

Juan Esteban Rivera Balaustre

Entrega Final: Trabajo de grado

Tutor: Jorge Andrés Alvarado

Pontificia Universidad Javeriana

Facultad de Ciencias

Junio 4, 2026

Índice - 2

1. Problema - 4
2. Objetivos - 4
    - 2.1. Objetivo general - 4
    - 2.2. Objetivos específicos - 4
        - 2.2.1. Análisis de sentimientos
        - 2.2.2. Percepciones sensoriales
        - 2.2.3. Sesgos semánticos de género
3. Justificación e Importancia - 5
4. Marco Teórico - 6
    - 4.1. Definición de conceptos clave y bases teórico-conceptuales - 6
        - 4.1.1. Procesamiento del lenguaje natural y embeddings semánticos - 6
    - 4.2. Emociones y análisis de sentimientos - 6
        - 4.2.1. Modelo empleado en este estudio - 7
    - 4.3. Sesgos implícitos de género en embeddings semánticos - 8
        - 4.3.1 Embeddings empleados - 8
        - 4.3.2 Vectores de referencia y scores de sesgo por noticia - 9
        - 4.3.3 Estudios Previos Sesgos de género embeddings semánticos - 9
    - 4.4. Percepciones sensoriales de Lancaster - 10
        - 4.4.1 Aplicación concreta de Lancaster Sensorimotor Norms - 11
        - 4.4.2. Estudios previos Percepción sensorial y sensoriomotoras - 12
5. Metodología - 12
    - 5.1.1. Fuentes de información y scraping - 12
    - 5.2. Análisis de sentimientos con robertuito - 14
    - 5.3. Preparación del corpus traducido para análisis sensorial y semántico - 14
        - 5.3.1. Traducción al inglés - 14
        - 5.3.2. Estandarización post-traducción - 15
        - 5.3.3. Tokenización y lematización en inglés - 15
    - 5.4. Aplicación de las Normas Sensoriales de Lancaster -16
        - 5.4.1. Procedimiento de correspondencia léxica - 16
    - 5.5. Análisis de sesgos semánticos - 17
        - 5.5.1. Construcción del eje de género - 18
        - 5.5.2. Vectores de referencia de género - 18

        - 5.5.3. Extracción de vocabulario por noticia - 18
        - 5.5.4. Cálculo de scores de sesgo de género - 18
    - 5.6. Análisis estadístico - 19
        - 5.6.1. Análisis estadístico del sentimiento - 19
        - 5.6.2. Análisis estadístico de Lancaster -29
        - 5.6.3. Análisis estadístico del sesgo de género - 20
6. Resultados
    - 6.1.1. Web Scraping - 20
    - 6.1.2. Descripción del dataset - 22
    - 6.2. Análisis de Sentimientos - 22
        - 6.2.1. Análisis inferencial del sentimiento — Chi-cuadrado de contingencia - 23
    - 6.3. Aplicación de las Normas Sensoriales de Lancaster - 24
        - 6.3.1. Preprocesamiento y cobertura léxica - 24
        - 6.3.2. Jerarquía sensorial global del corpus - 24
        - 6.3.3. Distribución por dimensión individual - 25
        - 6.3.4. Correlaciones entre dimensiones sensoriales - 26
        - 6.3.5. Análisis inferencial de las percepciones sensoriales (Lancaster) - 27
    - 6.4. Análisis de Sesgos Semánticos de Género - 34
        - 6.4.1. Construcción del eje de género y vectores de referencia - 34
        - 6.4.2. Obtención de los scores por noticia - 34
        - 6.4.3. Estadística inferencial - 36
7. Análisis y Conclusiones - 38
    - 7.1. Síntesis de hallazgos principales - 38
    - 7.2. La huella sensorial del deporte: el cuerpo como lenguaje - 39
    - 7.3. El sesgo semántico de género - 40
    - 7.4. Limitaciones del estudio - 41
    - 7.5. Líneas de investigación futura - 42

Anexos - 43

Referencias - 44

Título del proyecto: Análisis de sentimientos, percepciones sensoriales y sesgos semánticos de género en la cobertura mediática de los Juegos Olímpicos de París 2024.

1. Problema

Históricamente, la cobertura mediática y el análisis deportivo han favorecido hacer comparaciones de rendimiento, resultados y estadísticas, prestando menos atención a cómo el público percibe los deportes, eventos y atletas de manera emocional y sensorial. Esta brecha es particularmente preocupante con respecto a cuánto varían dichas percepciones dependiendo del deporte particular que practique el atleta, lo que afecta en gran medida tanto la equidad mediática en términos de representación como la construcción de narrativas que moldean la opinión pública, los expertos deportivos y la promoción de diferentes disciplinas olímpicas.

Basándose en los sentimientos y percepciones sensoriales evocadas por eventos clave en los Juegos Olímpicos de París 2024, este trabajo realiza un análisis computacional de los sentimientos y respuestas sensoriales provocadas por esos eventos para cerrar estas brechas. Para ello, se utilizan estrategias de procesamiento de lenguaje natural (PLN) para leer noticias deportivas especializadas en español con el fin de identificar patrones emocionales, asociaciones sensoriales e implícitas semánticas basadas en la disciplina de los atletas cubiertos.

2. Objetivos

2.1. Objetivo general

Explorar los sentimientos y percepciones sensoriales evocadas por las narrativas mediáticas de los Juegos Olímpicos de París 2024, analizando de manera diferenciada las representaciones emocionales, sensoriales y semánticas en torno a los deportistas y disciplinas olímpicas.

2.2. Objetivos específicos

2.2.1. Analizar las emociones predominantes (positivas, negativas y neutras) evocadas en la cobertura mediática de momentos clave (victorias, derrotas, controversias e hitos históricos) mediante técnicas automáticas de análisis de sentimientos.

2.2.2. Caracterizar las percepciones sensoriales evocadas en la cobertura mediática de disciplinas olímpicas, utilizando normas sensoriales para identificar patrones diferenciales en la evocación de sentidos como la vista, el oído, el tacto, la interocepción y la motricidad.

2.2.3. Evaluar la existencia de asociaciones semánticas implícitas en la representación mediática de los deportes olímpicos, mediante el análisis de sesgos semánticos implícitos en el vocabulario periodístico en relación con el eje de género construido a partir de embeddings semánticos

3. Justificación e Importancia

Los Juegos Olímpicos constituyen un evento deportivo de escala global en el cual no solo se celebra la competencia atlética, sino que también se configuran narrativas que evocan emociones, valores culturales y representaciones sensoriales del cuerpo en movimiento. Analizar cómo los medios deportivos construyen estas narrativas permite comprender de manera más profunda el impacto que tienen en la percepción pública de los deportes y los deportistas.

La exploración de las emociones predominantes en las noticias deportivas, así como de las percepciones sensoriales evocadas mediante el lenguaje, ofrece una contribución significativa a los campos de la comunicación deportiva, la psicología cognitiva y el procesamiento de lenguaje natural (PLN). El uso de técnicas computacionales como el análisis de sentimientos, las normas sensoriales de Lancaster y los modelos de asociación semántica permite abordar este fenómeno de manera sistemática y a gran escala, superando las limitaciones propias de los estudios cualitativos tradicionales.

Este trabajo también busca aportar evidencia empírica sobre posibles sesgos en la representación mediática, en particular aquellos relacionados con las asociaciones implícitas de género. El análisis no parte de clasificar noticias por el sexo del atleta, sino de detectar si el vocabulario implícito de cada cobertura se orienta hacia el polo masculino o femenino del espacio semántico, revelando patrones de género arraigados en el propio discurso periodístico.

La dimensión de género cobra relevancia adicional a la luz de los lineamientos de organismos internacionales como ONU Mujeres, cuya oficina para América Latina y el Caribe promueve el deporte como herramienta para acelerar el progreso hacia la igualdad de género y el empoderamiento de mujeres y niñas, reconociendo su papel en la ruptura de

estereotipos y en la construcción de espacios de liderazgo femenino (ONU Mujeres, 2024). En este sentido, detectar patrones narrativos implícitos en la cobertura olímpica contribuye directamente a los esfuerzos por identificar y visibilizar desigualdades en el ecosistema mediático deportivo.

Finalmente, los hallazgos de este estudio pueden resultar de interés para medios de comunicación, organizaciones deportivas y entidades promotoras de equidad de género, al ofrecer un marco metodológico replicable para el análisis de narrativas en eventos deportivos de alta visibilidad global.

4. Marco Teórico

4.1. Definición de conceptos clave y bases teórico-conceptuales

4.1.1 Procesamiento del lenguaje natural y embeddings semánticos

El Procesamiento de Lenguaje Natural (PLN) es un área interdisciplinaria de la inteligencia artificial que busca permitir que las máquinas entiendan, interpreten y generen lenguaje humano. A diferencia de los datos estructurados tradicionales (como números o códigos), el lenguaje es ambiguo, contextual y rico en matices, lo que hace necesario modelarlo de formas que los computadores puedan procesar (Jurafsky & Martin, 2009).

Una de las técnicas centrales del PLN es la creación de embeddings semánticos. Un embedding es una representación de palabras como vectores numéricos en un espacio multidimensional. Por ejemplo, tendrán vectores cercanos entre sí para "fútbol" y "deporte" y mucho más alejados entre sí para "fútbol" y "manzana". Estos vectores no expresan significado específicamente, sino que lo aprenden a través de palabras compartidas encontradas en grandes flujos de texto. En este estudio, el PLN es esencial ya que permite un estudio sistemático y a gran escala de esas características lingüísticas en el corpus de 723 artículos de noticias, lo cual no sería posible al escrutar manualmente dichos datos. Específicamente, las metodologías computacionales para identificar temas emocionales, sensoriales y semánticos incrustados en el lenguaje periodístico serían difíciles de identificar sin la ayuda de herramientas automatizadas.

4.2 Emociones y análisis de sentimientos

El análisis de sentimientos es un método de procesamiento de lenguaje natural que detecta y clasifica el contenido emocional del lenguaje escrito. Dicha carga emocional suele

clasificarse en tres polaridades: positiva, negativa o neutral (Pang & Lee, 2008). Depende de la idea de que el lenguaje no solo es descriptivo, sino también evaluativo. Hay ciertos adjetivos, adverbios, estructuras gramaticales e incluso el tono general de una narrativa que llevan emociones subyacentes, que el lector percibe implícitamente. Los modelos computacionales modernos que se basan en arquitecturas Transformer pueden capturar automáticamente estos matices asignando probabilidades a cada categoría emocional y produciendo una distribución interpretable sobre las tres categorías: positiva, negativa y neutral.

4.2.1 Modelo empleado en este estudio

En esta investigación se emplea el modelo pysentimiento/robertuito-sentiment-analysis (Pérez et al., 2021), una variante del modelo mT5 (Multilingual T5) específicamente ajustada para textos en español. Este modelo resulta especialmente adecuado para el corpus analizado por tres razones. Primero, fue entrenado sobre textos periodísticos y de redes sociales en español, lo que lo hace especialmente sensible a los matices emocionales del lenguaje deportivo formal. Segundo, procesa secuencias completas de texto respetando la estructura gramatical original, garantizando un análisis contextualizado a nivel de artículo completo. Tercero, produce una distribución de probabilidad sobre las tres categorías emocionales: positiva, negativa y neutra, de la cual se extrae la categoría con mayor probabilidad como variable de sentimiento por noticia.

Es importante señalar que el proceso de fine-tuning de robertuito no fue realizado dentro del presente estudio, sino que corresponde al entrenamiento original del modelo. En este trabajo se utiliza como herramienta pre entrenada para la clasificación automática de sentimientos sobre el corpus de noticias olímpicas.

4.3 Sesgos implícitos de género en embeddings semánticos

El sesgo lingüístico implícito es una asociación no intencionada que se manifiesta en cómo se utilizan las palabras, y esto incorpora estereotipos sociales incluso sin intención explícita. Cuando estas propiedades existen en enormes cantidades de texto, los modelos de lenguaje entrenados para aprender representaciones vectoriales de ellos naturalmente adoptarán tales patrones y los reproducirán en el espacio semántico resultante.

Bolukbasi et al. (2016) demostraron que las incrustaciones semánticas contienen una dirección geométrica identificable que captura la distinción de género en el lenguaje. Su

propuesta consiste en construir un eje de género aplicando la Descomposición en Valores Singulares (SVD) en la matriz formada por las diferencias vectoriales de pares de palabras con carga de género explícita she–he, woman–man o mother–father. Formalmente, para cada par se calcula el vector diferencia:

d_i = v_masculino − v_femenino

Sobre la matriz D = [d_1, d_2, ..., d□] centrada se aplica SVD determina que el primer vector singular sea el eje de género g que describe la tendencia de mayor desviación en el espacio semántico masculino-femenino. Al completar este eje, es posible imponer cualquier palabra de cualquier tipo sobre él: cuando se crea, el resultado:

$$\text{bias(w)} = \frac{w}{g \,.\, x}$$

donde valores positivos indican proximidad al polo masculino y valores negativos al polo femenino.

4.3.1 Embeddings empleados

Los embeddings utilizados en este estudio provienen del modelo en core_web_lg de spaCy, que implementa vectores GloVe de 300 dimensiones entrenados sobre grandes corpus en inglés. Matemáticamente, cada palabra queda representada como un vector u, y la similitud semántica entre dos palabras se mide mediante la similitud coseno:

$$SIM(u, v) = \frac{(u \cdot v)}{(\|u\| \cdot \|v\|)}$$

Este valor oscila entre -1 y 1, donde valores cercanos a 1 indican alta similitud semántica.

4.3.2 Vectores de referencia y scores de sesgo por noticia

A partir del eje de género g se construyen dos vectores de referencia: el centroide femenino (VEC_FEM) y el centroide masculino (VEC_MAS), calculados proyectando las 218 palabras gender-specific del Apéndice F de Bolukbasi et al. (2016) sobre g y promediando los vectores de cada subconjunto resultante.

Para cuantificar el sesgo de cada noticia se calculan dos scores complementarios sobre el vocabulario de adjetivos y verbos más frecuentes de cada texto (denominado w) , conformado por los 20 tokens de mayor frecuencia con categoría gramatical ADJ o VERB

disponibles en el modelo. El Score Coseno se define como la diferencia entre la similitud coseno promedio de w con VEC_MAS y su similitud coseno promedio con VEC_FEM:

$$S_{\cos}(d) = \cos(w, VEC_{\text{MAS}}) - \cos(w, VEC_{\text{FEM}})$$

El Score de Proyección se define como el promedio de las proyecciones de cada vector de palabra normalizado sobre el eje de género donde g es el eje de género y n es el número de palabras válidas del vocabulario de la noticia.:

$$S_{proy}(d) = \frac{1}{n} \sum_{i=1}^{n} \frac{\vec{w}_i}{||\vec{w}_i||} \cdot \vec{g}'$$

Ambos scores son complementarios: el Score Coseno captura la proximidad global del vocabulario al centroide de referencia de cada polo, y el Score de Proyección mide directamente la posición de cada palabra sobre el eje construido por SVD. Reportar ambos de forma conjunta permite fortalecer la interpretación y verificar la consistencia del patrón detectado. Un valor positivo en ambos Scores indica que el vocabulario de la noticia se orienta hacia el polo masculino del eje de género y un valor negativo hacia el polo femenino.

4.3.3 Estudios Previos Sesgos de género en embeddings semánticos

Caliskan, Bryson y Narayanan (2017) introdujeron el Word Embedding Association Test (WEAT), demostrando que los modelos de lenguaje entrenados en grandes corpus textuales reflejan sesgos sociales implícitos de manera cuantificable. El WEAT ha sido replicado y extendido en múltiples estudios posteriores, validando su uso para analizar asociaciones implícitas entre categorías sociales como género y atributos como competencia o sociabilidad.

Bolukbasi et al. (2016) constituyen la referencia central del análisis de sesgo de género de este estudio. En su trabajo seminal demostraron que los embeddings semánticos entrenados sobre grandes corpus textuales contienen una dirección geométrica que codifica el género de forma cuantificable, y que esta dirección puede ser identificada mediante SVD sobre diferencias de pares de palabras con carga de género explícita. Su método permite no solo medir el sesgo sino también caracterizar qué palabras del vocabulario se posicionan más cerca de uno u otro polo, lo que lo hace especialmente útil para el análisis del lenguaje periodístico deportivo.

En el contexto específico del deporte, Cooky, Messner y Musto (2015) realizaron un análisis longitudinal de 25 años sobre la representación mediática del deporte femenino, encontrando una subrepresentación sistemática de las mujeres en los noticieros deportivos televisados, así como un tratamiento narrativo más estereotipado y menos enfocado en la competencia atlética. Este tipo de sesgo mediático es precisamente el que este estudio busca detectar desde una perspectiva cuantitativa y computacional, aplicando el método de Bolukbasi et al. (2016) sobre narrativas olímpicas en lugar de contenido televisivo

4.4 Percepciones sensoriales de Lancaster

El lenguaje humano no solo transmite información conceptual, sino que también puede activar representaciones sensoriales en la mente del receptor. Esta propiedad, llamada evocación sensorial, sugiere que algunas palabras activan sensaciones como la vista, el oído, el tacto, el gusto, el olfato, el movimiento corporal o las sensaciones internas del cuerpo.

Las Normas Sensoriomotoras de Lancaster (Lynott et al., 2020) son un recurso lexicográfico que proporciona puntuaciones cuantitativas de la fuerza con la que más de 40,000 palabras evocan experiencias sensoriales en diferentes modalidades. Estas puntuaciones provienen de evaluaciones humanas donde los participantes calificaron cada palabra en escalas específicas para las siguientes modalidades:

- **Sensoriales**: visual, auditiva, táctil (háptica), gustativa y olfativa.
- **Motoras**: pie/pierna (Foot/Leg), mano/brazo (Hand/Arm), cabeza (Head), boca (Mouth) y torso.
- **Interoceptiva**: relativa a la percepción interna del cuerpo, como hambre, esfuerzo o aceleración cardíaca.

En total, el recurso cubre 11 modalidades, cada una con puntuaciones en una escala de 0 (sin evocación) a 5 (evocación muy fuerte). Por ejemplo, el término "gol" obtendría puntuaciones altas en las dimensiones auditiva y motora, mientras que "elegancia" tendría mayor peso en la dimensión visual.

Formalmente, el corpus Lancaster está representado como una matriz numérica M de dimensiones n × m, donde n es el número total de palabras en la base léxica y m = 11 corresponde al número de modalidades sensoriales consideradas en este estudio:

$M = (w_{i,j})$ para $i = 1,...,n$ y $j = 1,...,11$

$$M = \begin{pmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,m} \\ w_{2,1} & w_{2,2} & \dots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \dots & w_{n,m} \end{pmatrix}$$

Donde cada entrada w_i,j representa la puntuación de evocación de la palabra i en la modalidad sensorial j. Este recurso, disponible públicamente y ampliamente validado, permite analizar de manera sistemática qué tipo de experiencias perceptuales privilegia el lenguaje deportivo en su cobertura de distintas disciplinas olímpicas.

4.4.1 Aplicación concreta de Lancaster Sensorimotor Norms

El puntaje sensorial promedio para cada modalidad m en una noticia se obtiene mediante:

$$S_M = \frac{1}{n} \sum_{i=1}^{n} s_i(\mathbf{w})$$

Donde n es el número de palabras de la noticia con correspondencia en el lexicón Lancaster, y s_i(w) es la puntuación de evocación de la palabra w en la modalidad W. El resultado es un vector de 11 dimensiones por noticia que resume su perfil sensorial completo.

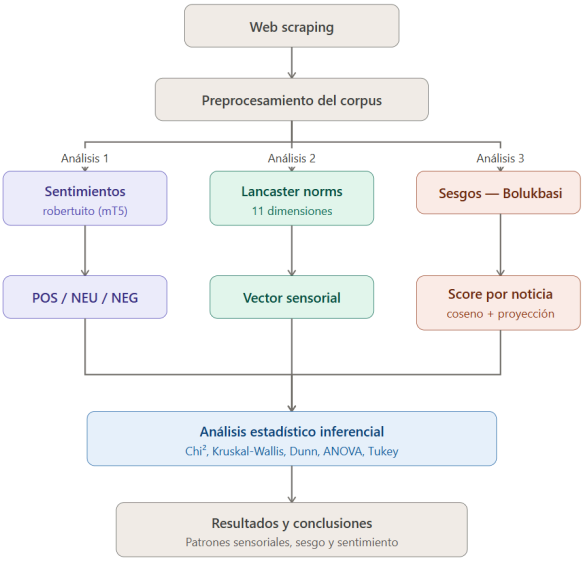
4.4.2 Estudios previos Percepción sensorial y normas sensoriomotoras

Lynott et al. (2020) publicaron las Lancaster Sensorimotor Norms, una base de datos que proporciona puntuaciones de fuerza asociativa en 11 modalidades sensoriales y motoras para más de 40.000 palabras del inglés. Este recurso, ampliamente validado y disponible públicamente, permite analizar de manera sistemática cómo el lenguaje evoca experiencias perceptuales específicas en el lector, y constituye el insumo principal del análisis sensorial de este trabajo.

Complementariamente, Grand et al. (2018) demostraron que los embeddings semánticos pueden recuperar atributos perceptuales implícitos como tamaño, textura o color, a partir de relaciones vectoriales, sin que dichos atributos aparezcan de forma explícita en el

texto. Este hallazgo refuerza la validez del enfoque adoptado en este estudio, en el que se asume que el lenguaje periodístico deportivo codifica representaciones perceptuales que pueden ser recuperadas computacionalmente.

5. Metodología



Este estudio se forma alrededor de tres análisis lingüísticos computacionales, que trabajan en paralelo sobre el mismo corpus de noticias. A partir de una única colección de corpus y una etapa de preprocesamiento, los textos se procesan simultáneamente a través de: (1) el modelo robertuito de análisis de sentimientos, que clasifica cada noticia como positiva, negativa o neutral; (2) análisis de percepción sensorial con las Normas de Lancaster, que produce un vector de 11 dimensiones por noticia; y (3) análisis de sesgo semántico de género utilizando el método de Bolukbasi et al. (2016), que cuantifica la orientación del vocabulario periodístico hacia el polo masculino o femenino del eje de género. Los resultados de estos tres análisis se integran a través de un paso final de análisis estadístico inferencial pruebas de Chi-cuadrado, ANOVA, Tukey, Kruskal-Wallis y Dunn que permite identificar patrones diferenciados según la disciplina deportiva cubierta.

5.1.1 Fuentes de información y scraping

Se recopiló una canalización automatizada de noticias deportivas relacionadas con los Juegos Olímpicos de París 2024 para construir el corpus de análisis que abarca los 15 días previos al evento, la duración oficial del evento y los 15 días posteriores a su cierre. La extracción se realizó utilizando las bibliotecas Playwright y BeautifulSoup en Python con técnicas de web scraping. Durante la extracción de datos se aplicó una selección de filtros

para calidad y relevancia. Técnicamente, se utilizaron selectores CSS para delimitar el contenido principal de cada artículo, excluyendo elementos estructurales de la página como menús de navegación, encabezados secundarios, pies de página y módulos de publicidad. Además, se aplicaron filtros de contenido para descartar artículos con menos de 100 palabras, principalmente correspondientes a galerías de imágenes o resultados tabulados sin texto narrativo, y artículos duplicados identificados por similitud de URL o título. Para cada noticia se recopilaron y almacenaron las siguientes variables:

Para cada noticia se recolectaron y almacenaron las siguientes variables:

- Título de la noticia
- Fecha de publicación
- Texto completo del contenido principal
- Fuente (sitio web de origen)
- URL de la noticia

En cuanto a la selección de fuentes, inicialmente se consideraron otros medios de habla hispana con cobertura olímpica, como Marca y El Mundo Deportivo. Para su cobertura, ambos fueron descartados como una posible razón del sesgo geográfico del conjunto de datos, ya que estaba relativamente enfocado en deportes españoles. Las cuatro fuentes seleccionadas fueron Olympics.com, ESPN Deportes, BBC Deportes y Diario AS; seleccionadas por su coherencia de cobertura institucional (Olympics.com), perspectiva latinoamericana (ESPN Deportes), orientación analítica internacional (BBC Deportes) y cobertura de disciplinas individuales (Diario AS), permitiendo mejorar y maximizar la diversidad narrativa de su corpus al mismo tiempo que se mantiene la uniformidad lingüística.

Se aplicó una función de detección de la lista producida por el Comité Olímpico Internacional (COI, 2024) basada en 37 términos relevantes para las disciplinas del programa oficial (París 2024) para evaluar la disciplina deportiva de cada noticia. Esto generó una fuente informativa en el sentido de que puede mostrar no solo la disciplina deportiva sino también el tipo y tamaño de eventos cubiertos en el evento. La lista completa del diccionario está en el Anexo A con el nombre de dominio de la disciplina y las variantes léxicas más comúnmente usadas en el periodismo deportivo en español (por ejemplo, "atletismo" tiene

muchas de sus variantes como "athletics", "track", "marathon", "jump"), para que podamos asegurar que la mayor parte de la cobertura encajará dentro de los textos con la mayor diversidad en términos de vocabulario.

Los pasos de detección tienen dos pasos sucesivos. (1) En el primer paso, se busca una mención directa de cualquiera de las palabras del diccionario en el título de la noticia, y si se encuentra una coincidencia, ese deporte se asigna de inmediato, tomando el título como la porción más representativa y breve del contenido. La segunda fase solo se aplica cuando el título no permite la identificación; se realiza un conteo de frecuencia para cada palabra del diccionario en el texto completo del artículo, y el deporte más referenciado se asigna al conteo de frecuencia más alto. Este método permitió la identificación de 723 noticias a través de 35 disciplinas olímpicas diversas. En los casos en que el conteo de frecuencias no arrojó un deporte dominante claro; ya sea porque la noticia mencionaba múltiples disciplinas con frecuencias similares, o porque cubría aspectos generales de los Juegos sin enfocarse en una sola disciplina , la noticia fue etiquetada bajo la categoría "múltiples deportes". Esta categoría concentró 69 noticias (9.5% del corpus) y fue excluida de todos los análisis inferenciales por su naturaleza heterogénea.

5.2. Análisis de sentimientos con robertuito

Para la clasificación emocional del corpus se aplicó el modelo pysentimiento/robertuito-sentiment-analysis sobre el texto completo de cada noticia en español, respetando su puntuación y estructura original. Como se describió en la Sección 4.1.2, este modelo es una variante del mT5 (Multilingual T5) ajustada específicamente para textos en español. El resultado de esta etapa es una variable de sentimiento asignada a cada noticia, utilizada como variable de agrupación en los análisis comparativos.

El resultado de esta etapa es una columna de sentimiento asignada a cada una de las 723 noticias del corpus, que constituye una de las variables de agrupación utilizadas en los análisis comparativos de la Sección 6.

5.3 Preparación del corpus traducido para análisis sensorial y semántico

5.3.1 Traducción al inglés

Dado que las Normas Sensoriales de Lancaster están disponibles únicamente en inglés, fue necesario traducir el corpus textual de noticias deportivas al inglés antes de

proceder con el análisis sensorial y semántico. Esta traducción se realizó después del análisis de sentimientos en español, preservando así la integridad lingüística original para esa etapa previa.

La traducción se llevó a cabo utilizando la librería googletrans en Python, que accede al servicio automatizado de Google Translate API. Cada texto completo fue traducido respetando la estructura de las oraciones, con el fin de preservar los significados originales y facilitar la correspondencia directa entre las palabras del corpus y las entradas del lexicón sensorial.

Por ejemplo, la oración en español "La atleta mostró una gran fuerza y determinación en la competencia" se traduce como "The athlete showed great strength and determination in the competition", manteniendo la correspondencia semántica necesaria para el análisis posterior.

5.3.2 Estandarización post-traducción

La traducción al inglés genera un nuevo corpus textual que requiere una etapa de estandarización antes del análisis semántico posterior. Esta estandarización incluyó los siguientes pasos:

- Conversión de todas las palabras a minúsculas, para evitar distinciones arbitrarias por capitalización.
- Eliminación de números y signos de puntuación que no aportan significado semántico relevante.
- Eliminación de caracteres especiales (emojis, menciones, símbolos) que podrían introducir ruido en el análisis.

Este proceso asegura la homogeneidad y consistencia del corpus en inglés, permitiendo su correcta integración con el lexicón de Lancaster y el modelo de embeddings en core_web_lg de spaCy utilizados en las etapas siguientes.

5.3.3 Tokenización y lematización en inglés

Sobre el corpus estandarizado se procedió a la tokenización y lematización utilizando la librería spaCy, que reduce cada palabra a su forma base o lemma considerando su categoría gramatical. A diferencia de enfoques más simples, spaCy permite integrar en un mismo

pipeline el análisis morfológico y el reconocimiento de entidades nombradas (NER), lo que resultó fundamental para garantizar la calidad léxica del corpus.

El proceso de filtrado aplicado sobre cada texto incluyó dos capas:

Primera capa: filtrado por categoría gramatical: se conservaron únicamente los tokens clasificados por spaCy como sustantivos, verbos, adjetivos y adverbios (categorías **NOUN, VERB, ADJ y ADV**). Se eliminaron los stopwords mediante la lista predeterminada de spaCy para inglés, así como los números, signos de puntuación y cualquier token cuya categoría gramatical no correspondiera a las cuatro categorías mencionadas.

Segunda capa: filtrado por entidades nombradas (NER): mediante el reconocimiento de entidades nombradas de spaCy se identificaron y excluyeron personas, organizaciones, países y lugares (etiquetas **PERSON, ORG, GPE y LOC**). Este paso es especialmente importante en el contexto deportivo, donde los nombres propios de atletas, equipos y sedes son extremadamente frecuentes pero no aportan información sensorial ni semántica generalizable.

Por ejemplo:

- "players" : "player" conservado (sustantivo)
- "running" : "run" conservado (verbo)
- "Simone Biles" excluido (persona — NER)
- "Paris" excluido (lugar — NER)
- "the", "a", "in" excluidos (stopwords)

Este procedimiento garantiza que los términos que ingresan al análisis de Lancaster y al Sesgo Semántico sean únicamente aquellos con contenido semántico sustantivo, maximizando la precisión de los perfiles sensoriales y de las asociaciones semánticas calculadas.

5.4 Aplicación de las Normas Sensoriales de Lancaster

5.4.1 Procedimiento de correspondencia léxica

Para cada noticia traducida y lematizada, se realizó un procedimiento de correspondencia palabra a palabra con las entradas del lexicón Lancaster. Específicamente:

- Se identificaron todos los tokens lematizados de la noticia que tuvieran correspondencia directa con una entrada del lexicón.
- Para cada token coincidente, se extrajeron sus puntuaciones en las 11 modalidades sensoriales.
- Se calculó el promedio de evocación para cada modalidad sobre el total de tokens coincidentes en la noticia, mediante:

$$S_M = \frac{1}{n} \sum_{i=1}^{n} s_i(\mathbf{w})$$

Donde n es el número de tokens con correspondencia en el lexicón y s_i(w) es la puntuación de la palabra w en la modalidad M.

El resultado final es un vector sensorial promedio de 11 dimensiones por noticia, que cuantifica la huella sensorial de cada texto y constituye el insumo directo para los análisis estadísticos comparativos entre disciplinas descritos en la Sección 5.6.

5.5 Análisis de sesgos semánticos

Antes de los análisis inferenciales del sesgo semántico y las percepciones sensoriales de Lancaster, el corpus se limitó a las disciplinas deportivas con mayor representación. En este sentido, las 8 disciplinas tenían un mínimo de 30 artículos de noticias cada una y la categoría de "deportes múltiples" se eliminó debido a su naturaleza heterogénea. Esta lista, TOP 8 durante todo el estudio, estaba compuesta por atletismo (n=72), boxeo (n=69), fútbol (n=56), lucha (n=47), remo (n=47), gimnasia (n=36), baloncesto (n=34) y natación (n=33), sumando un total de 394 artículos de noticias o el 54,5% del corpus. Este criterio no garantiza estabilidad estadística en sentido estricto, pero mejora la confiabilidad de las estimaciones al evitar incluir disciplinas con muestras demasiado pequeñas que comprometerían la potencia de las pruebas inferenciales.

Los sesgos semánticos se analizaron utilizando Bolukbasi et al. (2016), que proporciona una comprensión de esta dirección de género oculta en el lenguaje que puede ser identificada y cuantificada en términos de los vectores semánticos de las palabras. Al procesar directamente las incrustaciones de 300 dimensiones del modelo spaCy en core_web_glo (GloVe), la metodología prescinde de la necesidad de construcciones manuales de las listas de atributos.

5.5.1 Construcción del eje de género

El eje de género se construyó a partir de los 10 pares de palabras definitivos propuestos por Bolukbasi et al. (2016): (she–he, her–his, woman–man, mary–john, herself–himself, daughter–son, mother–father, gal–guy, girl–boy, female–male). Para cada par se calculó el vector diferencia (masculino − femenino), obteniendo una matriz de 10 vectores. Sobre esta matriz centrada se aplicó Descomposición en Valores Singulares (SVD), extrayendo el primer vector singular como eje de género **g**, que captura la dirección de mayor varianza en el espacio semántico masculino–femenino.

5.5.2 Vectores de referencia de género

Para construir los vectores de referencia femenino (VEC_FEM) y masculino (VEC_MAS) se utilizó la lista oficial de 218 palabras gender-specific del Apéndice F de Bolukbasi et al. (2016), disponible en el repositorio oficial de los autores y reproducida en el Anexo B del presente documento. Cada palabra fue proyectada sobre el eje de género **g**: las palabras con proyección negativa se clasificaron como femeninas y las de proyección positiva como masculinas. Con los dos subconjuntos resultantes se calcularon los centroides semánticos VEC_FEM y VEC_MAS respectivamente.

5.5.3 Extracción de vocabulario por noticia

Para cada noticia en el subconjunto TOP 8, extrajimos los adjetivos y verbos más comunes del texto traducido al inglés utilizando el modelo core_web_lg de spaCy. El algoritmo filtró las palabras vacías, los signos de puntuación y las palabras sin un vector en el modelo, reteniendo sólo los tokens que pertenecen a la categoría gramatical ADJ o VERB y que tienen una longitud de al menos 3 caracteres. La decisión de excluir los sustantivos del vocabulario analizado, más allá de los sustantivos propios ya filtrados por NER, es una razón metodológica específica del análisis de sesgo semántico y perfil sensorial. Los sustantivos comunes del lenguaje deportivo (partido, competencia, medalla, atleta, equipo, resultado) son compartidos de forma transversal por todas las disciplinas y por ambos polos del eje de género: se trata de un vocabulario neutro de referencia que no diferencia entre coberturas ni porta carga evaluativa o actitudinal. En cambio, los adjetivos y verbos son las categorías gramaticales que concentran el contenido evaluativo, descriptivo y estilístico del lenguaje periodístico: cómo se califica al atleta, qué acciones se destacan, qué atributos se enfatizan. Son estos los que codifican las orientaciones semánticas que este estudio busca detectar. Incluir sustantivos comunes entorpeceria la señal de sesgo al introducir ruido no informativo

para el análisis comparativo entre disciplinas. Los 20 tokens más frecuentes por noticia fueron seleccionados como un vocabulario representativo para el análisis de sesgo. El umbral de 20 tokens se determinó porque constituye el punto de estabilización de las puntuaciones: con incrementos adicionales de tokens, el valor de la puntuación varía en menos del 2% y la cobertura es adecuada incluso para las noticias más cortas (los tokens válidos mínimos después del filtrado son 23).

5.5.4 Cálculo de scores de sesgo de género

A partir del vocabulario extraído se calcularon dos scores de sesgo complementarios para cada noticia, como se describió en la Sección 4.1.3. El Score Coseno mide la diferencia entre la similitud coseno promedio del vocabulario con VEC_MAS y su similitud coseno promedio con VEC_FEM. El Score de Proyección mide la proyección promedio de los vectores normalizados del vocabulario sobre el eje de género g. En ambos casos, un valor positivo indica orientación hacia el polo masculino y un valor negativo hacia el polo femenino.

5.6. Análisis estadístico

Una vez obtenidas las tres métricas principales del estudio:sentimiento por noticia, vector sensorial de Lancaster y scores de sesgo de género. Se aplicaron análisis estadísticos diferenciados para cada una, seleccionados en función de la naturaleza de los datos y el cumplimiento de los supuestos requeridos.

5.6.1 Análisis estadístico del sentimiento

El sentimiento constituye una variable categórica nominal con tres categorías (positivo, negativo, neutro). Para evaluar si su distribución varía significativamente según la disciplina deportiva, se aplicó una tabla de contingencia con prueba Chi-cuadrado, que es la prueba adecuada para evaluar la asociación entre dos variables categóricas sin requerir supuestos de normalidad ni escala continua. Previo a la prueba se verificó el supuesto de frecuencias esperadas mínimas, confirmando que el 95.8% de las celdas presentaba frecuencias esperadas superiores a 5, cumpliendo así el criterio estándar para la validez del estadístico Chi-cuadrado.

5.6.2 Análisis estadístico de las percepciones sensoriales (Lancaster)

Las puntuaciones sensoriales de Lancaster producen 11 variables numéricas continuas por noticia. Dado que el estudio involucra múltiples variables dependientes, el primer paso

consistió en evaluar si era pertinente un análisis multivariado conjunto. Para ello se aplicaron dos pruebas previas: la prueba de esfericidad de Bartlett, que evalúa si las dimensiones están correlacionadas entre sí, y la medida de adecuación muestral KMO, que evalúa si la estructura de esas correlaciones es adecuada para un análisis multivariado. Los resultados de ambas pruebas determinaron si se procedía con MANOVA o con análisis univariados independientes por dimensión.

Una vez definida la estrategia analítica, se evaluó la normalidad de los residuos de cada dimensión mediante la prueba de Shapiro-Wilk, ajustando para cada una un modelo que predice la puntuación sensorial a partir de la variable deporte. Las dimensiones con residuos normales fueron analizadas mediante ANOVA unifactorial, verificando previamente la homogeneidad de varianzas con la prueba de Levene y aplicando comparaciones post-hoc de Tukey HSD. Las dimensiones con residuos no normales fueron analizadas mediante Kruskal-Wallis con comparaciones post-hoc de Dunn y corrección de Bonferroni.

5.6.3 Análisis estadístico del sesgo de género

Los scores de sesgo de género constituyen dos variables numéricas continuas por noticia. Dado que ambos scores se calculan sobre el mismo vocabulario de cada noticia, se evaluó primero su correlación para determinar si procede un análisis multivariado conjunto. En caso de correlación alta, se reportan ambos scores de forma descriptiva y se selecciona el más interpretable para el análisis inferencial principal. Para cada score se evaluó la normalidad de los residuos del modelo de medias por grupo mediante la prueba de Shapiro-Wilk. En caso de cumplirse el supuesto de normalidad se aplicó ANOVA unifactorial con prueba post-hoc de Tukey HSD, verificando previamente la homogeneidad de varianzas mediante la prueba de Levene.

Adicionalmente, para evaluar si el sesgo detectado es sistemáticamente masculino en todas las disciplinas y no simplemente positivo por azar , se aplicó una prueba t de una muestra por disciplina contrastando la hipótesis nula de que la media del score es igual a cero. Un resultado significativo con media positiva confirmaría que el vocabulario de esa disciplina se orienta hacia el polo masculino de forma estadísticamente robusta.

6. Resultados

6.1.1 Web Scraping

Utilizando técnicas de web scraping, se recopilaron 723 artículos de noticias deportivas de cuatro fuentes periodísticas en español: ESPN Deportes, BBC Deportes, Olympics.com y Diario AS. La recopilación incluyó los 15 días previos a los Juegos Olímpicos de París 2024, su duración oficial y los 15 días posteriores a su clausura.

En cuanto a la distribución por fuente, AS generó 350 artículos (48.3%), Olympics 194 (26.8%), ESPN 111 (15.3%) y BBC 70 (9.7%). En cuanto a la longitud promedio por artículo, BBC escribió los textos más largos (promedio de 1,014 palabras), seguido de Olympics con 965, ESPN con 459 y AS con 439.

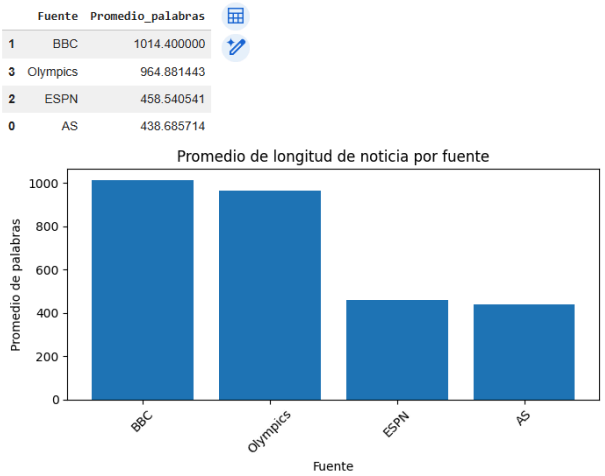


gráfico 1: promedio de longitudes de noticia por fuente

Posteriormente a la recolección, se aplicó el diccionario de datos sobre los títulos y contenidos de cada noticia para identificar la disciplina deportiva correspondiente. Este proceso permitió etiquetar las 723 noticias en 35 disciplinas olímpicas distintas. Las disciplinas con mayor representación en el corpus fueron atletismo (n=71), boxeo (n=64), remo (n=54), múltiples deportes (n=69) y fútbol (n=55), concentrando entre estas cinco categorías más del 40% del corpus total.

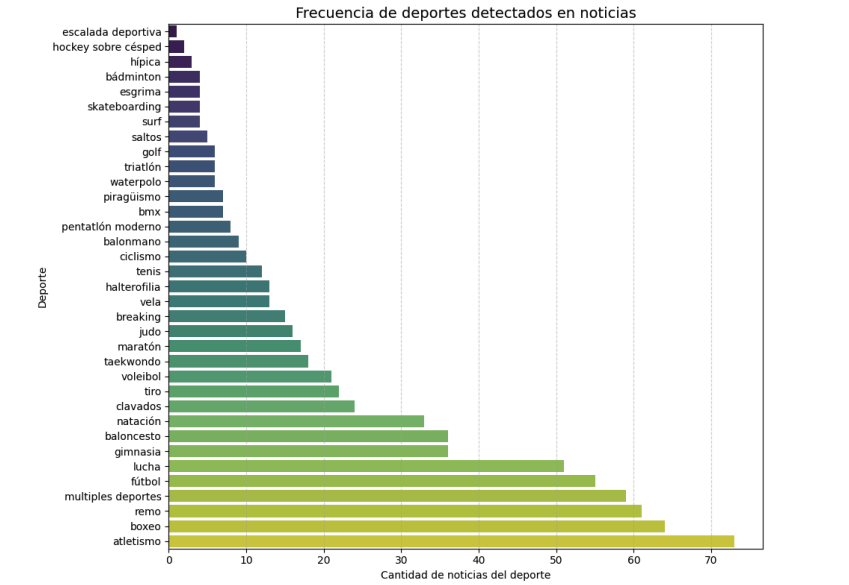


gráfico 2: distribución de las noticias en los deportes

A partir de esta distribución, se seleccionaron para el análisis inferencial las 8 disciplinas con mayor representación en el corpus y con un mínimo de 30 noticias cada una: atletismo (n=72), boxeo (n=69), fútbol (n=56), lucha (n=47), remo (n=47), gimnasia (n=36), baloncesto (n=34) y natación (n=33). En conjunto, estas disciplinas concentran 394 noticias, equivalentes al 54.5% del corpus total. La categoría "múltiples deportes" fue excluida del análisis inferencial por su naturaleza heterogénea, al agrupar coberturas que no corresponden a una sola disciplina olímpica.

6.1.2 Descripción del dataset

El dataset consolidado contiene seis variables principales por noticia: título, contenido textual completo, fecha de publicación, URL, fuente y número total de palabras. Teniendo al final un dataset de 723 filas y 6 columnas. La variable de contenido constituye la base para todos los análisis lingüísticos posteriores: sentimientos, percepciones sensoriales y sesgos semánticos. Mientras que la variable de deporte detectado actúa como variable de agrupación en los análisis comparativos.

La diferencia en extensión promedio entre fuentes; BBC y Olympics duplican en longitud a ESPN y AS, es un factor relevante a considerar en la interpretación de los resultados, ya que artículos más extensos ofrecen mayor cobertura léxica para los análisis de Lancaster y Sesgo semántico.

6.2. Análisis de Sentimientos (T5)

Se aplicó el modelo pysentimiento/robertuito-sentiment-analysis sobre el texto completo de las 723 noticias del corpus, clasificando cada artículo en una de tres categorías emocionales: positiva (POS), neutra (NEU) o negativa (NEG).

Distribución global de sentimientos

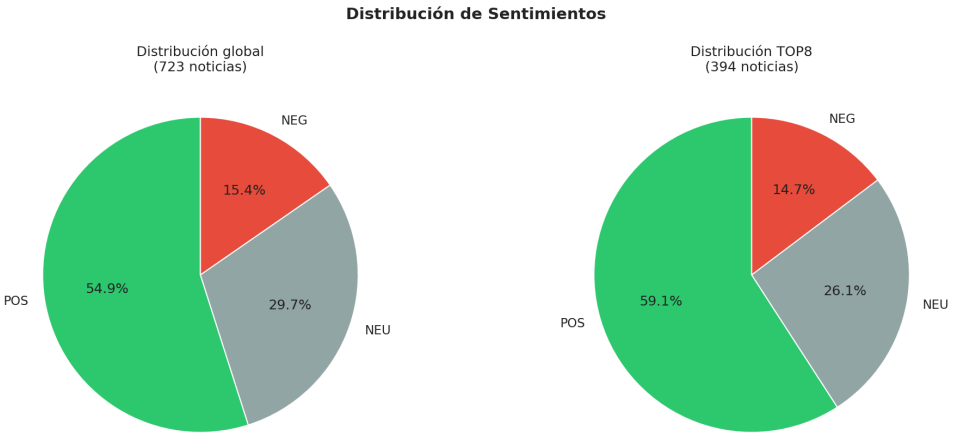


Gráfico 3: distribución de los sentimientos a través de las noticias

Los resultados muestran un predominio claro de la categoría positiva, que concentra el 54.9% de las noticias analizadas, seguida de la categoría neutra con 29.7% y la negativa con 15.4%.

Este patrón es coherente con la naturaleza del periodismo deportivo olímpico. Las narrativas de alto rendimiento tienden a favorecer el logro, la superación de desafíos y el éxito, lo que naturalmente resulta en un tono emocional positivo dominante en el discurso. Como celebración deportiva mundial, los Juegos Olímpicos atraen una gran cantidad de atención que se centra en victorias, récords, momentos históricos importantes y actuaciones excepcionales, todo impregnado de un tono positivo. El 29.7% del contenido neutral se encuentra dentro del elemento informativo y descriptivo del periodismo deportivo formal, donde una parte de la cobertura se presenta como un reporte de hechos y estadísticas sin carga evaluativa explícita. Aunque el elemento negativo (15.4%) sigue siendo menor de los tres, no es marginal en naturaleza. Ese porcentaje puede estar vinculado a derrotas o eliminaciones, lesiones, controversias o sanciones y críticas severas de rendimiento. Su tasa mucho más baja en comparación con el contenido positivo da credibilidad a la idea de que los discursos deportivos olímpicos tienden a presentar una narrativa de éxito y superación de obstáculos en lugar de conflicto o dificultad.

6.2.1 Análisis inferencial del sentimiento, Chi Cuadrado de contingencia

Deporte	NEG	NEU	POS	Total	% POS	% NEG
Atletismo	5	24	43	72	59.7%	6.9%
Baloncesto	2	12	20	34	58.8%	5.9%
Boxeo	15	17	37	69	53.6%	21.7%
Fútbol	12	11	33	56	58.9%	21.4%
Gimnasia	2	7	27	36	75.0%	5.6%
Lucha	9	11	27	47	57.4%	19.1%
Natación	5	9	19	33	57.6%	15.2%
Remo	8	12	27	47	57.4%	17.0%
Total	**58**	**103**	**233**	**394**	**59.1%**	**14.7%**

Tabla 1: resultados prueba chi cuadrado de contingencia

La prueba Chi-cuadrado no arroja una asociación estadísticamente significativa entre la disciplina deportiva y el sentimiento de la cobertura (χ^2=17.84, p=0.2142). Esto indica que la distribución emocional de las noticias no varía de manera sistemática y significativa entre deportes el tono positivo dominante es una característica transversal del corpus independientemente de la disciplina cubierta.

6.3. Aplicación de las Normas Sensoriales de Lancaster

6.3.1 Preprocesamiento y cobertura léxica

Tras la traducción, estandarización y lematización del corpus mediante spaCy, se obtuvieron en promedio 43.2 tokens por noticia, con un mínimo de 9 y un máximo de 127 tokens. El vocabulario único del corpus procesado comprende 3.492 términos, de los cuales 3.303 tienen correspondencia directa en el lexicón de Lancaster, representando una cobertura léxica del 94.6%. A nivel de noticia individual, la cobertura promedio fue de 97.8%, sin ninguna noticia por debajo del 50%, lo que garantiza la representatividad de los vectores sensoriales calculados.

Los 189 tokens sin correspondencia en el lexicón corresponden principalmente a términos deportivos especializados como repechaje, pentatlón o conade, cuya ausencia no afecta significativamente los perfiles sensoriales dada su baja frecuencia relativa.

6.3.2 Jerarquía sensorial global del corpus

Para contextualizar estos valores, el Gráfico 4 compara las medias del corpus deportivo contra las medias globales del lexicón Lancaster completo (39.707 palabras). El corpus de las noticias deportivas supera al lexicón general en 10 de las 11 dimensiones sensoriales, lo que sugiere descriptivamente que el lenguaje periodístico deportivo tiende a ser más rico en evocación sensorial que el lenguaje general. Las diferencias más pronunciadas se observan en las dimensiones motoras: Pie/Pierna (1.59x), Torso (1.40x) y Mano/Brazo (1.29x). La única dimensión por debajo del lexicón general es Gustativo (0.91x), lo esperado dado que el gusto es prácticamente irrelevante en las narrativas deportivas, aunque esta comparación no fue sometida a contraste estadístico formal y fue solo utilizada como un punto de partida para el análisis descriptivo.

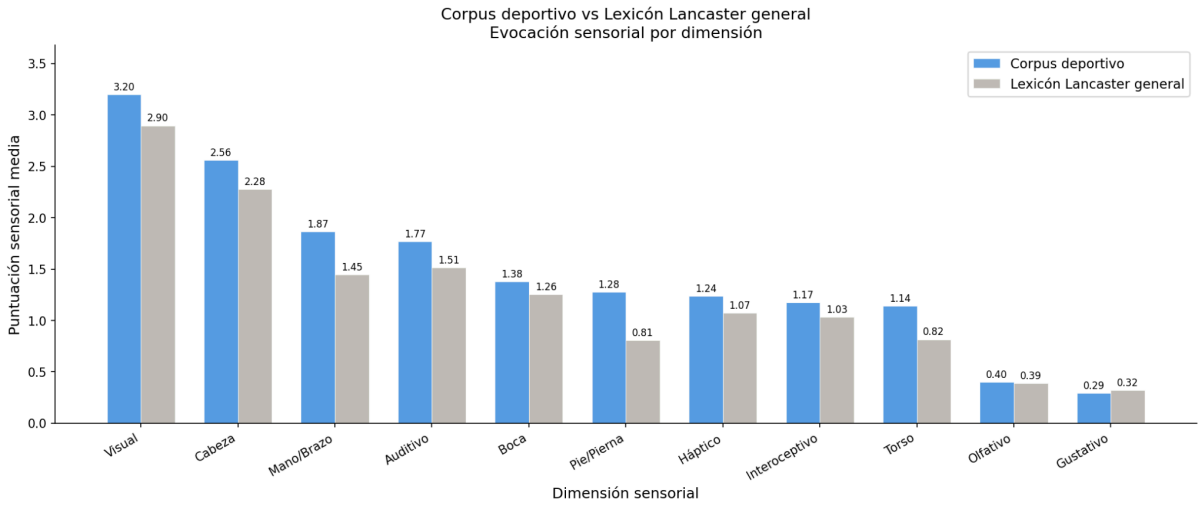


Gráfico 4 :comparación medias sensoriales del corpus deportivo contra el lexicón

6.3.3 Distribución por dimensión individual

El Gráfico 5 (histogramas por dimensión) permite observar la forma de la distribución de cada modalidad sensorial. La mayoría de las dimensiones presentan distribuciones aproximadamente normales, aunque con distintos grados de asimetría, resultado consistente con las pruebas de Shapiro-Wilk aplicadas sobre los residuos, donde 3 de 11 dimensiones no rechazaron la hipótesis de normalidad. La dimensión auditiva y la de cabeza son las más cercanas a la normalidad, mientras que dimensiones como pie/pierna y torso muestran mayor asimetría positiva, con una cola derecha más pronunciada que refleja la presencia de noticias con activación motora particularmente alta.

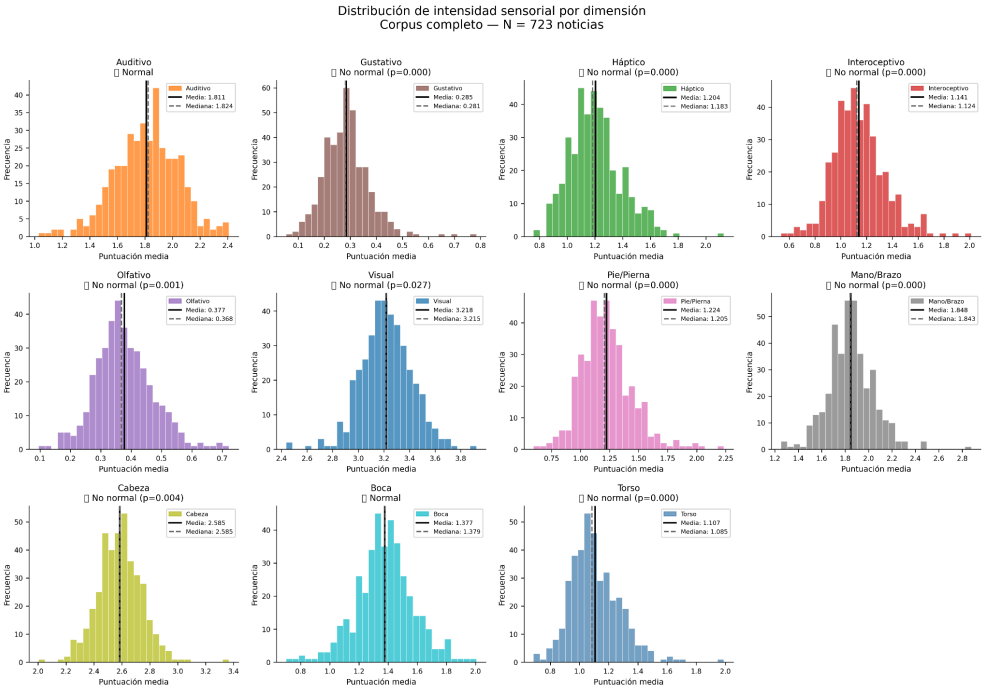


Gráfico 5: distribución de intensidad sensorial por cada dimensión

En la visualización, las dimensiones se distinguen mediante una paleta cromática específica: el auditivo se representa en naranja, el gustativo en marrón, el háptico en verde y el interoceptivo en rojo. Por su parte, el olfativo utiliza el color púrpura y el visual el azul. En cuanto a las dimensiones corporales, el pie/pierna aparece en rosa, la mano/brazo en gris, la cabeza en un tono amarillo oliva, la boca en celeste y el torso en un azul acero pálido.

6.3.4 Correlaciones entre dimensiones sensoriales

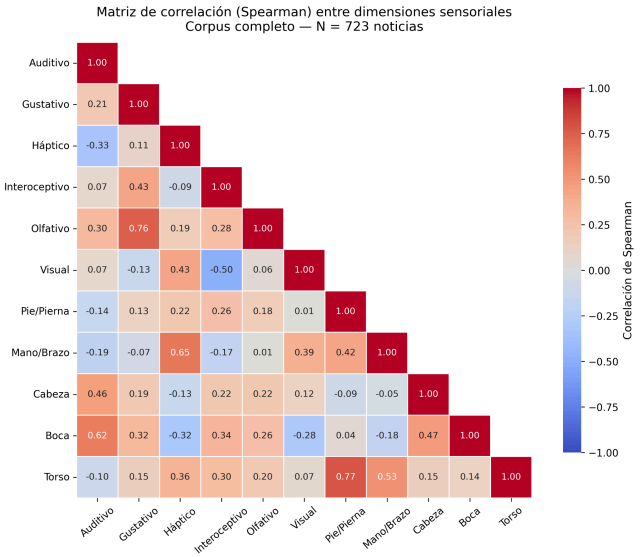


Gráfico 6: matriz de correlación de Spearman entre los sentidos

La matriz de correlación de Spearman revela una estructura clara en el lenguaje deportivo. Las correlaciones más fuertes se observan entre dimensiones motoras relacionadas: Pie/Pierna y Torso (ρ=0.77) y Mano/Brazo y Háptico (ρ=0.65), confirmando que el lenguaje deportivo construye descripciones corporales integradas donde el movimiento de una parte del cuerpo evoca naturalmente referencias al resto. En el extremo opuesto, la correlación más negativa se da entre Visual e Interoceptivo (ρ=-0.50), sugiriendo que las noticias con alta evocación visual tienden a tener menor referencia a sensaciones internas del cuerpo. Dentro de las dimensiones químico-sensoriales, Gustativo y Olfativo presentan una correlación notable (ρ=0.76). Esta asociación no es exclusiva del lenguaje deportivo, se observa también en otros ámbitos del lenguaje general, por lo que su presencia en este corpus confirma que el patrón se mantiene, sin poder atribuirse como característica propia del discurso periodístico deportivo. Finalmente, Auditivo y Boca muestran una correlación moderada (ρ=0.62), reflejando que las referencias sonoras del deporte tienden a acompañarse de menciones a la expresión oral de los protagonistas.

6.3.5 Análisis inferencial de las percepciones sensoriales (Lancaster)

Resultados Esfericidad de Bartlett y KMO

Antes de proceder con los análisis univariados, se evaluó si las 11 dimensiones sensoriales podían analizarse de forma conjunta mediante un MANOVA, para lo cual se aplicaron las pruebas de Bartlett y KMO

- Prueba de esfericidad de Bartlett: evaluó si las 11 dimensiones sensoriales son independientes entre sí. El resultado fue altamente significativo ($p < 0.001$), confirmando que las dimensiones están correlacionadas. Esto no descarta el MANOVA de hecho, la correlación entre variables dependientes es precisamente el supuesto que justifica su uso, sino que refuerza la pertinencia de un análisis multivariado conjunto si se cumplen los demás supuestos.

- Medida de adecuación muestral KMO: El valor KMO obtenido fue de 0.559, por debajo del umbral mínimo de 0.60 ampliamente aceptado en la literatura metodológica (Tabachnick & Fidell, 2013; Hair et al., 2018) , lo que indica que las correlaciones parciales entre dimensiones son relativamente altas en comparación con las correlaciones totales, reflejando una estructura de interdependencia heterogénea y poco cohesionada. Aunque la prueba de Bartlett confirmó correlaciones significativas entre dimensiones, el KMO bajo señala que esas correlaciones no responden a un patrón común estable que justifique tratarlas

como un conjunto multivariado integrado. Proceder con MANOVA en ese escenario comprometería la estabilidad e interpretabilidad del análisis conjunto. Por esta razón, se analizó cada dimensión de forma independiente mediante pruebas univariadas.

Resultados Shapiro Wilk

Para determinar qué prueba univariada aplicar a cada dimensión sensorial, se ajustó para cada una un modelo de medias por grupo en el que la puntuación sensorial de la dimensión es predicha por la variable deporte. Los residuos de este modelo, es decir las desviaciones de cada observación respecto a la media de su grupo, fueron sometidos a la prueba de Shapiro-Wilk para evaluar el supuesto de normalidad. Es importante destacar que lo que se evalúa es la normalidad de los residuos del modelo, no la normalidad de la variable sensorial directamente, ya que es sobre los residuos que descansa el supuesto paramétrico del ANOVA.

Dimensión	W	p-valor	Normal (α=0.0
Auditivo	0.9951	0.2470	**Sí**
Gustativo	0.9534	0.0000	**No**
Háptico	0.9799	0.0000	**No**
Interoceptivo	0.9772	0.0000	**No**
Olfativo	0.9876	0.0020	**No**
Visual	0.9933	0.0796	**Sí**
Pie/Pierna	0.9835	0.0002	**No**
Mano/Brazo	0.9764	0.0000	**No**
Cabeza	0.9868	0.0012	**No**
Boca	0.9933	0.0758	**Sí**
Torso	0.9685	0.0000	**No**

Tabla 2: resultados prueba Saphiro-Wilk sobre las 11 dimensiones

Los resultados se presentan en la Tabla 2. Únicamente tres dimensiones no rechazaron la hipótesis de normalidad de los residuos ($p > 0.05$): Auditivo ($W=0.9951$, $p=0.247$), Visual ($W=0.9933$, $p=0.0796$) y Boca ($W=0.9933$, $p=0.0758$). Para estas tres dimensiones se procede con ANOVA unifactorial, verificando previamente la homogeneidad de varianzas mediante la prueba de Levene. Las ocho dimensiones restantes: Gustativo, Háptico, Interoceptivo, Olfativo, Pie/Pierna, Mano/Brazo, Cabeza y Torso, presentaron residuos no normales ($p < 0.05$) y serán evaluadas mediante la prueba de Kruskal-Wallis con comparaciones post-hoc de Dunn y corrección de Bonferroni.

Resultados ANOVA

Dimensiones con residuos normales (Shapiro-Wilk α=0.05)

Dimensión	Levene	ANOVA	η²
Auditivo	W=0.9919 p=0.4366 varianzas homogéneas	F=3.9126 p=0.0004 **Significativo**	η² = 0.0663 (mediano)
Visual	W=1.4810 p=0.1724 varianzas homogéneas	F=1.5052 p=0.1639 **No significativo**	η² = 0.0266 (pequeño)
→ Tukey HSD omitido (ANOVA no significativo)			
Boca	W=1.3175 p=0.2405 varianzas homogéneas	F=3.6089 p=0.0009 **Significativo**	η² = 0.0614 (mediano)

Tabla 3: resultados ANOVA

Para las tres dimensiones con residuos normales se verificó previamente la homogeneidad de varianzas mediante la prueba de Levene, que resultó no significativa en los tres casos, confirmando varianzas homogéneas entre grupos. El ANOVA unifactorial arrojó resultados significativos para Auditivo y Boca, ambas con un tamaño del efecto mediano, lo que indica que existen diferencias estadísticamente relevantes en estas dimensiones entre las disciplinas deportivas analizadas. Para estas dos dimensiones se procedió a aplicar la prueba post-hoc de Tukey HSD para identificar los pares de deportes con diferencias significativas. La dimensión Visual, en cambio, no alcanzó significancia estadística y presentó un tamaño del efecto pequeño, por lo que la prueba post-hoc fue omitida, confirmando que la evocación visual es homogénea a través de todas las disciplinas del corpus.

Resultados Tukey

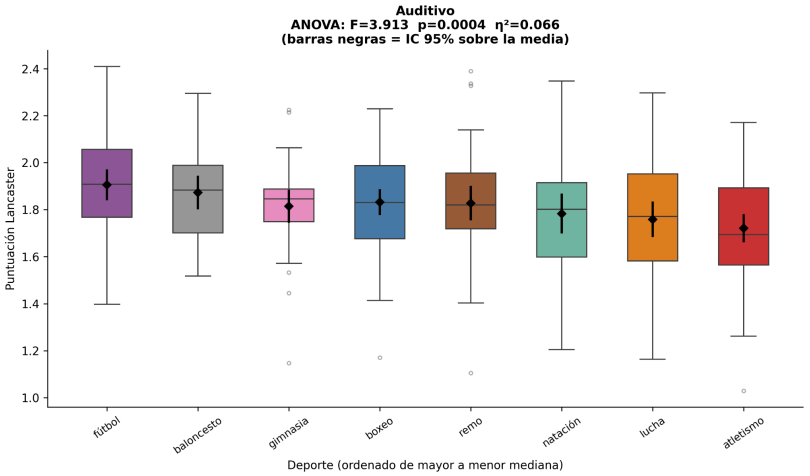


Gráfico 7: boxplot de resultados de Tukey de dimensión Auditiva

Para la dimensión Auditivo (F=3.913, p=0.0004, η²=0.066), las comparaciones post-hoc de Tukey HSD identificaron dos pares con diferencias estadísticamente significativas: fútbol presenta una evocación auditiva significativamente mayor que atletismo (Diff=0.185, p=0.0002), y baloncesto también supera a atletismo (Diff=0.152, p=0.031). El resto de los pares no alcanzó significancia estadística. Este resultado es coherente con la naturaleza colectiva del fútbol y el baloncesto, donde el sonido del ambiente, los cánticos y la comunicación entre jugadores son elementos centrales de la narrativa periodística, en contraste con el atletismo, deporte individual de ejecución física predominantemente silenciosa.

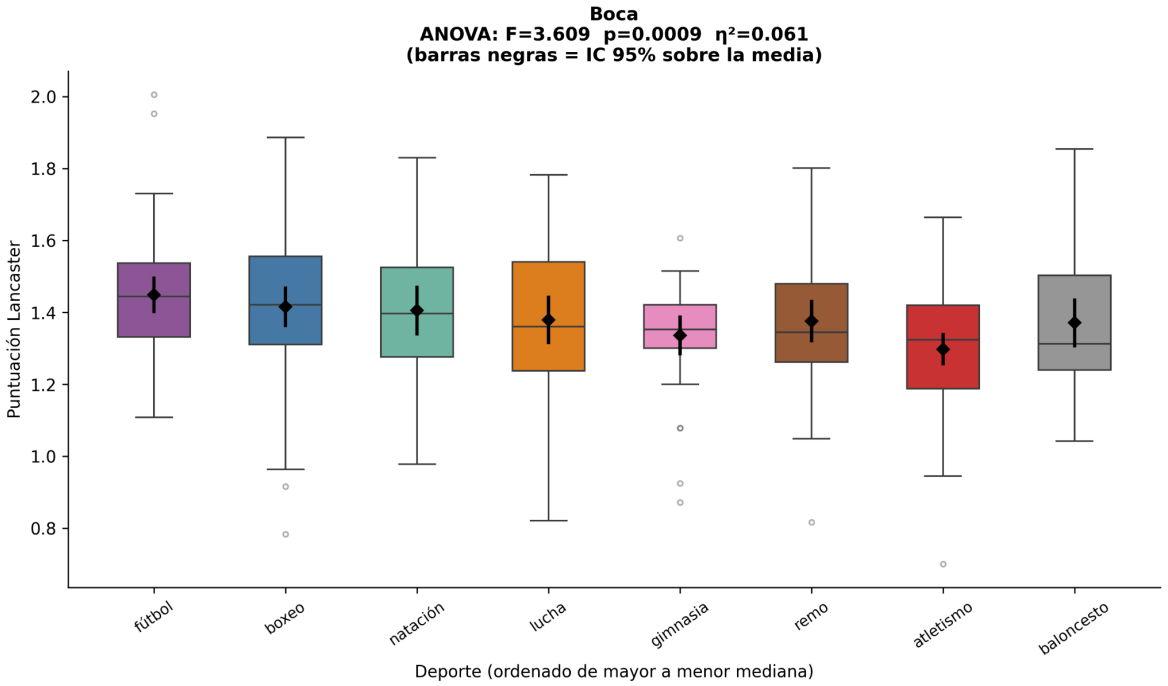


Gráfico 8: boxplot de resultados de Tukey de dimensión Bucal

Para la dimensión Boca (F=3.609, p=0.0009, η^2=0.061), las comparaciones post-hoc de Tukey HSD identificaron dos pares significativos: fútbol presenta mayor evocación de la dimensión bucal que atletismo (Diff=0.151, p=0.003), y boxeo también supera a atletismo en esta dimensión (Diff=0.117, p=0.008). En ambas dimensiones el atletismo aparece consistentemente como el deporte con menor evocación, lo cual es coherente con su naturaleza individual y de ejecución física silenciosa.

Resultados Kruskal-Wallis

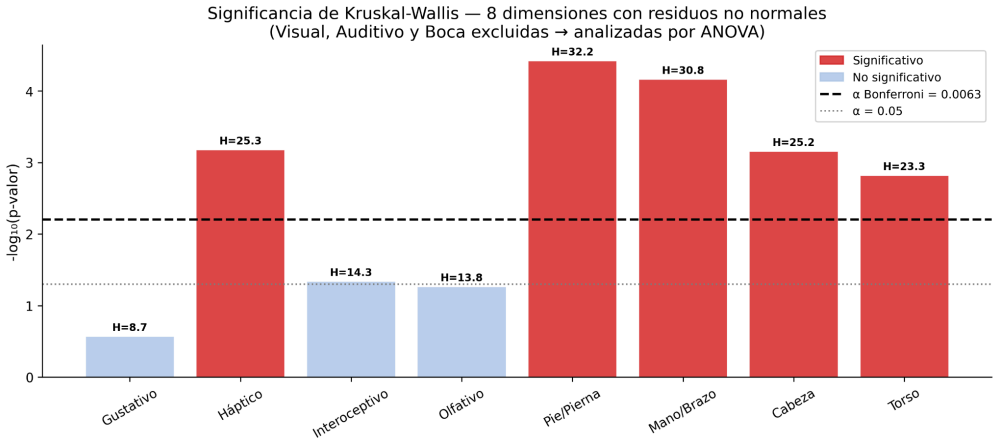


Gráfico 9: significancia KW por dimensión, $-\log_{10}(p)$

La prueba de Kruskal-Wallis se aplicó sobre las ocho dimensiones cuyos residuos no cumplieron el supuesto de normalidad: Gustativo, Háptico, Interoceptivo, Olfativo, Pie/Pierna, Mano/Brazo, Cabeza y Torso. Los resultados completos se presentan en el Gráfico 9.

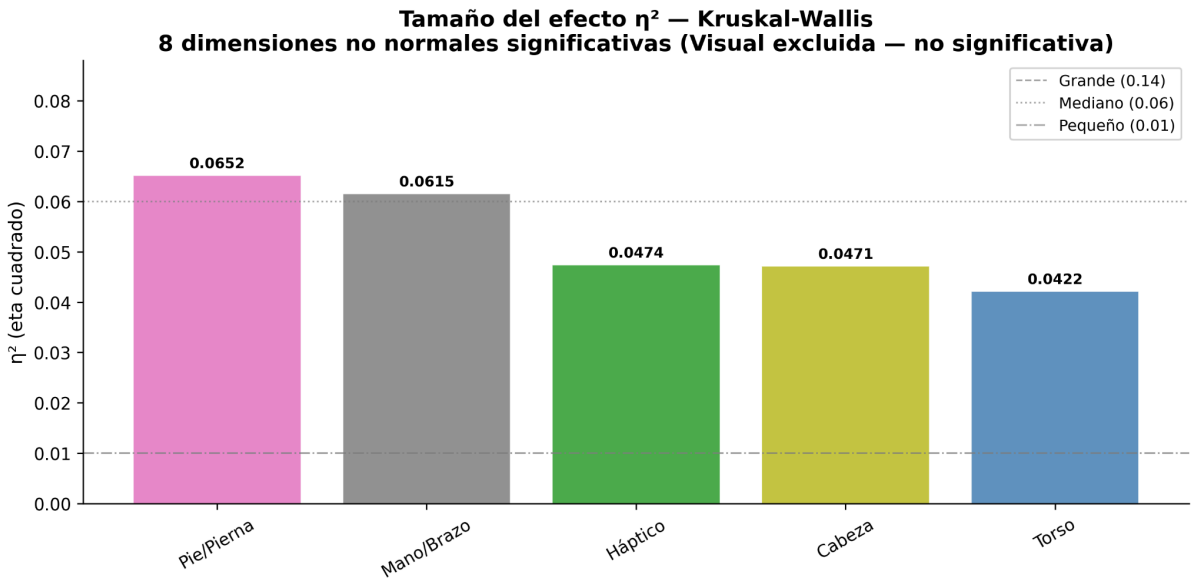


Gráfico 10: tamaños del efecto η² en las 5 dimensiones significativas

Cinco dimensiones superaron el umbral de Bonferroni (α=0.0063): Pie/Pierna (H=32.2, p<0.001, η²=0.065), Mano/Brazo (H=30.8, p<0.001, η²=0.062), Háptico (H=25.3, p<0.001, η²=0.047), Cabeza (H=25.2, p<0.001, η²=0.047) y Torso (H=23.3, p=0.0015, η²=0.042). Las tres dimensiones restantes — Gustativo (H=8.7), Interoceptivo (H=14.3) y Olfativo (H=13.8) — no superaron el umbral de Bonferroni, indicando que las diferencias entre deportes en estas modalidades no son estadísticamente significativas tras la corrección. Los tamaños del efecto de las cinco dimensiones significativas se presentan en el Gráfico 10

Las comparaciones post-hoc de Dunn con corrección de Bonferroni sobre las cinco dimensiones significativas identificaron los siguientes patrones diferenciados por disciplina:

En la dimensión **Háptico** (η²=0.047, efecto pequeño), la prueba de Kruskal-Wallis resultó globalmente significativa, con lucha presentando la mediana más alta del grupo, coherente con su naturaleza de contacto físico total. Sin embargo, ningún par específico superó el umbral de corrección de Bonferroni en las comparaciones post-hoc de Dunn, por lo que no es posible señalar diferencias estadísticamente robustas entre disciplinas concretas en esta dimensión. Todos los deportes comparten el grupo a en los grupos post-hoc.

En la dimensión **Pie/Pierna** (η²=0.065, efecto mediano), atletismo presenta la mediana más alta del grupo y se diferencia significativamente de boxeo (p<0.001), fútbol (p=0.001) y baloncesto (p=0.001). en contraste con el baloncesto donde predominan las referencias a manos y brazos. El caso del fútbol es particular: a pesar de ser un deporte de

pierna por definición, su puntuación en esta dimensión es relativamente baja, lo que puede explicarse porque el vocabulario periodístico futbolístico está fuertemente dominado por referencias auditivas y colectivas que diluyen la presencia relativa de términos motores del tren inferior. El boxeo, por su parte, concentra su léxico motor en la cabeza y los brazos, lo que reduce naturalmente su evocación de pie/pierna.

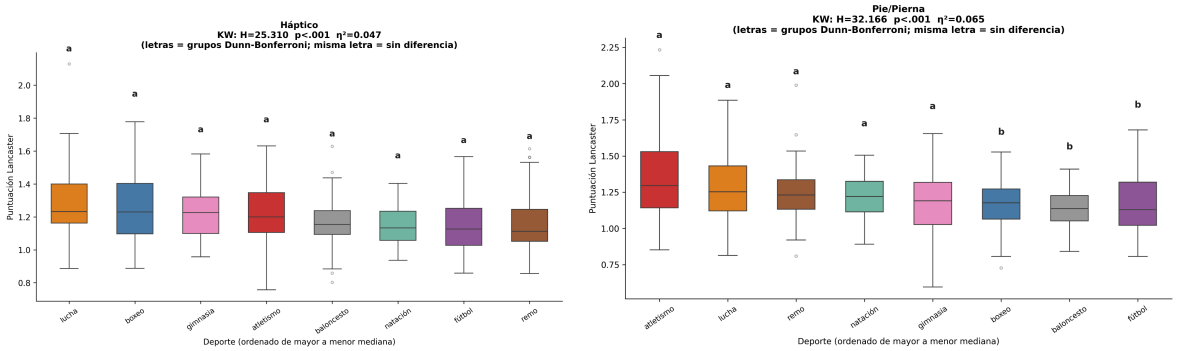


Gráfico 11 y 12: resultados KW de dimensiones Háptico y Pie/Pierna en los top 8 deportes

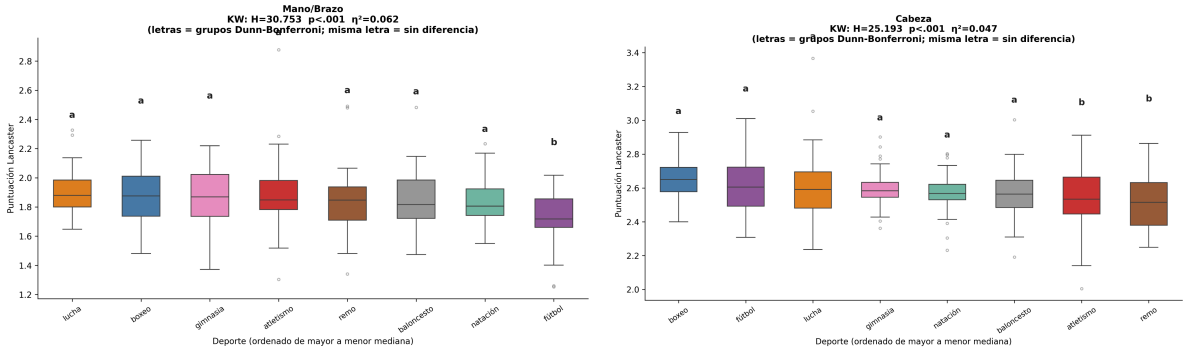


Gráfico 13 y 14: resultados KW de dimensiones Mano/Brazo y Cabeza en los top 8 deportes

En la dimensión **Mano/Brazo** (η^2=0.062, efecto mediano), lucha presenta la mediana más alta y se diferencia significativamente de fútbol ($p<0.001$). Adicionalmente, fútbol presenta la mediana más baja del grupo y se diferencia también de atletismo ($p<0.001$) y gimnasia ($p=0.002$). Esto refleja que el fútbol, por las propias reglas del juego, genera escasas referencias a las extremidades superiores en su cobertura periodística, mientras que la lucha, que es un deporte de agarre y forcejeo, y el atletismo concentran la mayor evocación de esta dimensión.

En la dimensión **Cabeza** (η^2=0.047, efecto pequeño), boxeo lidera con claridad y se diferencia significativamente de atletismo ($p<0.001$) y remo ($p=0.004$), ambos con letra b en los grupos post-hoc. La cabeza es el objetivo central de la acción en el boxeo y el lenguaje periodístico lo recoge inevitablemente en su léxico, mientras que atletismo y remo presentan las medianas más bajas en esta dimensión.

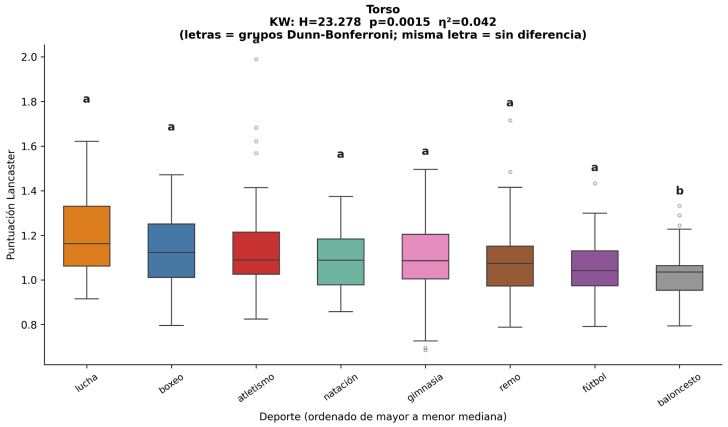


Gráfico 15: resultados KW de la dimensión Torso en los top 8 deportes

En la dimensión **Torso** (η^2=0.042, efecto pequeño), lucha presenta la mediana más alta seguida de boxeo y atletismo, y se diferencia significativamente de baloncesto ($p=0.003$), que registra la mediana más baja del grupo. Esto es coherente con deportes donde el torso es central en la ejecución técnica, en contraste con el baloncesto donde las referencias corporales se concentran en manos y brazos.

6.4 Análisis de Sesgos Semánticos

6.4.1 Construcción del eje de género y vectores de referencia

El eje de género se construyó a partir de los 10 pares definitivos de Bolukbasi et al. (2016) mediante SVD sobre las diferencias de vectores, obteniendo un primer componente principal que explica el 30.7% de la varianza total de las diferencias. A partir de la proyección de las 218 palabras gender-specific del Apéndice F sobre este eje, se clasificaron 98 palabras como femeninas y 107 como masculinas, descartando 13 por ausencia de vector en el modelo. Con estos dos subconjuntos se calcularon los vectores de referencia VEC FEM y VEC MAS que sirven de base para el cálculo de los scores por noticia.

6.4.2 Obtención de los scores por noticia

Para cada una de las 394 noticias del subconjunto TOP 8 se extrajeron los 20 adjetivos y verbos más frecuentes del texto traducido al inglés. Sobre este vocabulario se calcularon dos scores complementarios. El Score Coseno se obtiene como la diferencia entre la similitud coseno promedio del vocabulario con VEC MAS y su similitud coseno promedio con VEC FEM. El Score de Proyección se obtiene como el promedio de las proyecciones de cada vector de palabra normalizado sobre el eje de género. En ambos casos, un valor positivo indica orientación hacia el polo masculino y un valor negativo hacia el polo femenino. Ninguna noticia del corpus quedó sin score, confirmando cobertura total del vocabulario extraído en el modelo.

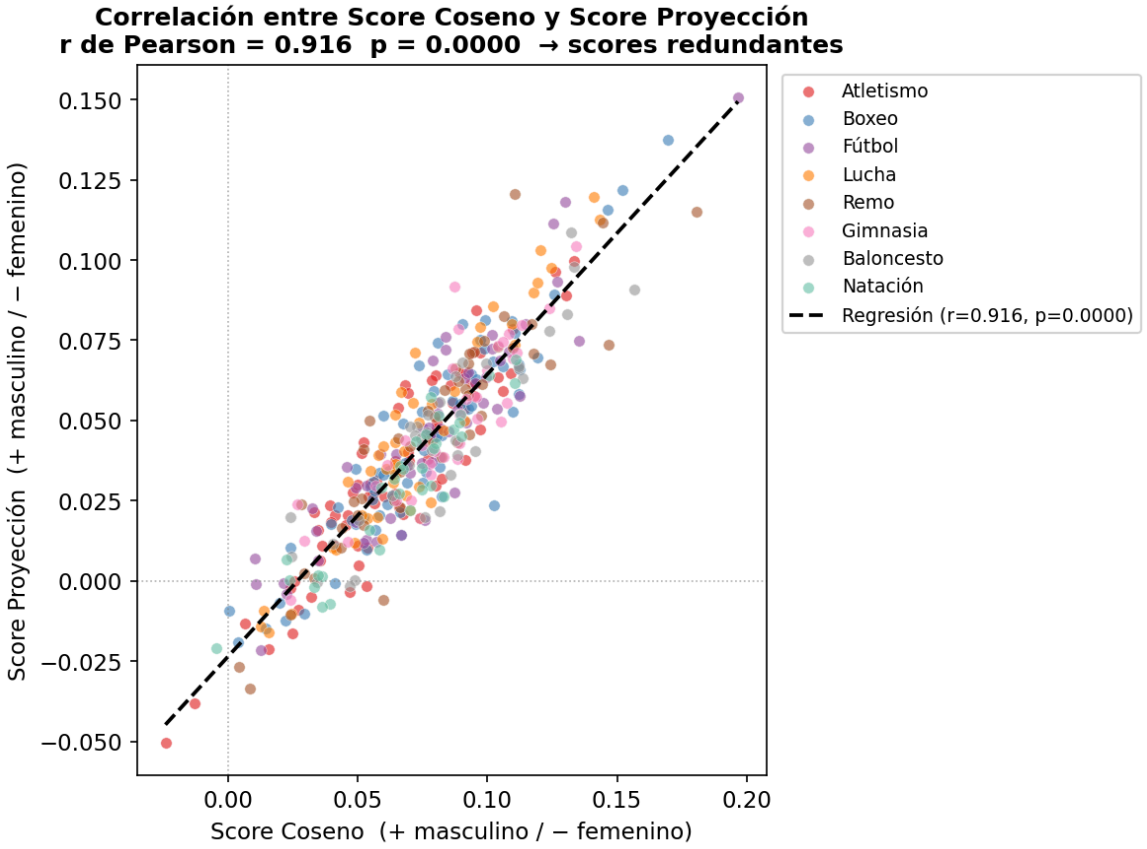


Gráfico 16: resultados correlación entre Score Coseno y Score de Proyección

Con el fin de evaluar si ambos scores miden la misma dimensión subyacente o capturan aspectos complementarios del sesgo de género, se calculó la correlación de Pearson entre el Score Coseno y el Score de Proyección sobre las 394 noticias del subconjunto TOP 8. El resultado arrojó una correlación de $r=0.916$ ($p<0.001$),confirmando que ambos scores son prácticamente redundantes: miden el mismo fenómeno desde dos aproximaciones matemáticas distintas pero llegan a las mismas conclusiones (Gráfico 16: scatterplot de

correlación entre scores). Dado este nivel de redundancia, proceder con un análisis multivariado conjunto como un MANOVA no estaría justificado, ya que el MANOVA asume que las variables dependientes capturan dimensiones distintas del mismo constructo, condición que no se cumple cuando r>0.9. En consecuencia, el análisis inferencial se realizó exclusivamente sobre el Score Coseno como medida principal, por ser el más directamente interpretable en términos de proximidad semántica a los centros de referencia VEC_MAS y VEC_FEM.

6.4.3 Estadística Inferencial

Supuesto Evaluado	Prueba Estadística	Estadístico (\	p-valor	Resultado (α=0.05)
Normalidad (sobre residuos)	Shapiro-Wilk	0.9928	0.0568	Residuos normales
Homogeneidad de Varianzas	Levene	0.8399	0.5547	Varianzas homogéneas

Tabla 4: resultados prueba Shapiro Wilk

La prueba de Shapiro-Wilk sobre los residuos del modelo de medias por grupo confirmó normalidad en Score Coseno habilitando el uso de ANOVA unifactorial. La prueba de Levene verificó la homogeneidad de varianzas. El ANOVA resultó estadísticamente significativo para el Score Coseno, con un tamaño pequeño del efecto.

Fuente	SS	df	MS	F	p
Entre grupos	0.016913	7	0.002416	2.4113	0.0200
Dentro grupos	0.386773	386	0.001002		
Total	0.403686				
Resumen de Resultados:	Tamaño del efecto (η^2): 0.0419 (efecto pequeño)				
Resultado de la Prueba F:	F(7, 386) = 2.4113, p = 0.0200				
Conclusión: Significativo					

Tabla 5: resultados ANOVA

Sesgo masculino sistemático: prueba t de una muestra

Antes de comparar disciplinas entre sí, se evaluó si el sesgo detectado es sistemáticamente masculino en todas las disciplinas —es decir, si las medias del Score Coseno son significativamente mayores que cero. Para ello se aplicó una prueba t de una muestra por disciplina contrastando H_0: $\mu = 0$, con corrección de Bonferroni (α=0.0063). Los

resultados, presentados en el Gráfico 17, son contundentes: los ocho deportes del TOP 8 presentan medias significativamente mayores que cero ($p<0.0063$ en todos los casos), con gimnasia registrando la media más alta ($\bar{x}=0.084$) y atletismo la más baja ($\bar{x}=0.065$). Este resultado confirma que el vocabulario periodístico olímpico se orienta hacia el polo masculino del eje de género de forma sistemática y estadísticamente robusta en todas las disciplinas analizadas, independientemente de la disciplina cubierta.

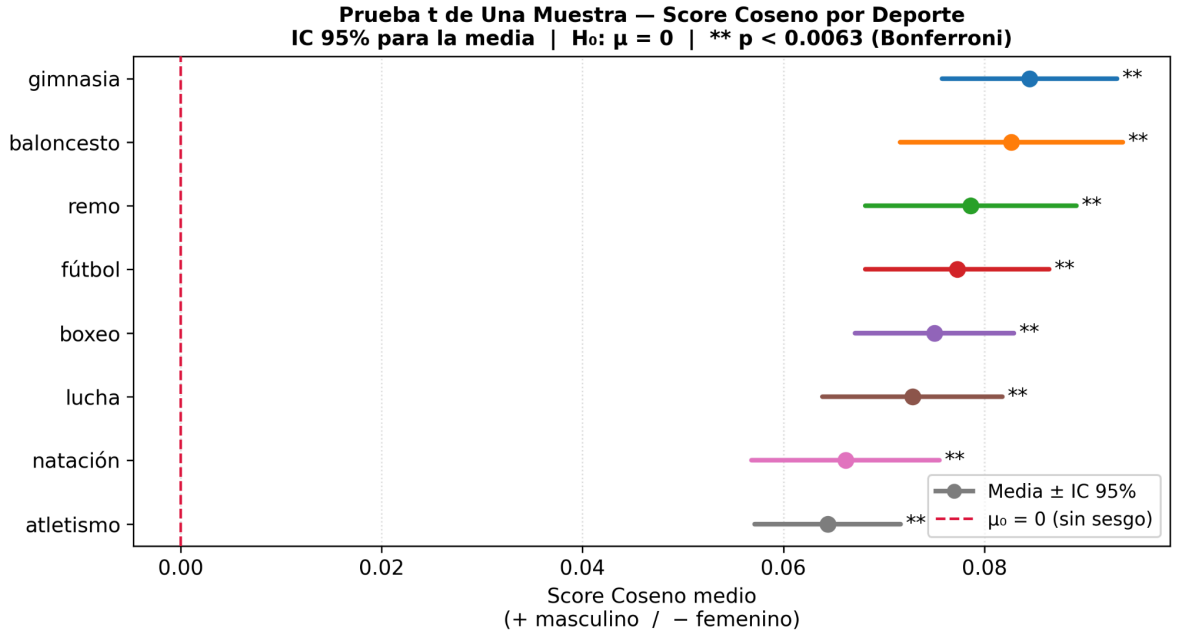


Gráfico 17: resultados Prueba t intervalos de confianza de una muestra

Diferencias entre disciplinas — ANOVA y Tukey HSD

Si bien todas las disciplinas presentan sesgo masculino confirmado, existen diferencias significativas en la magnitud de ese sesgo entre deportes. Los boxplots del Score Coseno por disciplina, ordenados de mayor a menor mediana, muestran que gimnasia presenta la mediana más alta del grupo (mediana=0.084), seguida de baloncesto y remo, mientras que atletismo registra la mediana más baja (mediana=0.065) con mayor dispersión hacia valores cercanos a cero (Gráfico 18).

Las comparaciones post-hoc de Tukey HSD identificaron una diferencia significativa entre gimnasia y atletismo (Diff=0.020, p=0.041), siendo gimnasia el deporte cuyo vocabulario se orienta más intensamente hacia el polo masculino del eje de género. Este resultado es contraintuitivo a primera vista, podría esperarse que la gimnasia, disciplina con fuerte componente estético y artístico, presenta menor sesgo masculino —, pero es coherente

con la naturaleza del eje de género construido mediante el método de Bolukbasi et al. (2016): el eje captura la orientación semántica implícita del vocabulario periodístico en el espacio de embeddings, no el estereotipo de género asociado a la disciplina. Este resultado contraintuitivo requiere cautela en su interpretación. Una hipótesis plausible es que la etiqueta 'gimnasia' en el corpus agrupa coberturas de distintas modalidades: gimnasia artística, rítmica y acrobática, cuyas narrativas periodísticas difieren considerablemente entre sí. Es posible que la mezcla de estas modalidades genere un perfil léxico que no refleja fielmente ninguna de las disciplinas por separado, lo que podría explicar parcialmente la posición atípica de gimnasia en el eje de género. Este hallazgo abre una línea de análisis futura que requeriría desagregar las noticias por modalidad específica para obtener una interpretación más precisa.

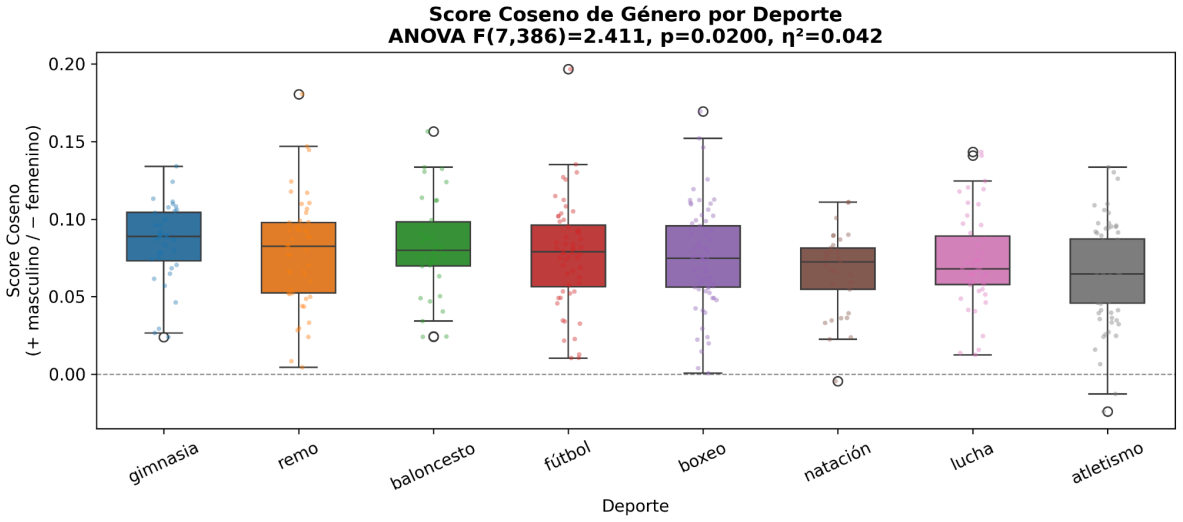


Gráfico 18: sesgo de género en los top 8 deportes según score coseno

7. Análisis y Conclusiones

El presente estudio analizó las narrativas mediáticas de los Juegos Olímpicos de París 2024 desde tres dimensiones complementarias: las emociones evocadas, las percepciones sensoriales activadas y los sesgos semánticos implícitos de género presentes en el lenguaje periodístico. Los hallazgos, obtenidos sobre un corpus de 723 noticias en español provenientes de cuatro fuentes especializadas, permiten trazar un retrato sistemático y cuantitativamente fundamentado del discurso deportivo olímpico.

7.1 Síntesis de hallazgos principales

En cuanto al sentimiento, la cobertura olímpica es predominantemente positiva (54.9%), lo que refleja la naturaleza celebratoria del evento. La prueba Chi-cuadrado no arrojó una asociación estadísticamente significativa entre la disciplina deportiva y el sentimiento de la cobertura (χ^2=17.84, gl=14, p=0.2142), indicando que el tono emocional positivo es una característica transversal del corpus independientemente de la disciplina cubierta. No se realizaron comparaciones post-hoc al no haber significancia estadística.

En cuanto a las percepciones sensoriales, el análisis muestra descriptivamente que el lenguaje periodístico deportivo tiende a ser más rico en evocación sensorial que el lenguaje general, con medias superiores al lexicón Lancaster en 10 de las 11 dimensiones y diferencias especialmente pronunciadas en las dimensiones motoras. Esta comparación es de carácter descriptivo y no fue sometida a contraste estadístico formal. Las pruebas inferenciales revelaron diferencias estadísticamente significativas entre deportes en cinco dimensiones motoras — Pie/Pierna, Mano/Brazo, Háptico, Cabeza y Torso — y en dos dimensiones evaluadas por ANOVA — Auditivo y Boca —, demostrando que cada disciplina genera una huella sensorial específica en el lenguaje que la narra. Las dimensiones Gustativo, Olfativo e Interoceptivo no mostraron diferencias significativas entre deportes, y la dimensión Visual, aunque es la más intensa del corpus (media=3.218), es homogénea a través de todas las disciplinas.

En cuanto al sesgo de género, el análisis mediante el método de Bolukbasi et al. (2016) produjo tres hallazgos principales. Primero, el vocabulario periodístico olímpico se orienta hacia el polo masculino del eje de género de forma sistemática y estadísticamente robusta en las ocho disciplinas del TOP 8 — confirmado mediante prueba t de una muestra con corrección de Bonferroni ($p<0.0063$ en todos los casos) —, lo que constituye evidencia cuantitativa de un sesgo masculino transversal en el discurso deportivo olímpico en español. Segundo, existen diferencias significativas en la magnitud de ese sesgo entre disciplinas (ANOVA $F(7,386)=2.411$, $p=0.020$, $\eta^2=0.042$), aunque el tamaño del efecto es pequeño. Tercero, gimnasia presenta el mayor score de sesgo masculino del grupo y difiere significativamente de atletismo — el deporte con menor score — según las comparaciones post-hoc de Tukey HSD (Diff=0.020, p=0.041).

7.2 La huella sensorial del deporte: el cuerpo como lenguaje

Uno de los hallazgos más relevantes de este estudio es que el periodismo deportivo no construye un lenguaje sensorial homogéneo, sino que cada disciplina genera un perfil

perceptual específico que refleja la naturaleza física de ese deporte. La alta cobertura léxica obtenida — 94.6% del vocabulario único del corpus con correspondencia en Lancaster, y 97.8% a nivel de noticia individual — garantiza que estos perfiles son representativos y no artefactos del procedimiento de mapeado.

El caso más ilustrativo es el contraste entre atletismo y fútbol. Atletismo presenta la evocación más alta en la dimensión Pie/Pierna y se diferencia significativamente de fútbol, boxeo y baloncesto en esta dimensión, coherente con un deporte donde la locomoción y la potencia del tren inferior son centrales. El fútbol, en cambio, lidera la dimensión auditiva y la dimensión bucal, lo que sugiere que la narrativa futbolística privilegia la experiencia sonora y colectiva del evento — el ruido de la hinchada, los cánticos, la comunicación entre jugadores — por encima de la descripción corporal detallada. Este contraste conecta con lo que Cooky et al. (2015) describen como diferencias narrativas entre deportes colectivos e individuales.

El boxeo lidera la evocación de cabeza con claridad, diferenciándose significativamente de atletismo y remo, coherente con un deporte donde la cabeza es el objetivo central de la acción. La lucha domina en las dimensiones Mano/Brazo y Torso, coherente con un deporte de contacto físico total donde el agarre y el forcejeo corporal son la esencia de la acción narrada. Finalmente, el fútbol presenta la mediana más baja en Mano/Brazo, consistente con un deporte donde las referencias a las extremidades superiores son escasas por las propias reglas del juego.

En la dimensión Háptica, si bien la prueba de Kruskal-Wallis resultó globalmente significativa, ningún par específico superó el umbral de corrección de Bonferroni, por lo que no es posible señalar diferencias robustas entre disciplinas concretas en esta dimensión.

7.3 El sesgo semántico de género

El hallazgo más relevante desde la perspectiva de los estudios de género es que el vocabulario periodístico olímpico se orienta consistentemente hacia el polo masculino del eje de género en todas las disciplinas analizadas, con confirmación estadística robusta en los ocho deportes del TOP 8. Este resultado, obtenido sobre un corpus de 723 noticias en español procedentes de cuatro fuentes especializadas, ofrece evidencia cuantitativa de un patrón que la literatura cualitativa había documentado de forma más fragmentaria (Cooky et al., 2015; Bruce, 2015).

Es fundamental subrayar que este sesgo no depende del sexo del atleta cubierto: el análisis no clasificó las noticias según el género del deportista protagonista. Lo que se detecta son patrones léxicos implícitos en el propio discurso periodístico, codificados en la posición de los adjetivos y verbos más frecuentes de cada noticia dentro del espacio semántico de género.

Este resultado contraintuitivo admite al menos dos hipótesis complementarias que vale la pena considerar. La primera es de naturaleza metodológica: la etiqueta "gimnasia" en el corpus agrupa coberturas de distintas modalidades — artística, rítmica y acrobática —, cuyas narrativas periodísticas difieren considerablemente. La gimnasia rítmica, por ejemplo, tiene una cobertura léxica marcadamente distinta a la artística, y su mezcla en una sola categoría puede generar un perfil semántico que no refleja fielmente ninguna de las dos disciplinas por separado.

La segunda hipótesis apunta a una limitación del propio modelo de embeddings. El vocabulario técnico asociado a la gimnasia — términos relacionados con física del movimiento, biomecánica y ejecución corporal como rotation, strength, balance, force o control — puede estar históricamente más próximo al polo masculino en el espacio vectorial de GloVe, no por una característica del discurso periodístico deportivo, sino porque el corpus sobre el que fue entrenado GloVe asocia ese tipo de vocabulario técnico y científico predominantemente con el género masculino. En este caso, el sesgo detectado sería parcialmente un artefacto del modelo de embeddings subyacente, más que un patrón exclusivo del lenguaje periodístico analizado.

Ambas hipótesis son compatibles entre sí y ninguna puede descartarse con los datos disponibles. Lo que sí puede afirmarse con seguridad es que el hallazgo de gimnasia requiere un análisis más granular — desagregando por modalidad y controlando el vocabulario técnico — antes de derivar conclusiones sustantivas sobre el sesgo de género en la cobertura de esta disciplina. En cualquier caso, dado que el tamaño del efecto del ANOVA es pequeño (η^2=0.042), las diferencias entre disciplinas deben interpretarse con cautela — lo verdaderamente relevante es el sesgo masculino sistemático que comparten todas las disciplinas, no las diferencias entre ellas.

7.4 Limitaciones del estudio

El presente estudio presenta varias limitaciones que deben considerarse en la interpretación de sus hallazgos.

En primer lugar, la traducción automática del corpus al inglés, necesaria para la aplicación de las Normas de Lancaster y el modelo de embeddings, introduce dos tipos de pérdida. Por un lado, pueden distorsionar matices semánticos propios del español periodístico deportivo. Por otro lado, y más relevante para el análisis de sesgo de género, la traducción elimina el género gramatical propio del español — los adjetivos "rápido" y "rápida" se traducen ambos como "fast" —, lo que impide que el sesgo de género morfológico del español influya en los scores calculados. Los patrones detectados reflejan exclusivamente el posicionamiento semántico del vocabulario en el espacio vectorial del inglés.

En segundo lugar, la restricción a cuatro fuentes periodísticas limita la diversidad editorial y geográfica del corpus. La sobrerrepresentación de AS (48.3% del corpus) puede sesgar los patrones observados hacia el estilo narrativo de esa publicación específica.

En tercer lugar, el análisis de sesgos semánticos captura orientaciones léxicas a nivel de corpus, no intenciones editoriales conscientes. La presencia de sesgo semántico no implica necesariamente que los periodistas tengan sesgos explícitos de género; puede reflejar convenciones lingüísticas sedimentadas en el idioma mismo o en las convenciones del género periodístico deportivo.

En cuarto lugar, el tamaño muestral por deporte en el subconjunto TOP 8 varía considerablemente — entre 33 noticias para natación y 72 para atletismo —, lo que afecta la potencia estadística de las comparaciones post-hoc y puede explicar que algunas diferencias entre disciplinas no alcanzan significancia estadística a pesar de que las tendencias son consistentes.

Finalmente, el análisis no incorpora el género del atleta protagonista como variable explícita. Esta decisión fue metodológicamente deliberada — el objetivo era detectar patrones léxicos del discurso, no clasificar la cobertura por sexo del deportista —, pero implica que no es posible distinguir si el sesgo semántico detectado se distribuye de forma diferencial entre noticias sobre atletas masculinos y femeninos.

7.5 Líneas de investigación futura

Los hallazgos de este estudio abren varias líneas de investigación que podrían profundizar y ampliar sus contribuciones.

La más inmediata es la incorporación del género del atleta como variable de análisis explícita, cruzando los perfiles sensoriales y los scores de sesgo con el sexo del deportista

protagonista de cada noticia. Esto permitiría responder directamente si las diferencias semánticas detectadas se asocian con el género del atleta cubierto o son independientes de él — pregunta que el presente estudio no puede responder pero que constituye su extensión natural más relevante.

Una segunda línea promisoria es la extensión temporal del análisis a ediciones olímpicas anteriores, lo que permitiría evaluar si los patrones de sesgo semántico han evolucionado o se han mantenido estables a lo largo del tiempo, en línea con el análisis longitudinal de Cooky et al. (2015).

Tercero, sería valioso desarrollar o adaptar un lexicón sensoriomotor en español, eliminando la necesidad de traducción y permitiendo un análisis más directo y culturalmente situado de las narrativas deportivas hispanohablantes. Esto resolvería simultáneamente la limitación de pérdida semántica en la traducción y la pérdida del género gramatical del español.

Finalmente, la metodología desarrollada en este estudio constituye un marco replicable que podría aplicarse a otros eventos deportivos de alta visibilidad, como los Juegos Paralímpicos, los Mundiales de fútbol o los Juegos Panamericanos, ampliando la evidencia empírica sobre los patrones narrativos del periodismo deportivo en español.

Anexo A — Diccionario de disciplinas olímpicas París 2024

El siguiente diccionario corresponde a las disciplinas del programa oficial de los Juegos Olímpicos de París 2024 (COI, 2024), utilizado para la detección automática de la disciplina deportiva en cada noticia del corpus:

Atletismo, Badminton, Baloncesto, Balonmano, Boxeo, Breaking, Ciclismo, Clavados, Escalada deportiva, Esgrima, Fútbol, Gimnasia artística, Gimnasia rítmica, Golf, Halterofilia, Hockey sobre césped, Hípica, Judo, Lucha, Maratón en aguas abiertas, Natación, Natación artística, Pentatlón moderno, Remo, Rugby, Skateboarding, Surf, Taekwondo, Tenis, Tenis de mesa, Tiro, Tiro con arco, Triatlón, Vela, Voleibol, Voleibol de playa, Waterpolo.

Anexo B — Lista de 218 palabras gender-specific (Bolukbasi et al., 2016)

La siguiente lista corresponde al Apéndice F de Bolukbasi et al. (2016), utilizada para construir los vectores de referencia VEC_FEM y VEC_MAS:

actress, actresses, aunt, aunts, bachelor, ballerina, barbershop, baritone, beard, beards, boyfriend, boyfriends, bride, brides, brother, brotherhood, brotherly, brothers, bull, businessman, businessmen, businesswoman, businesswomen, chairman, chairwoman, chick, chicks, child, childbirth, childcare, childhood, childless, childlike, children, convent, countess, dad, daddy, dads, daughter, daughters, duchess, duke, estrogen, ex-boyfriend, ex-girlfriend, fatherhood, fathers, fella, fellas, female, females, feminism, feminist, feminists, fiancee, fraternal, fraternities, fraternity, gal, gals, gentleman, gentlemen, girlfriend, girlfriends, god, goddess, godfather, godmother, granddaughter, grandfather, grandmother, grandson, guy, guys, he, her, hers, herself, him, himself, his, housewife, housewives, husband, husbands, king, kings, knight, knights, lad, ladies, lady, lord, lords, ma, ma'am, macho, maid, maids, mama, man, manhood, mankind, manly, mary, masculine, men, menopause, mistress, mom, mommy, moms, monastery, monk, monks, mother, motherhood, mothers, mr, mrs, ms, nephew, nephews, niece, nieces, nun, nuns, ovarian, paternity, penis, priest, priests, prince, princess, prostate, queen, queens, she, sir, sister, sisterhood, sisterly, sisters, son, sons, sorority, testosterone, uncle, uncles, uterus, vagina, waitress, waitresses, widow, widower, widowers, widows, wife, wives, witch, witches, wizard, woman, womanhood, womanly, women, brotherhood, fatherly, motherly, boyish, girlish, boyhood, girlhood, stepfather, stepmother, stepbrother, stepsister, stepson, stepdaughter, policeman, policewoman, fireman, firewoman, congressman, congresswoman, chairman, chairwoman, spokesman, spokeswoman, cameraman, camerawoman, craftsman, craftswoman, sportsmanship, workman, workwoman.

Referencias

Referencias citadas directamente en el texto

- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.
- Bruce, T. (2015). Assessing the sociology of sport: On media and representations of sportswomen. *International Review for the Sociology of Sport, 50*(4–5), 380–384.

- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
- Comité Olímpico Internacional. (2024). *Programa deportivo de los Juegos Olímpicos de París 2024*. https://olympics.com
- Cooky, C., Messner, M. A., & Musto, M. (2015). "It's dude time!": A quarter century of excluding women's sports in televised news and highlight shows. *Communication & Sport, 3*(3), 261–287.
- Grand, G., Blank, I., Pereira, F., & Fedorenko, E. (2018). Semantic projection: Recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.
- Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2018). *Multivariate data analysis* (8th ed.). Cengage Learning.
- Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing* (3rd ed.). Pearson.
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*, 1271–1291.
- ONU Mujeres. (2024). *El deporte como herramienta para la igualdad de género*. Oficina Regional para América Latina y el Caribe. https://lac.unwomen.org
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.
- Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks. *arXiv preprint arXiv:2106.09462*.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Referencias de consulta

- Arshamian, A., Majid, A., & Speed, L. J. (2022). Senses and emotion are frequently connected by language and semantics. Journal of Multisensory Research, 38, 114–130.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29, 4349–4357.
- Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. Journal of Personality and Social Psychology, 46(4), 735–754.
- Elling, A., & Knoppers, A. (2005). Sport, gender, and ethnicity: Practices of symbolic inclusion/exclusion. Journal of Youth and Adolescence, 34(3), 257–268.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.
- Koca, C. (2009). Gender interaction in coed physical education: A study in Turkey. Adolescence, 44(173), 165–185.
- Lewis, M., & Lupyan, G. (2020). Gender bias in word embeddings: A multidisciplinary analysis. Journal of Machine Learning Research.
- Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.
- ResearchGate. (2018). Schematic illustration of semantic projection. Word vectors in the category animals [Imagen 1]. Recuperado de

https://www.researchgate.net/figure/Schematic-illustration-of-semantic-projection-Word-vectors-in-the-category-animals_fig1_322950010

- Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.
- Schultz, J. (2005). Reading the catsuit: Serena Williams and the production of blackness at the 2002 US Open. Journal of Sport and Social Issues, 29(3), 338–357.
- Spence, C. (2020). Crossmodal correspondences: A tutorial review. Attention, Perception, & Psychophysics, 82, 451–472.
- Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Multisensory integration and the society for neuroscience: Then, now, and the future. Journal of Neuroscience, 34(46), 15188-15193.
- Strategic Platform. (n.d.). Proyección de los vectores con Word2Vec [Imagen 3]. Recuperado de https://strategicplatform.com/articulos/natural-language-processing
- Trail, G. T., & James, J. D. (2015). Sport consumer behavior. Fitness Information Technology.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141-188.
- Velasco, C., Salgado-Montejo, A., Marmolejo-Ramos, F., & Spence, C. (2016). On the correlation between taste and shape features. Food Quality and Preference, 52, 17-25.
- Walker, P. (2016). Cross-sensory correspondences: A theoretical framework and their relevance to music. Psychomusicology: Music, Mind, and Brain, 26(2), 103-116.
- Weinberg, R. S., & Gould, D. (2014). Foundations of Sport and Exercise Psychology. Human Kinet
- XLSTAT. (n.d.). Semantic differential chart: Excel tutorial [Imagen 4]. Recuperado de https://help.xlstat.com/6488-semantic-differential-chart-excel-tutorial