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 presenta y en la que está basado. 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.

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 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 17000 oraciones con un promedio de 708.3 oraciones por texto.

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, guardados igualmente como archivos de texto .txt dando un total de 5032 oraciones, con un promedio de 629 oraciones por texto.

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 EJEMPLO. Luego, los textos fueron divididos por salto de línea, para después poder eliminar las primeras dos líneas de texto y las últimas siete. Este paso es para evitar las pequeñas introducciones a cada video, que suelen ser todas iguales o similares, y los outro en donde se pide a las visitas a subscribirse al canal o a ver otros videos de las creadoras. Esto es importante ya que es posible que al momento de ver los resultados las expresiones como “*subscribe*”, “*welcome to*” o “*press the like button*” podrían causar ruido ya que, si bien demuestran cierto lenguaje utilizado exclusivamente por creadores de YouTube, no es necesariamente relevante para el presente estudio. 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, había antes aplanado la lista de oraciones para tener un string.

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, no es necesario eliminar comienzo ni final de las clases, por lo que inmediatamente son tokenizadas por oración, dando un total de 5032 oraciones la calcular el len() del corpus . A su vez, la lista de tokens fue aplanada solo para poder eliminar puntuación, ya que los textos ya se encontraban en minúscula.

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*). La nominalización será caracterizada por las terminaciones más comunes, estás siendo:

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

La voz pasiva será caracterizada por su estructura, en otras palabras, el verbo *to be* seguido de un participio pasado.

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)