*Data paper*

**Difference between academic and non-academic language in structure**

Yin Chun Chan s1154133@s.eduhk.hk\*, Kin Pang So and Man Huen Wong

*a Faculty of Humanities, The Education University of Hong Kong, Hong Kong Special Administration Region*

*b Faculty of Humanities, The Education University of Hong Kong, Hong Kong Special Administration Region*

*c Faculty of Humanities, The Education University of Hong Kong, Hong Kong Special Administration Region*

\*Corresponding author: first name, initials, surname; email address

**Author roles**

Conceptualization: Yin Chun Chan

Data Curation: Kin Pang So

Formal Analysis: Man Huen Wong

Investigation: Man Huen Wong

Methodology: Kin Pang So, Man Huen Wong

Project Administration: Yin Chun Chan

Software: Kin Pang So, Man Huen Wong

Resources: Man Huen Wong

Supervision: Yin Chun Chan

Validation: Man Huen Wong

Visualization: Yin Chun Chan, Kin Pang So

Writing -- Original text: Yin Chun Chan

Writing -- review & editing: Man Huen Wong

**Abstract**

This dataset investigates the variations in structure and delivery between academic and non-academic languages. It contains texts and audio/video recordings from a variety of sources, such as academic papers, lectures, news items, blog entries, and casual chats. Linguistic factors such as sentence complexity, word selection, jargon usage, and discourse markers are examined. Speech pace, intonation, and body language are among the delivery elements considered. The data is kept in a structured format (e.g .txt,. ipynb, .spv) with annotations for language characteristics. Researchers in languages, education, and communication can benefit greatly from this dataset. It may be used for developing language learning tools, improving communication skills, and analyzing the influence of language on diverse audiences.

Keywords: linguistic features; communication; lexical relation; context; structure

**(1) Overview**

***Repository location*** – <https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure-and-delivery.git>

***Context*** – This dataset was produced as part of a research project conducted by the course called Language Information Management (Course code: LIN3046) in The Education University of Hong Kong. The project aimed to explore the differences between academic and non-academic language in terms of structure. It involved collecting a range of texts and audio/video recordings from various contexts, including academic lectures, online TV shows, and public talks. The findings aim to contribute to the understanding of how language varies across different settings and to inform language learning.

**(2) Method**

***Steps*** –The dataset was created through a multi-step process. Firstly, we gathered YouTube videos, including university lecture, TED talk, and American online TV show. Secondly, using Python code to do transcription from YouTube video URL link to video recording, video to audio, and then audio to text. Thirdly, using those .txt file to do text annotation to analyze the structural feature between academic and non-academic English. Lastly, using the result from text annotation to do T-Test, for analyze and organize the language features base on the text annotation. Each text and recording were then categorized based on its context (academic vs. non-academic) to facilitate comparison.

***Sampling strategy*** – Texts were selected to represent a diverse range of academic and non-academic contexts, ensuring a comprehensive analysis of language features.

***Quality control*** – To ensure the integrity of the dataset, a quality control process was implemented. This included many rounds of review for the accuracy of annotations in linguistic features.

**(3) Dataset Description**

***Repository name*** – The name of the repository to which the data is uploaded. E.g., Figshare, Dataverse, etc.

***Object name*** – difference-between-academic-and-non-academic-language-in-structure

***Format names and versions*** – Text files (.txt), audio files (.mp3), video files (.mp4), Jupiter Notebook (.ipynb)

***Creation dates*** – 2025/4/16 to 2025/4/18

***Dataset creators*** – Yin Chun Chan, Kin Pang So, Man Huen Wong

***Language*** – English

***License*** – CC BY 4.0

***Publication date*** – 2025/4/18

**(4) Reuse Potential** – This study dataset offers great potential for developing instructional tools and doing communication strategy research. It can improve language learning across a wide range of populations by emphasizing the use of both academic and non-academic language styles. The dataset is also a great resource for scholars studying communication tactics in a variety of situations, including education, business, and media. However, it has drawbacks, such as biases in sample selection and context-specificity. Researchers should be cautious when interpreting and implementing findings, and future studies should broaden the dataset to include more sources and situations.

**Acknowledgements**

This research benefited from insights presented in the YouTube video Justice: What's The Right Thing To Do? Episode 01 "THE MORAL SIDE OF MURDER"' by Harvard University, published on 5th September , 2009 [https://www.youtube.com/watch?v=kBdfcR8hEY&list=PL72C62342291D5DAE&index=2](https://www.youtube.com/watch?v=kBdfcR8hEY&list=PL72C62342291D5DAE&index=2)). This video features perceived as distancing, academic language, when employed strategically, can captivate a receiver by signalling expertise and precision.

This research benefited from the explanations provided in the YouTube video '' by Ted talk, published on April, 2025 <https://www.ted.com/talks/carole_cadwalladr_this_is_what_a_digital_coup_looks_like?subtitle=en>. The video highlighted gestures are essential for grabbing and sustaining a receiver's attention.

The methodology used in this research was inspired by the techniques demonstrated in the YouTube video 'CHICKEN GIRLS | Season 1 | Marathon' by Brat TV, published on 11th March, 2019 <https://www.youtube.com/watch?v=CsK_msr0sVo>. This video helped us to enhance how non-academic language will engage receiver’s interest. builds a sense of connection and comprehension, thereby raising the receiver's attention and readiness to listen.

**Funding statement**

This research received no specific grant from any funding assistance from public, commercial, or non-profit sectors. All work was conducted independently and without external financial support.

**Competing interests**

The authors declare that they have no competing financial, professional, or personal interests that could have influenced the collection, analysis, or presentation of the data described in this data paper. The research was conducted in an academic setting with no external funding that could create a conflict of interest.

**References**

Brat TV. (2019, March 10). CHICKEN GIRLS | Season 1 | Marathon [Video]. YouTube. <https://www.youtube.com/watch?v=CsK_msr0sVo>

Cadwalladr, C. (n.d.). This is what a digital coup looks like [Video]. TED Talks. <https://www.ted.com/talks/carole_cadwalladr_this_is_what_a_digital_coup_looks_like>

Harvard University. (2009, September 4). *Justice: What’s the right thing to do? Episode 01 “THE MORAL SIDE OF MURDER”* [Video]. YouTube. <https://www.youtube.com/watch?v=kBdfcR-8hEY>

**Supplementary Files**

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Abstract.md>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Text_Annotation.ipynb>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Video1_sentence.txt>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/summary-video1.txt>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Video2-sentences.txt>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/summary-video2.txt>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Video3-sentences.txt>

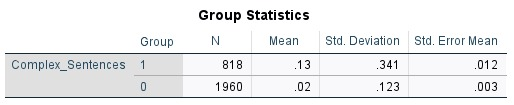
<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/summary-video3.txt>

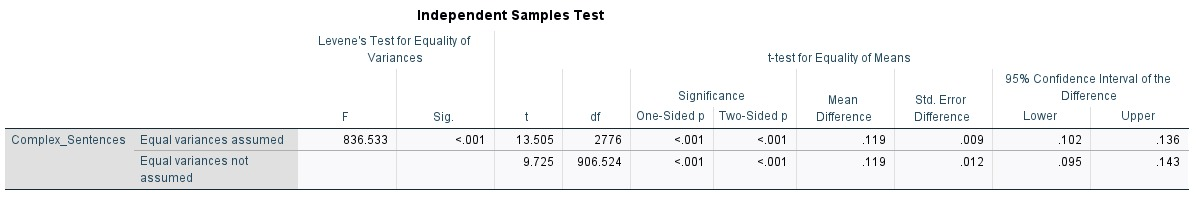
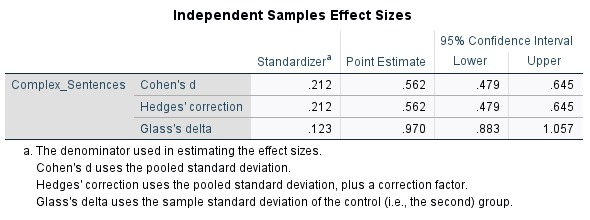
<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/main/Text_Annotation.ipynb>

<https://github.com/kaylawmh/difference-between-academic-and-non-academic-language-in-structure/blob/71acefe3c5e9fbec633447b74ce2765fa9e0047d/T%20test.spv>

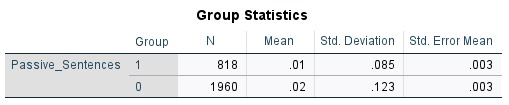
T-Test:

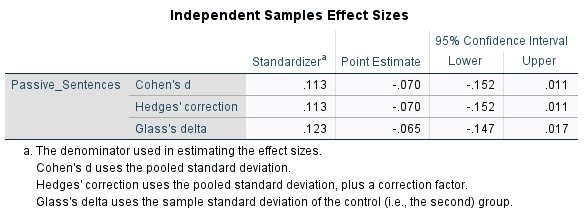
Complex Sentences:

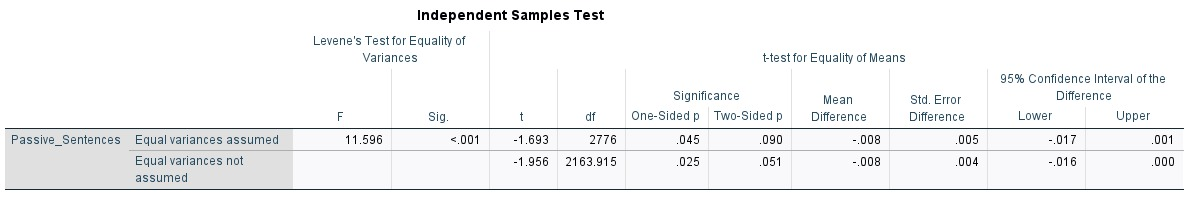




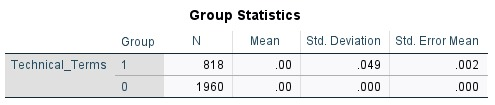
Passive Sentences:

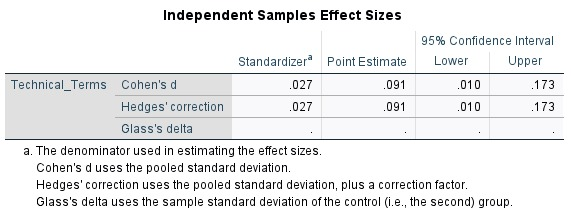


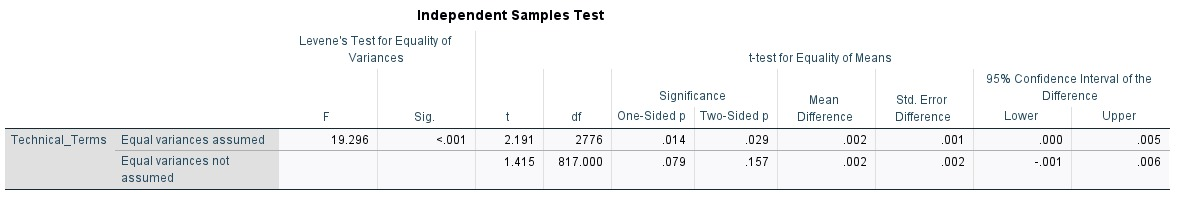




Technical terms:





Jargon TermsL

