enunciación

Cómo citar: Campo-Paredes, A., y Cataño, G. (2026). Frecuencia de uso de la forma de tratamiento marica en Twitter (X). *Enunciación*, *31*(1), e24252. https://doi.org/10.14483/22486798.24252

Artículo de investigación. Lenguaje, comunicación y cultura digital

Frecuencia de uso de la forma de tratamiento *marica* en Twitter (X)

Frequency of use of the nominal address form *marica* on Twitter (X)

Frequência de uso da forma de tratamento *marica* em Twitter (X)

Andrés Felipe Campo-Paredes[1] y Gloria Marcela Cataño García[2]

Highlights

- Las redes digitales son laboratorios lingüísticos en tiempo real.
- La FTN marica en los últimos veinte años no se adhiere a los modelos tradicionales de difusión lingüística basados en la proximidad geográfica.
- La migración y la conectividad en la red tienen prioridad sobre la mera proximidad de los territorios.
- La normalización por millón es clave en corpus heterogéneos digitales.
- El Big data y métodos computacionales pueden ser útiles para identificar variaciones diatópicas.

Resumen

Las formas de tratamiento nominal (FTN) son uno de los mayores campos de estudio dentro de la sociolingüística variacionista. Estas son conceptualizadas como elecciones léxicas que los hablantes utilizan para establecer relaciones sociales dentro del acto de habla. El siguiente artículo presenta el análisis de la frecuencia y la distribución geográfica de la forma de tratamiento nominal *marica* en un corpus panhispánico extraído de la plataforma de red social Twitter (ahora X). Para ello se utilizó un corpus de datos masivo (*big data*), compuesto por 122.356.443 tuits en español publicados entre 2009 y

[1] Universidad del Valle. Correo electrónico: andres.campo@correounivalle.edu.co
[2] Universidad del Valle. Correo electrónico: gloria.catano@correounivalle.edu.co

Como citar: Campo-Paredes, A., y Cataño, G. (2026). Frecuencia de uso de la forma de tratamiento marica en Twitter (X). *Enunciación*, *31*(1), e24252. https://doi.org/10.14483/22486798.24252

Fecha de postulación: 17 de octubre de 2025; fecha de aceptación: 16 de abril de 2026

2016, en 188 ciudades de 10 países hispanohablantes diferentes. Debido a la sobrerrepresentación de uso de la FTN, Colombia queda por fuera del análisis. La metodología empleó un diseño cuantitativo que combinaba enfoques de lingüística de corpus con análisis de variación sociopragmática digital. Los datos se procesaron utilizando librerías del lenguaje de programación Python. Los resultados identificaron un total de 8.567 apariciones de *marica*. Venezuela fue el epicentro, seguido por Estados Unidos, lo que demuestra tendencias de uso que trascienden la proximidad geográfica a Colombia. Los resultados indican que la circulación léxica en entornos digitales podría seguir patrones de conectividad de redes y flujos migratorios más que la proximidad geográfica, aunque los datos disponibles no permiten confirmarlo de manera definitiva. Se concluye que el análisis de frecuencias normalizadas es metodológicamente necesario para corpus heterogéneos y que las redes digitales pueden ser laboratorios de observación de la variación diatópica de las FTN en tiempo real y a nivel panhispánico.

Palabras clave: marica; dialectología; Twitter; lingüística de corpus; formas de tratamiento nominal

Abstract

Nominal forms of address (NFAs) are one of the major fields of study within variationist sociolinguistics. They are conceptualized as lexical choices that speakers use to establish social relationships within the speech act. The following article presents an analysis of the frequency and geographic distribution of the nominal form of address marica in a pan-Hispanic corpus extracted from the social media platform Twitter (now X). To this end, a big data corpus was used, comprising 122,356,443 Spanish-language tweets published between 2009 and 2016 in 188 cities across 10 different Spanish-speaking countries. Due to the overrepresentation of the NTF's use, Colombia was excluded from the analysis. The methodology employed a quantitative design that combined corpus linguistics approaches with digital sociopragmatic variation analysis. The data were processed using Python programming language libraries. The results identified a total of 8,567 occurrences of marica. Venezuela was the epicenter, followed by the United States, demonstrating usage trends that transcend geographical proximity to Colombia. The

results suggest that lexical circulation in digital environments may follow patterns of network connectivity and migratory flows rather than geographical proximity, although the available data do not allow for a definitive confirmation of this. It is concluded that normalized frequency analysis is methodologically necessary for heterogeneous corpora and that digital networks can serve as laboratories for observing the diatopic variation of FTNs in real time and on a pan-Hispanic scale.

Keywords: *marica*; dialectology; Twitter; corpus linguistics; nominal address form

Resumo

As formas de tratamento nominal (FTN) constituem um dos principais campos de estudo da sociolinguística variacional. Elas são conceituadas como escolhas léxicas que os falantes utilizam para estabelecer relações sociais no ato de fala. O artigo a seguir apresenta a análise da frequência e da distribuição geográfica da forma de tratamento nominal "marica" em um corpus pan-hispânico extraído da plataforma de rede social Twitter (agora X). Para isso, foi utilizado um corpus de big data, composto por 122.356.443 tuítes em espanhol publicados entre 2009 e 2016, em 188 cidades de 10 países de língua espanhola diferentes. Devido à super-representação do uso da FTN, a Colômbia ficou de fora da análise. A metodologia empregou um desenho quantitativo que combinava abordagens da linguística de corpus com análise de variação sociopragmática digital. Os dados foram processados utilizando bibliotecas da linguagem de programação Python. Os resultados identificaram um total de 8.567 ocorrências de marica. A Venezuela foi o epicentro, seguida pelos Estados Unidos, o que demonstra tendências de uso que transcendem a proximidade geográfica com a Colômbia. Os resultados indicam que a circulação lexical em ambientes digitais poderia seguir padrões de conectividade de redes e fluxos migratórios, em vez da proximidade geográfica, embora os dados disponíveis não permitam confirmar isso de forma definitiva. Conclui-se que a análise de frequências normalizadas é metodologicamente necessária para corpora heterogêneos e que as redes digitais podem servir como laboratórios de observação da variação diatópica das FTN em tempo real e em nível pan-hispânico.
Palavras-chave: *marica*; dialetologia; Twitter; linguística de corpus; formas de tratamento nominal

Introducción

Las formas de tratamiento nominal (FTN) son elecciones léxicas y pronominales que los hablantes emplean para interactuar en distintos contextos sociales y de identidad (Brown y Levinson, 1987). Las formas de tratamiento pueden expresar, según Brown y Levinson (1987) y Hummel et al. (2010), solidaridad, distancia e incluso intimidad entre los hablantes. Esto muestra cómo el lenguaje manifiesta y perpetúa normas sociales y culturales más amplias. En español, las formas nominales de tratamiento muestran las maneras en que el léxico adquiere sentido social a través de marcadores lingüísticos y elementos pragmáticos. El estudio de estas formas tiene además una dimensión dialectológica cuando se analiza su variación y distribución entre distintas comunidades de habla hispanohablantes, lo que permite identificar una tendencia de uso diferencial según la variedad geográfica (Moreno Fernández, 2009).

El objetivo de este artículo es analizar la frecuencia de uso y la distribución geográfica de la FTN *marica* en un corpus panhispánico extraído de la red social Twitter (ahora X). Este estudio combina medidas de frecuencia y dispersión con la elaboración de mapas geolingüísticos para analizar cómo *marica* funciona como una forma de tratamiento nominal en contextos digitales. Si bien la FTN *marica* se ha documentado en el español colombiano como marcador de camaradería y anticortesía, entendida esta como recurso de afiliación identitaria entre pares, especialmente jóvenes (Carantón González, 2024; Zimmermann, 2005), su uso trasciende esta variedad, puesto que ha sido estudiada también en el marco del colectivo LGTBI+ hispanohablante (Navarro-Carrascosa, 2021, 2023) y en otras variedades peninsulares y latinoamericanas. El estudio no presupone que Colombia sea el centro de difusión de *marica* hacia otras variedades, sino que adopta una perspectiva pluricéntrica que reconoce trayectorias históricas y sociopragmáticas propias en cada comunidad de habla. Así, su circulación transnacional, frecuencia y dispersión pueden variar entre los distintos países hispanohablantes.

Marco teórico

El marco teórico de este estudio se sitúa en la intersección de la sociolingüística variacionista, la pragmática y la lingüística de corpus, con el propósito de analizar la frecuencia de uso de la Forma de Tratamiento Nominal (FTN) *marica* en la red social

Twitter (ahora X) como un fenómeno de variación diatópica en la comunidad hispanohablante.

Para esto, se aborda, en primer lugar, la teoría de las formas de tratamiento y su estudio en el campo de la sociolingüística. En segundo lugar, se presenta un estado de la cuestión en relación con el término *marica* en diferentes variedades del español, especialmente en su evolución semántico-pragmática. En tercer lugar, se explora la relación entre el discurso digital y la construcción de identidades. Finalmente, se delimitan las herramientas metodológicas de la lingüística de corpus que facilitan el análisis cuantitativo de la frecuencia y dispersión de este elemento léxico.

Sociolingüística y formas de tratamiento nominal

Las formas de tratamiento han sido un campo de estudio fundamental dentro de la sociolingüística, ya que constituyen un mecanismo central para la gestión de las relaciones interpersonales. Autores clásicos como Brown y Levinson (1987) las enmarcan dentro de la teoría de la cortesía y las conciben como estrategias para mitigar actos que amenazan a la imagen pública. Estas formas, ya sean pronominales como *tú*, *usted*, *vos* o nominales como *señor*, *amigos* o *doctor*, son entendidas, como lo expresa Gholami (2021), como elecciones léxicas y gramaticales que los hablantes utilizan para establecer relaciones sociales, interpretar y negociar jerarquías y construir identidades.

Desde la perspectiva variacionista, Hummel et al. (2010) destacan que el estudio de las FTN permite explorar cómo factores como la edad, el género, la clase social y el contexto comunicativo inciden en su elección y uso. En este sentido, las FTN son recursos pragmáticos que reflejan las normas culturales y dinámicas de solidaridad, distancia o respeto. No obstante, como lo señala la *Nueva Gramática de la Lengua Española* (RAE, 2009), el registro de las FTN es no solo amplio y abierto, sino heterogéneo. Por lo tanto, es necesario justificar la extensión del concepto a términos que, como *marica*, no son consustancialmente relacionales, como el caso de *señor* o *doctor*, sino que adquieren funciones de tratamiento en el discurso o los actos de habla, particularmente en aquellos cuyo contexto es informal y entre pares, como lo expone Bravo (2009). Esta investigación adopta una perspectiva amplia, considerando *marica* como una FTN que, en contextos

específicos sociopragmáticos, funciona como un marcador de afiliación, solidaridad o, en otros casos, de anticortesía.

El caso de *marica* en el español: evaluación semántico-pragmática y variación diatópica

La FTN *marica* presenta una evolución semántico-pragmática compleja que la convierte en un caso ejemplar para el estudio de la variación lingüística. Desde una perspectiva etimológica, Corominas y Pascual (1980) documentan el término *marica* y sus formas derivadas con el sentido de 'afeminado', asociado originalmente a una connotación peyorativa. Este significado despectivo y homofóbico persiste en múltiples contextos, pero coexiste con resignificaciones que, como lo expone Navarro-Carrascosa (2021, 2023), desde la lingüística *queer* se han analizado como procesos de reapropiación política.

Investigaciones previas como la de Méndez Vallejo (2014) han dado cuenta de esta polifuncionalidad en distintas variedades del español colombiano. Por su parte, Rincón Martínez (2021) muestra cómo *marica* se desempeña como un marcador de camaradería, solidaridad y humor entre pares, con el uso limitado como insulto en ciertos contextos de comunicación. Este proceso se entiende a través del concepto de *anticortesía* de Zimmermann (2005), que se refiere a formas abstractas de interacción donde los participantes utilizan insultos, apodos despectivos y expletivos para construir un sentido de afiliación e identidad grupal, especialmente en círculos de hombres jóvenes. A diferencia de la descortesía convencional, cuya función es dañar la imagen del interlocutor, la anticortesía funciona en un marco de complicidad entre pares, donde el uso de términos aparentemente ofensivos refuerza la pertenencia al grupo y la confianza mutua. En este sentido, el término *marica* no es un insulto en todos los contextos, lo que explica su uso en la comunicación digital. Estudios similares en Venezuela, como el de Gutiérrez-Rivas (2016), explican aún más el fenómeno y sugieren no solo un marco empírico compartido para el español caribeño, sino que también identifican los patrones sociales asociados con el uso del término en los registros informales del lenguaje. En Estados Unidos, Cashman (2017), Engra Minaya (2024) y Navarro-Carrascosa (2021, 2023) estudian el término y observan que, si bien en ciertos contextos el término *marica*

es insultante, existen contextos donde el término fomenta la solidaridad en grupo, particularmente dentro de la comunidad LGBTQ+ y en el discurso informal juvenil.

Sin embargo, el fenómeno de la resignificación de términos de tratamiento no es exclusivo de *marica*; se inscribe en un marco más amplio del español en el que los vocativos, inicialmente con función de insulto o términos despectivos, pasan a funcionar como marcadores de solidaridad entre pares. En este sentido, Bravo (2009) presenta un marco analítico que permite comprender cómo las prácticas de cortesía y descortesía están ancladas en las particularidades socioculturales de cada unidad de habla. Este marco resulta relevante para el análisis de FTN como marica, que tiene un valor pragmático peyorativo, afiliativo o anticortés y solo puede determinarse a partir del contexto relacional y sociocultural en el que se produce su uso.

Discurso digital, pragmática y construcción de identidades

Los espacios digitales han actuado como catalizadores de los procesos de resignificación semántica y pragmática. Plataformas como X (antes Twitter) y Facebook proveen condiciones de visibilidad, repetición e inmediatez transnacional que permiten la difusión y reinterpretación de las FTN, así como lo plantean en sus investigaciones Tagg et al. (2022) y Jebaselvi et al. (2023). En este entorno digital, el lenguaje no es solo un vehículo de comunicación, sino un medio central para la construcción activa de identidades y afiliaciones comunitarias.

Como lo señalan Page (2012) y Amer (2024), las redes sociales componen espacios de afiliación donde los usuarios negocian su pertenencia a grupos a través de prácticas lingüísticas compartidas. Estas comunidades en línea crean dinámicas de grupo que favorecen la adopción de formas lingüísticas específicas, entre ellas las FTN, como marcadores de diferenciación frente a otros grupos.

Desde el punto de vista pragmático, las formas de tratamiento como *marica* pueden funcionar como marcadores identitarios, recursos de conexión afectiva o herramientas humorísticas que implican pertenencia a determinadas comunidades y colectivos. Como lo exponen Arrieta y Avendaño (2018), en la digitalidad, el uso de *marica* puede manifestarse como un indicador de juventud, pertenencia a comunidades urbanas o, incluso, de afinidad a determinadas identidades. Esto último muestra que esta FTN trasciende su significado original peyorativo y se incorpora con otros significados a

las distintas comunidades de hablantes. Sin embargo, no se deja de reconocer su naturaleza polifuncional, ya que puede operar como insulto homofóbico o como marcador de solidaridad o camaradería. Esto implica que su interpretación depende en gran medida del contexto interaccional y de la relación entre los participantes.

Perspectivas cuantitativas para el estudio de la variación: frecuencia y dispersión

A continuación, se presentan los fundamentos teóricos y conceptuales que sustentan el enfoque cuantitativo adoptado para esta investigación. La conceptualización y diferenciación entre términos como *frecuencia*, *dispersión* y *normalización* no se limita solo al conjunto de técnicas de recolección y análisis de datos, sino que responde a principios epistemológicos de la lingüística de corpus y la sociolingüística computacional.

El estudio de las FTN en el discurso digital, de acuerdo con Baker y Egbert (2021) y Gries (2022), se beneficia de una metodología que combine enfoques cualitativos y cuantitativos. Como lo expone Gries (2022), la lingüística de corpus y la sociolingüística computacional ofrecen herramientas para medir la frecuencia y dispersión de un término. Esto permite identificar patrones de uso a gran escala que no serían evidentes mediante el análisis cualitativo de casos aislados.

Conceptos como *frecuencia* y *dispersión* son fundamentales para este estudio. Como expone Gries (2009), el conteo de frecuencia permite identificar qué términos son sobresalientes en un corpus. Las medidas de dispersión, por su parte, muestran si su uso está generalizado en toda la muestra o si, por el contrario, se ha restringido a subgrupos específicos (por ejemplo, a una región geográfica). Dado que el tamaño de los textos en un corpus puede variar, es necesario recurrir a medidas de frecuencia relativa que permitan hacer comparaciones válidas.

La normalización por millón de palabras (*frequency per million words* o *fpmw*) es una de las unidades estandarizadas más utilizadas. Su cálculo consiste en dividir el número de ocurrencias de un ítem por el tamaño total del corpus y multiplicar el resultado por un millón, según la siguiente fórmula:

$$fpmw = (nt / nw) \times 1.000.000$$

donde *nt* es el número de *tokens* (apariciones) del término específico que se analiza y *nw* es el número total de palabras en el corpus completo.

Esta normalización permite contrastar de manera confiable la frecuencia de uso de *marica* entre países con diferente volumen de datos. Si bien esta metodología cuantitativa es pertinente para identificar tendencias de difusión de FTN y contrastar regiones con alta o baja frecuencia, por sí sola no permite evidenciar las funciones pragmáticas. Así, no se puede reconocer, por ejemplo, si el uso de cada ocurrencia es peyorativo o afiliativo, lo que es una limitación que se considera en la interpretación de los resultados.

Metodología

Diseño y tipo de estudio

El estudio se llevó a cabo utilizando un diseño metodológico cuantitativo (Creswell, 2014), con enfoque descriptivo interpretativo. Este diseño permitió a los investigadores medir y comparar la frecuencia de uso del tratamiento nominal de *marica* en diferentes comunidades hispanohablantes. Asimismo, se trata de un análisis sincrónico, ya que la información recogida responde a un periodo delimitado.

Corpus del estudio

El corpus utilizado en esta investigación proviene de un corpus de datos masivos extraído de la red social Twitter (ahora X). El corpus, construido por Jiménez et al. (2018), está compuesto por tuits publicados en 333 ciudades de 21 países de habla hispana, producidos entre 2009 y 2016. El criterio para seleccionar las ciudades del país fue que estas contaran con más de 100.000 habitantes. La recolección se realizó a través de la API de red social, estableciendo un radio de 15 millas alrededor de cada ciudad para garantizar la precisión geográfica. De este proceso inicial se constituyeron 247.928.418 tuits. La composición y el refinamiento del corpus, que implicaron la eliminación de retuits, duplicados y otros, se describen en detalle en Jiménez et al. (2018).

La selección de los países del estudio responde a decisiones metodológicas. México, Centroamérica y el Caribe hispanohablante fueron excluidos para mantener la coherencia geolingüística del análisis, que está centrado en Sudamérica como unidad; su inclusión habría implicado ejes de análisis adicionales que no eran coherentes con los objetivos del estudio. Colombia, por su parte, presenta un gran uso de la FTN *marica* (Asqueta Corbellini, 2008; Carantón González, 2024). Por lo tanto, su inclusión en la

investigación habría generado una sobrerrepresentación del término, y habría sesgado el análisis comparativo por su alta frecuencia en el corpus (Cataño García y Campo-Paredes, 2025). Asimismo, la incorporación de España y Estados Unidos responde al interés de contrastar tres grandes sectores de la hispanidad con trayectorias históricas y sociolingüísticas distintas. En el caso de Estados Unidos, a pesar de que el inglés es la lengua oficial del país desde 2025, 41.460.427 personas mayores de 5 años usaban el español en el entorno doméstico en 2018 (Instituto Cervantes, 2020). Este alto número de hablantes justifica su inclusión como país de análisis. Al final, la filtración de datos arrojó 122.356.443 tuits en español, de 188 ciudades de 10 países hispanohablantes, a saber: Argentina, Bolivia, Chile, Ecuador, EE. UU., España, Paraguay, Perú, Uruguay y Venezuela. Estos 122.356.443 tuits son la muestra de este estudio. La tabla 1 describe los datos cuantitativos del corpus.

Tabla 1. *Descripción cuantitativa del corpus de esta investigación*

País	**Número de ciudades**	**Número de tuits extraídos**
Argentina	26	26.933.107
Bolivia	8	289.683
Chile	24	15.291.490
Ecuador	10	4.350.877
EE. UU.	35	6.172.521
España	36	43.766.154
Paraguay	6	1.230.288
Perú	14	3.296.368
Uruguay	7	4.252.022
Venezuela	22	16.773.933
Total	**188**	**122.356.443**

Nota. Elaboración propia

Procesamiento y análisis de datos

A continuación, se presenta el paso a paso del procesamiento de los datos:

1. Los datos principales fueron archivos TSV (Tab-Separated Values). Para facilitar su manejo, estos fueron convertidos y agrupados en una hoja de cálculo de Excel.
2. Una vez organizados, los datos se procesaron con el lenguaje de programación Python versión 3.14.0. Se utilizó este lenguaje debido a su gran capacidad para manejar un corpus de *big data* de redes sociales, tal como lo manifiestan autores como Seyidova y Shakhayev (2023), Moreno-Ortiz y García-Gámez (2023) y Surya et al. (2020). En este *software* se hizo uso de librerías como Pandas, NLTK, Matplotlib y Seaborn, lo que permitió la tokenización, normalización y visualización de frecuencias.
3. Depuración del corpus: se eliminaron mensajes en otros idiomas o cadenas de caracteres sin sentido. A su vez, se implementaron filtros que permitieran guardar solo registros con contenido del tuit y metadatos geográficos.
4. Para organizar esta información, se implementó una tríada de *scripts* o "códigos", descritos a continuación: Código 1: Clasificación general por país y ciudad. Código 2: conteo de ocurrencias léxicas. Código 3: visualización gráfica. Para ello, se utilizaron las librerías Matplotlib y Seaborn.

Categorías de análisis

La búsqueda se restringió a ocurrencias exactas de *marica*, insensibles a mayúsculas y minúsculas y con delimitadores de palabra para evitar falsos positivos. Formas relacionadas como *marico*, *maricón* o variantes escriturales como *marik* quedaron fuera del análisis por constituir unidades léxicas distintas con trayectorias pragmáticas propias, cuyo estudio desborda los objetivos de este trabajo. El análisis de los datos se llevó a cabo en función de las siguientes dos categorías establecidas previamente:

a. Frecuencia de uso: se hizo la cuantificación absoluta de la forma de tratamiento nominal *marica* en los países del corpus. Los datos se normalizaron por frecuencia por millón.
b. Distribución geográfica (variación diatópica): se llevó a cabo la comparación y cuantificación de usos entre países panhispánicos desde una perspectiva

dialectológica, entendiendo la variación diatópica como eje central del análisis. A partir de la distribución geográfica, se elaboró un mapa geolingüístico por medio de la herramienta ArcGIS.

Validez y confiabilidad

La validez se aseguró mediante la representatividad lingüística y geográfica del conjunto de datos. El corpus se construyó a partir de un corpus panhispánico principal de *big data* (Jiménez et al., 2018) que contiene más de 120 millones de tuits. Los criterios de inclusión (tuis escritos en español, de cuentas públicas y geolocalizados dentro de un radio de 15 millas de cada ciudad) representan el discurso digital auténtico de varias ubicaciones. Además, las frecuencias normalizadas por millón de palabras aplicadas al corpus permitieron la comparabilidad y disminuyeron los sesgos de muestreo existentes.

La confiabilidad, en segundo lugar, fue alcanzable mediante la implementación de métodos computacionales estandarizados y reproducibles en el lenguaje de programación Python. Las fases de limpieza de datos, tokenización y conteo de frecuencias fueron automatizadas con las librerías *Pandas*, *NLTK* y *Matplotlib*, *librerías* útiles para el análisis de lenguaje computacional, lo que redujo el margen de error humano, como indica Batta (2024). Por último, el uso de un corpus validado previamente por Jiménez et al. (2018) y la posibilidad de que distintos investigadores puedan replicar los criterios de filtrado que se usaron en este estudio permiten asegurar la reproducibilidad del estudio.

Consideraciones éticas

Esta investigación fue de carácter lingüístico y sociocultural, y no estaba orientada a la evaluación de individuos, por lo que fue eximido el consentimiento informado formal. Sin embargo, el estudio se desarrolló conforme a los principios éticos de la investigación, especialmente aquellos relacionados con la protección de la privacidad y el uso responsable de datos en entornos digitales, según el marco de habeas data (Ley 1581 de 2012 de Colombia). La información analizada proviene de publicaciones disponibles de manera pública en Twitter (ahora X), y a su vez, no se socializan datos sensibles ni se usan para fines distintos a la investigación, sin llevar a cabo una caracterización individual ni contacto directo con los usuarios autores de los tuits. Además, se siguieron las

recomendaciones de la Association of Internet Researchers (AoIR, 2019) y de comités éticos en investigación digital.

Resultados y análisis

El análisis de la FTN *marica* se organizó en dos dimensiones: frecuencia de uso y distribución geográfica. La primera dimensión estableció la magnitud del uso mediante una cuantificación tanto absoluta como normalizada; la segunda brindó evidencia de los patrones de dispersión y concentración en diferentes países de habla hispana, situando la forma dentro del espectro geolingüístico. Dada la gran variación en el tamaño de los corpus entre países (que oscila entre 289.683 tuits en Bolivia y 43.766.154 en España), la normalización de la frecuencia por millón de palabras fue necesaria para llevar a cabo comparaciones estadísticas más reales entre los países del corpus (Gries, 2020; Paquot y Bestgen, 2023). Este enfoque metodológico se ajusta a las prácticas actuales en lingüística de corpus, que enfatizan las frecuencias normalizadas como estándar para la comparación entre corpus (Egbert y Biber, 2019; Weisser, 2023). A continuación, se presentan los resultados.

Frecuencia de uso de *marica*

El primer eje de análisis corresponde a la frecuencia de uso del término *marica* en el corpus. Se calcularon tanto las frecuencias absolutas como las frecuencias normalizadas para el término en todos los países del corpus. Dada la considerable variación en el tamaño de los corpus, la frecuencia normalizada brinda una medida más precisa de la intensidad de uso, ya que controla las diferencias en el tamaño de la muestra y permite una comparación directa entre países (Brezina, 2018; Cantos, 2013 y Almela-Sánchez, 2018). Las frecuencias, organizadas de mayor a menor tasa normalizada, se pueden ver en la tabla 2.

Tabla 2. *Frecuencia de la forma de tratamiento "marica" por país*

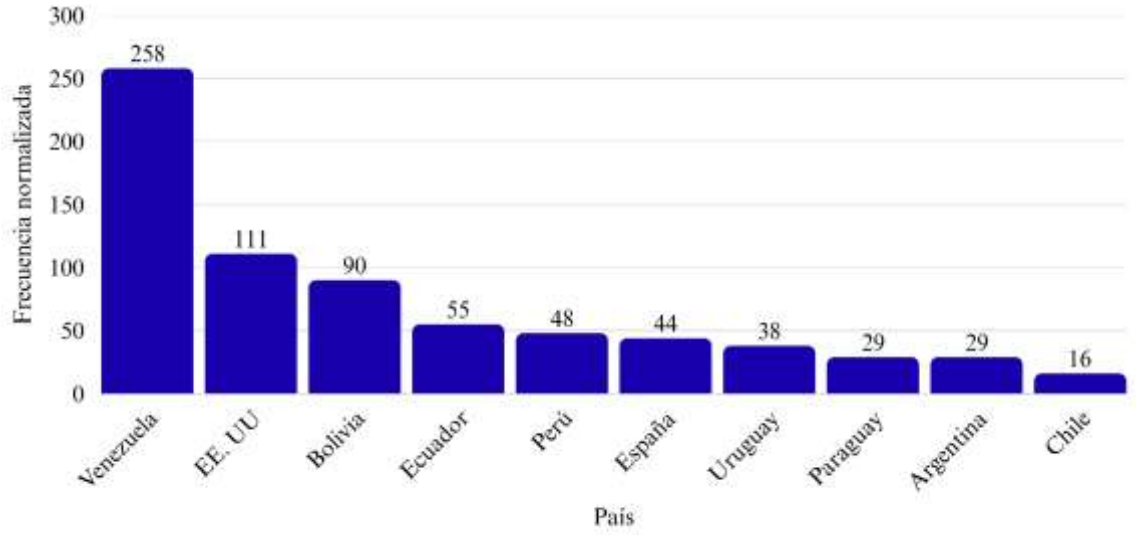
País	Número de tuits extraídos	Apariciones de *marica*	Frecuencia por millón	Representación en la muestra
Venezuela	16.773.933	4.320	257,54	50,40%
EE. UU.	6.172.521	687	111,3	8,00%

Bolivia	289.683	26	89,75	0,30%
Ecuador	4.350.877	241	55,39	2,80%
Perú	3.296.368	158	47,93	1,80%
España	43.766.154	1.920	43,87	22,40%
Uruguay	4.252.022	161	37,86	1,90%
Paraguay	1.230.288	36	29,26	0,40%
Argentina	26.933.107	773	28,7	9,00%
Chile	15.291.490	245	16,02	2,90%
Total	**122.356.443**	**8.567**	**70,02**	**100,00%**

Nota. Elaboración propia.

La tabla 2 presenta las frecuencias absolutas y normalizadas de la forma de tratamiento *marica* en el corpus panhispánico. En total, se identificaron 8.567 ocurrencias léxicas, distribuidas de manera desigual entre los países analizados. La frecuencia normalizada muestra tendencias diferentes en comparación con las que sugieren los recuentos absolutos por sí solos, lo que demuestra la importancia de la normalización del tamaño del corpus en el análisis sociolingüístico comparativo (Brezina, 2018; Paquot y Bestgen, 2023). Asimismo, investigaciones recientes sobre las redes sociales en español demuestran que las frecuencias normalizadas son necesarias para identificar tendencias de variación frente a artefactos de muestreo (Gonçalves et al., 2018; Huang et al., 2020).

Figura 1. *Distribución de las frecuencias normalizadas por millón*



Nota. Elaboración propia. Los valores representados corresponden a frecuencias normalizadas redondeadas al entero más cercano.

Como se puede observar, Venezuela es el país con mayor uso de la FTN, con 257,54 ocurrencias por millón de tuits, más del doble que cualquier otro país. Esto representa un gran uso que no se le puede atribuir simplemente al tamaño del corpus, ya que la contribución de Venezuela (16 773 933 tuits, el 13,7 % del corpus) es moderada en comparación con las muestras más grandes de España y Argentina. No obstante, esta interpretación aplica con mayor fuerza a corpus de tamaño moderado o grande. En casos como Bolivia, con apenas el 0,2 % del corpus, la frecuencia normalizada alta debe interpretarse con cautela, ya que puede reflejar sobrerrepresentación estadística más que un uso generalizado.

Estados Unidos ocupa la segunda posición con 111,30 por millón, un resultado de relevancia dado el contexto nacional anglófono, pues, a pesar de que el inglés es la lengua oficial de ese territorio desde marzo de 2025, son muchos los hispanohablantes que residen en el país (Instituto Cervantes, 2020). Esta tasa, procedente de un corpus de 6.172.521 tuits (el 5,0 % del total), podría indicar un uso muy concentrado dentro de las comunidades diaspóricas de habla hispana, donde *marica* puede desempeñar funciones de marcado de identidad en el discurso mediado digitalmente, aunque sin análisis cualitativo del corpus es imposible determinar si predominan los usos afiliativos, peyorativos o de reapropiación identitaria. Esta limitación se discute en §4.4.

En relación con Bolivia, no se esperaba la frecuencia normalizada de 89,75 por millón. A pesar de tener solamente 26 ocurrencias y de sumar 289.683 tuits (0,2% del corpus), los usuarios bolivianos de Twitter (ahora X) parecen hacer más uso de la FTN *marica* que la mayoría de los países. No obstante, las 26 ocurrencias absolutas han de tomarse con precaución. Si bien podrían mostrar un auténtico uso concentrado, un fenómeno común en una comunidad digital que no tiene una extensión muy amplia, puede que también haya una sobrerrepresentación estadística como resultado de que la muestra es muy pequeña.

Ambas interpretaciones son plausibles y no mutuamente excluyentes. Entre las hipótesis que podrían explicar este resultado se encuentran la proximidad geográfica con países de alta frecuencia como Colombia, Venezuela y Ecuador, posibles intercambios migratorios regionales o la concentración del uso de *marica* entre determinados segmentos demográficos activos en Twitter (ahora X) durante el período analizado. Cabe señalar que el español boliviano se ha caracterizado por rasgos distintivos influenciados por el contacto con el quechua y el aimara (Lipski, 2018), lo que hace aún más necesario explorar estas hipótesis en estudios futuros con análisis cualitativo.

Un grupo intermedio con tasas de uso moderadas incluye a Ecuador (55,39 por millón), Perú (47,93 por millón), España (43,87 por millón) y Uruguay (37,86 por millón). El caso de España llama la atención, porque, aunque cuenta con el mayor volumen de datos de un solo país (43.766.154 tuits, el 35,8% del total) y ocupa el segundo lugar en frecuencia absoluta (1.920 ocurrencias, el 22,4% del total), su tasa normalizada la posiciona en sexto lugar en todo el corpus. El dominio de España en términos absolutos proviene principalmente de su gran tamaño de muestra y no del uso absoluto del corpus. Las 1.920 ocurrencias del corpus de España provienen de un corpus 2,6 veces más grande que el de Venezuela; sin embargo, los usuarios de este país emplean *marica* con 5,9 veces más frecuencia que en España. Resultados de esta naturaleza son útiles para analizar la dispersión transnacional de la terminología en contextos digitales, en consonancia con trabajos publicados recientemente que afirman la reconfiguración de las redes sociales y las fronteras geográficas de la diversidad lingüística (Grieve et al., 2019).

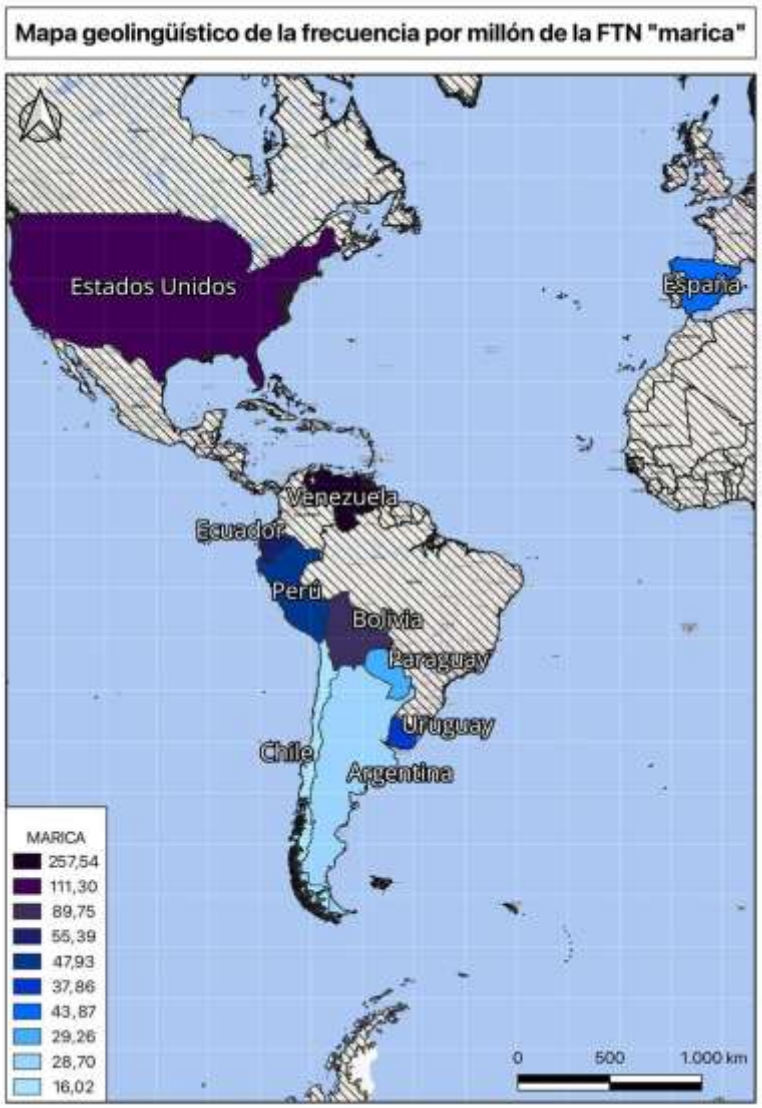
Por último, el grupo con las frecuencias más bajas está compuesto por Paraguay (29,26 por millón), Argentina (28,70 por millón) y Chile (16,02 por millón). Argentina tiene el segundo mayor volumen de datos (26.933.107 tuits, el 22,0% del total) y ocupa

el tercer lugar en frecuencia absoluta (773 ocurrencias, el 9,0% del total); pero su tasa normalizada la posiciona en noveno lugar entre los diez países. Asimismo, con una tasa de 16,02 por millón, Chile se encuentra en la parte inferior de la distribución y tiene una tasa de uso que es baja para un país que está geográficamente cerca de otros en Sudamérica donde el término se usa con mucha mayor libertad. Trabajos recientes sobre la variación dialectal del español en las redes sociales han identificado patrones similares de adopción diferencial entre regiones vecinas (Moreno Fernández, 2021).

Distribución geográfica

La distribución geográfica de la forma nominal de tratamiento *marica* demuestra que su uso no se concentra en un solo país o región específica, sino que presenta diferentes tendencias en función de los contextos panhispánicos del corpus. A continuación, se brinda una visión geográfica comparativa de Sudamérica, España y Estados Unidos, destacando tanto las zonas de mayor concentración como aquellas con menor presencia.

Figura 2. *Distribución de frecuencias de la FTN* marica *en un mapa geolingüístico*



Nota. Elaboración propia haciendo uso del *software* ArcGIS

El mapa de distribución geográfica muestra un uso desigual de la forma nominal de tratamiento *marica* dentro del corpus. La representación cromática indica los puntos más altos y bajos de frecuencia normalizada, permitiendo así establecer contrastes entre los diez países de habla hispana analizados, distribuidos en tres grandes áreas geográficas, a saber: América del Sur, España y Estados Unidos. Los tonos más oscuros muestran una mayor frecuencia por millón, mientras que los tonos más claros muestran una menor frecuencia. Los casos presentados a continuación son los más destacados derivados del análisis.

Distribución en cuatro niveles: uso muy alto, alto, moderado y bajo

Los resultados muestran cuatro grupos distintos de intensidad de uso, establecidos a partir de la distribución por cuartiles de las frecuencias normalizadas (Q1 = 29,07; mediana = 40,87; Q3 = 72,53). El grupo de uso muy alto (>Q3) está representado por Venezuela (257,54 por millón), Estados Unidos (111,30 por millón) y Bolivia (89,75 por millón), y tiene tasas de uso muy por encima del resto de países. El grupo que presenta un alto uso (Q3–Q2) lo forman Ecuador (55,39 por millón) y Perú (47,93 por millón), cifras notables en relación al número de muestras. El grupo de uso moderado (Q2–Q1) lo componen España (43,87 por millón) y Uruguay (37,86 por millón). El grupo de bajo uso (≤Q1) incluye Paraguay (29,26 por millón), Argentina (28,70 por millón) y Chile (16,02 por millón). Este sistema de cuatro niveles indica diferentes etapas de adopción, siguiendo lo que plantea la teoría de difusión de innovaciones (Rogers, 2003), aunque con unas variaciones respecto a la circulación digital.

De esta manera, la distribución pone en discusión los modelos tradicionales de difusión geográfica, puesto que los países no se distribuyen según un gradiente de distancia con respecto a ningún centro irradiador único. En cambio, la intensidad de uso parece seguir patrones de conectividad de redes y flujos migratorios, más que la proximidad geográfica (Eisenstein, 2018; Tarrade et al., 2024). Es importante señalar que los factores que explican la distribución pueden diferir según el país: mientras que para Venezuela y Ecuador la proximidad geográfica y los intercambios migratorios con Colombia son hipótesis plausibles, para España el uso de *marica* puede corresponder a una trayectoria histórica autóctona que no puede reducirse a influencia externa (Navarro-Carrascosa, 2021, 2023).

Venezuela: uso autóctono e influencia bidireccional

La posición de Venezuela como centro de máximo uso (257,54 por millón, 4.320 ocurrencias absolutas) no puede atribuirse al tamaño del corpus: los 16.773.933 tuits de Venezuela representan solo el 13,7 % del total, menos de la mitad de la contribución de España. La tasa normalizada es 5,9 veces superior a la de España, 9,0 veces superior a la de Argentina y 16,1 veces superior a la de Chile.

El uso venezolano de *marica* no se explica únicamente por la influencia colombiana. El español venezolano tiene su propia tradición de usar FTN como *marico*

(Gutiérrez-Rivas, 2016). Esto sugiere que la alta frecuencia del término puede estar relacionada con las dinámicas internas de la variedad junto con el continuo lingüístico entre Colombia y Venezuela, una consecuencia de proximidades dialectales y cercanía territorial histórica (Auer y Hinskens, 2005; Ramírez y Mendoza, 2020).

Además, algunos estudios sociolingüísticos han evidenciado una influencia lingüística bidireccional del español venezolano, que presenta numerosas características del español colombiano, especialmente entre los jóvenes urbanos (Obediente Sosa, 2019). El período de recolección de datos (2009-2016) coincide con un aumento en los flujos migratorios y en el uso de las redes sociales, lo que favorece que el contexto difunda el uso de la forma de tratamiento nominal (Krogstad et al., 2019).

España: uso con trayectoria histórica

La posición de España (43,87 por millón, 1.920 ocurrencias absolutas, 22,4 % del total) no puede atribuirse a influencia externa. *Marica* es una forma con uso autóctono documentado en el español peninsular desde al menos los años noventa (Navarro-Carrascosa, 2021, 2023; Engra Minaya, 2024), con trayectorias pragmáticas propias que incluyen tanto usos peyorativos como procesos de reapropiación identitaria, especialmente dentro del colectivo LGTBI+.

Una explicación plausible para su frecuencia moderada puede ser el uso de *marica* en España, puesto que esta FTN y derivados como *maricón* pueden percibirse socialmente como despectivos en amplios sectores del país (Engra Minaya, 2024). Este tabú puede limitar el uso de la palabra como un término cariñoso en comparación con cómo otros dialectos usan la palabra como jerga prosocial (Zimmermann, 2005). La posible influencia de la migración colombiana en algunos contextos urbanos de alta concentración no puede descartarse, pero los datos de este estudio no permiten cuantificar ni confirmar tal influencia.

Este resultado va en contra de los modelos tradicionales de difusión basados en la proximidad geográfica. La teoría de redes sociales (Milroy y Milroy, 1985) sugiere que el medio digital puede actuar, en cierta medida, como un reemplazo de la proximidad. En este sentido, el resultado reconoce la existencia de un fenómeno en el espacio digital panhispánico que crea lo que se puede analizar como comunidades lingüísticas

'translocales' (un término popularizado por Blommaert, 2010) que circulan y redescriben elementos de manera 'localmente distante' (Eisenstein, 2018; Tagg y Seargeant, 2021).

Comunidades diaspóricas de Estados Unidos

Las 687 apariciones en Estados Unidos también son llamativos. Los datos de 6.172.521 tuits (el 5,0 % del corpus) producen una frecuencia normalizada de 111,30 por millón, la segunda más alta del corpus a nivel mundial, lo que sugiere tasas de uso más altas que reflejan la dinámica de mantenimiento del idioma en las comunidades diaspóricas.

Las investigaciones sobre el español heredado demuestran que los registros informales, los marcadores de solidaridad y las características lingüísticas vinculadas a la identidad muestran una mayor resistencia que las estructuras gramaticales formales en contextos bilingües (Otheguy y Stern, 2023; Valdés, 2001). Las formas de tratamiento como *marica*, que tienen funciones pragmáticas relacionadas con la solidaridad dentro del grupo y la identidad cultural, pueden ser resistentes al desgaste cuando el español sirve como marcador de identidad etnolingüística (Fishman, 1991; Potowski, 2023).

Esto ejemplifica los "espacios de translingüismo" (García y Wei, 2014), en los que los usuarios bilingües despliegan recursos lingüísticos para negociar identidades complejas. Twitter (ahora X) puede animar a los hablantes nativos a emplear formas dialectales como reclamo de autenticidad, lo que explica la frecuencia relativamente alta observada (Androutsopoulos y Juffermans, 2019; Lee, 2021).

Bolivia y Ecuador: uso intensivo en comunidades más pequeñas

Bolivia (89,75 por millón, 26 ocurrencias absolutas) y Ecuador (55,39 por millón, 241 ocurrencias absolutas) muestran un uso intensivo a pesar de sus contribuciones relativamente pequeñas al corpus. El caso de Bolivia es relevante, puesto que, con solo 289.683 tuits (0,2 % del corpus), los usuarios bolivianos muestran la tercera tasa de uso más alta a nivel mundial. La frecuencia normalizada de Bolivia debe interpretarse con atención dado el reducido tamaño de su muestra. Entre las hipótesis posibles se encuentran la proximidad geográfica con países de alta frecuencia, posibles intercambios migratorios regionales y las características propias del español boliviano, marcado

históricamente por el contacto con el quechua y el aimara (Lipski, 2018). A pesar de ello, ninguna de estas hipótesis puede ser confirmada con los datos recogidos.

El resultado de Ecuador (55,39 por millón) tiene múltiples interpretaciones: bien sea por su cercanía geográfica a Colombia, bien sea por los intercambios históricos migratorios, que pueden ser considerados mecanismos de circulación léxica. Esto es respaldado por trabajos sobre la influencia colombiana en las zonas limítrofes ecuatorianas y en sus grandes ciudades (Haboud y de la Vega, 2021; Jijón, 2022). No obstante, en este caso sería necesario un análisis cualitativo para confirmar si el uso de *marica* en Ecuador responde a funciones afiliativas similares a las documentadas en Colombia.

El Cono Sur

Paraguay (29,26 por millón), Argentina (28,70) y Chile (16,02) forman un grupo distintivo en el extremo inferior de la distribución. La posición de Argentina llama la atención dada su enorme contribución al corpus (26.933.107 tuits, el 22,0 % del total). A pesar de esta muestra sustancial, Argentina solo produce 773 ocurrencias, lo que da una frecuencia normalizada de 28,70, la novena de diez países. Para Argentina, el bajo uso del término *marica* puede explicarse por la existencia de expresiones locales que comparten rasgos de similitud con *marica*. Por ejemplo, el término *boludo*, que podría cumplir funciones pragmáticas similares y está documentado en el español rioplatense (Bravo, 2009). Así, *marica* no se usaría como sustituto de otros términos locales, especialmente en el caso del español rioplatense.

Chile ocupa una posición aún más baja, con 16,02 usos del término marica por millón de tuits, el menor uso en todo el corpus. Chile tiene un total de 15.294.490 tuits, que representan el 12,5% de todo el corpus, y produjo 245 tuits que contienen el término *marica*, lo que significa que el término se usa muy poco. El español chileno se caracteriza por rasgos fonológicos distintivos y una diferenciación léxica sustancial, lo que contribuye a una fuerte identidad lingüística regional (Sadowsky y Aninao, 2019), con lo que expresiones locales como *huevón* podrían cumplir en Chile las funciones que *marica* desempeña en otras variedades.

El contexto sociolingüístico de Paraguay difiere debido al bilingüismo generalizado entre el guaraní y el español, una ecología de contacto que moldea la

estructura y el léxico del español paraguayo (Palacios y Pfänder, 2019). Entre los factores que contribuyen al bajo uso se encuentran la limitada infraestructura digital durante el periodo 2009-2016, los factores demográficos que afectaron la adopción de Twitter y la distancia geográfica combinada con identidades regionales propias (Kerswill y Williams, 2000; Meyerhoff, 2018).

Limitaciones del estudio

El estudio presenta algunas limitaciones. En primer lugar, la falta de análisis cualitativo: sin él, el contenido de los tuits no brinda una base para determinar el valor pragmático de *marica* en cada país. Por ejemplo, en un país, la menor frecuencia de uso puede demostrar una falta de uso o una dominancia excesiva de un uso peyorativo que es socialmente tabú en lo político y social. En segundo lugar, el análisis de la estratificación social es limitado por la falta de metadatos sociodemográficos. Así, no es posible concluir si existe un grupo de edad, género o socioeconómico que predomine en el uso de *marica*. Como tercer punto, centrarse en Twitter (ahora X) limita la extrapolación de los hallazgos o resultados a otras redes sociales o incluso el habla oral (Bruns, 2020; Nguyen et al., 2020; Sloan et al., 2019). Sin embargo, vale la pena resaltar que la construcción de corpus similares es mucho más difícil ahora que en el pasado debido a las restricciones en las API impuestas por X en 2023, lo que aumenta el valor documental del estudio. En cuarto lugar, el período de tiempo en el que fueron recogidos los datos (2009-2016) sugiere que los patrones analizados podrían no representar los usos actuales del término. No obstante, el estudio hace contribuciones al campo de estudio: prueba de manera metódica la aplicabilidad de los métodos computacionales a conjuntos de datos grandes (Brezina, 2018), plantea interrogantes teóricos sobre la circulación de FTN dentro del contexto digital panhispánico (Eisenstein, 2018) y brinda una cuantificación comparativa de la distribución de la FTN *marica* a nivel panhispánico (Cataño García y Campo-Paredes, 2025).

Conclusiones

Este estudio muestra que la distribución en el entorno digital de la FTN *marica* en los últimos veinte años no se adhiere a los modelos tradicionales de difusión lingüística

basados en la proximidad geográfica. Aunque los datos no permiten llegar a conclusiones definitivas, apoyan la hipótesis de que la migración y la conectividad en la red tienen prioridad sobre la mera proximidad de los territorios. Esto plantea que futuros estudios puedan hacer uso de metodologías computacionales mixtas.

Los resultados de la investigación demostraron que los corpus de gran tamaño y los métodos computacionales pueden ser útiles para identificar variaciones diatópicas a nivel hispánico que no se podían analizar tan fácilmente. De igual forma, estos resultados convierten las redes digitales en laboratorios en tiempo real para el estudio del uso de la lengua y del cambio lingüístico. Los hallazgos tienen también implicaciones para las teorías de difusión de lenguas en la era de las tecnologías e internet. Todo lo anterior sugiere que deben realizarse análisis cualitativos y mixtos para comprender el significado y los valores pragmáticos de un fenómeno lingüístico.

Disponibilidad de los datos

Por razones éticas y legales, los datos que respaldan los resultados de este estudio no pueden compartirse públicamente. El corpus pertenece a terceros y los datos no son de acceso público.

Reconocimientos

Este artículo es producto de la investigación “Frecuencia de uso y funciones pragmáticas de las fórmulas de tratamiento nominal *parce* y *marica* en un corpus de *Big Data* extraído de la red social X para España, EE. UU. y Suramérica”, presentada para optar al título de Magíster en Lingüística Panhispánica de la Universidad de la Sabana, Colombia. Expresamos nuestros sinceros agradecimientos al Dr. Sergio Jiménez, del Instituto Caro y Cuervo, por brindarnos acceso al corpus base de este estudio.

Referencias

Amer, M. (2024). Linguistic Landscapes of Social Media Discourse: Exploring

Language Practices and Identities on Jordanian Online Platforms. *Theory and Practice in Language Studies*. https://doi.org/10.17507/tpls.1411.01

Androutsopoulos, J., y Juffermans, K. (2019). Digital language practices in superdiversity: Introduction. *Discourse, Context y Media,* 30, 100287. doi.org/10.1016/j.dcm.2014.08.002

Arrieta, L. E. y Avendaño G. S. (2018). El discurso del tuit: un análisis lingüístico, sociodiscursivo y sociopragmático. Cuadernos de Lingüística Hispánica, (32), 107-130.

Asqueta Corbellini, M. C. (2008). Discurso y variación: el caso de marica en el habla de los estudiantes universitarios. *Lenguaje*, *36*(2), 551-572. https://doi.org/10.25100/lenguaje.v36i2.4876

Association of Internet Researchers. (2019). Internet research: Ethical guidelines 3.0. https://aoir.org/reports/ethics3.pdf

Auer, P., y Hinskens, F. (2005). The role of interpersonal accommodation in a theory of language change. In P. Auer, F. Hinskens, y P. Kerswill (Eds.), *Dialect change: Convergence and divergence in European languages* (pp. 335-357). Cambridge University Press.

Baker, P., y Egbert, J. (Eds.). (2021). *Triangulating methodological approaches in corpus linguistic research.* Routledge.

Batta, V. (2024). Human Language Data Processing using NLTK. *International Journal of Advanced Research in Science, Communication and Technology*. https://doi.org/10.48175/ijarsct-17685

Biber, D., Conrad, S., y Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press.

Bravo, D. (2009). Pragmática, sociopragmática y pragmática sociocultural del discurso de la cortesía. Una introducción. In D. Bravo, N. Hernández-Flores, y A. Cordisco (Eds.), *Enseñanza de la pragmática en español LE* (pp. 31-68). Editorial Dunken.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide.* Cambridge University Press.

Brown, P., y Levinson, S. C. (1987). *Politeness: Some universals in language usage.*

Cambridge University Press.

Bruns, A. (2020). Big social data approaches in internet studies: The case of Twitter. In J. Hunsinger, M. M. Allen, y L. Klastrup (Eds.), *Second international handbook of internet research* (pp. 51–67). Springer. https://doi.org/10.1007/978-94-024-1555-1_3

Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Equinox Publishing.

Carantón González, I. (2024). *El cambio semántico del lexema marica: estudio lexicológico-histórico y propuesta metodológica* [Trabajo de grado de pregrado]. Universidad de Antioquia. https://bibliotecadigital.udea.edu.co/server/api/core/bitstreams/e2df9065-5a84-4194-9a15-33a0457846d1/content

Cashman, H. R. (2017). Queer, Latinx, and bilingual: Narrative resources in the negotiation of identities. Routledge.

Cataño García, G. M., y Campo-Paredes, A. F. (2025). *Frecuencia de uso y funciones pragmáticas de las fórmulas de tratamiento nominal parce y marica en un corpus de Big Data extraído de la red social X para España, EE. UU. y Suramérica* [Tesis de maestría]. Universidad de La Sabana.

Corominas, J., y Pascual, J. A. (1980). *Diccionario crítico etimológico castellano e hispánico.* Gredos.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Egbert, J., y Biber, D. (2019). Incorporating text dispersion into corpus-based research: A survey of issues and challenges. *Corpora,* 14(1), 77-105.

Eisenstein, J. (2018). Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, y D. Watt (Eds.), *The handbook of dialectology* (pp. 368-383). Wiley Blackwell.

Engra Minaya, S. (2024). Identidades sociolingüísticas y reapropiación: análisis sociocultural de *maricón* y *bollera* en un corpus de Twitter. *MariCorners: Revista de Estudios Interdisciplinares LGTBIA+ y queer*, *1*(1), 235–266. https://doi.org/10.24197/mcreilq.1.2024.235-266

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations.*

Multilingual Matters.

García, O., y Wei, L. (2014). *Translanguaging: Language, bilingualism and education.* Palgrave Macmillan.

Gholami, L. (2021). Incidental reactive focus on form in language classes: Learners' formulaic versus nonformulaic errors, their treatment, and effectiveness in communicative interactions. *Foreign Language Annals*. https://doi.org/10.1111/flan.12546

Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J., y Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PLoS ONE,* 13(5), e0197741. https://doi.org/10.1371/journal.pone.0197741

Grieve, J., Nini, A., y Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence,* 2, 11. https://doi.org/10.3389/frai.2019.00011

Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction.* Routledge.

Gries, S. Th. (2020). Analyzing dispersion. In M. Paquot y S. Th. Gries (Eds.), A practical handbook of corpus linguistics (pp. 99-118). *Springer*. https://doi.org/10.1007/978-3-030-46216-1_5

Gries, S. (2022). Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics*. https://doi.org/10.1016/j.rmal.2021.100002

Gutiérrez-Rivas, C. (2016). La palabra *marico* como nueva forma de tratamiento nominal anticortés en el habla de jóvenes universitarios de Caracas: un estudio desde la perspectiva de los hablantes. *Logos (La Serena)*, *26*(1), 03-22.

Instituto Cervantes. (2020). *El español: una lengua viva. Informe 2020*. https://cvc.cervantes.es/lengua/anuario/anuario_20/informes_ic/p04.htm

Haboud, M., y de la Vega, E. (2021). Language attitudes and language policies in Ecuador. *International Journal of the Sociology of Language*, 2021(269), 157-176. https://doi.org/10.1515/ijsl-2020-0094

Huang, Y., Guo, D., Kasakoff, A., y Grieve, J. (2020). Understanding US regional

linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59, 244-255. https://doi.org/10.1016/j.compenvurbsys.2016.07.010

Hummel, M., Kluge, B., y Vázquez Laslop, M. E. (Eds.). (2010). Formas y fórmulas de tratamiento en el mundo hispánico. *El Colegio de México/Karl-Franzens-Universität Graz.*

Jebaselvi, A., Mohanraj, K., Thangamani, A., y Kumar, R. (2023). The Impact of Social Media on the Evolution of Language and Communication Trends. *Shanlax International Journal of English*. https://doi.org/10.34293/english.v12i1.6725

Jijón, J. F. (2022). *Lengua y sociedad en el Ecuador andino*. Abya-Yala.

Jiménez, S., Dueñas, G., Gelbukh, A., Rodriguez-Diaz, C. A., y Mancera, S. (2018). Automatic detection of regional words for Pan-Hispanic Spanish on Twitter. En *Lecture Notes in Computer Science* (pp. 403–414). Springer. https://doi.org/10.1007/978-3-030-03928-8_33

Kerswill, P., y Williams, A. (2000). Creating a new town koine: Children and language change in Milton Keynes. *Language in Society,* 29(1), 65-115. https://doi.org/10.1017/S0047404500001020

Krogstad, J. M., Passel, J. S., y Cohn, D. (2019). *5 facts about illegal immigration in the U.S.* Pew Research Center.

Lee, C. (2021). *Multilingualism online.* Routledge.

Lipski, J. M. (2018). Afro-Bolivian Spanish: The Iberian connection. *Journal of Ibero-Romance Creoles*, 8(1), 141-183

Méndez Vallejo, D. C. (2014). The M word: face and politeness in Colombian Spanish. *Dialectologia: Revista Electrónica, 12*, 89-108.

Meyerhoff, M. (2018). *Introducing sociolinguistics (3rd ed.*). Routledge.

Milroy, J., y Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, 21(2), 339-384. https://doi.org/10.1017/S0022226700010306

Moreno Fernández, F. (2009). *Principios de sociolingüística y sociología del lenguaje.* Ariel.

Moreno Fernández, F. (2021). *Variedades de la lengua española*. Routledge.

Moreno-Ortiz, A., y García-Gámez, M. (2023). Strategies for the Analysis of Large

Social Media Corpora: Sampling and Keyword Extraction Methods. *Corpus Pragmatics*, 1 - 25. https://doi.org/10.1007/s41701-023-00143-0

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., y Winters, J. (2020). How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence,* 3, 62. https://doi.org/10.3389/frai.2020.00062

Navarro-Carrascosa, C. (2021). *Análisis pragmalingüístico de las formas nominales de tratamiento en la comunidad de habla LGTBI* [Tesis doctoral, Universitat de València].

Navarro-Carrascosa, C. (2023). *Lingüística queer hispánica. Las formas nominales de tratamiento de la comunidad de habla LGTBI.* Peter Lang.

Obediente Sosa, E. (2019). El español de Venezuela: Características generales. *Lingüística y Literatura*, 75, 17-42.

Otheguy, R., y Stern, N. (2023). *Spanish and immigration: Current U.S. Spanish.* Cambridge University Press.

Page, R. (2012). *Stories and social media: Identities and interaction.* Routledge.

Palacios, A., y Pfänder, S. (2019). Español y guaraní: Contacto de lenguas en Paraguay. In C. Parodi, A. M. Ortíz López, y M. Lacorte (Eds.), *Manual de Lingüística Hispánica* (pp. 393-415). De Gruyter.

Paquot, M., y Bestgen, Y. (2023). Distinctiveness in learner corpora. In P. Baker y J. Egbert (Eds.), *Triangulating methodological approaches* (pp. 142-155). Routledge.

Potowski, K. (Ed.). (2023). *The Routledge handbook of Spanish as a minoritized language*. Routledge.

Real Academia Española. (2009). *Nueva gramática de la lengua española*. Espasa.

Rincón Martínez, J. A. (2021). *"Marica" como marcador de solidaridad en el habla juvenil de Bogotá*. Forma y Función, 34(1), 45–68.

Rogers, E. M. (2003). *Diffusion of innovations (5th ed.).* Free Press.

Sadowsky, S., y Aninao, T. (2019). *Fonética del español chileno*. Editorial Universidad de Concepción.

Seyidova, N., y Shakhayev, V. (2023). Python for big data analytics in social media research. *International Journal of Computer Science and Network Security,* 23(4), 89-102.

Sloan, L., Morgan, J., Burnap, P., y Williams, M. (2019). Who tweets? Deriving

demographic characteristics. *PLoS ONE,* 10(3), e0115545. https://doi.org/10.1371/journal.pone.0115545

Surya, K., Kumar, S. P., y Rao, T. S. (2020). Social network analysis using Python: Data mining and visualization. *International Journal of Engineering and Advanced Technology,* 9(3), 1234-1240.

Tagg, C., y Seargeant, P. (2021). *The language of social media (2nd ed.)*. Palgrave Macmillan.

Tagg, C., Seargeant, P., y Brown, A. (2022). *Taking offence on social media: Conviviality and communication*. Palgrave Macmillan.

Tarrade, L., Chevrot, J.-P., y Magué, J.-P. (2024). How position in the network determines the fate of lexical innovations on Twitter. *PLOS Complex Systems*, *1*(1), e0000005. https://doi.org/10.1371/journal.pcsy.0000005

Valdés, G. (2001). Heritage language students: Profiles and possibilities. En J. K. Peyton, D. A. Ranard y S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 37–77). Center for Applied Linguistics y Delta Systems.

Weisser, M. (2023). *Corpus linguistics for pragmatics: A guide for research (2nd ed.)*. Routledge.

Zimmermann, K. (2005). Construcción de la identidad y anticortesía verbal. En D. Bravo (Ed.), *Estudios de la (des)cortesía en español* (pp. 245-271). Dunken.