JOURNAL OF LINGUISTICS, CULTURE AND COMMUNICATION
Vol.01, No.02, 2023: December pages:91-99

YouTube Transcripts Word Frequency Measure

*[1]Vincent Smith, [2]Michael Garrett, [3]Austin Harwood, [4]James Shamblin
[1,2,3,4] University of Charleston
Email: vincentsmith@ucwv.edu,
*corresponding author

Submission Track:
Received: 03-08-2023, Final Revision: 03-09-2023, Available Online: 01-12-2023

ABSTRACT

Many YouTube videos provide written audio transcripts which provide information on the language used on YouTube. One important measure relating to language usage is word frequency. Using student-developed software and libraries in R, Python, and Microsoft Excel, the transcripts of one million YouTube videos from the YouTube-8M data set were scraped and analyzed. The word frequency of the YouTube data set was shown to correlate with commonly used word frequency measures from established studies, such as the subtitle word frequency and the HAL word frequency.

Keywords*: Youtube, Youtube Transcript, Word Frequency Measure.*

INTRODUCTION

Developed in 2005, YouTube has amassed over 2 billion unique users and is the second most visited website (Ceci, 2022). YouTube is the third most popular video and entertainment channel, and one of the most accessible applications (Ceci, 2022). In YouTube's infant stage, it was seen as solely an entertainment site, but as time progressed, YouTube became more than just entertainment (Cicconet, 2013). YouTube Channels such as Crash Course and MinutePhysics began uploading short, lecture videos to educate the public (Cicconet, 2013). As YouTube continued to grow, more categories were developed; these categories help to organize the massive number of daily uploads to the platform (McCallister,

2022). Since YouTube is a worldwide platform with uploads in many categories, content is diverse in terms of culture and languages (Mohan & Punathambekar, 2019).

With such a vast array of content originating from all corners of the globe due to YouTube's worldwide reach, the diversity extends to culture and languages and to various linguistic measures, such as those seen in the English Lexicon Project. First published in 2007, the English Lexicon Project continues provide information on linguistic measures with the latest update in 2019 (Balota et al., 2007). One important linguistic measure available in the English Lexicon Project is word frequency; the English Lexicon Project includes original measures related to word frequency and measures reported by other studies (Balota et al., 2007). Due to the skewness of word frequency measures, a logarithmic transformation is often used in practice (Balota et al., 2007). Word frequency is essential in visual word recognition (Balota et al., 2007). There are many different measures of word frequency; the best measure for word frequency changes based on application (Brysbaert et al., 2018; Johns et al., 2016). Additionally, word frequency is a measure that is used in many applications involving machine learning and natural language processing (Zhao et al., 2020). Developing more accurate word frequency measures could also make voice assistants, text generation, text summarization, and other natural language processing applications more accurate.

This research aligns with the current theoretical framework of basic linguistic theory, which attempts to statistically describe language through data (Dryer, 2006). Basic linguistic theory combines previous functional and theoretical frameworks about linguistics through history (Dryer, 2006). This theoretical framework is considered a descriptive framework which is used to describe and collect data associated with language (Dryer, 2006). One article merges linguistic theory with word frequency, exploring high-frequency terms in foreign language education by analyzing data from 30 German textbooks. It contrasts their core vocabulary with a reference wordlist and CEFR-based corpus, unveiling frequency-practicality gaps and core vocabulary discrepancies across the texts (Başaran, 2022).

RESEARCH METHOD

The YouTube-8M dataset lists 8.6 million YouTube videos gathered and published in June of 2019 (Abu-El-Haija et al., 2016). The videos in this data set were all uploaded at varying times before mid-2019 (Abu-El-Haija et al., 2016). The YouTube-8M dataset aims to maintain a diverse distribution of video topics (Abu-El-Haija et al., 2016). Each video in this data set has a unique randomly generated video ID associated with the video (Abu-El-Haija et al., 2016).

Video IDs from the dataset were used alongside web scraping tools to attempt to retrieve transcripts from one million videos on YouTube. The one million videos were collected alphabetically based on video ID. If transcripts were not available for the video, no data was collected. The software was developed using the coding languages Python and R and the integrated development environments PyCharm and RStudio. The scraped datasets were then stored in CSV files.

Many of the transcripts collected included non-words that were marked by the word "foreign" appearing in the transcript in the data set. To clean the datasets, a spam filter that utilized Bayes Theorem was developed using Python to remove inaccurate transcripts. Each transcript had to have a minimum amount of twenty words to be considered for the filter. A transcript containing less than twenty words was removed from the data set. The spam filter was trained on 725 transcripts containing "foreign". Each of these transcripts was manually labeled as "spam" if it was a transcript including primarily non-words or "not spam" if the word foreign was adequately used in the video. If the cell's spam probability was above 90%, it would be flagged as spam and removed from the CSV file. In total, 177,586 transcripts were collected. After the data cleaning, 25,470 transcripts did not fit the criteria, and there were 152,116 transcripts remaining.

After using the spam filter, the data set was cleaned in Excel to remove any markers in the transcript placed by YouTube to indicate that music was playing. This eliminated music videos, individuals singing karaoke, and individuals performing cover songs. After

this, a copy of the dataset was created. Using Microsoft Excel, each individual word was split into its own cell delimited by a space or common punctuation. A program was then developed in R to create a list of each individual word used. The individual words were then entered into the English Lexicon Project website. This website filtered out non-words and provided measurements of word frequency on each of the words found in the transcripts. This list of words was then imported back into R and used to measure the count of each word in the original cleaned list of transcripts. Ultimately, this developed a YouTube word frequency measure that was used for statistical analysis. This YouTube word frequency measure contained a total of 80,892,245 words and 56,583 unique words. Code and data for this project can be found at: https://www.dropbox.com/sh/qiiw2d5amy364p3/AACIUmWavMCNJpZBGo_w5urRa?dl=0

RESULT & DISCUSSION

The study aimed to compare the newly developed YouTube word frequency to other major word frequency measures, such as the HAL frequency and the subtitle word frequency. Since each measure was created with different criteria using different samples, some words do not appear in some of the frequency measures.

The HAL (Hyperspace Analogue to Language) corpus contains approximately 131 million words from approximately 3,000 newsgroups (Brysbaert & New, 2009). The large number of newsgroups that contributed to the corpus provided a wide variety of topics for the text (Brysbaert & New, 2009). The HAL word frequency is determined by the number of times a word appeared in the corpus (Brysbaert & New, 2009). Since the measure for HAL word frequency is often skewed, a logarithmic transformation of the HAL word frequency measure is often used and was used in this analysis (Balota et al., 2007).

The subtitle word frequency (SUBTLEX-US) was created to measure better words used daily in informal interactions (Brysbaert & New, 2009). The subtitle word frequency

measure was considered a better predictor of processing times (naming and lexical decision) than other measures (Brysbaert & New, 2009).

A correlation matrix was created using R, as found in Table 1. The results of the correlation matrix showed a significant correlation with $p < 0.001$ for the YouTube word frequency measure and each of the already established measures of word frequency. The Pearson correlation coefficients were both high, suggesting a strong correlation. The sample size for each correlation varied. The HAL and YouTube frequencies both included plural and possessive words; thus, any missing data from HAL was assumed to be 0. Since the subtitle word frequency did not provide information on plural and possessive words, any missing data was removed from the analysis instead of assuming 0 as not to skew the results. Despite this difference in sample size, both analyses still had large sample sizes.

Table 1.
Correlation Analyses

Variable	Measure	YouTube Word Frequency
HAL Frequency	Correlation Coefficient	.94
	Significance	<.001
	Sample Size	55298
Subtitle Frequency	Correlation Coefficient	.89
	Significance	<.001
	Sample Size	45845

CONCLUSION

The high correlations between YouTube word frequency and the other established word frequency measures suggest that this measure could be useful in practice. The YouTube word frequency measure could be used in applications such as social media analysis, video analysis, or conversational speech analysis to see how well it performs in comparison to

other popular measures. The best usage cases of this word frequency measure will likely be found through future applied research.

Many words appear in the YouTube word frequency measure that do not appear in other word frequency measures with examples such as accumulatively, aircraftman, and anchorite. This shows that language used on YouTube includes words that other word frequency measures do not consider. The higher correlation was found between HAL frequency and the YouTube word frequency. This is a surprising result since films (from which subtitles were collected) and YouTube videos are the same form of media. This could be related to how each data set was measured since the HAL and current studies had more similar methodologies. Additionally, the sample size for the HAL and YouTube correlation was higher.

Some interesting single-word usage comparisons exist between the YouTube word frequency data set and the other data sets. For example, the word *balusters* appears once in the HAL frequency, and there is no data in the subtitle word frequency, yet it appears 87 times in the YouTube word frequency. This could suggest that instructional home improvement videos or other similar videos that include *balusters* are popular while this word is not used as frequently in other cases. Other cases exist such as the word *whippoorwill* appearing 1,373 times in the HAL word frequency and once in the YouTube word frequency.

The results of this research align with basic linguistic theory, which statistically describes language through data (Dryer, 2006). The high correlation result was expected, but the measures' differences may prove helpful in application. By introducing more descriptive information on language, fields such as natural language processing can apply this analysis for applications.

Limitations and Future Work:

Due to restrictions of time and processing power, all 8 million videos were not used as a sample; however, one million videos still constituted a huge sample. This project could continue with approximately eight times the current sample size in the end. Another obstacle

was 82.24% of videos did not have transcripts. Additionally, although many of the functions used are available in Microsoft Excel, the limits of Microsoft Excel were exceeded many times due to the size of the data set and R was used instead.

Another complication was obtaining an accurate measurement for the words *true* and *false* in the HAL study. The result from the English Lexicon Project is that the HAL study showed 0 occurrences of either word which is not accurate in a sample that large. This was the only noticeable issue in using the English Lexicon Project. Additionally, the subtitle frequency did not include plural forms and words with apostrophes; however, the HAL frequency and YouTube frequency did contain these words. Unlike some other word frequency measures, this data set distinguishes between plural and non-plural instances of words with different spellings. For example, the words *sanctuary* and *sanctuaries* have a different frequencies in this data set.

This project used a subset of a fixed dataset of 8-million videos. Every video in this data set was uploaded before 2019. Future work could consider scraping data sets of more recent data, considering videos during a certain time, or relating to certain topics. This project developed a YouTube word frequency measure but does not apply it to specific cases. As previously mentioned, applications of this measure could be used for research in social media analysis, video analysis, or conversational speech analysis. This measure could also be compared to other similar measures of word frequency.

REFERENCES

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint. https://doi.org/10.48550/arXiv.1609.08675

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods, 39*(3), 445-459. https://doi.org/10.3758/bf03193014

Başaran, B. (2022). Word frequency and vocabulary for the language demands of textbooks. *International Journal of Curriculum and Instruction*, *14*(1), 1029-1051.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990. https://doi.org/10.3758/brm.41.4.977

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45-50. http://doi.org/10.1177/0963721417727521

Ceci, L. (2022). YouTube – Statistics & Facts. *Statista*. https://www.statista.com/topics/2019/youtube/#topicHeader wrapper

Cicconet, M. (2013, April 7). *YouTube is not just a site for entertainment, but education*. Washington Square News. https://nyunews.com/2013/04/07/cicconet-13/

Dryer, M. S. (2006). Descriptive theories, explanatory theories, and basic linguistic theory. *Trends in linguistics studies and monographs, 167*, 207.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic bulletin & review*, *23*(4), 1214-1220. https://doi.org/10.3758/s13423-015-0980-7

Mohan, S., & Punathambekar, A. (2019). Localizing YouTube: Language, cultural regions, and digital platforms. *International Journal of Cultural Studies*, *22*(3), 317-333. https://doi.org/10.1177/1367877918794681

Zhao, K., Shi, N., Sa, Z., Wang, H. X., Lu, C. H., & Xu, X. Y. (2020). Text mining and analysis of treatise on febrile diseases based on natural language processing. *World Journal of Traditional Chinese Medicine, 6*(1), 67. https://doi.org/10.4103/wjtcm.wjtcm_28_19

Mosek, E. (2017). *Team flow: The missing piece in performance* [Doctoral dissertation, Victoria University]. Victoria University Research Repository. http://vuir.vu.edu.au/35038/

Perry, S. M. (Ed.). (2018). Maximizing social science research through publicly accessible data sets. *IGI Global*. https://doi.org/10.4018/978-1-5225-3616-1

Ruxton, C. (2016). Tea: Hydration and other health benefits. *Primary Health Care*, 26(8), 34-42. https://doi.org/10.7748/phc.2016.e1162

Shah, T. H. (2018). Big data analytics in higher education. In S. M. Perry (Ed.), Maximizing social science research through publicly accessible data sets (pp. 38-61). IGI Global. https://doi.org/10.4018/978-1-5225-3616-1