Tarea 1. Análisis de texto

Helena Patricia Carrillo Soto
Facultad de Ciencias Fisico-Matemáticas, UANL

Maestría Ciencia de Datos: Procesamiento y Clasificación de Datos

Resumen

Se analizaron dos traducciones y una adaptación de la novela inglesa *Orgullo y Prejuicio* de dicho análisis se obtienen y comparan estadísticas descriptivas de texto.

Keywords: Palabras; Frecuencias; Análisis de texto

1. Introducción

Análisis a realizar En este artículo se analizará un análisis estadístico de dos diferentes traducciones y una adaptación de la famosa novela inglesa *Orgullo y Prejuicio* de la autora Jane Austen. Se analizaron las palabras totales, párrafos totales y una comparativa de las palabras más utilizadas por cada traducción. Es necesario aclarar que el tercer texto a utilizar se considera una adaptación ya que se eliminan pasajes a elección del adaptador para hacer la historia más fácil de digerir sin comprometer la esencia de la historia.

Obra elegida *Orgullo y Prejuicio* publicada en 1813 se ha convertido en una de las mayores referentes de la literatura inglesa con el paso del tiempo. Se considera una de las primeras comedias románticas publicadas y ha tenido múltiples adaptaciones literarias y cinematográficas a lo largo de la historia .

Please refer to your figures as: Figure 1, Figure 2, etc.

2. Descripción de los datos

Las traducciones a analizar son las realizadas por:

- ***Ana María Rodriguez*** Editorial Penguin Clásicos.
- ***Marta Salís*** Editorial Alba.
- ***Lourdes Íñiguez*** Editorial Clásicos a medida.

3. Metodología

Para realizar el análisis de las traducciones se limpian los textos quitando las introducciones de cada una de las traducciones así como las notas de los editores y traductores que se encuentran al final.

Se obtienen las palabras claves de cada párrafo al tiempo que se eliminan *stop-words* las cuales son palabras que se consideran de no valor para el texto.

Para esto se utiliza la librería *nltk* y se agregaron algunas palabras a las *stop-words* de la librería después de un análisis inicial.

4. Resultados

Los primeros estadísticos que se obtienen son el número de palabras y párrafos de cada una de las traducciones. Los resultados se pueden observar en el cuadro 1.

De los resultados de dicho cuadro observamos que las traducciones de Ana María Rodriguez y Marta Salís tienen longitudes muy similares mientras que la de Lourdes Íñiguez es considerablemente más pequeña esto seguramente se debe a que esta última es considerada una adaptación del libro más que una traducción. Esta adaptación se hizo con el objetivo de hacer la novela más breve por lo que es consistente con los resultados.

Cuadro 1: Número de palabras y párrafos por traducción/adaptación.

Traductor	Número de palabras	Número de parrafos
Ana María Rodriguez	640,758	2,224
Marta Salís	685,406	2,225
Lourdes Íñiguez	174,096	671

Después de este análisis se revisaron las palabras que más se repiten en cada una de las traducciones. Se grafican el top quince de cada uno de los casos.

Para la traducción de Ana María Rodriguez se tienen los resultados en la figura 1. En esta gráfica podemos observar que los nombres de varios personajes forman parte del top. También obervamos títulos como *mr*, *mrs* y *miss*.Esto nos da una idea de que esta novela esta basada fuertenemte en sus personajes y las relaciones entre ellos.

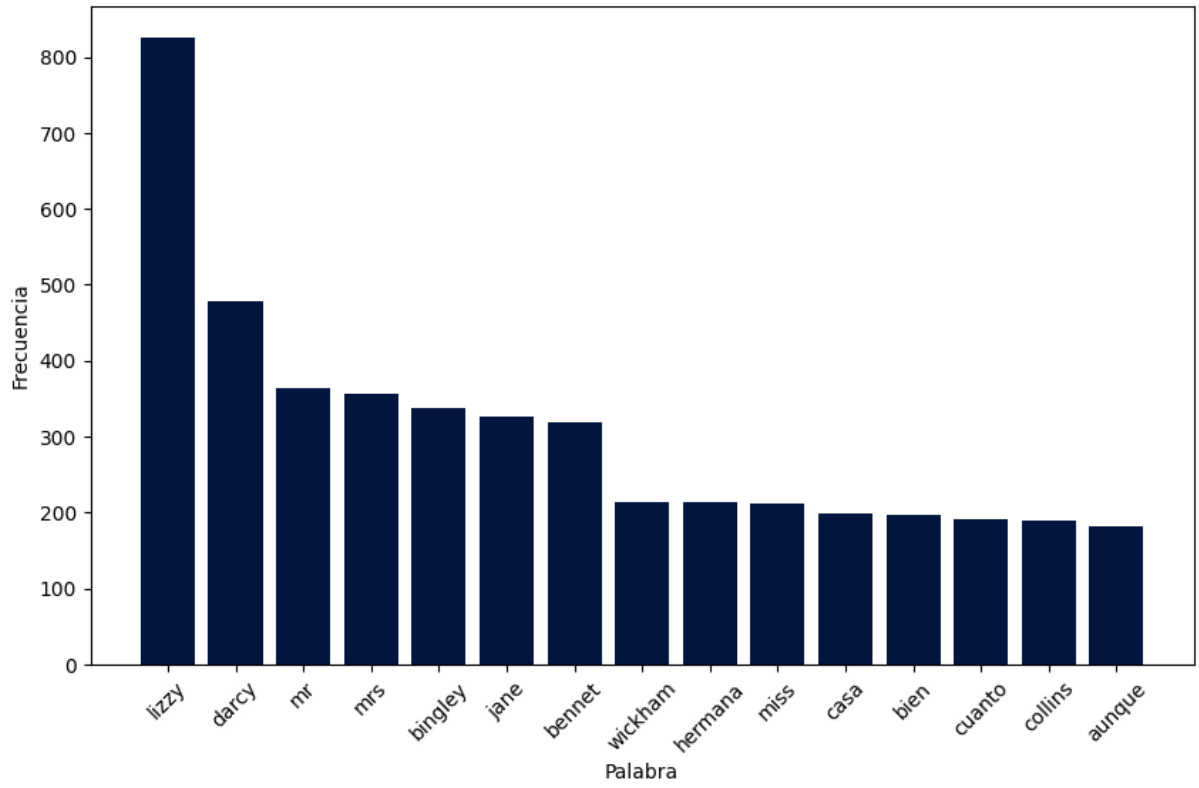

Figura 1: Palabras más repetidas en traducción de Ana María Rodriguez

De forma similar, en la traducción de Marta Salís los nombres de personajes también conforman gran parte del top palabras más usadas. Esto lo podemos observar en la figura 2. Un punto interesante de señalar es que los títulos en esta traducción estan en español y no en inglés: *señor* en lugar de *mr*, *señora* en lugar de *msr* y *señorita* en lugar de *miss*.

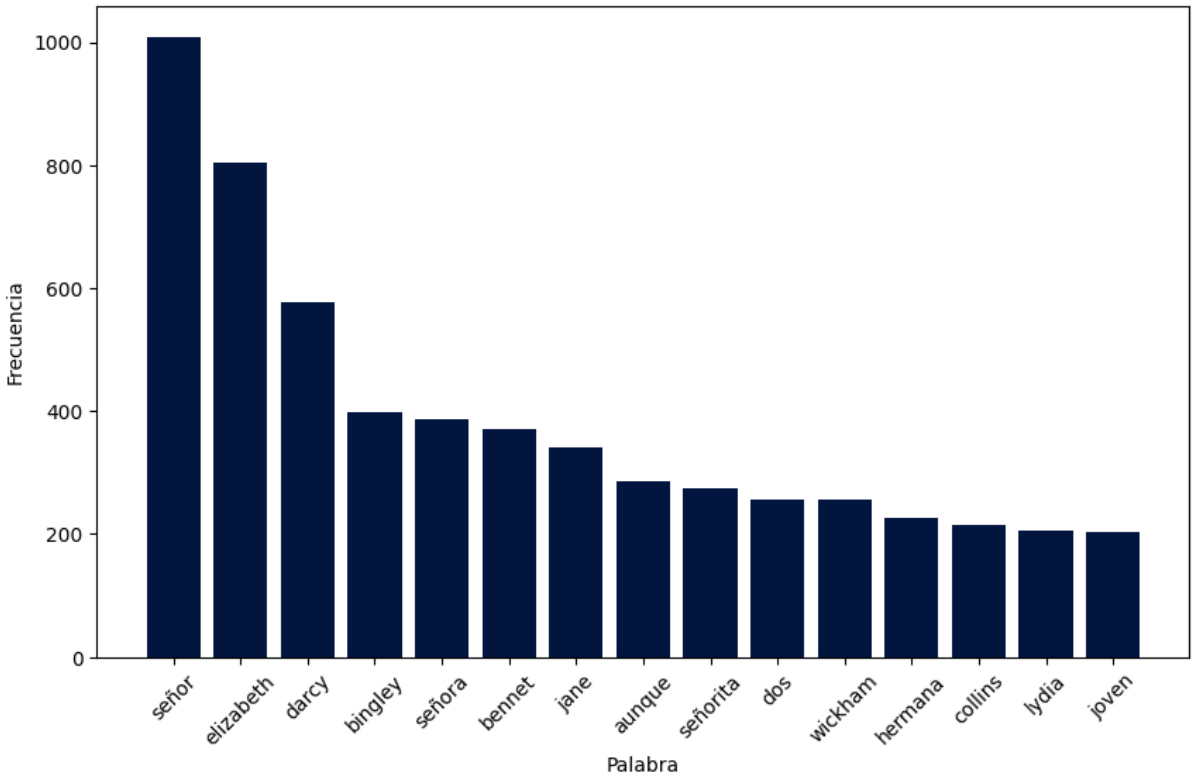

Figura 2: Palabras más repetidas en traducción de Marta Salís

Para la adaptación de Lourdes Íñiguez se tiene una situación similar a las anteriores. Se puede observar el top en la figura 3. Esto puede ser una señal de que a pesar de no ser una traducción directa sino una adaptación, se conserva en el texto la esencia de la novela original, a menos en lo que a frecuencia se refiere.

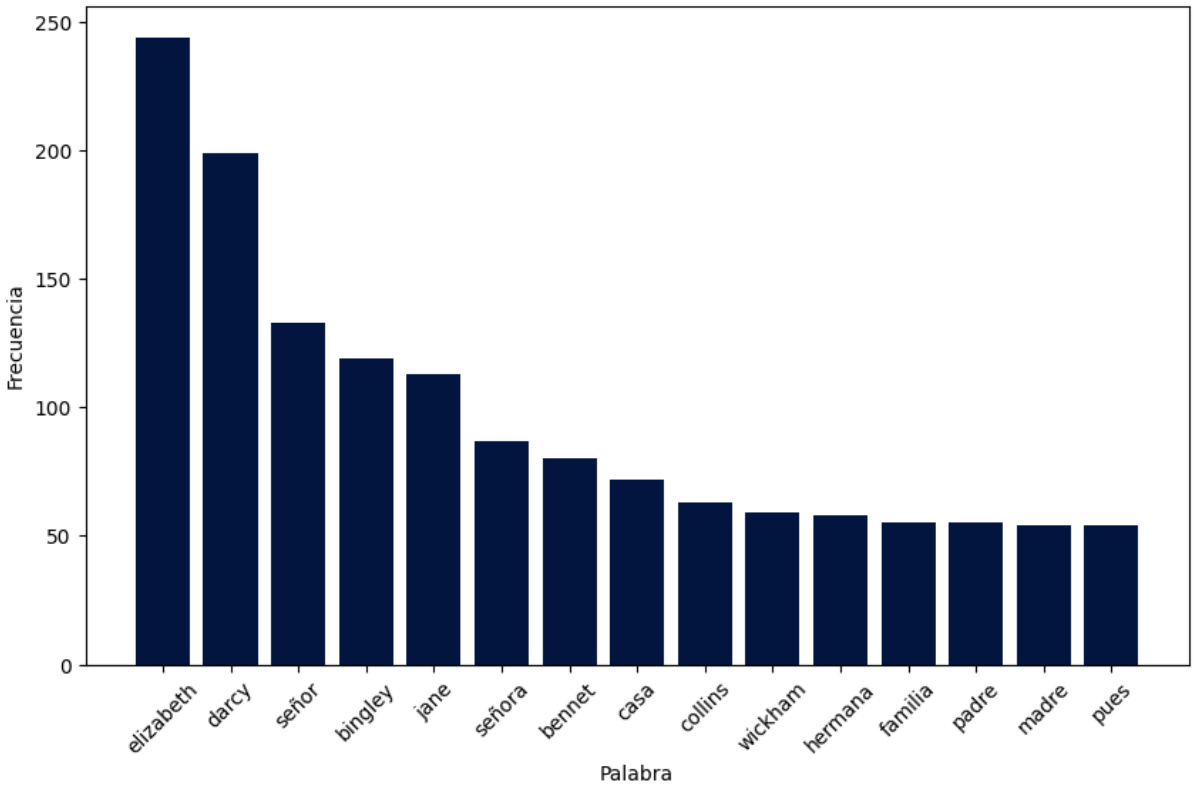

Figura 3: Palabras más repetidas en adaptación de Lourdes Íñiguez

Las tres gráficas tienen distribuciones bastante similares y podemos ver que los protagonistas son los más mencionados por los textos.

Otra manera de visualizar las palabras más frecuentes de los tres textos es a través de nubes de palabras. Estas se pueden encontrar en las figuras 4, 5 y 6.

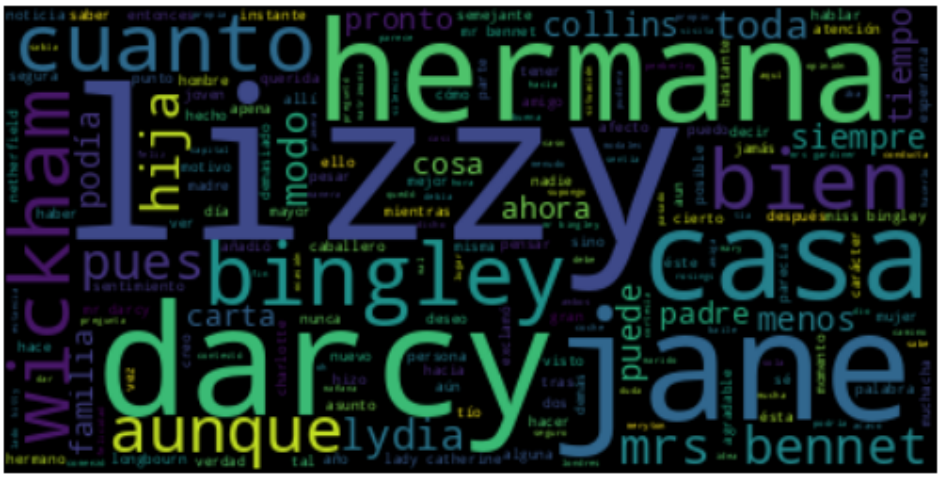

Figura 4: Nube de palabras de traducción de Ana María Rodriguez

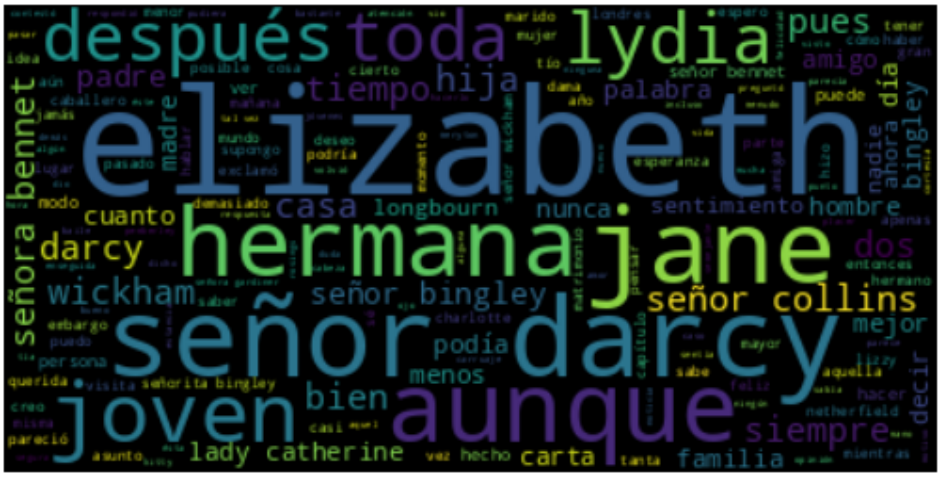

Figura 5: Nube de palabras de traducción de Marta Salís

Figura 6: Nube de palabras de adaptación de Lourdes Íñiguez

También se decidió realizar una comparación del top diez de palabras más usadas que todos los textos compartan entre sí. Estos resultados se pueden ver en la figura 7.

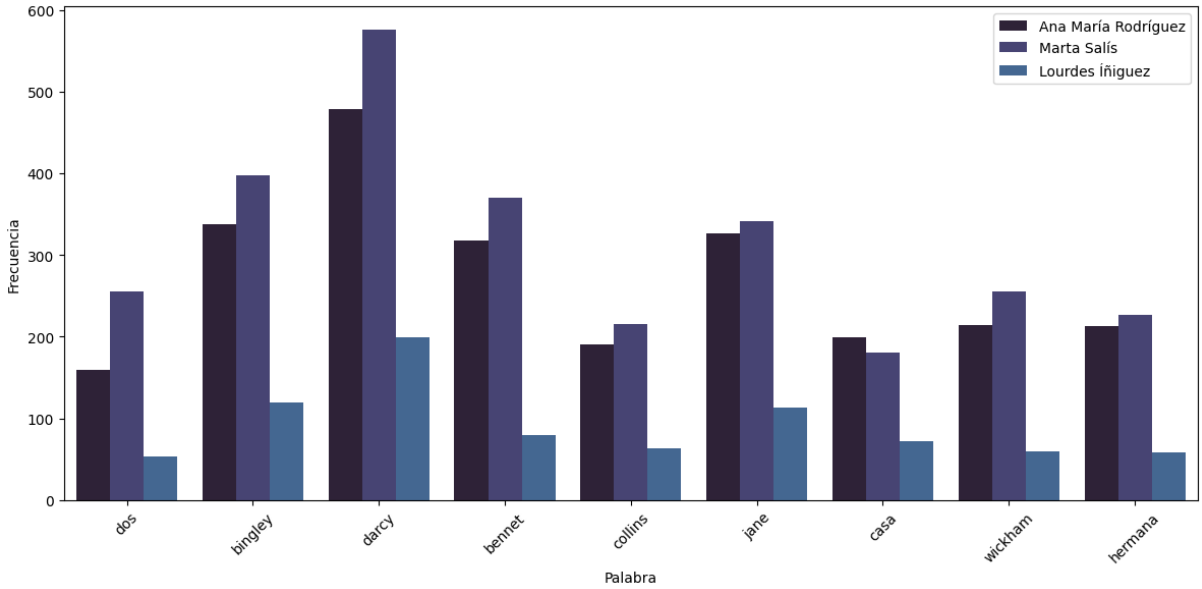

Figura 7: Frecuencia de palabras en común entre los tres textos

Esta última gráfica muestra los resultados en función al total de repeticiones medida que puede afectar al texto de la adaptación ya que esta es significativamente más corta en su totalidad. Por este motivo se decidió efectuar nuevamente el análisis pero en esta ocasión tomando la proporción de frecuencia que tienen estas palabras respecto a las otras palabras claves de sus propios textos. Los resultados se pueden observar en la figura 8.

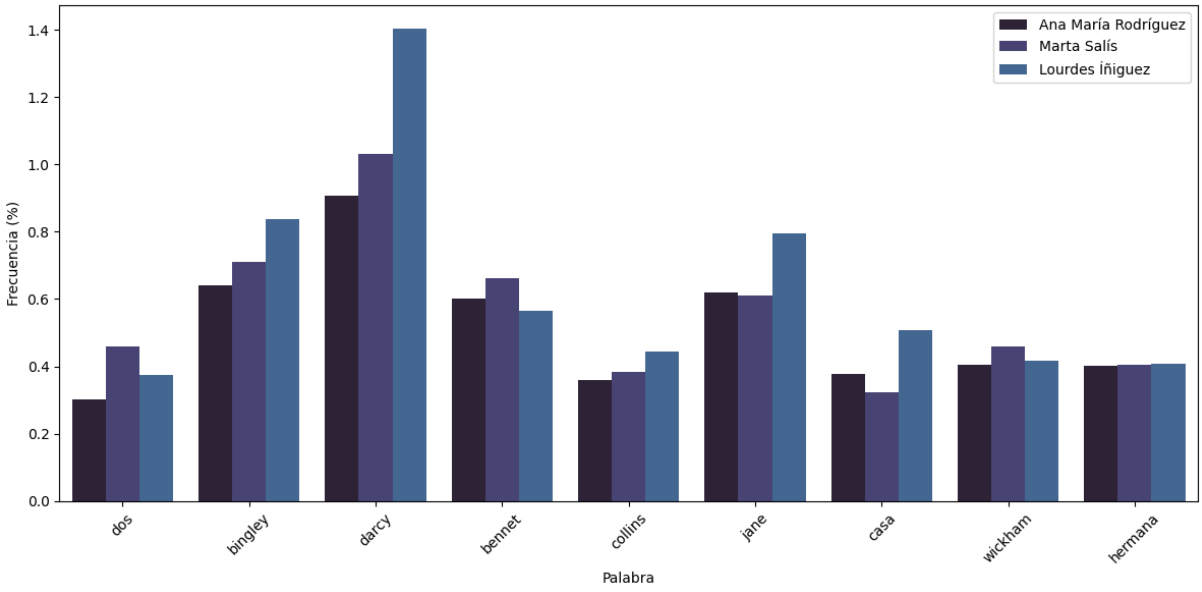

Figura 8: Frecuencia proporcional de palabras en común entre los tres textos

La gráfica es significativamente distinta una vez que se toman en cuenta las proporciones de los textos. Así podemos observar por ejemplo que aunque el nombre de *Darcy* se repite en total menos veces en el tercer texto a comparación de los otros dos este aun tiene la proporción mayor en cuanto a palabras clave totales de cada uno.

5. Conclusiones

En base a los análisis estadísticos efectuados no se encuentran diferencias notables entre los tres textos analizados salvo por la longitud de la adaptación la cuál, acorde a su objetivo, presenta un número menor considerable de palabras y párrafos pero se muestran indicios de que se conserva la esencia de la novela al comparar la frecuencia de las palabras con los otros textos.