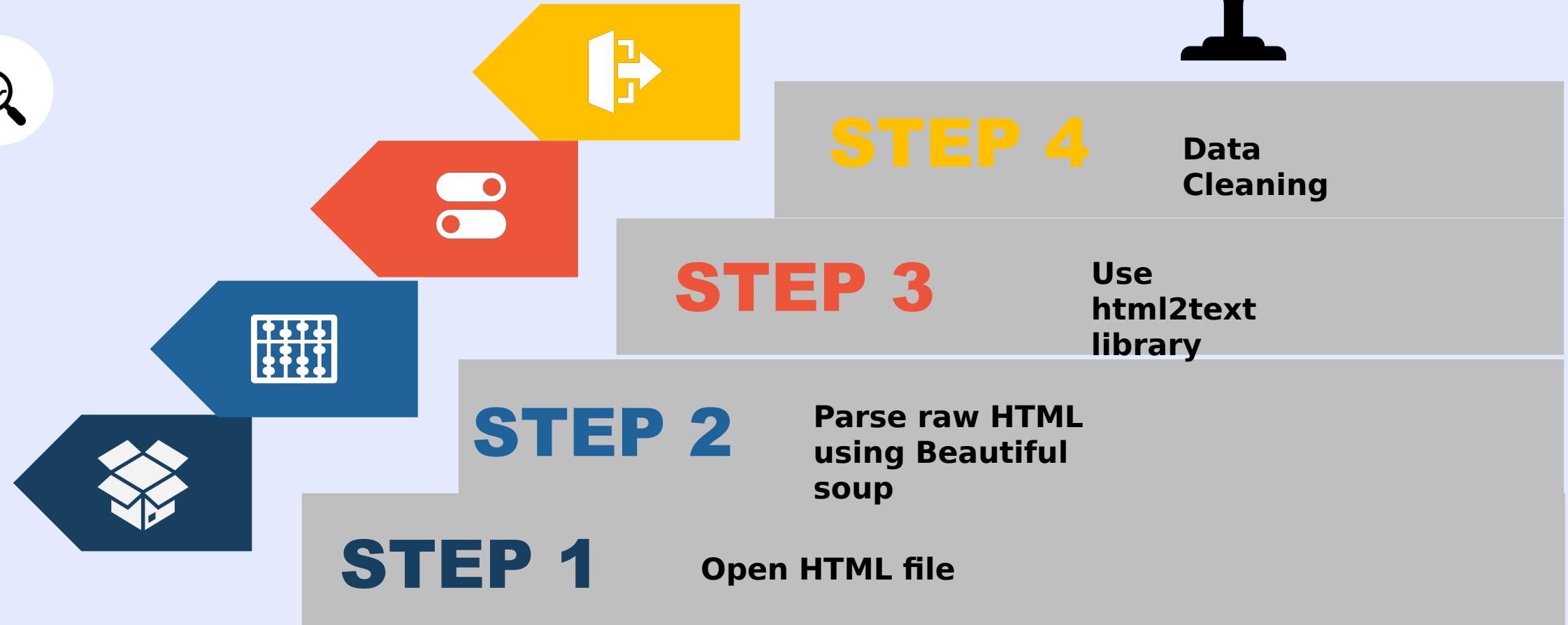


IDENTIFICATION OF
MAJOR CHANGES



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 change

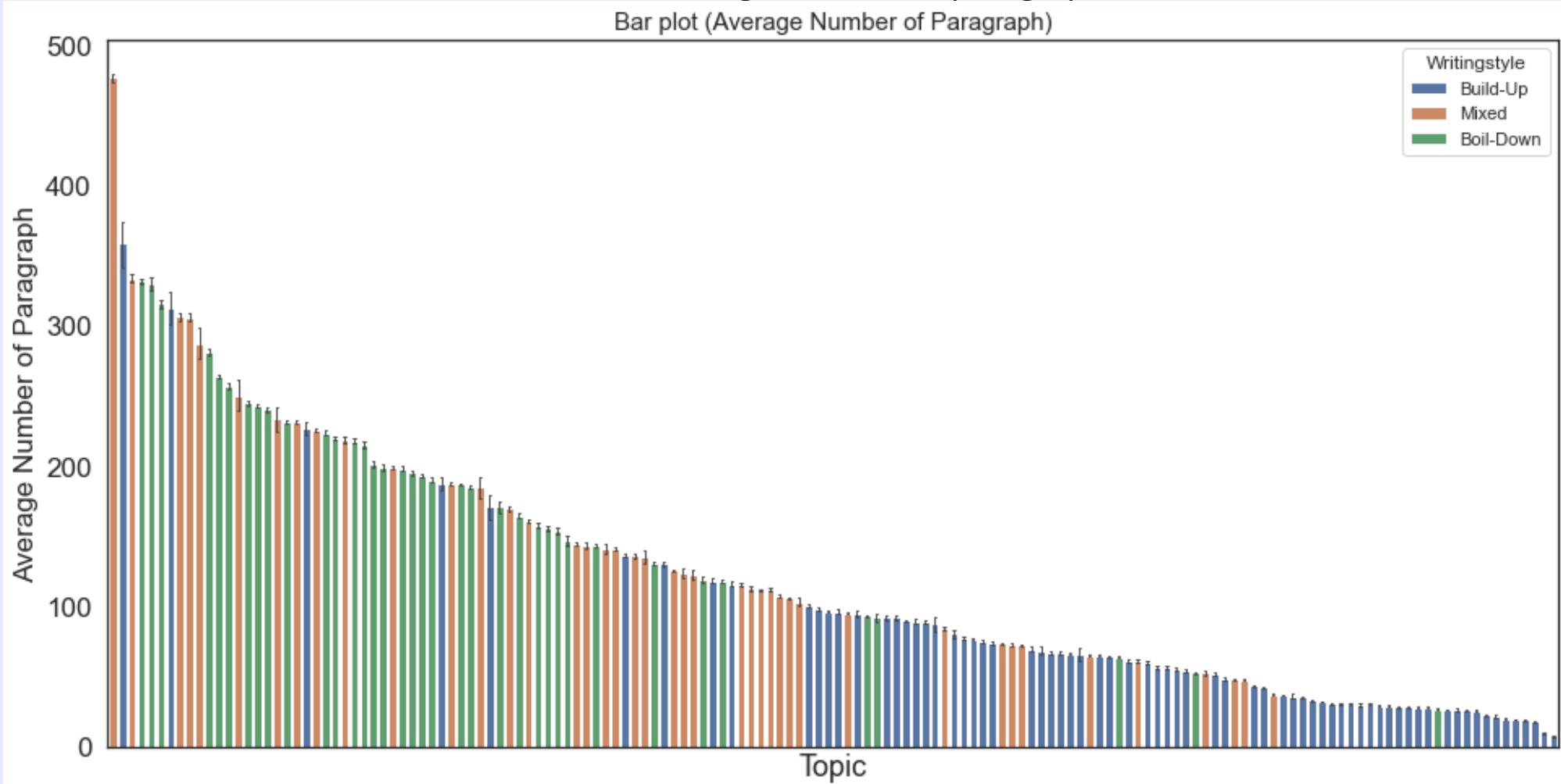
DATA PREPROCESSING



DATA ANALYSIS

Average number of paragraphs.

Bar plot (Average Number of Paragraph)

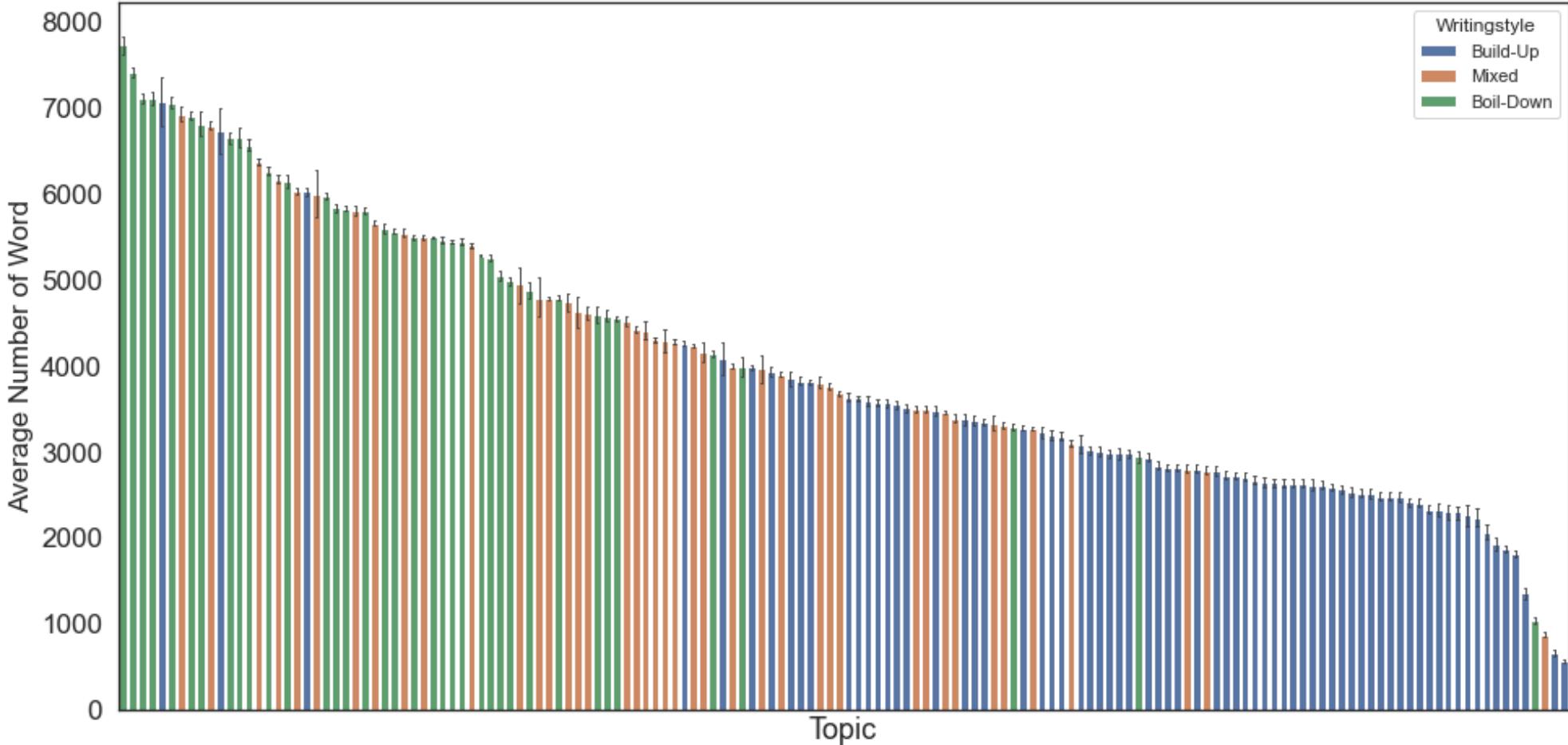


DATA PREPROCESSING

DATA ANALYSIS

Average number of words.

Bar plot (Average Number of Word)

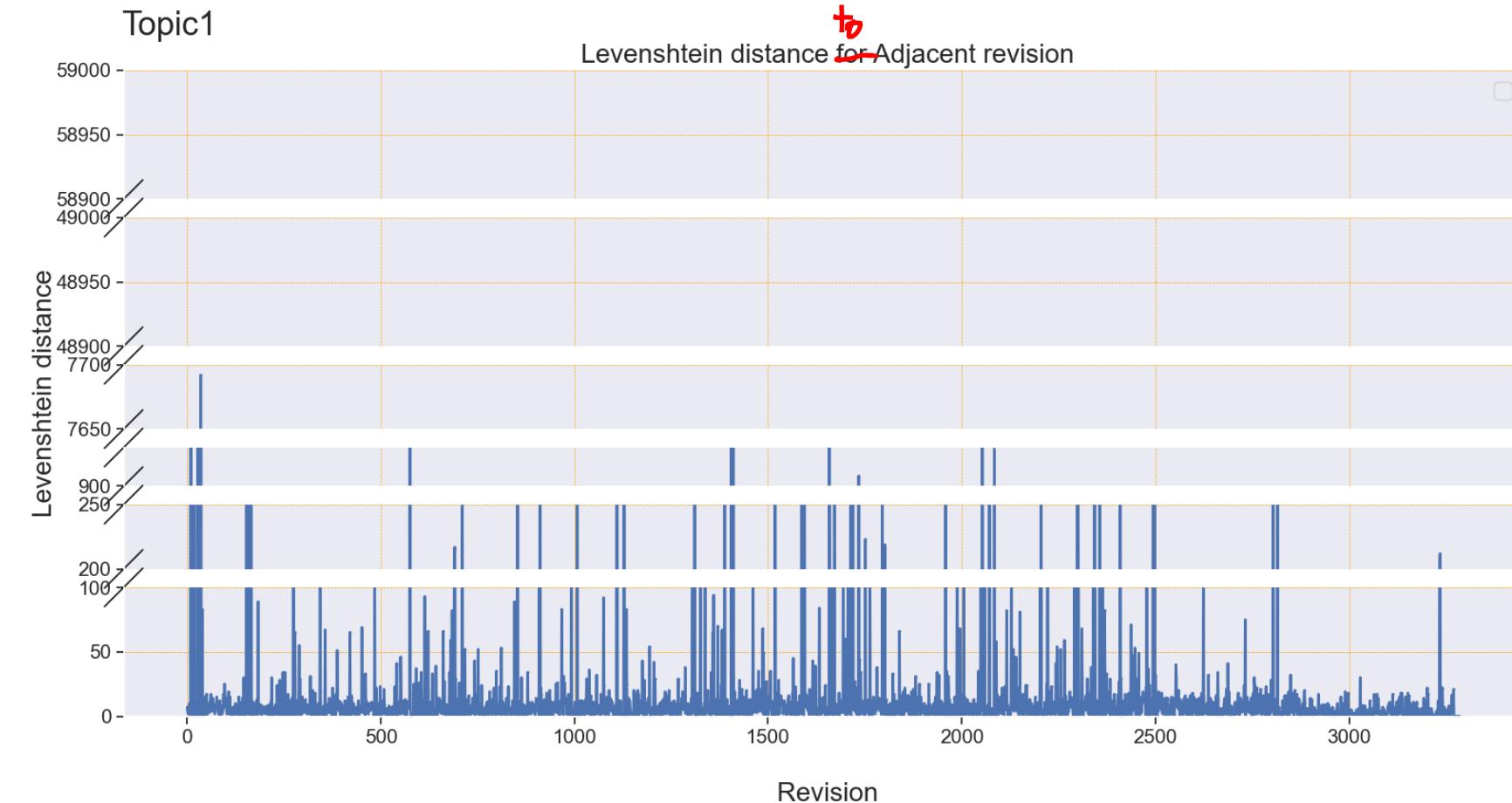


DATA PREPROCESSING

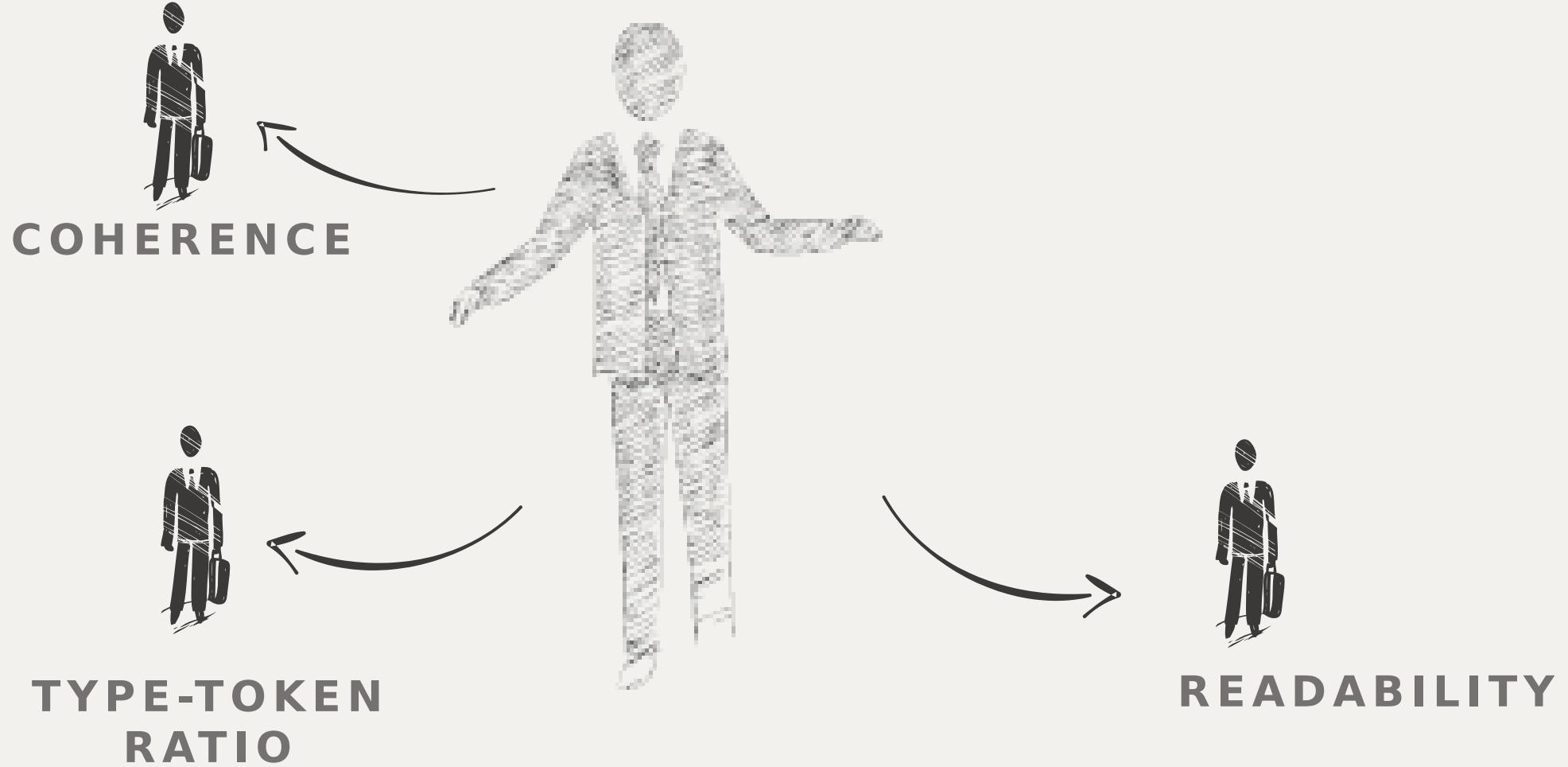


DATA ANALYSIS

The Levenshtein distance.



TEXT QUALITY MEASURES



TEXT QUALITY MEASURES

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."



1	[PRP] [VBP] [DT] [NN] [VBN] [NNP]	I have a friend called Bob .
2	[PRP] [VBZ] [VBG] [NN]	He loves playing basketball .
3	[PRP] [RB] [VBP] [VBG] [NN]	I also love playing basketball .
4	[PRP] [VBP] [NN] [RB] [RB]	We play basketball together sometimes .



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

TEXT QUALITY MEASURES

COHERENCE (ENTITY GRID)

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

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

TEXT QUALITY MEASURES

READABILITY

- ❖ Flesch reading Ease Formula.
$$206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})$$
- ❖ Textstat python library.

Data: List of clean Text

Result: Readability_Score

```
1 Text ← [cleaned_text];      /* [cleaned_text] is lists of
   preprocessed paragraphs [p1,p2,p3,...] where P1,P2,p3
   are paragraph */
2 for i = 0 to len(Text) do
3   text = Text[i];
   /* Check empty element to avoid zero exception */
4   if len(text) > 0 then
5     | Readability_Score ← flesch_reading_ease(text) ;
6   else
7     | Readability_Score ← 0 ;
8   end
9 end
```

TEXT QUALITY MEASURES



TYPE-TOKEN RATIO

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

I'd remove the Counter line, it's redundant with the table below

Type=11---Total Token=15

TTR Score=0.7333333333333333 so many significant digits aren't necessary

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?

On these slides, can you make the circle around the non-active RQs gray instead of colored? (I.e. here, only RQ1 would be green, RQ2 and RQ3 would be gray)

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

Editing types mean how the second revision ~~is~~ edited compared to first revision
~~the~~

Change the ordering:- analyse pairs of adjacent revisions



maybe add figure with two actual adjacent revisions

02

During analysing dataset we identify ~~different~~ editing types

give the number instead of "different"

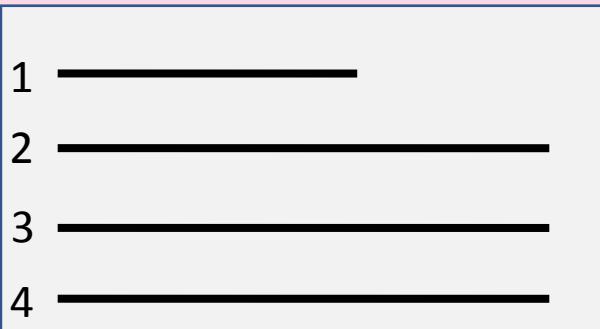
03

We analyse across pairs of adjacent revisions

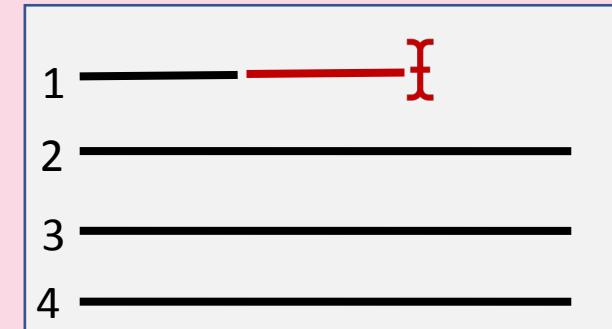
EDITING TYPES

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

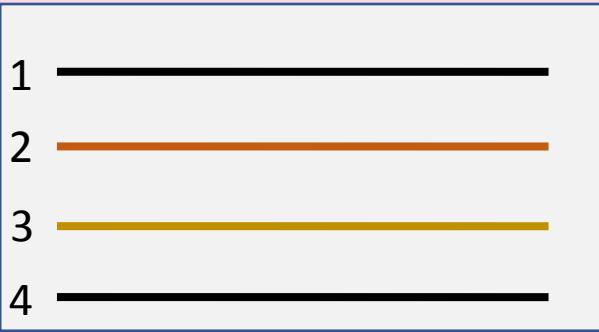
Oh Wow! Did you catch the latest news from Indiana? The State with the motto "The Crossroads of America" is not just a great place to watch motor races like the Indy 500, it's also a great place

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.

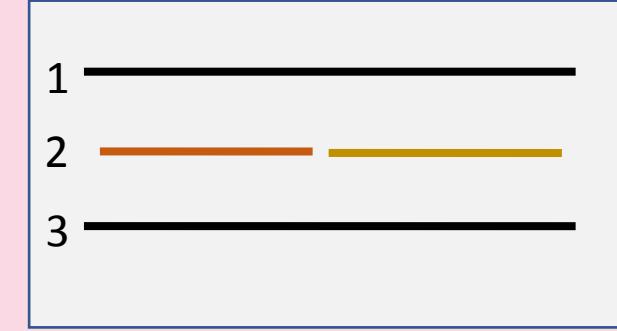
EDITING TYPES

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.

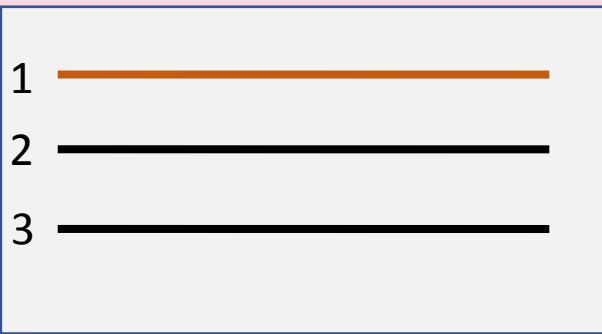
What a fantastic place to locate a Casino. No wonder the designers chose to add a touch of wonderland to the magic state when they built French Lick resort and Casino on 8670 W. State Road 56
French Lick, IN 47432 (Map it)

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,

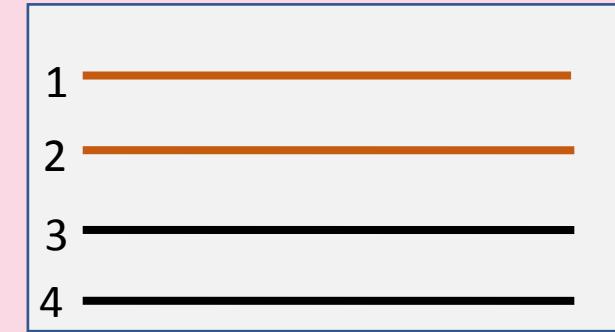
EDITING TYPES

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

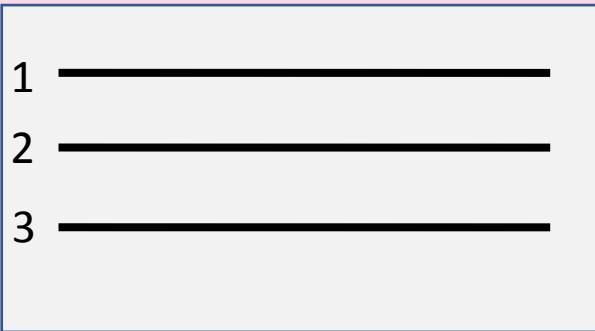
French Lick Resort has embarked on an incredible \$500 million historic restoration and expansion. In addition to the restoration of the French Lick Springs Hotel, including 443 fully renovated guest rooms, a new event center, exciting retail shops, and fully restored public areas, we are proud to announce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

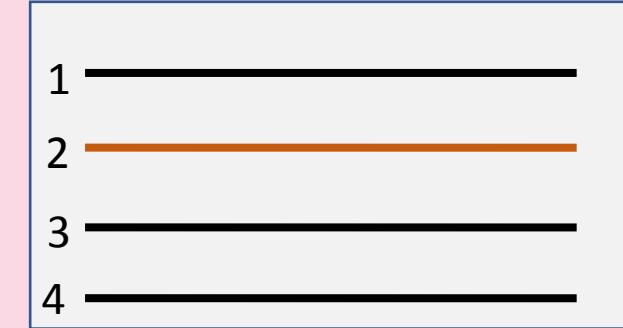
EDITING TYPES

EXAMPLE OF EDITING TYPES

Block
Insertion



Before



After

French Lick Indiana got its name from early French settlers and the mineral licks in 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 Taggart 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.

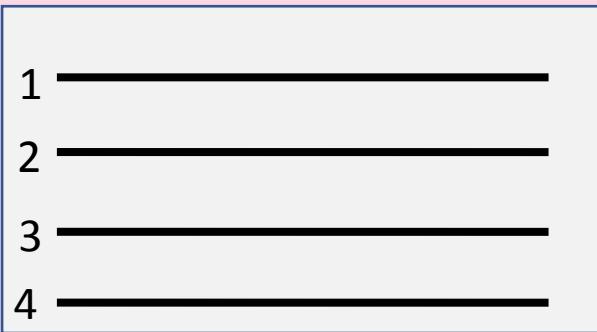
French Lick Resort and Casino. Write an advertising brochure for the French Lick Resort and Casino in Indiana. Interesting things could be the history of the casino, discounted packages for staying at the resort, are there close by other casinos, what could be job opportunities, etc.

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.

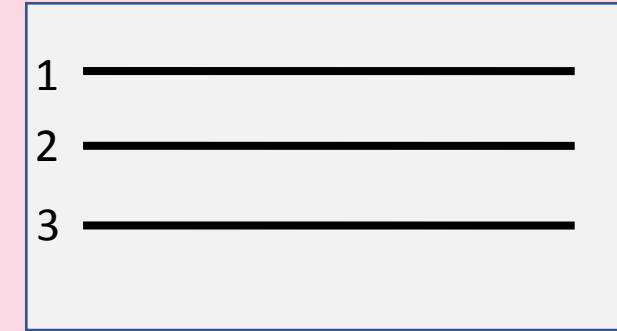
EDITING TYPES

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

01

Number of characters and ~~the number of paragraphs~~ plays a vital role

02

We check the ~~number of characters and number of paragraphs in adjacent revisions.~~



Block deletion

number of paragraphs
 $>$ number of paragraphs
number of characters $>$ number of characters



1

Block edit

number of paragraphs $=$ number of paragraphs
number of characters $=!$ number of characters



2

Block merged

number of paragraphs $>$ number of paragraphs
number of characters $=$ number of characters

Block insertion

number of paragraphs $<$
number of paragraphs
number of characters $<$ number of characters



5



4



3



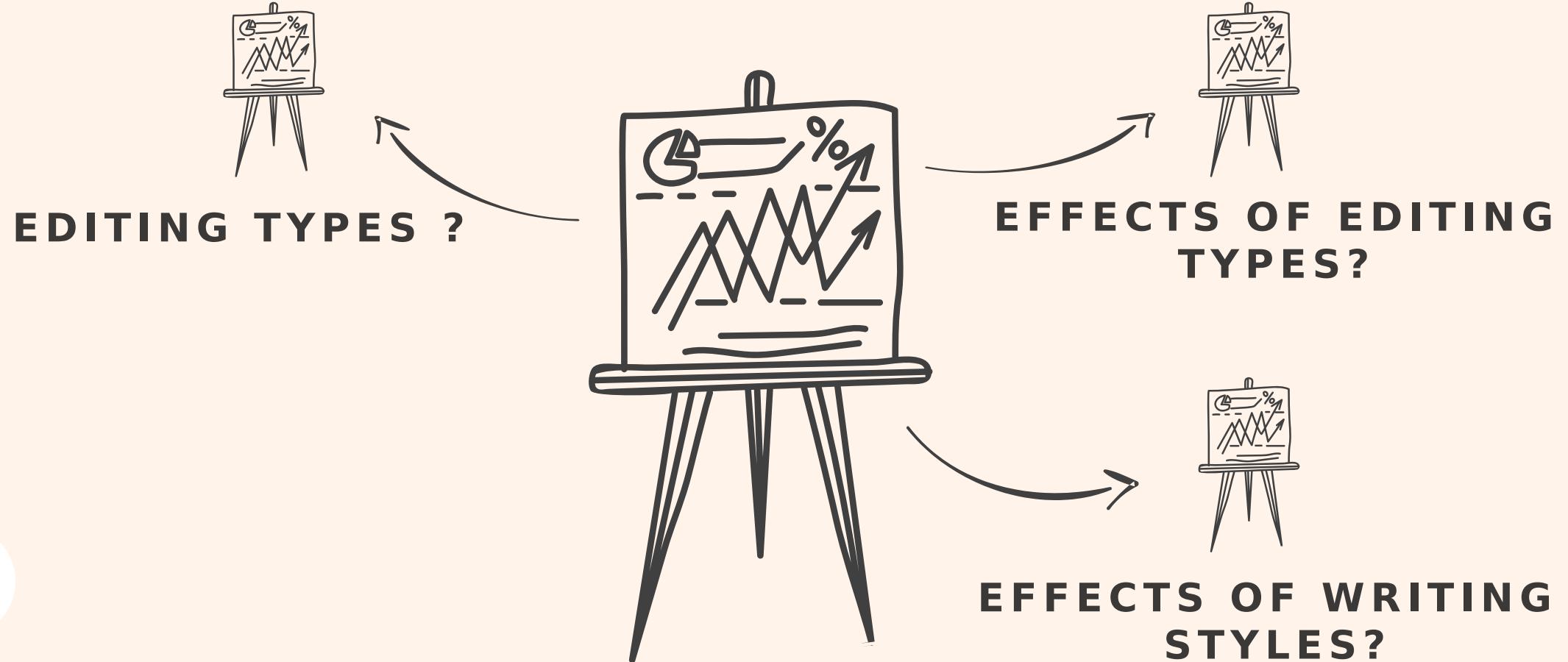
Block split

number of paragraphs $<$ number of paragraphs
number of characters $=$ number of characters

these operators comparing the same quantity

a) "paragraphs ↑"

RESULT AND ANALYSIS



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 edit	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

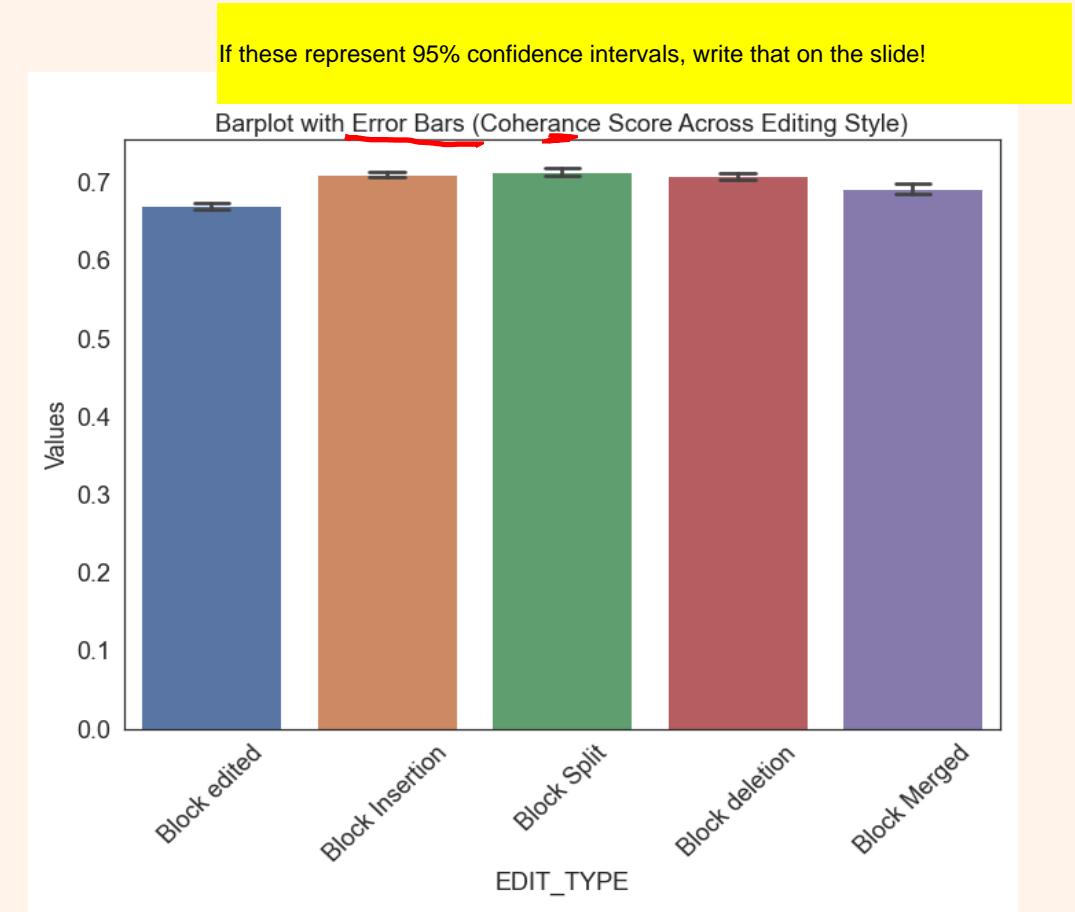
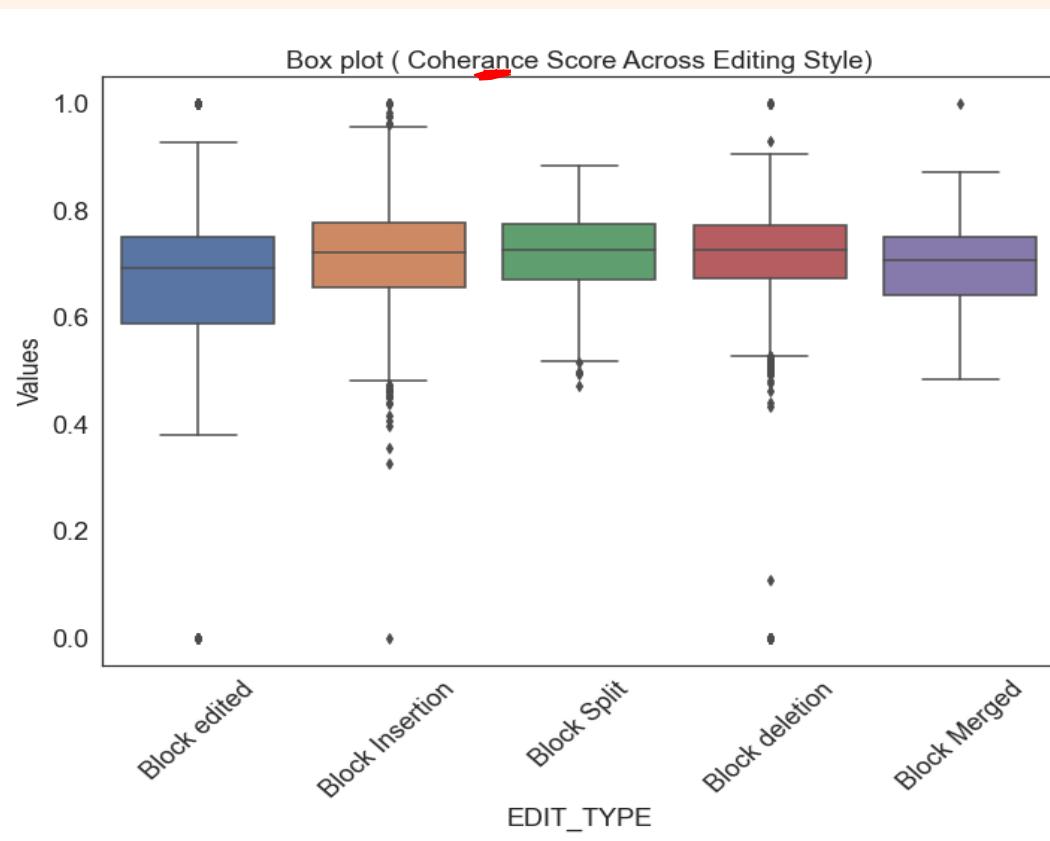
How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3

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

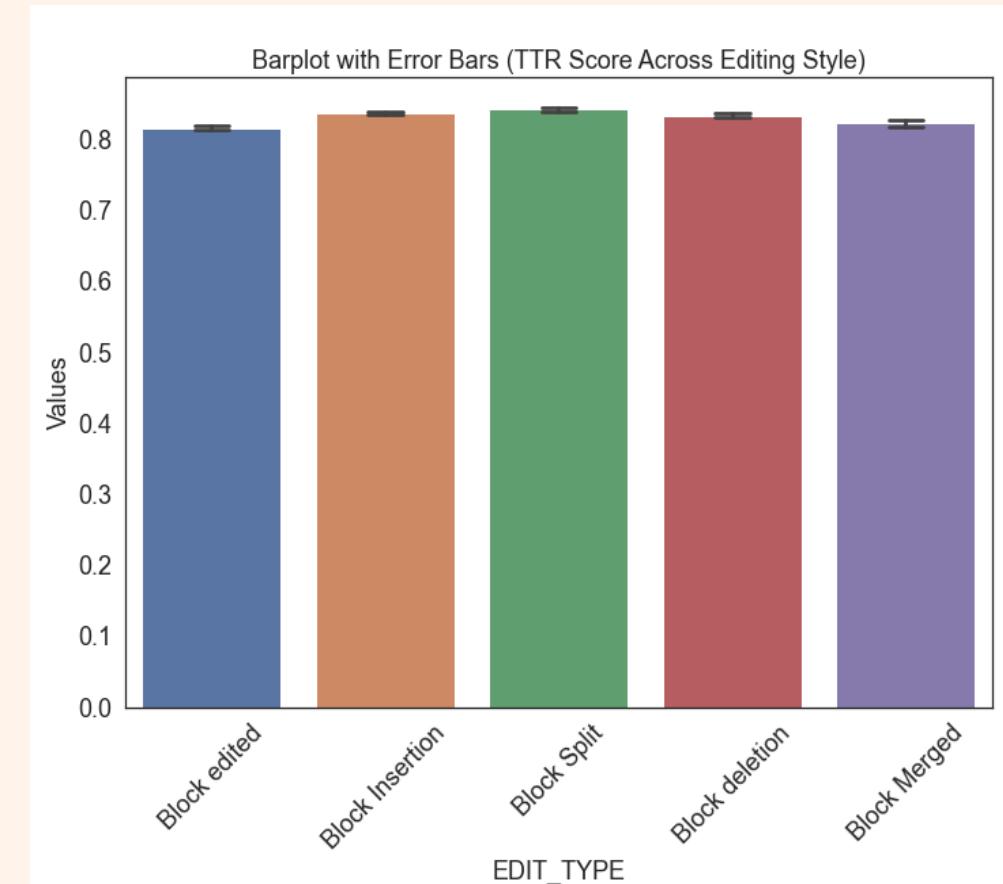
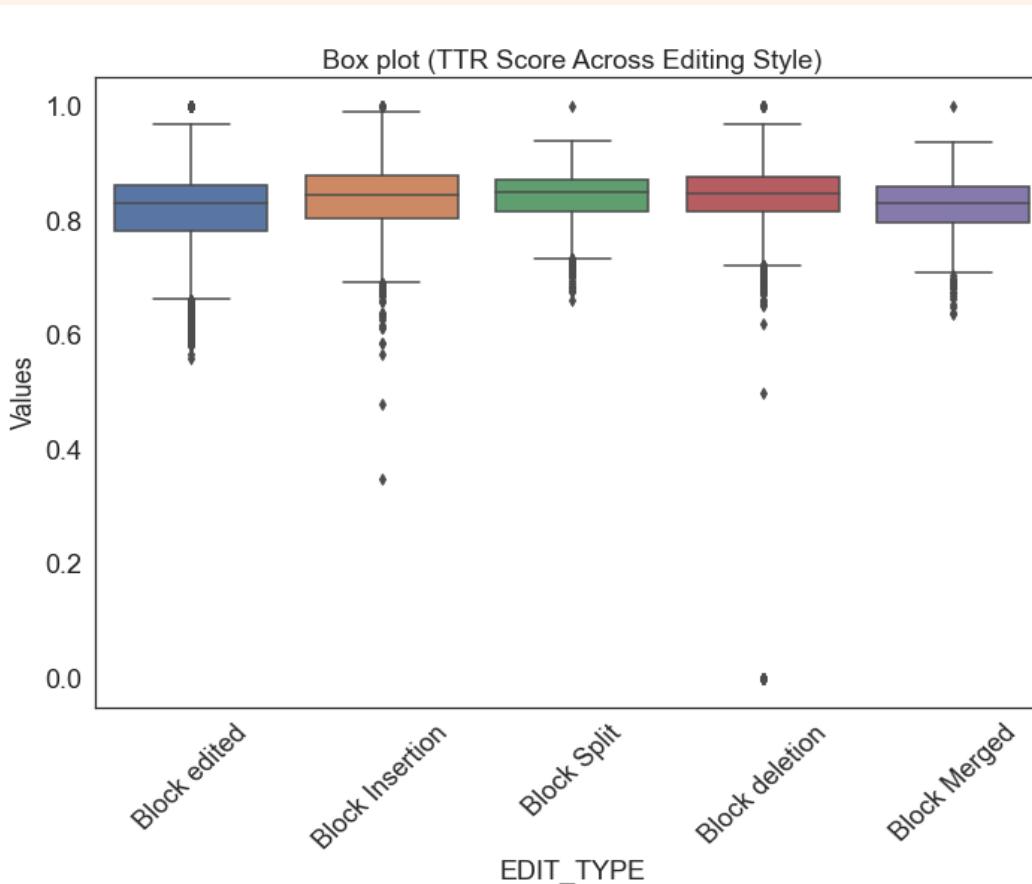
RESULT AND ANALYSIS

RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY



RESULT AND ANALYSIS

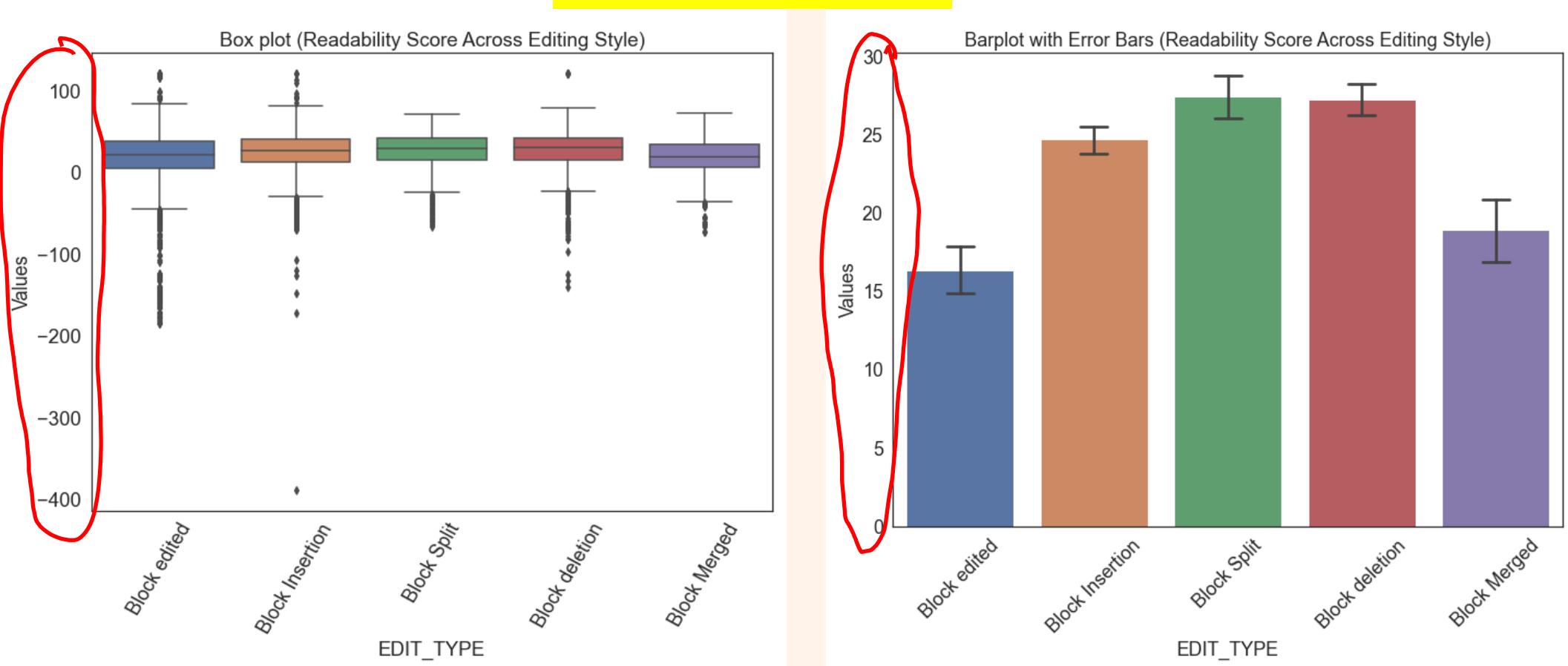
RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY



RESULT AND ANALYSIS

RQ2: EFFECTS OF EDITING TYPES ON TEXT QUALITY

why are the scales so different here?



RESULT AND ANALYSIS

RESEARCH QUESTION RQ3

RQ1

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

RQ2

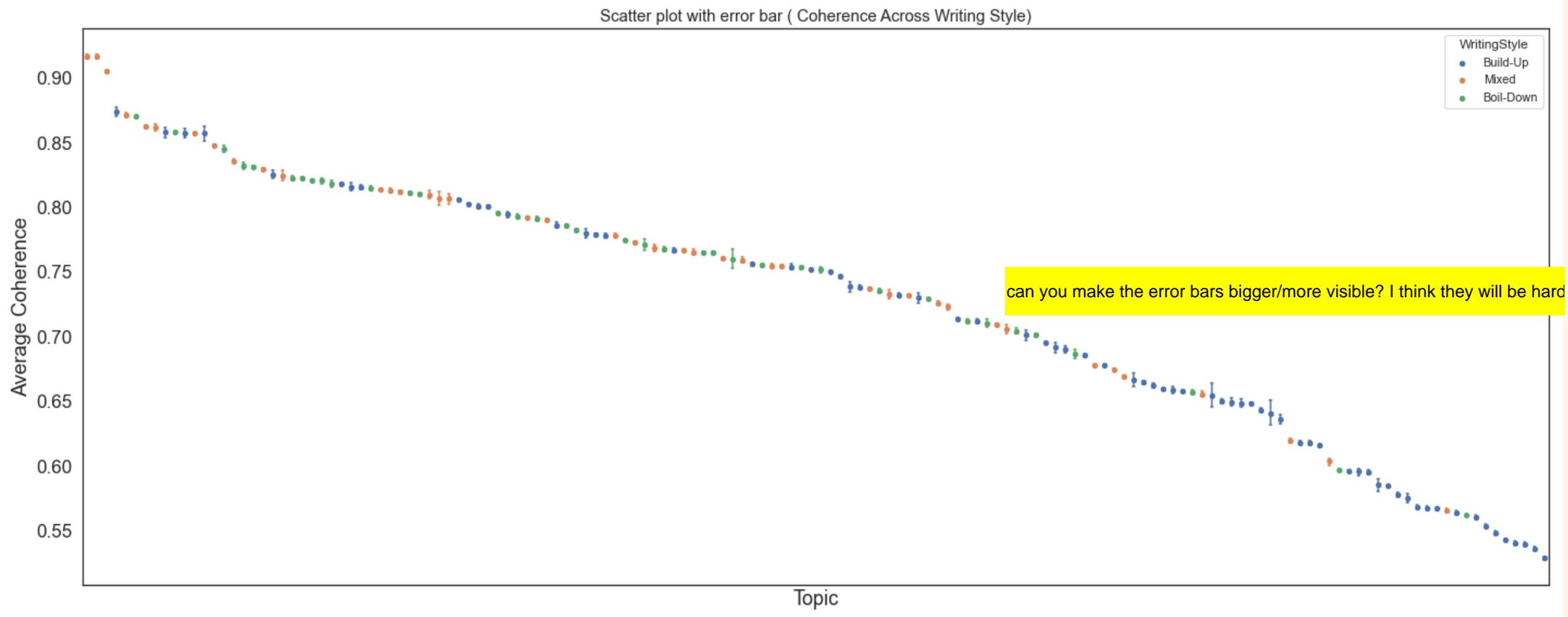
How do different types of editing affect Coherence, Type-Token ratio, and Readability?

RQ3

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

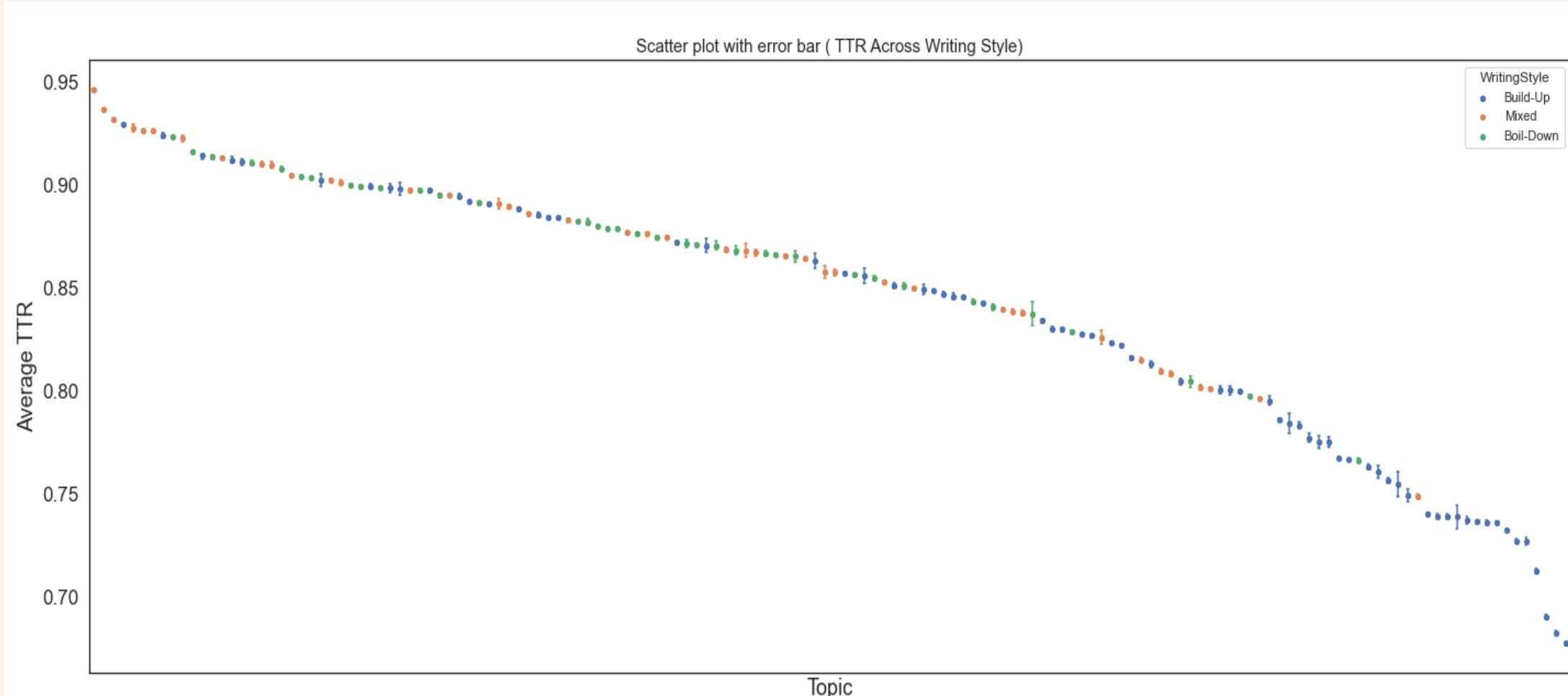
RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



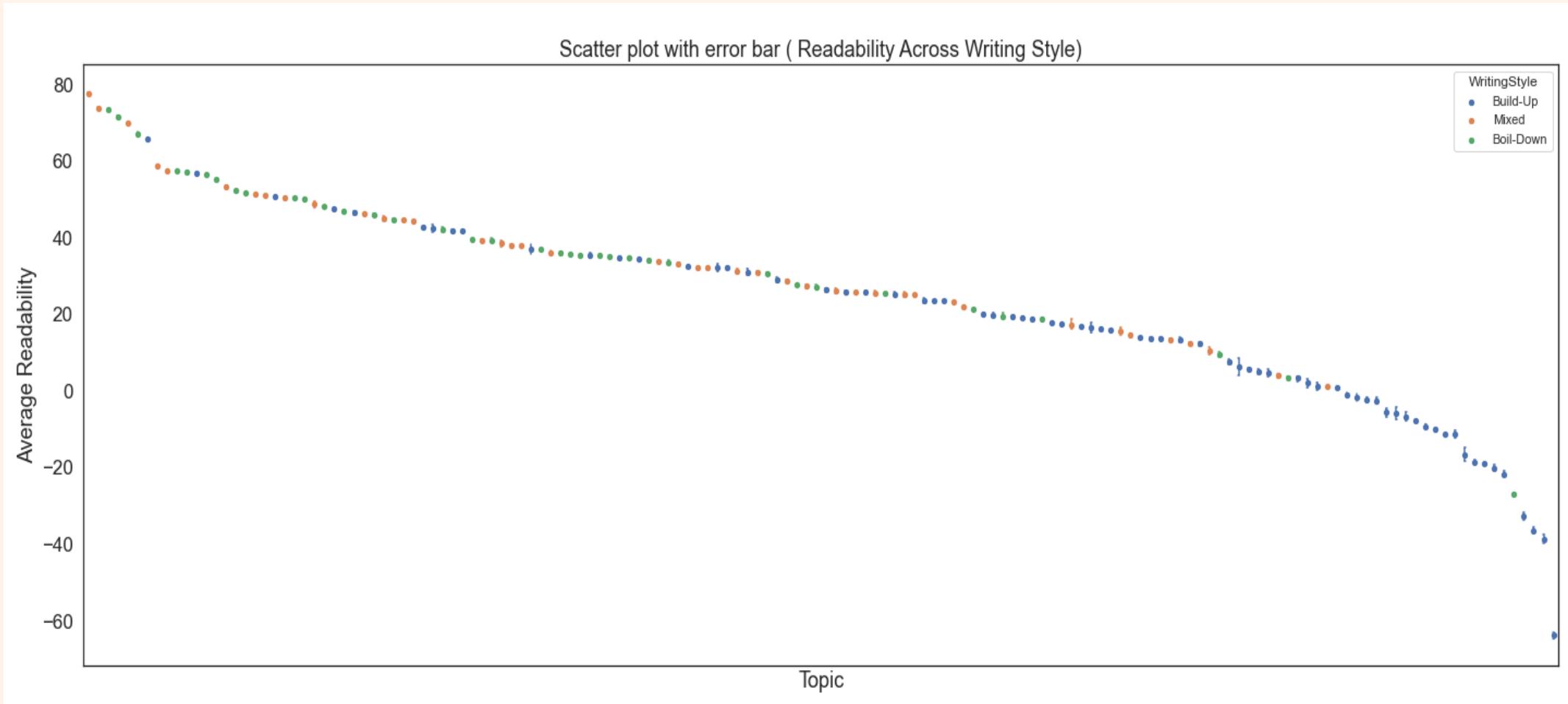
RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



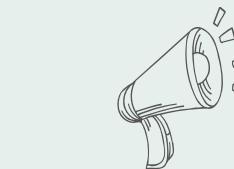
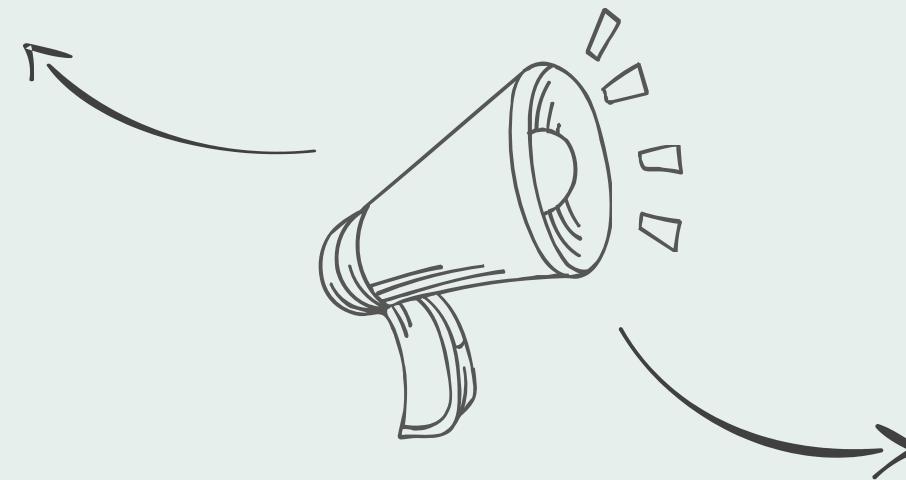
RESULT AND ANALYSIS

RQ3: EFFECTS OF WRITING STYLES ON TEXT QUALITY



CONCLUSION

MAIN FINDINGS



FUTURE WORK



CONCLUSION

MAIN FINDINGS

RQ1

~~Different editing types ;~~ name them



RQ
3

Coherence, TTR, and readability are higher in mixed writing

RQ
2

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

RQ
3

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

CONCLUSION

FUTURE WORK

01

~~Investigate and explain the lower coherence in Build-up writing.~~

These are so similar that they should be the same point? Also, I'd again avoid full ser

02

~~Investigate the reason and explain the reason for lower TTR and Readability in Build-up writing.~~

03

~~Investigate how edits to the essay that introduce new entities in particular affect the coherence score.~~

04

How coherence is affected by cohesion.

Thank you!