SAS® GLOBAL FORUM 2019

USERS PROGRAM

APRIL 28 – MAY 1, 2019 | DALLAS, TX

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Text Analytics is the process of examining large collections of written resources to generate new information; transforming unstructured text into structured data helps us find meaningful insights from the text. It is a subgroup of Natural Language Processing (NLP). Statistical methods, rule-based modeling, and machine learning techniques are applied in text analytics allowing for the extraction of topics, keywords, semantics, and sentiments from the raw text in an effort to categorize terms.

ROHIT BANERJEE

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

- Prologue

- The Idea

Short Term

- Quantifying the interpretability through different indices
- Using the peer reviewed journals to create a framework
- Providing a Proof of Concept

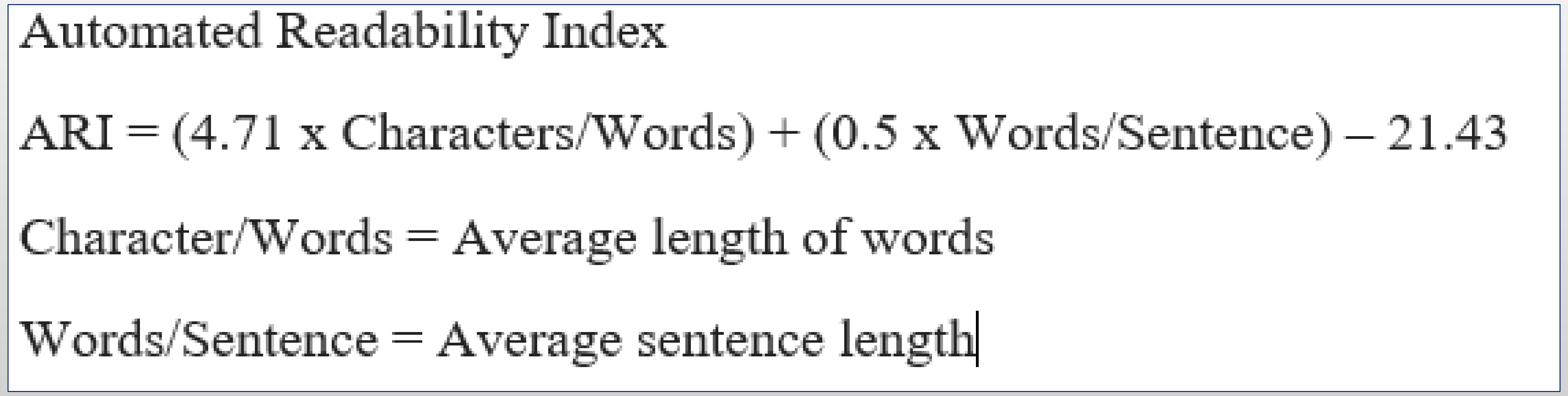
Long Term

- Implement the idea in college level education
- Work on creating a SAS node

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion



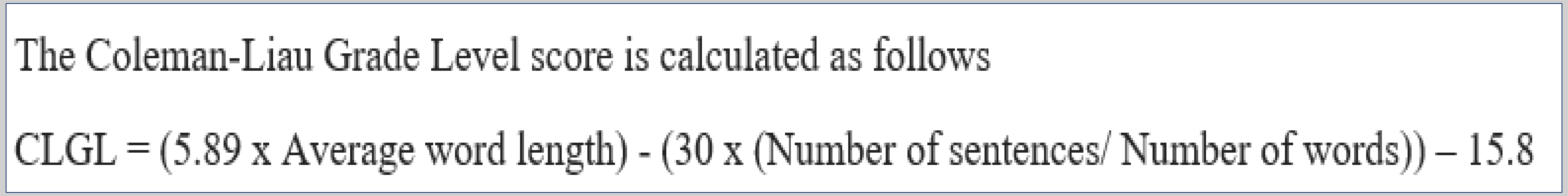
Automated Readability Index

ARI = (4.71 x Characters/Words) + (0.5 x Words/Sentence) – 21.43

Character/Words = Average length of words

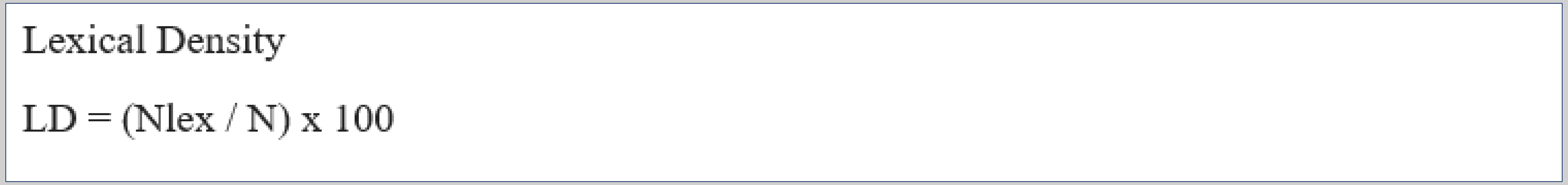
Words/Sentence = Average sentence length

- 13 marketing papers published in International Journal of Research in Marketing
- The ARI test is developed and used by the US Army to understand technical documents
- The Coleman-Liau grading is more suitable for 4th grade to college level texts



The Coleman-Liau Grade Level score is calculated as follows

CLGL = (5.89 x Average word length) - (30 x (Number of sentences/ Number of words)) – 15.8



Lexical Density

LD = (Nlex / N) x 100

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Assumption

- The written piece is in the English language
- Proper grammatical structure and punctuations are used in an orderly manner
- The documents should not contain Greek numeric

The Corpus

13 abstracts from Marketing papers were sampled from American Marketing Association of year 2017

3500 words

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

17

The average ARI and CLGL value

23

The average sentence length

0.56

The average Lexical Density

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Example

“I have a house in west Provo. I like the view from the house. We have lived there since November. We also have a cat that I like very much. We were in an accident a few months ago. We hit a deer that was crossing the street at night. I felt sorry for the deer, but it cost a lot of money to repair the car.”

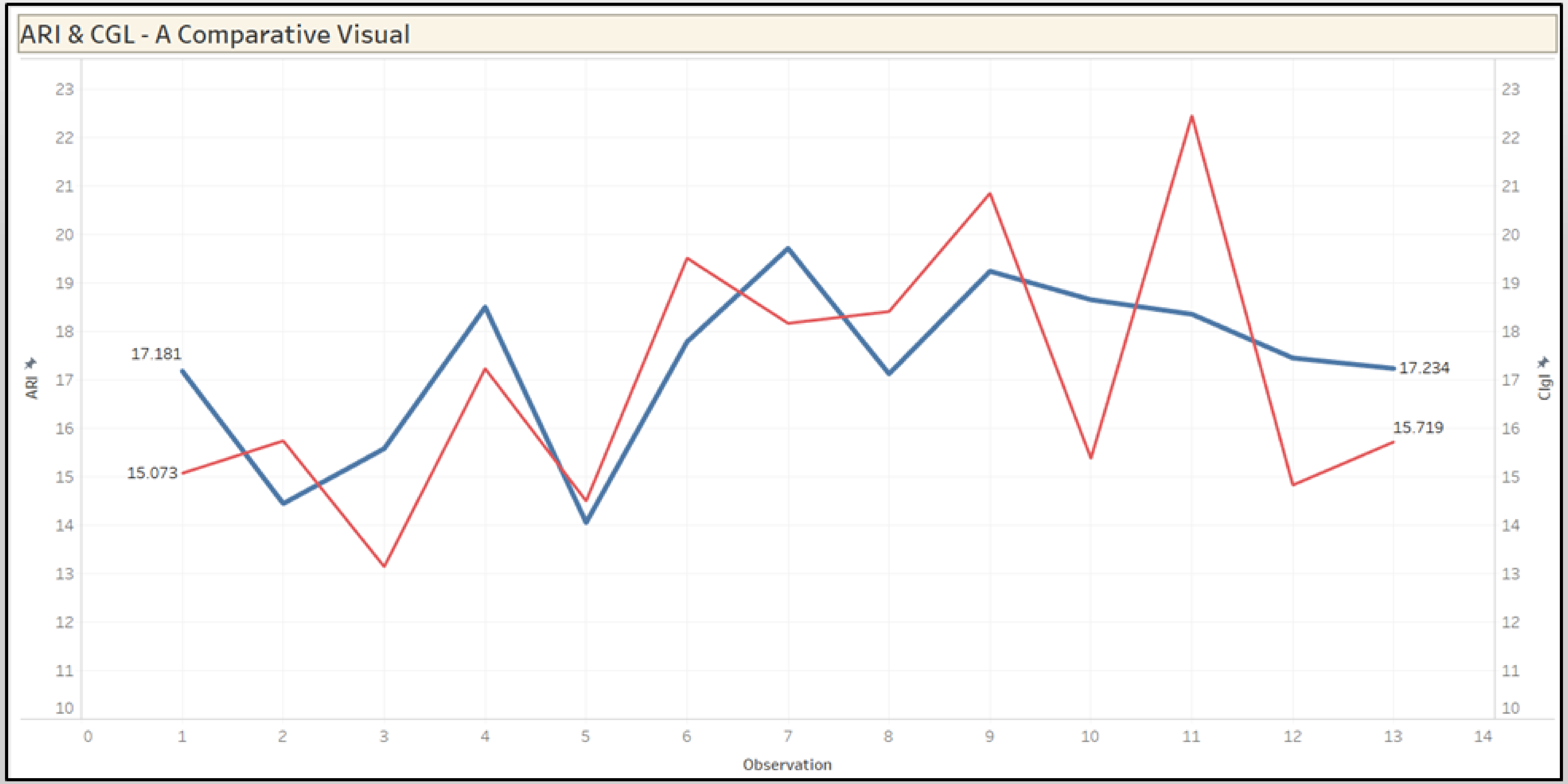
-0.26 is the ARI

1.55 is the CLGL

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

Abstract
Introduction
Methods
Results 1
Results 2
Results 3
Conclusion

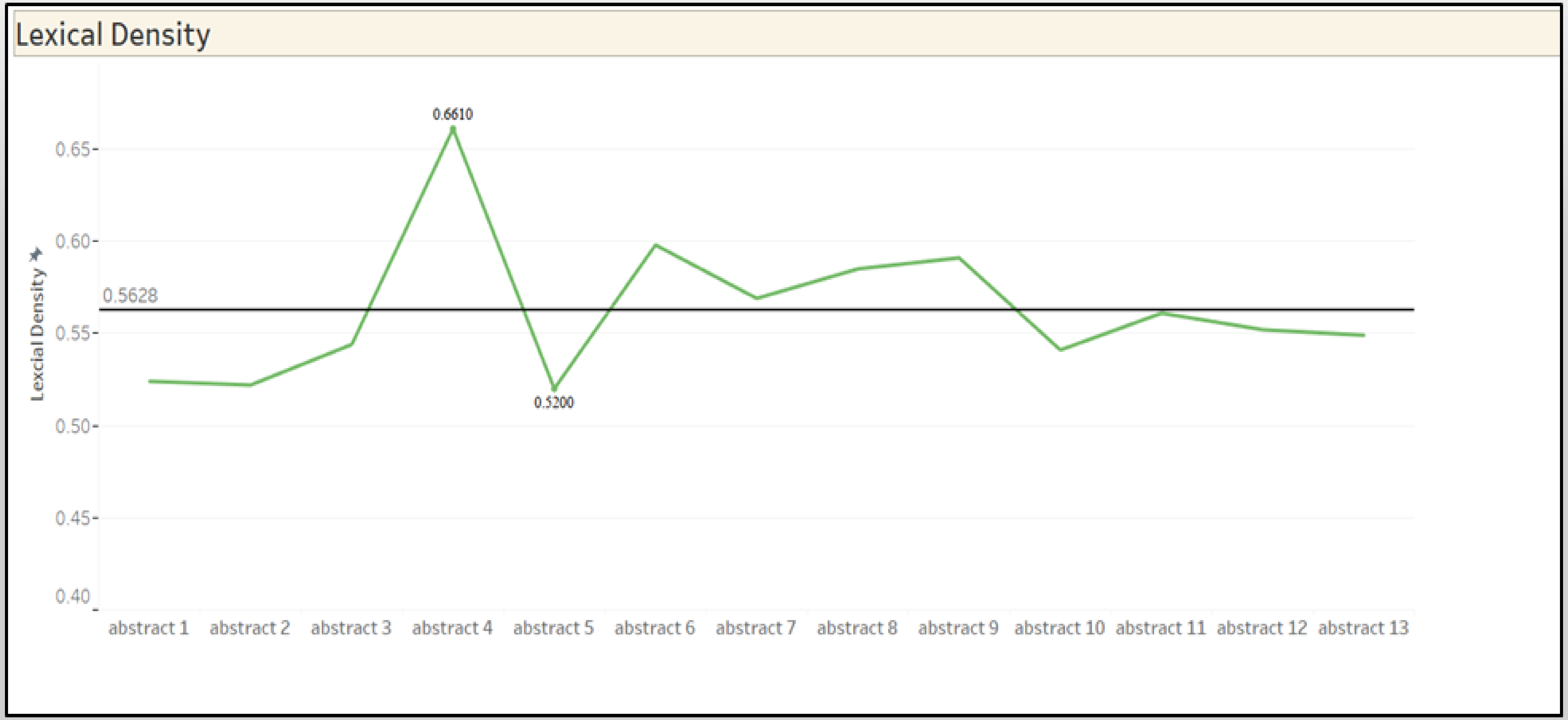


A comparative image of both of the index described in the paper

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

- Abstract
- Introduction
- Methods
- Results 1
- Results 2
- Results 3
- Results 4
- Conclusion



Lexical density with range and average value

In Search Of Perfect Writing: A project on linguistics through SAS

Rohit Banerjee

- Tracking the writing quality of a set of students over their college tenure
- Finding the important words that are very specific to the subjects
- Considering the Lexical diversity of a text along with lexical density

Abstract
Introduction
Methods
Results 1
Results 2
Results 3
Results 4
Conclusion

References

Text Analytics: the convergence of Big Data and Artificial Intelligence - Antonio Moreno, Teófilo Redondo

Lexical diversity and lexical density in speech and writing: a developmental perspective - Victoria Johansson
Dickens vs. Hemingway: Text Analysis and Readability Statistics in Base SAS® Jessica Hampton

Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) For Navy Enlisted Personnel - J. Peter Kincaid

#SASGF

SAS® GLOBAL FORUM 2019

APRIL 28 – MAY 1, 2019 | DALLAS, TX

Kay Bailey Hutchison Convention Center