



Análisis Estadístico de Redes Sociales: Taller #4

Relaciones entre palabras: Caso 03 de la JEP

Valentina Cardona Saldaña

1 Introducción

La Jurisdicción Especial para la Paz (JEP) se encuentra inmersa en un arduo proceso de investigación y esclarecimiento de los episodios más críticos del conflicto armado colombiano, agrupados en 11 macrocasos. En este informe, se analizará el **Caso 03**, que aborda específicamente los delitos perpetrados por la fuerza pública, agentes del Estado en colaboración con grupos paramilitares, o terceros civiles en el contexto del conflicto. La lista de reproducción relacionada con versiones, declaraciones e información relevante sobre este caso se encuentra en el siguiente enlace: https://www.youtube.com/playlist?list=PLbtegW3d3L4Id_lrAgBRDGf-k_bwEYH6_. Así mismo, este análisis es basado en el realizado por el Juan Camilo Sosa en <https://rpubs.com/jstats1702/946584> (Sosa, n.d.)

2 Procesamiento de datos

La recolección de datos se lleva a cabo mediante la extracción de las transcripciones de los vídeos en la lista de reproducción de YouTube del caso 03. Debido a las limitaciones del procesador de mi computadora, no fue posible extraer los datos utilizando las alternativas sistematizadas proporcionadas por el profesor. Sin embargo, se optó por utilizar la librería **Selenium** para realizar el *Web Scraping* y automatizar el procedimiento mencionado anteriormente. El Script que contiene esta extracción se denomina **redes - taller 4 - valentina cardona.ipynb**.

En total, se lograron extraer exitosamente 438 transcripciones de un conjunto de 438 vídeos de la lista de reproducción. Las seis zonas priorizadas (Meta, Casanare, Huila, Antioquia, Norte de Santander, Costa Caribe) por la JEP en el desarrollo del Caso 03 fueron seleccionadas tras contrastar la información proporcionada por entidades gubernamentales y la sociedad civil con las versiones ofrecidas por los comparecientes (*Caso 03: Asesinatos y desapariciones forzadas presentados como bajas en combate por agentes del Estado*, n.d.). Para efectos del análisis, se realizó una clasificación de los vídeos, a partir de sus títulos, en las zonas o "Subcasos" (ver Tabla 1). Como resultado de este proceso, se observó que el subcaso con mayor número de vídeos es el Huila (178), mientras que solo se encontró un vídeo para Norte de Santander y ninguno para el Meta. Es importante destacar que los vídeos que no mencionaban explícitamente uno de los subcasos principales pero sí hacían referencia a ciertos municipios fueron clasificados según su ubicación geográfica. Por ejemplo, aquellos relacionados con Barranquilla fueron incluidos en la Costa Caribe. Además, los vídeos relacionados con el pueblo indígena Arhuaco fueron agrupados en la región de Costa Caribe debido a las áreas que habitualmente ocupan (*Ijku - Arhuaco*, n.d.).

Table 1: Número de vídeos por subcasos

Subcaso	Número de vídeos
Antioquia	14
Casanare	101
Costa Caribe	34
Huila	178
Norte de Santander	1
Otros: Audiencias sin clasificación	67
Otros: Ruedas de prensa, respuesta a preguntas y otros vídeos informativos	43
Total	438

Los vídeos que no pudieron ser clasificados carecían de indicaciones en sus títulos sobre los subcasos o eran de naturaleza informativa, como ruedas de prensa o respuestas a preguntas. De un total de 110 vídeos, 43 correspondían a ruedas de prensa, respuestas a preguntas y contenido informativo, mientras que 67 eran de audiencias que no pudieron ser clasificadas según su título. Para el análisis de las zonas priorizadas centrado en los discursos relacionados con los delitos perpetrados por la fuerza pública, agentes del Estado en colaboración con grupos paramilitares o terceros, se excluyeron estos vídeos del análisis, al igual que el vídeo de Norte de Santander dado que carece de variedad de discursos, por tanto, no brinda mayor información. En consecuencia, el análisis se realizó sobre 327 transcripciones de vídeos.

3 Análisis de datos

3.1 Tokenización

Después de tokenizar las transcripciones de los vídeos, se obtuvo un total de 11.313.312 filas de palabras. Tras eliminar 101.893 números, 7.451.836 palabras vacías (*stop words*) y las *muletillas* "eh" repetida 112.089 y "digamos" repetida 19.204, se redujo el conjunto a 3.628.290 palabras para su análisis. Es importante destacar que, si bien las cifras (números) son fundamentales para esclarecer la verdad y brindar reparación a las víctimas del conflicto, carecen de relevancia en un análisis de palabras sin un contexto claro. Por lo tanto, se optó por excluirlos. Además, como parte del proceso de normalización del texto, se eliminaron minúsculas, signos de puntuación y acentos.

La Tabla 2 presenta la distribución de palabras asociadas a cada subcaso dentro del Caso 03, reflejando una tendencia similar a la observada en la Tabla 1: a mayor cantidad de vídeos (y, por ende, información), mayor extensión del discurso.

Table 2: Número de palabras (*tokens*) por subcasos

Subcaso	Número de vídeos
Antioquia	91265
Casanare	1.049.410
Costa Caribe	312.466
Huila	2.175.149
Total	3.628.290

3.2 Frecuencia de palabras

A nivel general, al calcular la frecuencia de las palabras en todas las audiencias, se observa que el término más utilizado es "señor" (con 56.577 apariciones), seguido de palabras asociadas a las fuerzas armadas como "batallón" (con 41.907 apariciones) y "comandante" (con 33.007 apariciones). Este análisis sugiere la presencia predominante de formas formales de dirigirse a las personas ("señor", "doctor"), así como términos

relacionados con el ejército ("batallón", "comandante", "magistrado", "operaciones", "brigada") y con la búsqueda de la verdad ("hechos", "información").

Table 3: *tokens* más frecuentes

Token	Frecuencia
señor	56.577
batallon	41.907
comandante	33.007
magistrado	22.698
hechos	22.466
informacion	22.416
doctor	22.082
operaciones	21.330
brigada	19.570
personas	17.293

Considerando la disparidad en la distribución de palabras entre los distintos subcasos, resulta relevante profundizar en el análisis de la frecuencia de palabras por cada uno de ellos. La Figura 1 ilustra que, en todas las zonas prioritarias, la palabra más recurrente fue "Señor", mostrando una tendencia similar a la tabla mencionada previamente. Sin embargo, se destaca una discrepancia notable en el caso de Antioquia, donde la narrativa parece diferir de los demás subcasos. En este contexto, se observa un mayor uso de términos como "víctimas", seguido de "gracias". Además, se identifican otras palabras de interés, como "justicia" y "nombre".

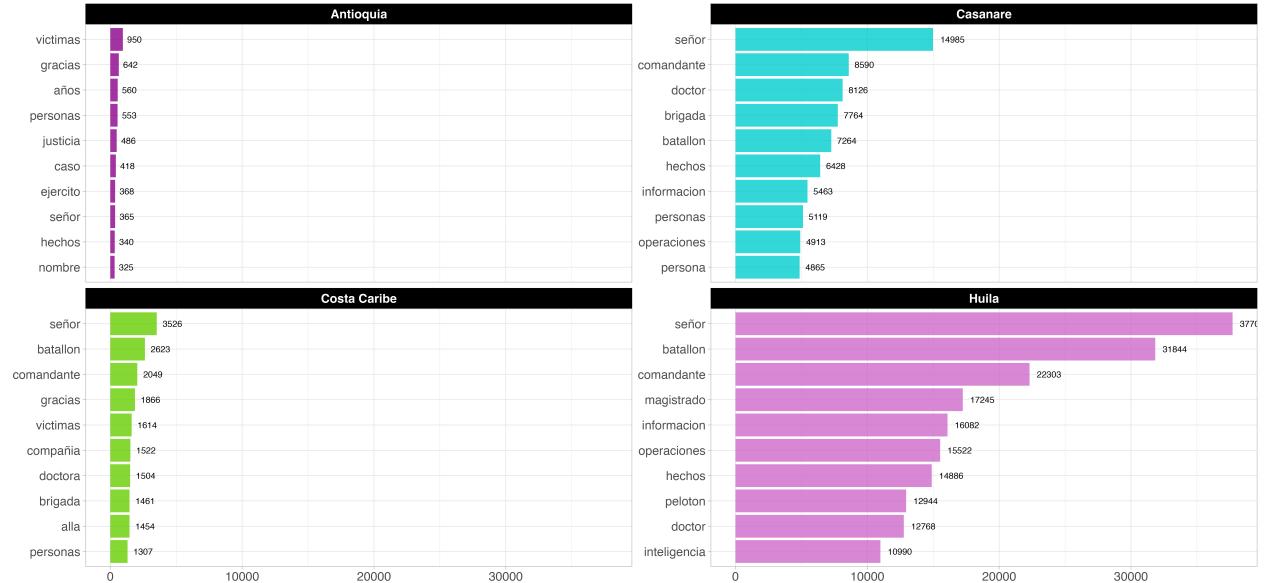


Figure 1: Top 10 palabras más frecuentes por subcaso

Además, en la Figura 2 se presenta una nube de palabras para cada subcaso, donde se muestran las 100 palabras más frecuentes. Las tendencias observadas previamente se mantienen en general. En Casanare, Costa Caribe y Huila, se observa una mayor presencia de términos jurídicos ("magistrado", "civil") y militares ("operación", "combate", "soldados"), y comienzan a surgir nombres de personas más específicos. En el caso de Antioquia, se destacan palabras como "casa", "familia", "vida" y, por primera vez, "paz". Considero importante destacar palabras como "dolor", "recuerda" y "perdón".



Figure 2: Top 50 palabras más frecuentes por subcaso

Por otro lado, es importante destacar que la proporción de palabras compartidas entre los diferentes subcasos, en relación con el total de palabras analizadas, alcanza el 11.3%. Este porcentaje resulta significativo considerando la extensión del corpus en estudio. Además, como parte del análisis, se calcularon las correlaciones entre los subcasos. Se observa que el coeficiente de correlación más alto se registra entre Casanare y Huila, alcanzando un valor de 0.921, y el coeficiente más bajo se presenta entre Huila y Antioquia, con un valor de 0.475. Estos hallazgos revelan patrones interesantes en la distribución y utilización del lenguaje en los diferentes contextos examinados. La alta proporción de palabras compartidas entre subcasos sugiere la existencia de temas y preocupaciones comunes que atraviesan diversas regiones afectadas por el conflicto.

Table 4: Coeficientes de correlación

	Antioquia	Casanare	Costa Caribe	Huila
Antioquia	1	0.584	0.635	0.475
Casanare	0.584	1	0.883	0.921
Costa Caribe	0.635	0.883	1	0.858
Huila	0.475	0.921	0.858	1

3.3 Análisis de sentimiento

Para llevar a cabo el análisis de sentimientos, se emplea un diccionario específico en español obtenido de la fuente <https://www.kaggle.com/datasets/ratman/sentiment-lexicons-for-81-languages>. Utilizando este diccionario, se pueden asignar polaridades a las palabras en función de su carga afectiva, lo que permite cuantificar el tono emocional de los discursos en cada subcaso del Caso 03. Identificar y comprender estas palabras clave es fundamental para captar la complejidad de las experiencias y narrativas presentes en cada subcaso del Caso 03.

Entre las palabras con connotaciones positivas, "gracias" emerge como una de las más frecuentes en cada subcaso, lo que sugiere una expresión de aprecio o reconocimiento en los discursos analizados. Además, otras palabras relevantes que reflejan una carga emocional positiva incluyen "justicia", "paz" e "inteligencia". Estos términos resaltan la importancia de conceptos como la equidad, la tranquilidad y la sabiduría en el discurso en torno al conflicto.

Por otro lado, en el espectro negativo, se encuentran palabras como "baja", "muerte", "enemigo" y "conflicto". Estas palabras transmiten un sentido de pérdida, hostilidad y adversidad, lo que indica la presencia de narrativas relacionadas con la violencia y la confrontación en los discursos analizados.

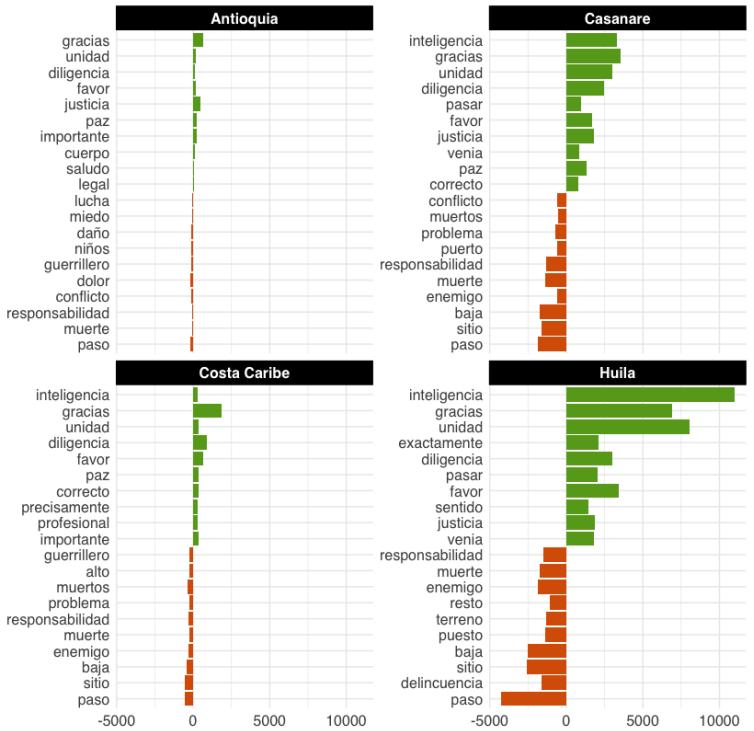


Figure 3: Top 10 Conteo por sentimiento

3.4 Redes

Al tokenizar por secuencias de 2 palabras y crear bigramas para cada subcaso, se generó una base inicial de 11.313.297 bigramas. Tras normalizar esta base siguiendo el mismo proceso utilizado para un solo token, se redujo a 366.818 bigramas. A partir de estos bigramas, se construyeron grafos utilizando un umbral de 2 frecuencias por par de palabras. La Figura 4 muestra las redes correspondientes a cada subcaso, donde el tamaño de los vértices indica su centralidad de intermediación. Este enfoque permite identificar las palabras que funcionan como puntos de conexión clave dentro de la red de palabras utilizadas en los discursos sobre el conflicto. El tamaño de los vértices permite identificar las palabras con mayor centralidad de intermediación; esto puede ayudar a revelar los conceptos o temas centrales que conectan diferentes aspectos del discurso sobre el conflicto.

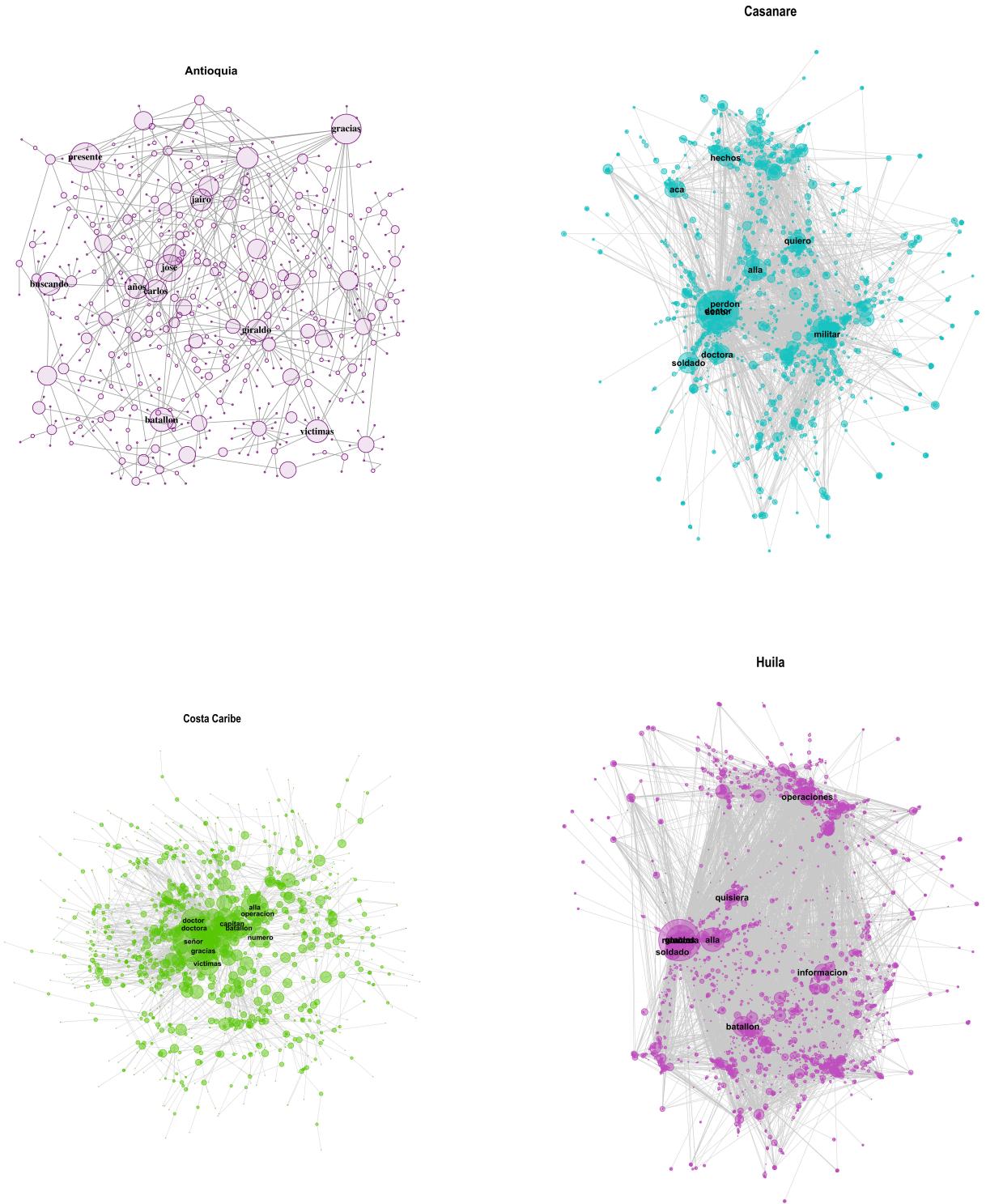


Figure 4: Red de palabras por subcaso

En la figura anterior, se presentan las redes por subcaso. Se destaca que en el discurso de todos los subcasos se hace un uso frecuente de formas formales o denominaciones de cargo (ej. "doctor", "capitán", entre otros), así como de nombres propios. Esto indica que estas formas de referencia se emplean con regularidad a lo

largo de los discursos. Además, se observa una centralidad de intermediación incluso en palabras que ya hemos identificado como frecuentes, como "víctimas", "batallón" y "operación".

Este hallazgo subraya la importancia y la prominencia de ciertos términos y conceptos dentro de los discursos sobre el conflicto en cada subcaso. Las formas formales y las denominaciones de cargo pueden reflejar la estructura jerárquica y la formalidad del lenguaje utilizado en estas audiencias. Por otro lado, la centralidad de intermediación de palabras como "víctimas", "batallón" y "operación" sugiere su relevancia como temas recurrentes y puntos focales en los discursos analizados. Estos resultados proporcionan una visión más detallada de la dinámica lingüística y temática presente en cada subcaso, lo que contribuye a una comprensión más completa de los discursos sobre el conflicto.

4 Limitaciones

Es cierto que este análisis presenta algunas limitaciones que deben ser consideradas con precaución. En primer lugar, la precisión de la traducción automática de YouTube, de la cual se extrajeron los datos, puede ser variable y, en ocasiones, poco exacta. Esto podría afectar la calidad de las transcripciones y, por ende, la interpretación de los resultados obtenidos.

Además, el proceso de *web scraping* utilizado para recopilar los datos puede ser complejo y difícil de replicar, lo que puede limitar la accesibilidad y la robustez del análisis. Asimismo, la clasificación de los videos en subcasos basada únicamente en los títulos puede ser imperfecta, lo que podría resultar en la exclusión involuntaria de algunos videos relevantes y, por lo tanto, en una representación incompleta de los discursos presentes en el conjunto de datos.

References

- Caso 03: Asesinatos y desapariciones forzadas presentados como bajas en combate por agentes del estado.* (n.d.). Sitio web. Retrieved from <https://www.jep.gov.co/macrocasos/caso03.html> (Accedido en abril de 2024)
- Iku - arhuaco.* (n.d.). Sitio web. Retrieved from <https://www.onic.org.co/pueblos/110-arhuaco> (Accedido en abril de 2024)
- Sosa, J. C. (n.d.). *Relaciones entre palabras*. Sitio web. Retrieved from <https://rpubs.com/jstats1702/946584> (Accedido en abril de 2024)