Link preprocesamiento corpus YouTube:

<https://github.com/no-you-shouldnt/Trabajo_final/blob/main/YT%20Essay%20Corpus.ipynb>

Link preprocesamiento corpus Académico:

<https://github.com/no-you-shouldnt/Trabajo_final/blob/main/Academic%20Corpus.ipynb>

En el último tiempo la utilización del lenguaje académico se ha cuestionado debido a la discriminación que representa y que utiliza como base. Se entiende como lenguaje académico como el registro utilizado en la escolarización y para referirse a la ciencia, normalmente basado en ideas basadas en el prescriptivismo lingüístico tomando como superior el habla de quienes se encuentran al poder (normalmente refiriéndose a personas blancas provenientes del norte global). Por esto, es interesante estudiar como este lenguaje sigue siendo utilizado cuando quienes pueden acceder al contenido pueden ser personas de diferentes orígenes y experiencias, no necesariamente relacionados a la academia, pero aun así interesados en los temas de los cuales se habla en estos videos, o como los creadores, pudiendo optar por un lenguaje más coloquial y accesible, siguen prefiriendo la variedad con mayor prestigio. En el presente trabajo se plantea comparar video ensayos (*video essays* en inglés) publicados gratuitamente y de libre disposición en la plataforma de YouTube con el corpus de lenguaje académico hablado, BASE, de clases universitarias en Reino Unido en áreas de humanidades y ciencias sociales.

Si bien no existe una definición precisa, en general se opta por referirse al lenguaje académico como el registro utilizado para hablar de la ciencia y utilizado en la escolarización REFERNCE. El problema se presenta debido a que este registro es considerado más refinado y apropiado y en otras formas superior que otros registros de uso cotidiano REFERENCE. Esta superioridad es naturalizada por los enseñantes, aprendientes y sociedad en general causando que especialmente variedades de por sí minorizadas sea aún más discriminadas debido a una supuesta desviación de esta norma establecida. Esta discriminación se puede ver presente en como se puede pensar que las personas que no manejan la variedad académica del lenguaje no pueden comprender la ciencia o que estudiantes inmigrantes o cuya lengua materna no sea la lengua mayoritaria poseen una discapacidad comunicativa debido a que no han adquirido este registro en su totalidad, pero ignorando totalmente otras habilidades comunicativas propias del multilingüismo REFERENCE.

Debido a esta naturalización de la superioridad del lenguaje académico para hablar sobre ciertos temas, es de esperarse que incluso para videos de difusión masiva a través de redes sociales como YouTube sigan optando por el uso de este registro. En ningún caso se está asumiendo que los creadores de los videos buscan discriminar o consideran que su contenido es superior a esos que no utilizan el lenguaje académico, pero es interesante ver cómo se presenta en esta comunidad, especialmente en la comunidad de comentaristas y ensayistas audiovisuales de la plataforma.

Para poder comprobar la utilización del lenguaje académico en la plataforma de YouTube en la comunidad de video ensayos, se compararán dos corpus. Uno proveniente de la red social y otro del corpus BASE de lenguaje académico en clases de universidad. Corpus creado por las universidades de Warwick y Reading.

“*The recordings and transcriptions used in this study come 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”* (BASE, n.d.)

El corpus de YouTube consiste en los subtítulos de los videos hechos vía transcripción manual[[1]](#footnote-1) proporcionados por las mismas creadoras, los cuales fueron copias, pegados en bloc de notas e guardados como archivos de texto .txt. Se seleccionaron ocho creadoras provenientes de diferentes países de habla inglesa. Todas las creadoras se identifican como mujer y se encuentran en sus 20s o 30s. Se seleccionaron tres videos de cada creadora, dando un total de veinticuatro textos. De éstos se seleccionaron temas relativamente similares (belleza, reseña de una serie de televisión o película, temas relacionados a la comunidad LGBT+ o el fenómeno de “*cancelling*”). Esta preselección se hizo de manera de que el vocabulario utilizado por las creadoras sea relativamente similar. Este corpus posee un total de 17002 oraciones con un promedio de 708.4 oraciones por texto. Este corpus posee un total de 146763 palabras, dando un promedio de 8.6 palabras por oración y 6115.1 palabras por texto. De estas palabras podemos encontrar 11506 palabras únicas, equivalente al 7.8% del número total.

En cuanto al corpus BASE, se seleccionaron clases que tuvieran relación a los temas abordados en el corpus de YouTube, de manera que de nuevo el vocabulario sea lo más consistente posible. Por esto se seleccionaron clases de sociología (*sociology*), política (*politics*) y de estudios de televisión y cine (*film and televisión studies*), dando un total de ocho clases. Cada clase es una transcripción textual de lo que se ha discutido en ella, en forma de dialogo, guardados igualmente como archivos de texto .txt con un total de 5032 oraciones, con un promedio de 629 oraciones por texto. También posee un total de 67236 palabras, de las cuales, sin embargo, se eliminaron los códigos de quien era el hablante en el dialogo y muletillas como “*er*”, además de espacios en blanco, dando un total de 64738 palabras. Así dando un promedio de 12.9 palabras por oración y 8092.25 palabras por texto. De este total de palabras, 5964 son palabras únicas, lo cual es aproximadamente el 9.2% de las palabras totales del corpus.

El preprocesamiento de cada corpus fue ligeramente distinto debido a sus diferentes características.

Al corpus de YouTube, en primer lugar, se le fueron eliminados todos los paréntesis y corchetes con su contenido, de manera de evitar la interferencia de la descripción de la música o sonidos del video como por ejemplo *“(upbeat electronic music)”* para describir la música de fondo, “[ \_\_ ]” para describir una pausa en el discurso o *“(laughs)”* para describir una acción en particular. Tras esto, los textos fueron divididos en oraciones con la función sent\_tokenize de la librería NLTK, dando la cantidad de 17000 oraciones. Finalmente se ponen todas las palabras en minúscula y se eliminan los signos de puntuación, habiendo antes aplanado la lista de oraciones para tener un string. Luego las palabras son tokenizadas una vez más, esta vez en palabras. Tras esto se fueron eliminados los espacios en blanco, así dando el total de 146763 palabras. Finalmente, se realizó un POS tagging gracias a las funciones .text. .pos\_ y .lemma\_ de SpaCy.

El preprocesamiento del corpus BASE, si bien bastante similar, posee algunas diferencias significativas dado la diferencia entre ambos corpus. Si bien comenzamos de la misma manera, eliminando paréntesis y corchetes con su contenido y dividiendo el string por salto de línea, para que luego sean tokenizadas por oración, dando un total de 5032 oraciones la calcular el len() del corpus, al contrario del corpus anteriorez, la lista de tokens fue aplanada solo para poder eliminar puntuación, ya que los textos ya se encontraban en minúscula. Al ser tokenizado de nuevo, además de simplemente eliminar espacios en blanco, se eliminaron palabras que se repetían constantemente en el corpus, pero causaban interferencia al momento de analizar. Como lo son la muletilla “er” o los códigos otorgados a los hablantes del dialogo, como “nm0063”, “nf0069” o “nm1166”.

De manera de poder analizar la similitud de estos corpus se tomarán como rasgos la nominalización y la voz pasiva, además de ver el número de palabras utilizadas de una lista de palabras comunes en el lenguaje académico (*Academic Word List*). Se optó por estos rasgos ya que no solo son mencionados repetidas veces como cualidades caracterizadoras del registro académico en inglés (REREFRENCE), sino que además se incluyen dentro de los rasgos deseados al analizar textos académicos provenientes de la prueba TOEFL (REFERENCE). La nominalización será caracterizada por las terminaciones más comunes, estás siendo:

* -ibility
* -ity
* -ness
* -tion
* -sion
* -al
* -ment
* -ing

Para esto se creó una pequeña función como ciclo *if* que separa cada oración que incluye una palabra que posea una de estas terminaciones y agregándolas a una nueva lista. Con esto se entregaron 754 oraciones con palabras nominalizadas en el corpus de YouTube, las cuales equivalen a un 4,44% del total de oraciones del corpus. A su vez, el corpus Académico entrega 324 oraciones, equivalentes al 6,43% del corpus total. Con esto podemos apreciar una diferencia de un 1,99% entre ambos corpus. Sin embargo, al momento de analizar los resultados encontrados por esta función, se encontró un nivel de error considerable. Se analizó manualmente el 10% de las oraciones de cada corpus viendo si las oraciones marcadas eran efectivamente nominalizaciones o no. El resultado no solo fue decidor al momento de ver que el margen de error era demasiado elevado, sino que la diferencia de error entre ambos corpus también era considerablemente diferente. El corpus de YouTube presentó un 57.33% de error, cuando el corpus académico solo presentó un 28.13%.

Para solucionar esto primero se agregaron las terminaciones plurales a la lista de terminaciones, ya que, si bien no cambiarían el margen de error, se estaba ignorando gran parte de las posibles nominalizaciones, lo que hizo aumentar la cantidad de oraciones del corpus de YouTube a 883 y a 358 el corpus académico. Luego se creó una lista de stopwords comunes que generan ruido al dividir por terminaciones (tales son las palabras terminadas en “*thing*” como “*something*”,” *anything*” y “*thing*” misma) además de cambiar el código un poco, agregando otro ciclo *for*. Tales son las palabras terminadas en “*thing*” como “*something*”,” *anything*” y “*thing*” misma. Inesperadamente, tras este cambio el número se nominalizaciones creció de nuevo, quedando en 888 y 359 respectivamente. Sorpresivo también, es que el margen de error no se vio afectado de gran manera, pasando del 57.33% original en el corpus de YouTube a un 55.68%, y de un 28.13% a un 33.33% en el corpus Académico.

Para seguir refinando esta función se buscará separar las palabras que posean POS tag NOUN, además de hacer coincidir las terminaciones y agrandar la lista de palabras que causan ruido. Una vez que el margen de error no sea tan considerable, se procederá a trabajar en la detección de voz pasiva y de comparar palabras en la *Academic Word List*.

La voz pasiva será caracterizada por su estructura, en otras palabras, el verbo *to be* seguido de un participio pasado. Para esto se espera clasificar las palabras según su sintaxis a través de la herramienta tag de la librería SPACY para luego encontrar esta estructura de *be + participio* gracias a los LEMA y DEP de la misma librería.

La *Academic Word List* se utilizará para contar el número de veces que estas palabras coinciden en ambos corpus para luego ser comparados. Está demás decir que la importancia de este resultado no será el número de palabras en sí, sino el porcentaje referente a la totalidad de las palabras.

1. Se optó por la transcripción manual ya que la automática entregada por YouTube mismo comente errores que pueden ser significantes para el análisis de los textos. [↑](#footnote-ref-1)