Do-It-Yourself Forensic Linguistic techniques for authorship identification, ghostwriting and plagiarism detection

Olumide Popoola, FHEA

Queen Mary University of London

olu@outliar.blog / @oepopoola

European Network for Academic Integrity Webinar 09/06/2023
https://www.academicintegrity.eu/wp/enai-monthly-webinars/

About your presenter

- Forensic/Investigative Linguist
  - Deception Detection
    - Investigative interview analysis
    - Fake Amazon reviews
    - Fake News
    - Fake Essays
    - Fake Research
    - Outliar.blog
- Education Developer
  - Assessment Integrity
  - Social Justice Pedagogy

What is forensic linguistics?

Language of evidence

Language to support the delivery of justice

Language of law

Detecting academic misconduct

Plagiarism

- Intrinsic
- Extrinsic
- Intentional vs. Unintentional

Ghostwriting

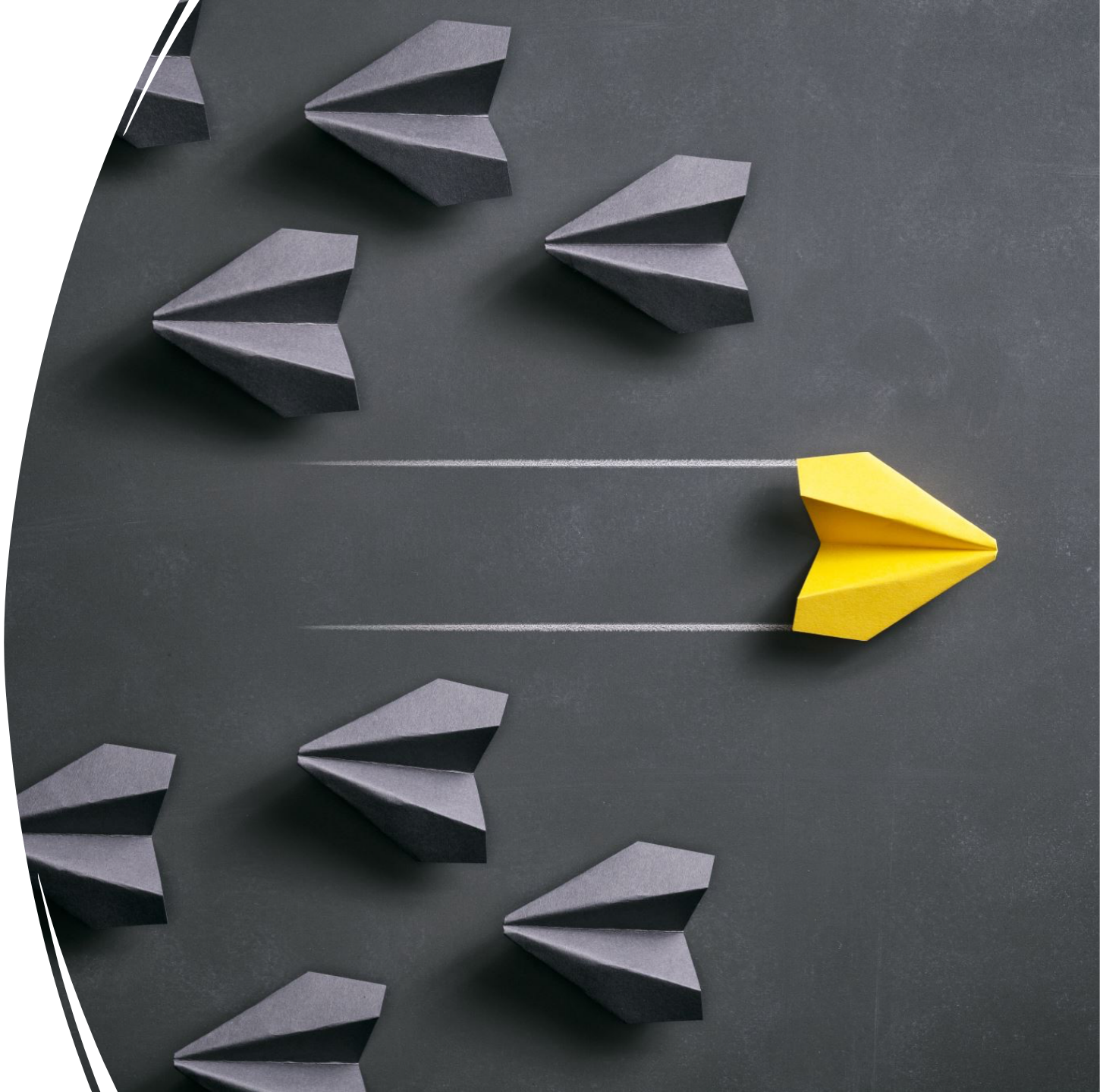
- Contract cheating
  - Essay Mills
  - Generative AI

What language might be used as *evidence* of misconduct such as plagiarism/ ghostwriting?

- Language that can identify authorship
- Language that indicates capability
  - Language proficiency
  - Subject knowledge
- Language that indicates mental state (mens rea)
  - Intention
    - Manipulation
    - Obfuscation
    - Revision

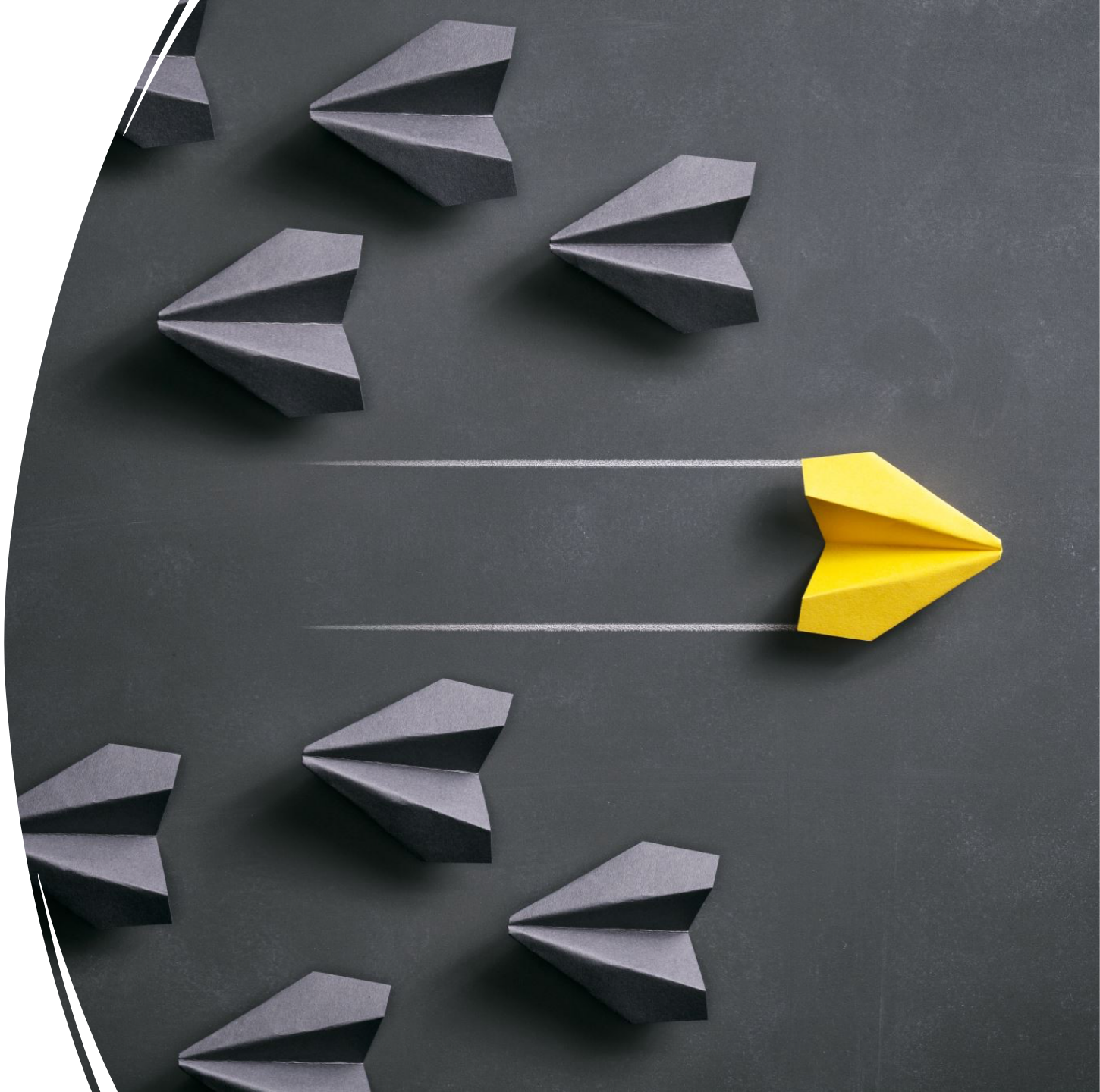
Linguistic feature engineering

Which 'bits of language' can be used to identify the author of a text?



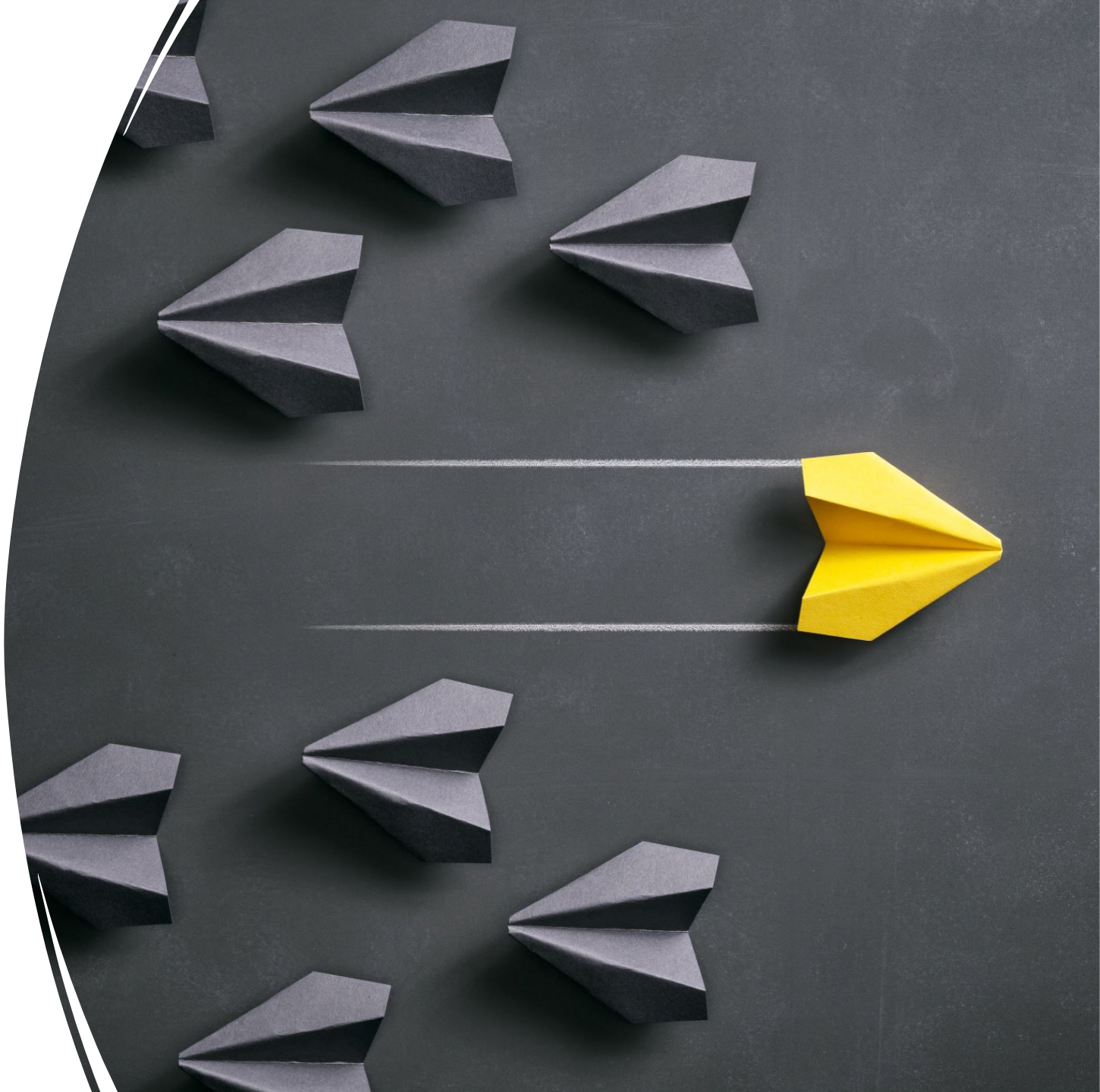
Linguistic feature engineering

The comparative evaluation of numerous shape of Passive Vane Vortex Generator is the main goal that is performed in this paper. We have taken special shapes of vortex generator like gothic, rectangle and triangle shapes for the analysis based on different angle of attack. The principal causes of aerodynamic drag are the separation of flow near the aerodynamic object's rear end. To control the flow separation we install a device called vortex generator.



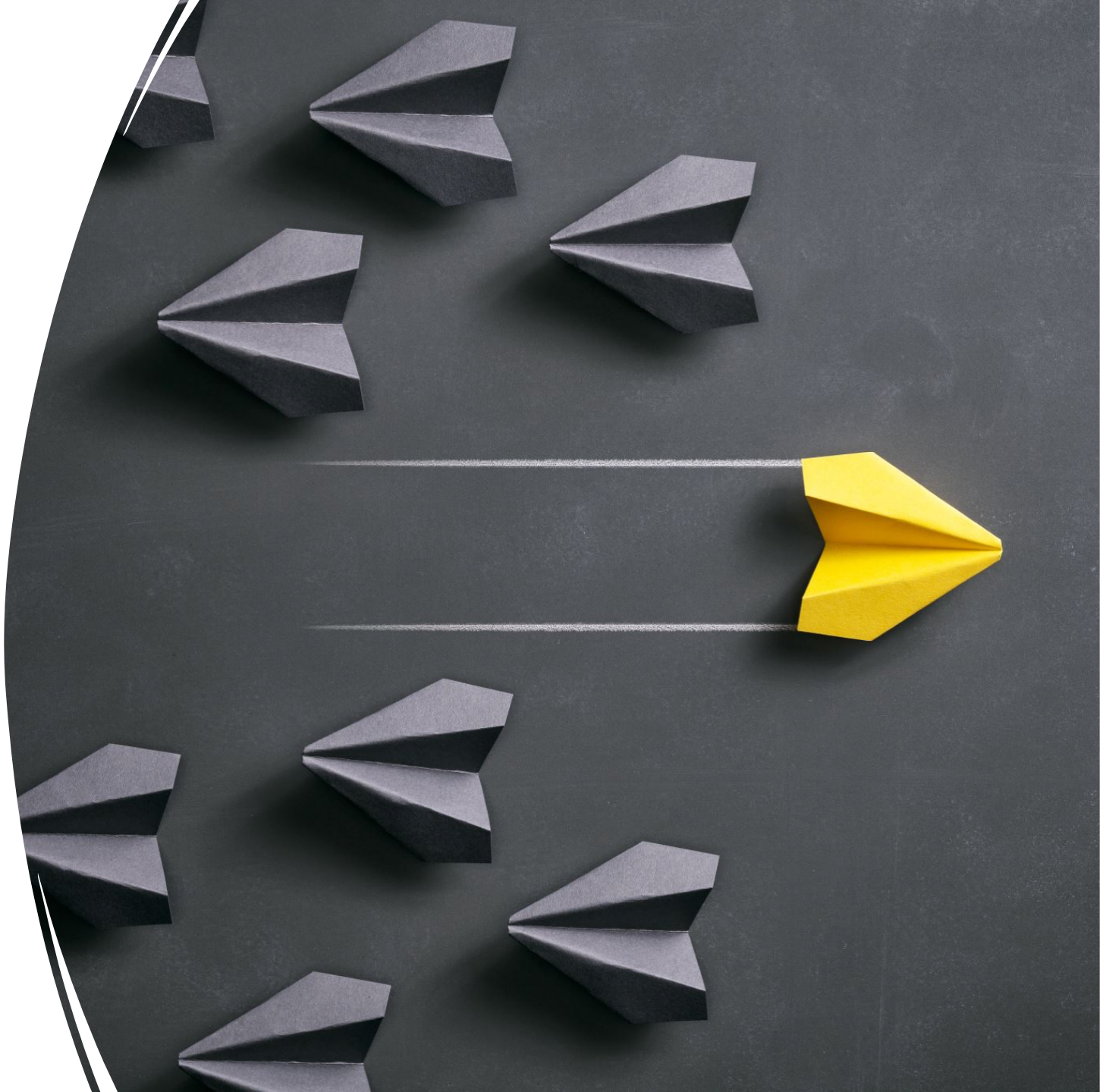
Linguistic feature engineering

The comparative evaluation of numerous shape of Passive Vane Vortex Generator is the main goal that is performed in this paper. We have taken special shapes of vortex generator like gothic, rectangle and triangle shapes for the analysis based on different angle of attack. The principal causes of aerodynamic drag are the separation of flow near the aerodynamic object's rear end. To control the flow separation we install a device called vortex generator.



Linguistic feature engineering

The comparative evaluation of numerous shape of Passive Vane Vortex Generator is the main goal that is performed in this paper. We have taken special shapes of vortex generator like gothic, rectangle and triangle shapes for the analysis based on different angle of attack. The principal causes of aerodynamic drag are the separation of flow near the aerodynamic object's rear end. To control the flow separation we install a device called vortex generator.



Linguistic *features* used to identify potential ghostwriting (Popoola, 2023)

- Lexical sophistication

Luo (1996) noted that **strategic alliances** can be thought of as an organising framework where **partnership** and **relationships** facilitate the knowledge and **capabilities** required to sustain an international growth strategy. Perceptions of **strategic alliances** from a Chinese perspective have also been explored with Dong and Glaister (2007) exploring the **cultural** differences from a Chinese perspective.

Linguistic *features* used to identify potential ghostwriting(Popoola, 2023)

- Vague reference

It would be up to the Canadian All Reds to show that **they** did have goodwill in the goods and the logo; **they** would also have to prove that there had been some form of false representation, **whether it** was intentional or not, to the public, by virtue of the goods being offered by John. For **this** to be the case, **it** will be necessary for **them** to show that there is a likelihood that the public would be deceived, but it has been established that the standard is not that of a 'moron in a hurry', but rather the public at large. The court will determine whether or not there is a similarity in terms of the goods. **This** may result in a difference of opinion in terms of whether or not the scarves without the words ''All Reds' on them would be deemed passing off, in comparison to the ones without the words on the scarves.

Linguistic *features* used to identify potential ghostwriting(Popoola, 2023)

- Adverbials

Moreover, it has been stated that Brexit has high probabilities of affecting not **only** the United Kingdom but **also** the rest of the EU economy through various transmission channels, for instance, uncertainty, trade, investment, as well as migration. **In addition,** it is evident that in the near term, the major effect of Brexit is heightened uncertainty, both political and economic. **Accordingly**, these issues are likely to slow investment growth and private consumption, **as well as** affect foreign trade, **primarily** in the United Kingdom; even though other EU member states also are likely to be **adversely** affected by Brexit. **Also,** Brexit has caused unexpected exchange rate fluctuations, as well as financial market instability.

Linguistic *features* used to identify potential ghostwriting(Popoola, 2023)

- Summarising nouns

This piece has shown how essential it is that **the** approach to care is adapted to **the** individuals' need to reduce distress and enhance their quality of care. Implementation of **the** butterfly scheme was helpful to a degree in **this** particular scenario but I also recognise that not all staff adapted their practice because of this. **This** piece has demonstrated **the** complexity of delivering care for a person with a communication difficulty and highlights that provision of care is largely influenced by personal attitudes and beliefs towards care delivery. **This** piece has illustrated **the** importance of not using medical jargon when communicating with patients, particularly those with Dementia as **this** could exacerbate confusion and cause distress.

Linguistic *features* used to identify potential ghostwriting(Popoola, 2023)

- Lexical Density/Sparsity

Low lexical sparsity Sample I (commercial)	High lexical sparsity Sample J (student)
Research **has therefore** suggested **that another** significant benefit **of** breastfeeding **may be that it** acts **as a** protective factor **against** obesity **in** childhood. Kramer **was the** first **to** report **that** breastfeeding **may** result **in a** "significantly reduced" risk **of** obesity **in** children (1981, p. 4). **In the** next two decades, **a** number **of** similar studies **also** suggested **an** association **between** breastfeeding **and a** reduction **in the** risk **of** childhood obesity. **In the** mid-2000s **this** research **was** collated **into** three seminal meta-analyses **which** concluded **that**, overall, breastfeeding **for the** first six months **did** reduce **the** risk **of** childhood obesity.	Difficulties **in** gaining admission **to** inpatient beds (i.e. inefficient bed management **or** insufficient bed capacity) **The** congestion **in the** emergency department. Incorrect retention **of** patient beds. Need **for** improving different administrative processes associated **with** patient flow arises **for** efficient **and** effective management **of** hospital beds **and** other resources. **The** effective management **of** hospital beds is essential **if the** growing demand **of** inpatient beds is **to** be met. **With the** limited supply **of the** medical resources **and** excess **of** demand, **the** hospital beds are **in** short supply.

Linguistic *features* used to identify potential academic misconduct (Popoola, 2023)

- Informality

> Well, the main reason that most people **don't** like sales is because of having to deal with rejection. No one likes to be rejected but if **you're** in a sale, **that's** all part of the game. The more rejections you get, the closer to a sale you will be. Now just because you expect your sales people or yourself to go out there and make those sales calls like a machine, it **doesn't** mean motivation should be neglected. If you are a sales person, take the time to read and listen to motivation material. By **doing** this, you will constantly be feeding your mind with positive and encouraging thoughts that will help you get through those days where everyone prospect seems to be in a bad mood.

Linguistic *features* used to identify potential academic misconduct (Popoola, 2023)

- Colloquialisms

Basic steps in the money laundering process are showing below, Placement: In this step large **amount** of **black money** placed into the financial system, used to **buy high dollar** goods or smuggled out of the country. This idea is to transform the **cash** as quickly as possible into other types of assets and thus avoid detection. Cash deposited into bank often with complicity of staff or **mixed** with proceeds of legitimate business. In placement process cash are physically transported out of the country. Cash is used to **buy high** value goods, properties or business assets.

Proprietary Tools

- **Turnitin**
  - Similarity Report
    - String matching
  - Authorship Investigate
    - Some stylometric features…[find them]
      - Computational = Black box for novices.
      - Features not related to learning, marking or the student.
- **AI Detectors (**Weber-Wulff, Foltýnek *et al. forthcoming)*
    - Perplexity
    - Burstiness

DIY Tools

Microsoft Editor for Word

- Flesch-Kincaid Readability Metrics
  - Word length
  - Sentence length

- Basic measure of text complexity
- Different readability metrics
  - *ARTE – Automatic Readability Tool for English* (Choi and Crossley, 2022)

DIY Tools

AntWordProfiler (Laurence Anthony, 2022)
https://www.laurenceanthony.net/software/antwordprofiler/

- Vocabulary analysis of a text using predefined word lists
  - Academic Word List (AWL)
  - General Service List (GSL)
  - Subject specific lists
  - Can create own word lists
- Language Level
- Domain knowledge

DIY Tools

Similarity Texter (Kalaidopolou, 2016)
https://people.f4.htw-berlin.de/~weberwu/simtexter/app.html

- Side-by-side text comparison tool
- Useful for patchwriting (Howard, 1992) detection
- Unacknowledged patchwriting is good example of *intentional* plagiarism

Now let’s try!

Case 1: Ghostwriter?

Text A

The comparative evaluation of numerous shape of Passive Vane Vortex Generator is the main goal that is performed in this paper. We have taken special shapes of vortex generator like gothic, rectangle and triangle shape the analysis based on different angle of attack. The principal causes of aerodynamic drag are the separation of flow near the aerodynamic object' rear end. To control the flow separation we install a device called vortex generator.

Rationale :

A vortex generator is an aerodynamic fin-like device attached on the wing surface. When the aircraft is in motion relative to the air, the vortex g generates vortex ,by doing this some part of the slow moving boundary lay over the wing eliminates, delay flow separation and improve the effectiver of wings.

Vortex generators may be the perfect option for reducing flow separation. Each of these tiny components generates a spinning wake that gives energ into the wing's boundary layer. The result is lower stall speed and higher critical angle of attack. The advantages and disadvantage of the performar of the device have to shape precisely.

The vortex generator, in the flow around the wings control the boundary layer.Tarbulant boundary is more resistance to separate. Therefore it is possible to fly at lower speed and higher angle of attack. It is really import to properly place the vortex generator. They should be placed precisely in t boundary layer transition zone.

Text B

Abstract

This report investigates the effects of the Vortex Generators on the aerodynamic performance of the NACA 4415 aerofoil. Simscale software is used to carry out the simulation. At the upper surface of the wing, triangular vortex generators were added. Vortex Generators (VGs) were used to reduce drag force, increase lift force, and delay boundary separation at the profile trailing edge. It was found that the Vortex Generators are a cheap and effective way to improve wing performance at a range of high attack angles.

Motivation

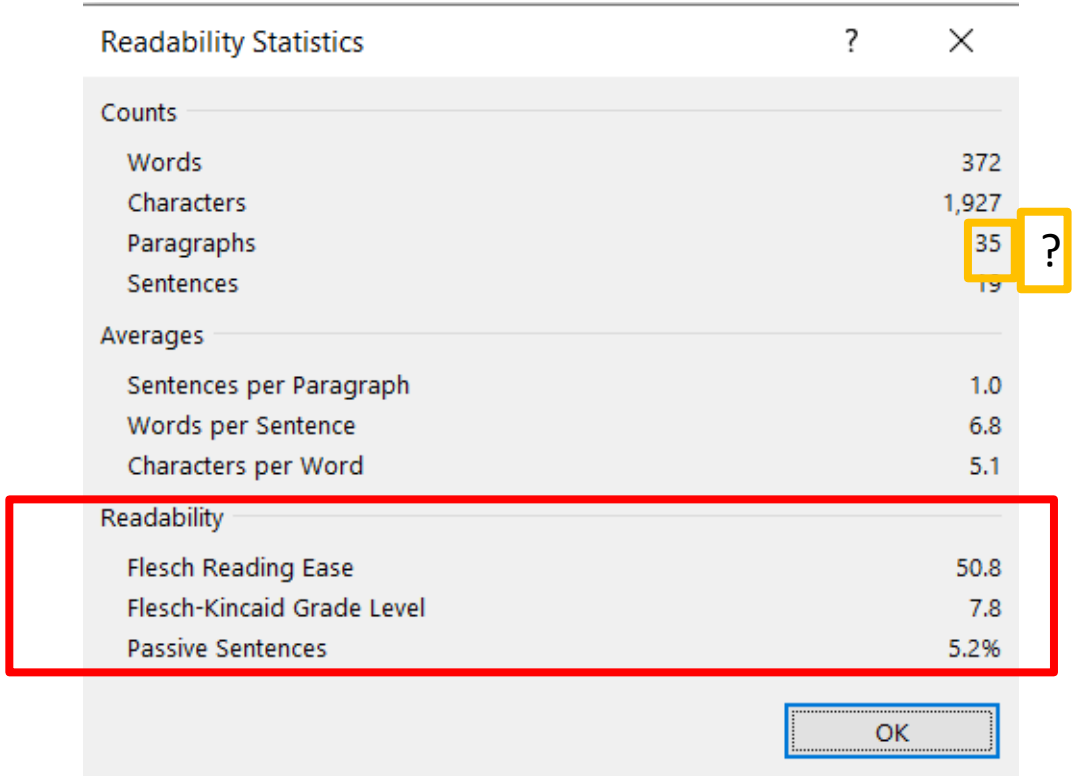
With the rising oil price and the environmental issue about harmful exhaust, particulate, and greenhouse gases, reducing fuel consumption is a key issue for vehicle manufactures and those who use them. Large progress was made in engine performance, aerodynamic drag, and other areas over the past few decades, but there is still room for further advancement. This research addresses a basic aerodynamic problem: how to control boundary layer flow separation on the surface of the wing a process that in most cases results in a sudden loss of lift, wing stall, and drag. To avoid separation during takeoff and landing, a device called Vortex Generators is used on the upper surface of the wing.

Literature Review

Flow control device, such as Blowing and suction [Kametani.2015], Vortex Generators (VGs) [Li, X.2019], Synthetic jets [Ziade.2018], and Flexible walls [Yang, H.2018], used Vortex Generators to improve the aerodynamic performance of wind turbine blade. These flow control methods were compared by Johnson [2010] and Barlas [2010]. Vortex Generators are an effective device to enhance the aerodynamic efficiency of the blade, according to Wang [2017] and Lin [2002]. To improve wing flow control, Taylor [1947] was the first to use Vortex Generators . [Khoshvaght. 2015, Skullong.2016]The fundamental theory to control flow separation with Vortex Generator is that a concentrated vortex is formed downstream when the fluid moves through a Vortex Generator. Because of the concentrated vortex, the

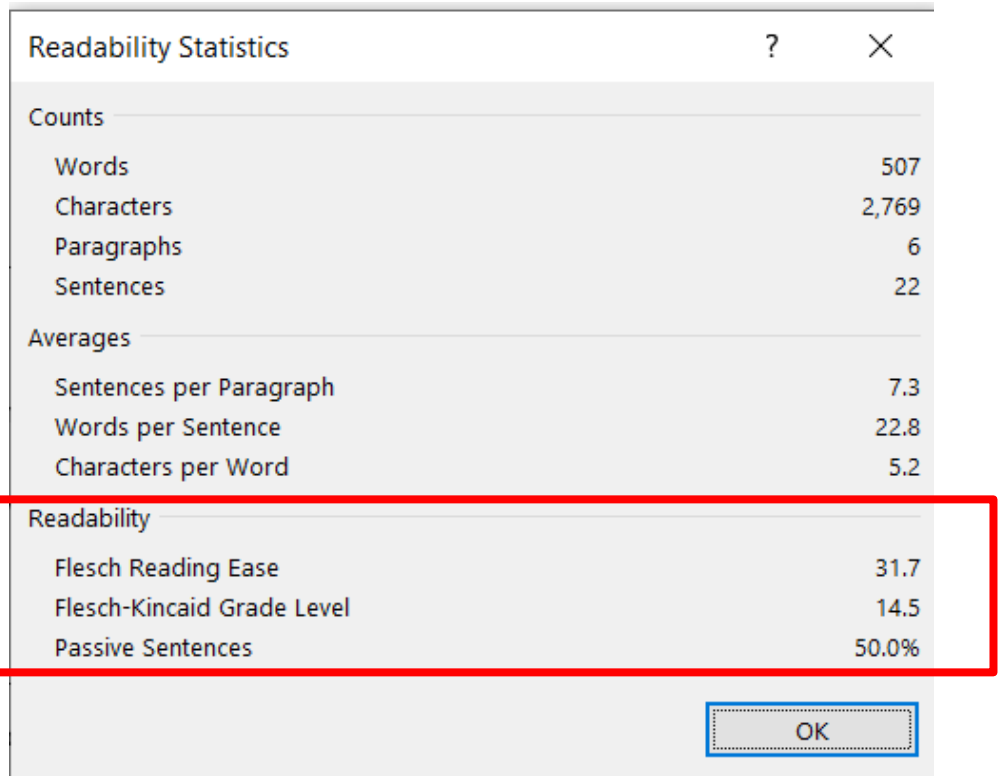
Readability: *Microsoft Editor for Word* output

Student Text A



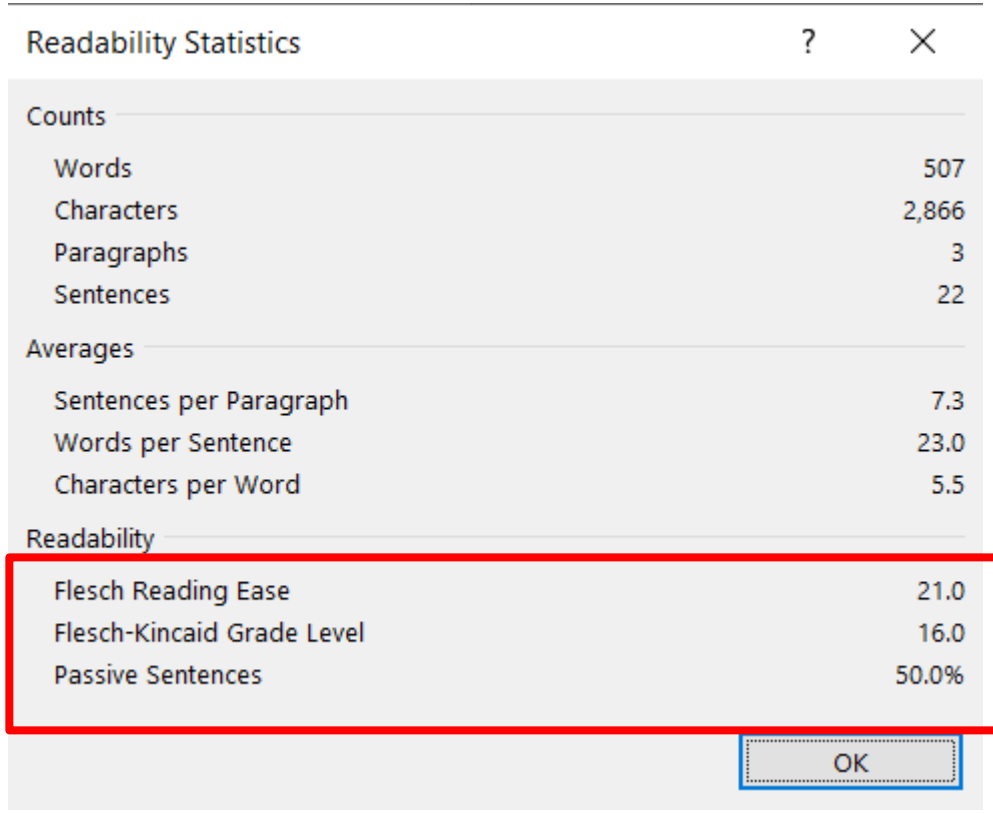
?

Student Text B1

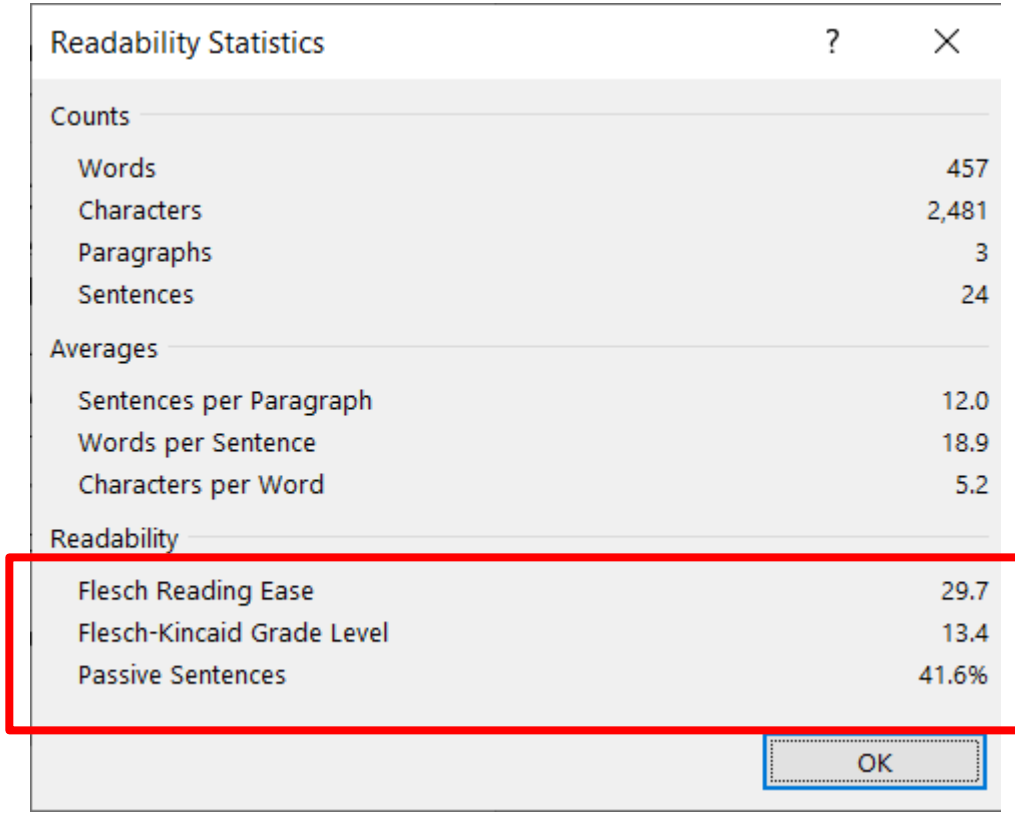


Readability: *Microsoft Editor for Word* output

Student Text B2

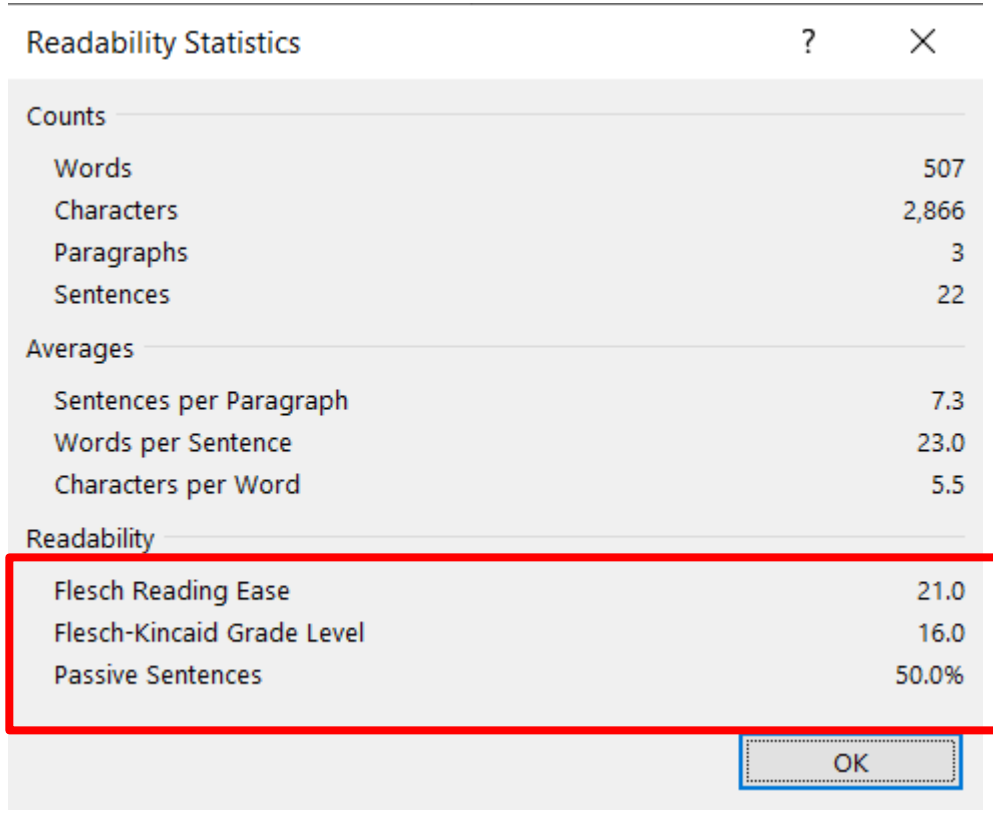


Student Text B3

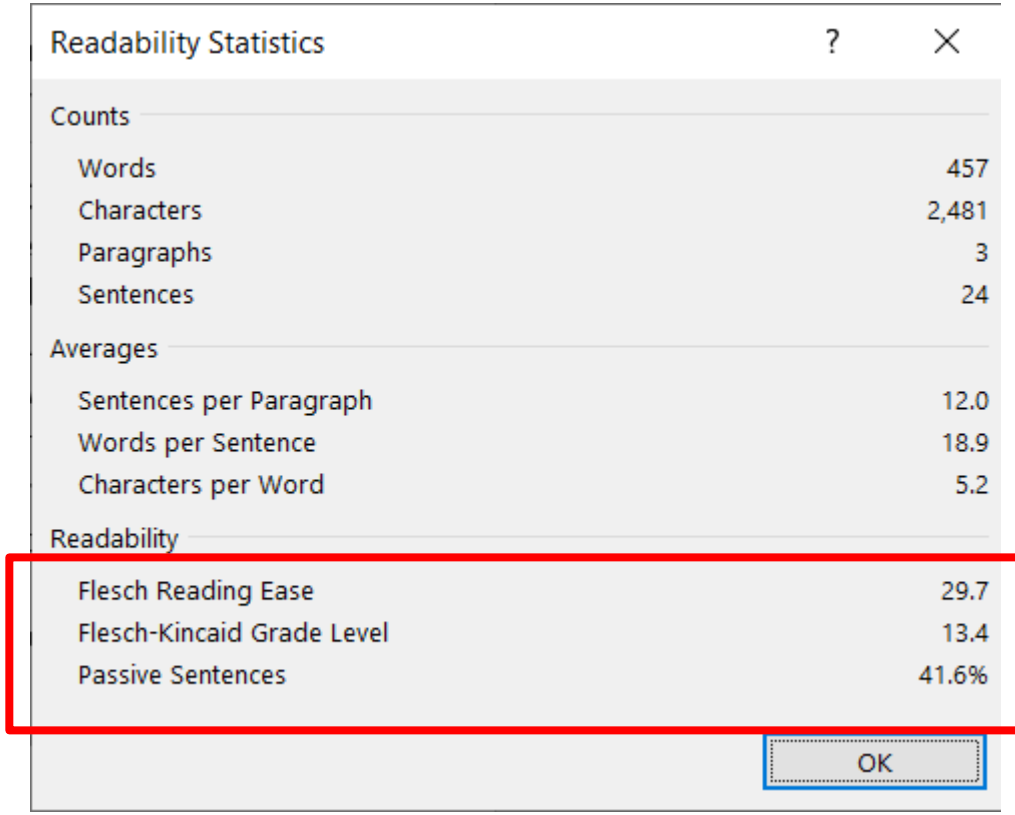


Readability: *Microsoft Editor for Word* output

Student Text B2

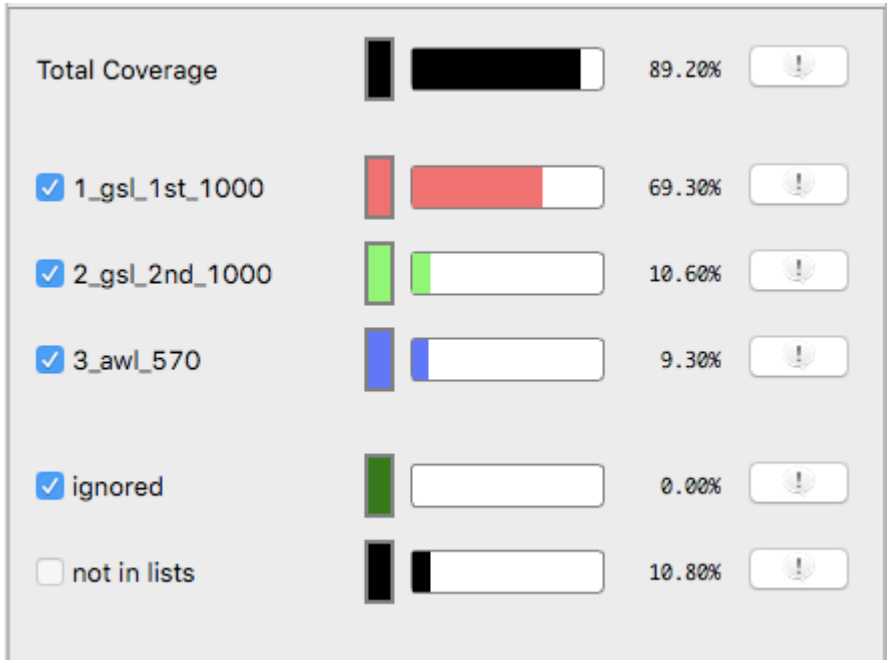


Student Text B3

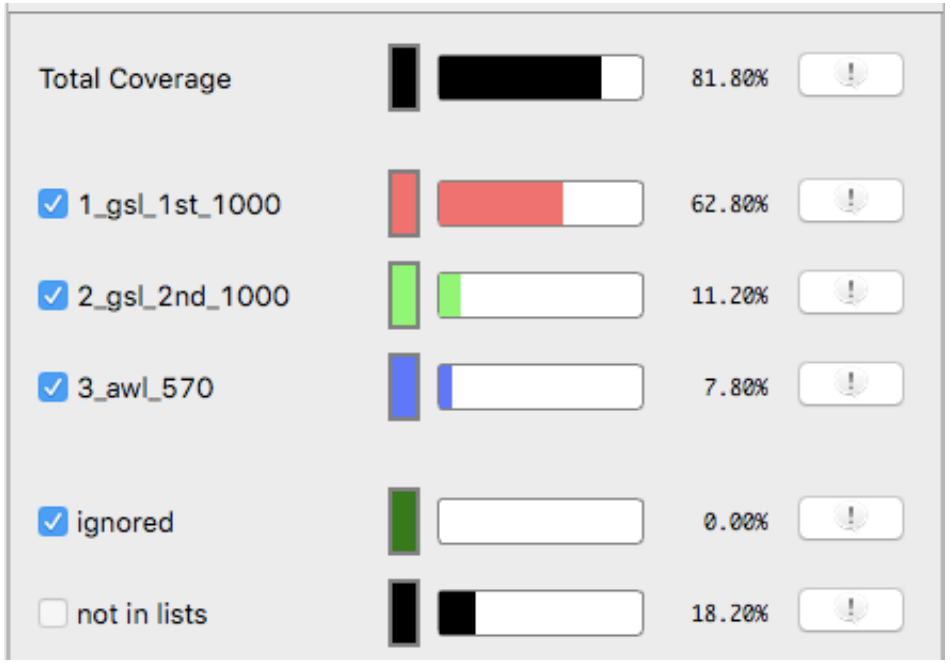


Vocabulary Analysis: *AntWordProfiler* output

Student Text A

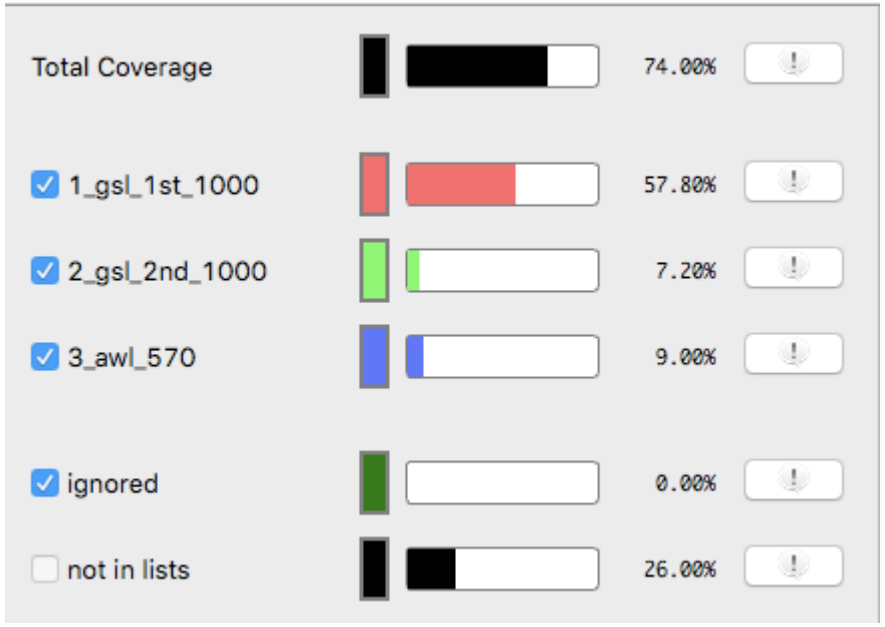


Student Text B1

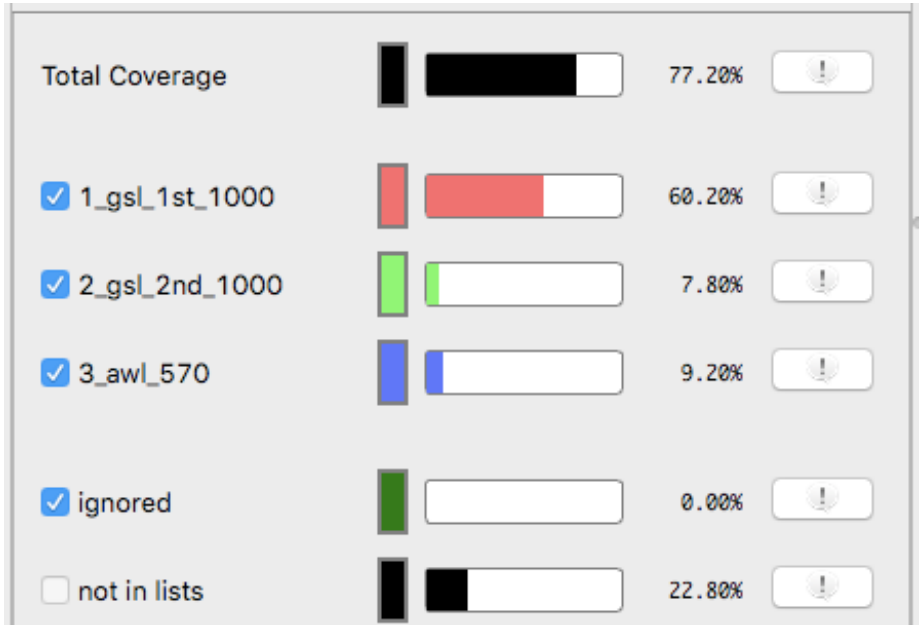


Vocabulary Analysis: *AntWordProfiler* output

Student Text B2



Student Text B3



Vocabulary Range: *AntWordProfiler*

	Text A	Text B (Overall)	Text B1	Text B2	Text B3
GSL 1st 1000	69.3%	60.3%	62.8	57.8%	60.2%
GSL 2nd 1000	10.6%	8.8%	11.2	7.2%	7.8%
AWL	9.3%	8.7%	7.8%	9.0%	9.2%
Not in lists	10.8%	22.3%	18.2%	26.0%	22.8%

Now let’s try!

Case 2: Intentional plagiarism?

I combine commitment to research with my passion for teaching. I am a Teaching Fellow in the Networks group at Kings College. My comprehensive background on computer networks and systems qualifies me to teach advanced classes in such domains, as well as introductory computer science classes such as data structures, algorithms, network architectures and protocols. As researcher, I am looking at new ways to improve networks management and performance. 21st Century Networks course highly matches my interests and background as it focusses on new trends and solutions in relations to computer networks design and implementation, covering aspects of engineering and architectures.

My teaching style is highly interactive, aimed at engaging students with frequent questions during lectures, while my main goal is to push students to explore real-world problems, provide them with the necessary skills to understand the trade-offs when moving from theory to practice, and to enable them to effectively communicate with others. In this regard, the content of the module as well as its structure highly matches my teaching philosophy. This is mainly because the course focusses on showing how communications networks are evolving over the time, touching upon real-world examples.

I couple my research ambition with my passion for teaching. I am a Teaching Fellow in the School of Engineering at the Kings College. My vast experience and research work in distributed systems, machine learning and computer networks qualify me to teach advanced computer science topics in these areas. As a systems researcher, I always seek
new ways to improve and optimize systems that our applications rely on to fulfil their goals. The Machine Learning Systems module matches my interests and background as it focuses on new techniques and methods in relation to systems supporting machine learning applications, covering their aspects of design, engineering and architecture.

I strive to maintain a highly interactive teaching style by engaging students with frequent questions during lectures, labs or group-based activities. Moreover, other key goals are to enable the students to effectively present ideas and communicate with others, provide them with the necessary skills to understand the trade-offs when moving from theory to practice, and, most importantly, push them to explore real-world problems.

Side-by-side comparison: *Similarity Texter*

TEXT: Plain text input

I combine commitment to research with my passion for teaching. I am a Teaching Fellow in the Networks group at Kings College. My comprehensive background on computer networks and systems qualifies me to teach advanced classes in such domains, as well as introductory computer science classes such as data structures, algorithms, network architectures and protocols. As researcher, I am looking at new ways to improve networks management and performance. 21st Century Networks course highly matches my interests and background as it focusses on new trends and solutions in relations to computer networks design and implementation, covering aspects of engineering and architectures.

My teaching style is highly interactive, aimed at engaging students with frequent questions during lectures, while my main goal is to push students to explore real-world problems, provide them with the necessary skills to understand the trade-offs when moving from theory to practice, and to enable them to effectively communicate with others. In this regard, the content of the module as well as its structure highly matches my teaching philosophy. This is mainly because the course focusses on showing how communications networks are evolving over the time, touching upon real-world examples.

TEXT: Plain text input

I couple my research ambition with my passion for teaching. I am a Teaching Fellow in the School of Engineering at the Kings College. My vast experience and research work in distributed systems, machine learning and computer networks qualify me to teach advanced computer science topics in these areas. As a systems researcher, I always seek
new ways to improve and optimize systems that our applications rely on to fulfil their goals. The Machine Learning Systems module matches my interests and background as it focuses on new techniques and methods in relation to systems supporting machine learning applications, covering their aspects of design, engineering and architecture.

I strive to maintain a highly interactive teaching style by engaging students with frequent questions during lectures, labs or group-based activities. Moreover, other key goals are to enable the students to effectively present ideas and communicate with others, provide them with the necessary skills to understand the trade-offs when moving from
theory to practice, and, most importantly, push them to explore real-world problems.

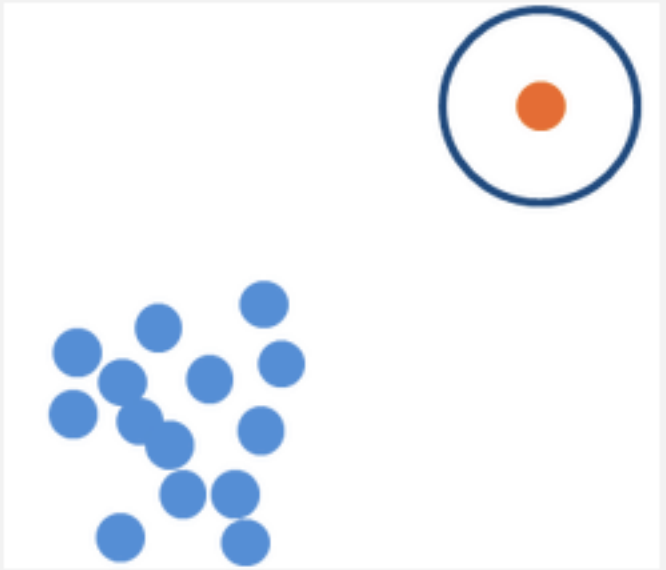
References

Choi, J.S. and Crossley, S.A. (2022), July. Advances in Readability Research: A New Readability Web App for English. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 1-5). IEEE.

Howard, R. (1992). A plagiarism pentimento. *Journal of teaching writing*, *11*(2), pp.233-45.

Popoola, O. (2023). Decision Support for Marker Detection of Contract Cheating: An Investigative Corpus Linguistic Approach. In: Bjelobaba, S., Foltýnek, T., Glendinning, I., Krásničan, V., Dlabolová, D.H. (eds) Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Ethics and Integrity in Educational Contexts, vol 4. Springer, Cham. Available at: https://rdcu.be/da6Zm

Weber-Wulff, D., Foltýnek, T., Anohina-Naumeca, A., Bjelobaba, S., Guerrero-Dib, J., Šigut, J., Popoola, O., Waddington, L. (forthcoming) *Testing of Detection Tools for AI-Generated Text*



OutLiar – The Deception Blog

LINGUISTIC ANALYSIS OF DECEPTION, DISINFORMATION AND DOUBLESPEAK. ALLEGEDLY.

Email: olu@outliar.blog Twitter: @oepopoola |Website: http://outliar.blog