Paula Badilla Vásquez
Análisis Computacional de Datos Lingüísticos
Magister en Lingüística
Universidad de Chile

The Use of Academic Language in Social Media: A Dive into YouTube Essays' Register

1. Introduction

The current text has as an objective the creation of a corpus of YouTube essays and use it to quantify the differences in register versus a corpus comprised of university lectures. This is in order to evidence how, despite talking about topics that should and pretend to be shared among the general public, the barrier of academic language works as of to gatekeep this information. It is believed that the YouTube videos will resemble the language of a more academic context, such as the university lessons in question.

Though there is no consensus on the definition of academic language, in general, this notion is defined as the language needed to succeed in school and academics (Bunch & Martin, 2020; Flores & Rosa, 2015; Jensen & Thompson, 2020; Schleppegrell, 2009; Thompson & Watkins, 2021). The issue at hand comes in the form of how the teaching and its apparent prestige naturalizes certain ideas about language. Some of which can be how academic language is much more appropriate than other varieties, sometimes even suggesting a certain superiority (Flores & Rosa, 2015; Thompson & Watkins, 2021). The naturalization of such ideas result in discrimination against any variety that deviates from this standard, sometimes even implying that individuals who cannot understand or produce it have some sort of communication or speech disability, especially when it comes to bilinguals (Thompson & Watkins, 2021).

Due to this naturalization, it is expected that even in social media where information is being shared to the masses is still perpetuating academic language as their register of choice. Therefore, sometimes creating a barrier from the general public, unintentionally (allegedly) gatekeeping this information. There is intention to accuse these creators of discriminating or feeling superior for using academic language. Yet, it is of interest to see how they opt for this specific register, especially the video essayist and commentators community in YouTube, which will be the focus of the current study.

Academic has many different features, however only two will be researched, at least for the moment being. The first one is the lexicon. There are several words that are considered more academic and usually found almost exclusively in academia. The other is nominalization. This is process in which the information of a whole clause is condensed in a single noun. Its goal is to pack more information into sentences, which is seem as functional for developing theories and explanations (Jensen & Thompson, 2020).

Computational resources have been used to carry out this study due to how it facilitates the study of multiple corpora at the same time. It streamlines and automatizes the analysis portion, making it the ideal way to draw conclusions from the existing texts.

2. Methodology

2.1. Corpus Creation

The first objective of this research was to create a corpus of YouTube video essays in order to then carry out a comparison between this corpus and an academic one. In order to do this, YouTube videos were found that contained the following characteristics:

- Manually written subtitles. This is because the transcriptions will be used to create the corpus itself. Furthermore, even though automatic close captions are becoming more accurate, there are still some considerable errors, reason why it was decided to use the ones that had been passed through a human filter beforehand.
- YouTube channels which contents focuses on video essays or video commentary[1]. This is to make sure the vocabulary used is as similar as possible among all the videos. Because of this, similar topics have also been chosen.

With this in mind, eight channels were selected, taking three videos from each one, giving a total of twenty-four video transcripts. All of the creators come from an English-speaking country, are between 20 or 30 years of age, and identify as female. As stated before, similar topics were chosen too. These were: beauty, a review of a film or series, a topic related to LGBTQ+ rights (mainly trans rights), and the phenomenon of *cancelling* or *cancel culture*.

After this selection, the subtitles of each video were pasted into the notepad app of Windows, as so to create a .txt file containing each video separately. These can be found on GitHub at https://github.com/no-you-shouldnt/The-Use-of-Academic-Language-in-Social-Media/tree/main/YTessays .

2.2 Corpora Preprocessing and Description

As explained before, the YouTube corpus will be compared to an academic corpus. The latter corpus comes from the British Academic Spoken English (BASE) corpus. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. Eight classes from three courses that relate to the topics of the YouTube corpus were selected, so, again, the lexicon would be as similar as possible. The courses in question come from: Sociology, Film and Television Studies, and Politics. The entire corpus can be found at http://www.reading.ac.uk/acadepts/ll/base_corpus/ and the selected classes at https://github.com/no-you-shouldnt/The-Use-of-Academic-Language-in-Social-Media/tree/main/Academic .

Each corpus was cleaned of any parenthesis and their contents, divided by lines, tokenized by sentences and words, put all in lower case, and stripped of any remaining punctuation. This was done with different Python libraries, including: spaCy, and NLTK (Natural Language ToolKit). This was carried out both by text and as general corpora. After the preprocessing, it was possible to get the following information out of them, as seen in Table 1, 2 and 3. Due to the nature of the

[1] With this, I do not mean comments on video nor people reacting to other people's videos, but to a genre that focuses on commenting on current topics (which can sometimes include other people's videos or YouTube comments, but it is not a necessity).

BASE corpus (being comprised of non-scripted speech events) the number of sentences was not counted.

Because of a similar issue, one of the YouTube transcripts was taken off the study for being incompatible with the methodology after the preprocessing step.

Table 1. General Corpus

Corpus	**Tokens**	**Types**	**Sentences**	**Average sentence per text**	**Average words per sentence**	**Text**
YouTube	142029	17593	7355	319.78	5.17	23(+1)
BASE	71135	9243	-	-	-	8

Table 2. Per Text – BASE

Tokens	**Types**	**Percentage of types within the tokens (%)**
8732	1159	13.27
8753	1022	11.68
8768	1235	14.09
9940	1154	11.61
13901	1713	12.32
8239	1128	13.69
5267	884	16.78
7535	948	12.58

Table 3. Per Text – YouTube

Tokens	**Types**	**Percentage of types within the tokens (%)**	**Sentences**
6448	1301	20.18	322
20874	2303	11.03	1041
16843	2277	13.52	915
2089	425	20.34	73
2998	529	17.65	170
2792	576	20.63	128
8217	1496	18.21	259
7271	1251	17.21	264
3523	660	18.73	170
6488	1253	19.31	333
5241	1013	19.33	272
8595	1507	17.53	367
8649	1441	16.66	369
8296	1256	15.14	424

7519	1532	20.38	310
10654	1488	13.97	425
6639	1086	16.36	222
9705	1423	14.66	317
3206	796	24.83	173
8006	1033	12.90	318
4441	867	19.52	122
4947	926	18.72	201
3566	937	23.47	160

There is a third corpus that is going to be used in order to analyze the lexicon of each corpus. This is the Academic Word List (AWL) created by the Victoria University of Wellington, developed by Coxhead (2000). This corpus can be accessed with the following link https://www.wgtn.ac.nz/lals/resources/academicwordlist . It includes 570 words, none of them being included within the top 2000 most common. In order to use this corpus, the AWL most frequent words in sublists were used. All subheadings were taken out and the string was tokenized into words and added into a list.

The preprocessing and analysis stages can be found on the following links. At https://github.com/no-you-shouldnt/The-Use-of-Academic-Language-in-Social-Media/blob/main/YouTube-code.ipynb for the YouTube corpus and https://github.com/no-you-shouldnt/The-Use-of-Academic-Language-in-Social-Media/blob/main/BASE-code.ipynb for the BASE corpus

2.2. **Nominalization**

Nominalizations were chosen due to repeatedly appearing as a desirable feature of academic language (Bunch & Martin, 2020; Jensen & Thompson, 2020; Schleppegrell, 2009; Thompson & Watkins, 2021). It has also been marked as a desired feature in TOEFL exams, to highlight the students' academic abilities (Biber et al., 2004). In order to find the number of nominalizations in the YouTube and Academic corpora, the following steps were taken:

- The most common endings for nominalization were identified. These being "*-ibility*", "*-ity*", "*-ness*", "*-tion*", "*-sion*", "*-al*", "*-ment*" y "*-ing*". Once this was done, the sentences that contained words that possessed these were separated and added into a list. In order to include plurals, the words were previously lemmatized.
- A percentage of this list was manually analyzed in order to see the accuracy of the code. With this, certain concepts were found and added to a stop words list. These focused mainly on words that end in "thing" (i.e. something, anything, nothing, etc.) and common non-nominalizations that had the same endings (such as normal, special, original, etc.).
- Finally, a separate list of only the nouns present in the corpora was created. This was made by identifying the POS tag of each word, and the ones that were tagged as

NOUN were added to said list. This was done with spaCy. The corpora passed through all the steps again to end with one list of words that encompassed all of the characteristics.

The second and third steps were repeated until a 89% of accuracy was achieve on average between both corpora.

2.3. Academic Word List

Having the words already tokenized, the list was compared to the words of each text. It is important to note that the word "comments" was erased when analyzing the YouTube corpus due to the nature of the YouTube genre itself, though interesting in itself, it would not represent the information being researched.

3. Results

3.1. Nominalizations

After going through the analysis, the number of nouns, nominalization, and the relation between them can be found in Tables 4 and 5. It can be appreciated that in most of the cases, the BASE corpus had a higher average (18.2%) of nominalizations than the YouTube corpus (10.78%).

Table 4. BASE Corpus

N° Nouns	N° Nominalization	% Nouns which are nominalization
1470	120	8.16
1434	249	17.36
1491	233	15.63
1529	191	12.49
2910	471	16.19
1652	342	20.70
1102	318	28.86
1363	357	26.19
	Average	18.20

Table 5. YouTube Corpus

N° Nouns	N° Nominalization	% Nouns which are nominalization
1043	84	8.05
2891	377	13.04
2848	338	11.87
312	38	12.18

382	34	8.91
382	26	6.81
1292	147	11.38
1095	81	7.40
438	41	9.36
968	115	11.88
804	78	9.70
1296	135	10.42
1389	136	9.79
1486	234	15.75
1526	253	16.58
1729	164	9.49
1077	95	8.82
1594	125	7.84
504	62	12.30
953	104	10.91
618	83	13.43
579	75	12.95
474	43	9.07
	Average	10.78

3.2. Academic Word List

In order to estimate how many words of the AWL were used on each text for each corpus, the percentage of tokens that are included in the Academic World List was calculated. These results, with the respected average per corpus, can be found on Tables 6 and 7.

Table 6. BASE Corpus

% of tokens which are Academic Words	
1.41	
3.29	
3.43	
1.89	
3.11	
4.24	
5.94	
4.30	
Average	3.45

Table 7. YouTube Corpus

% of tokens which are Academic Words	
1.27	
1.45	
1.73	
1.96	
0.73	
0.64	
1.24	
1.25	
2.90	
1.50	
2.04	
1.61	
1.58	
3.48	
3.17	
2.54	
2.12	
2.22	
2.65	
1.24	
1.40	
1.50	
1.26	
Average	1.80

4. Discussion & Further Research

It is hard to draw conclusions from the data gathered so far. Though, there are some differences, it is impossible to tell whether they are significant or not without a third corpus. Nevertheless, these distinctions still seem to indicate that the YouTube corpus is not as close to the academic register as originally thought. However, it is likely that it is still not on a colloquial level, but somewhere in between.

It can be said that the BASE corpus uses a more academic lexicon, since more of its words are nominalizations and in the AWL, which was expect. Yet, surprisingly, the YouTube corpus presents a bigger number of types in comparison to its tokens, which can contribute to the complexity of understanding of what is being shared.

As previously stated, in order to get more conclusive results, another corpus is needed, one that is composed of more colloquial speech. For this, a movie corpus is proposed. It will go through the same process and could work as another point of reference. If the YouTube corpus is closer to it, then it is not as academic as expected. In the case it resembles the BASE corpus, then it can be assumed that it is using a more academic register. There is also the case in which it may be place right on the middle, which could reflect how academic language has certain prestige but due to the nature of the videos and social media, it is still somewhat colloquial.

Moreover, more features should be added to the analysis. Passive voice and readability are good options to continue with the study.

Finally, though harder to achieve, it would be interesting to speak with the YouTube creators themselves, so as to get their own opinion about the use of the language. Ask why they use a particular register when they film their videos, what is their opinion about academic language in general, and maybe if they think that the language they use could attract or marginalized a certain audience.

References

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. In *TOEFL Monograph Series*.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O&#x27;Reilly Media, Inc."

Bunch, G. C., & Martin, D. (2020). From “academic language” to the “language of ideas”: a disciplinary perspective on using language in K-12 settings. *Language and Education*, *0*(0), 1–18. https://doi.org/10.1080/09500782.2020.1842443

Coxhead, Averil (2000) A New Academic Word List. TESOL Quarterly, 34(2): 213-238.

Flores, N., & Rosa, J. (2015). Undoing appropriateness: Racioling uistic ideologies and language diversity in education. *Harvard Educational Review*, *85*(2), 149–171. https://doi.org/10.17763/0017-8055.85.2.149

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Jensen, B., & Thompson, G. A. (2020). Equity in teaching academic language—an interdisciplinary approach. In *Theory into Practice* (Vol. 59, Issue 1, pp. 1–7). Routledge. https://doi.org/10.1080/00405841.2019.1665417

Schleppegrell, M. J. (2009). Language in academic subject areas and classroom instruction: What is academic language and how can we teach it? In *National Research Council Workshop on the Role of Language in School Learning: Implications for Closing the Achievement Gap* (pp. 1–39).

Thompson, G. A., & Watkins, K. (2021). Academic language: is this really (functionally) necessary? *Language and Education*, *0*(0), 1–17. https://doi.org/10.1080/09500782.2021.1896537

University of Reading; University of Warwick, The British Academic Spoken English (BASE) corpus < http://www.reading.ac.uk/acadepts/ll/base_corpus/ >