Identificación y análisis de patrones del parroquialismo en múltiples idiomas.

Trabajo de grado

Victoria Sofía Chavarro Diaz

Tutor: Jorge Andrés Alvarado Valencia

Pontificia Universidad Javeriana
Facultad de ciencias
Pregrado en Ciencia de datos
Bogotá D.C
2025

Resumen

Este proyecto aborda el estudio del parroquialismo lingüístico, entendido como la tendencia de los hablantes a priorizar referencias geográficas culturalmente cercanas. Se analizaron 14 idiomas representativos de diversas regiones del mundo, con el objetivo de ofrecer una comparación global de este fenómeno.

Mediante técnicas de procesamiento de lenguaje natural y word embeddings multilingües, se desarrollaron dos índices complementarios: uno basado en la frecuencia relativa de menciones locales frente a externas, útil para explorar sus causas y asociaciones; y otro que mide la desviación respecto a un escenario ideal de cobertura geográfica proporcional a la población mundial, facilitando la visualización de desequilibrios territoriales. Juntos, estos índices proporcionan una medida robusta y versátil del fenómeno.

Estas métricas permitieron no solo identificar variaciones en el grado de parroquialismo entre idiomas, sino también vincular esos niveles con factores estructurales como desigualdad, confianza social, migración y apertura cultural. Además, se generaron mapas comparativos y cartogramas que ilustran de manera accesible las diferencias de importancia territorial en el discurso multilingüe.

Los resultados permiten comprender mejor el fenómeno del parroquialismo lingüístico y explorar qué factores influyen en su aparición y magnitud. Este análisis ofrece así una base sólida para futuras investigaciones sobre sesgos geográficos y dinámicas culturales en el uso global del lenguaje.

Tabla de contenido

Resumen 2
1. Introducción 8
2. Objetivos 9
Objetivos específicos 9
3. Marco Teórico 10
3.1 Parroquialismo y parroquialismo nacional 10
3.1.1 Características del parroquialismo nacional 11
3.2 Procesamiento de lenguaje natural (PLN) 12
3.2.1 Hipótesis distribucional y representación vectorial del lenguaje 12
3.3 Redes neuronales en el PLN 13
3.4 Arquitecturas del modelo Word2Vec 14
3.4.1 CBOW (Continuous Bag of Words) 14
3.4.2 Skip-Gram 15
3.4.3 Fundamento estadístico del modelo Word2Vec 15
3.4.4 Negative Sampling 16
3.4.4.1 Ley de Zipf 17
3.4.5 FastText y la representación mediante subpalabras 18
3.5 Modelos de regresión regularizada: Lasso, Ridge y Elastic Net 19
3.6 Tecnologías complementarias en el análisis semántico multilingüe 21
3.6.1 API de Google translate 21
3.6.2 BERT: Representaciones contextuales del lenguaje 21
3.6.3 Visualización geográfica: Mapas estáticos, interactivos y cartogramas 23
3.7 Orígenes y explicaciones del parroquialismo 24
3.7.1 Estrategias adaptativas y antecedentes históricos 25
3.7.2 Teorías modernas 26
3.8 Aplicaciones relevantes y casos de estudio 27
4. Metodología 28
4.1 Construcción e integración de bases de datos 29
4.1.1 Recolección de información para la base de datos de las ciudades 29
4.1.2 Traducción Multilingüe y Evaluación de Precisión 30
4.1.2.1 Caso del Idioma Indonés 32
4.1.3 Creación y Enriquecimiento de la Base de Datos de Países 36
4.1.3.1 Integración de variables socioeconómicas y culturales 36
4.1.3.1.1 Trust Index 37
4.1.3.1.2 Corruption Index 38
4.1.3.1.3 Competition Index 39

4.1.3.1.4 Migration Index 39
4.1.3.1.5 Openness Index 40
4.1.3.2 Incorporación de Indicadores Adicionales 40
4.2 Análisis de frecuencia de ciudades en word embeddings multilingües 42
4.2.1 Procesamiento de frecuencia y manejo de ciudades de múltiples palabras 43
4.3 Construcción de la métrica de parroquialismo y visualización de resultados 45
4.3.1 Creación del primer índice de parroquialismo 46
4.3.2 Creación del segundo índice de parroquialismo 47
4.3.3 Creación del tercer índice de parroquialismo 49
4.3.4 Análisis de índices con variables explicativas 52
4.3.4.1 Modelos de Regresión Regularizada para Identificación de Variables Relevantes 55
4.3.5 Representatividad y cobertura de los datos 58
4.3.6 Visualización De Resultados 59
4.3.6.1 Mapas coropléticos estáticos 59
4.3.6.2 Mapa interactivo con leaflet 60
4.3.6.3 Cartogramas por continente 60
4.3.3 Validación analíticae interpretación de resultados 62
5. Resultados y análisis 63
5.1 Resultados de los Índices de Parroquialismo 64
5.1.1 Resultados del índice RMSE por idioma 64
5.1.2 Resultados del índice new por idioma 68
5.1.3 Resultados de comparación e integración de índices 70
5.1.4 Resultados del LASSO, Ridge y Elastic Net 73
5.2 Resultados de las visualizaciones 74
5.2.1 Mapas estáticos y dinámicos 75
5.2.2 Cartogramas por continente 78
5.2.2.1 América 78
5.2.2.2 África 80
5.2.2.3 Europa y Asia 82
5.2.2.4 Análisis de patrones comunes 83
5.2.2.5 Limitaciones del análisis y sugerencias para futuras investigaciones 85
5.3 Síntesis Integrada 86
6. Conclusiones 87
Referencias 91

Índice de figuras

Figura 1. Metodología general del proyecto 31
Figura 2. Primera etapa metodología 32
Figura 3. Representación multilingüe de los nombres de ciudades en la base de datos 34
Figura 4. Preguntas índices de confianza WVS 40
Figura 5. Preguntas para el índice de apertura WVS 42
Figura 6. Segunda etapa de la metodología 44
Figura 7. Prefijo usado para los diferentes idiomas 46
Figura 8. Idiomas con tokenizador específico 47
Figura 9. Posiciones de ciudades en los idiomas 48
Figura 10. Tercera etapa de la metodología 48
Figura 11. Dataframe inicial con primer índice 50
Figura 12. Dataframe actualizado con integración del segundo índice 51
Figura 13. Rankings de paises por idioma 52
Figura 14. Orden de idiomas por índice 54
Figura 15. Correlaciones entre índices 55
Figura 16. Correlación indicelog vs. variables explicativas 56
Figura 17. . Correlación índice new vs. variables explicativas 56
Figura 18. . Correlación índice RMSE vs. variables explicativas 56
Figura 19. Representatividad de los datos 62
Figura 20. Resultados del índice new por idioma 72
Figura 21. Índice new por idioma con promedios 72
Figura 22. Correlaciones del índice new vs. variables 73
Figura 23. Promedio entre índices 75
Figura 24. Ranking final de idiomas 76
Figura 25. Resultados regresión regularizada 77
Figura 26. Mapa escenario ideal 79
Figura 27. Mapa vietnamita 80
Figura 28. Mapa chino 80
Figura 29. Mapa inglés 81
Figura 30. Mapa suajili 81
Figura 31. Mapa coreano 81
Figura 32. Mapa finés 82
Figura 33. Mapa japonés 82
Figura 34. Mapa tamil 82
Figura 35. Mapa árabe 83
Figura 36. Mapa turco 83
Figura 37. Mapa portugués 84
Figura 38. Mapa francés 84
Figura 39. Mapa indonesio 84
Figura 40. Mapa español 84
Figura 41. Ejemplo de mapas interactivos (caso español) 86
Figura 42. Cartograma América (español) 88
Figura 43. Cartograma América (coreano) 88
Figura 44. Cartograma América (indonesio) 88
Figura 45. Cartograma América (turco) 88
Figura 46. Cartograma África (árabe) 90
Figura 47. Cartograma África (suajili) 90
Figura 48. Cartograma África (portugués) 90
Figura 49. Cartograma África (chino) 90
Figura 50. Cartograma África (viertnamita) 90

Figura 51. Cartograma Europa y Asia (finés) 92
Figura 52. Cartograma Europa y Asia (tamil) 92
Figura 53. Cartograma Europa y Asia (francés) 92
Figura 54. Cartograma Europa y Asia (japonés) 92
Figura 55. Cartograma Europa y Asia (inglés) 92
Figura 56. Resumen índices e indicadores por idioma 96

Índice de tablas

Tabla 1. Idiomas seleccionados con sus familias....33
Tabla 2. Niveles de precisión en la traducción de nombres de ciudades....35
Tabla 3. Representatividad de ciudades por continente....37
Tabla 4. Metadatos de la base de datos de las ciudades....37
Tabla 5. Metadatos de la base de datos de los países....44
Tabla 6. Resultados del índice RMSE por idioma....68
Tabla 7. Rankings más altos y bajos del francés....69
Tabla 8. Rankings más altos y bajos del español....69
Tabla 9. Rankings más altos y bajos del inglés....70
Tabla 10. Rankings más altos y bajos del vietnamita....70

1. Introducción

1.1 Planteamiento del problema y justificación

En un mundo cada vez más conectado, donde las fronteras físicas y culturales se distorsionan gracias a los avances en comunicación y movilidad, entender cómo las personas representan su entorno sigue siendo un desafío importante. Aunque se habla de globalización, los sentidos de pertenencia al propio grupo siguen siendo fuertes, y uno de los ejemplos más claros de esto es el parroquialismo nacional, entendido como la tendencia a cooperar más con miembros del propio país que con personas extranjeras.

Este fenómeno tiene implicaciones profundas, ya que, puede limitar la cooperación necesaria para enfrentar retos globales como las pandemias, el cambio climático o la gestión de bienes públicos internacionales. Aunque algunos estudios sugerían que el nivel de parroquialismo podía variar bastante entre culturas, un experimento reciente en 42 países y con más de 18.000 participantes mostró que este sesgo es generalizado y ocurre con una intensidad similar en todo el mundo, sin importar la distancia cultural o el nivel de desarrollo institucional [1]. Esto refuerza la idea de que el parroquialismo nacional es un rasgo humano bastante estable, y no un fenómeno exclusivo de ciertas sociedades.

A pesar de su relevancia, todavía hay vacíos importantes en su estudio. La mayoría de las investigaciones previas se han enfocado en comportamientos observables, como decisiones de las personas en juegos económicos, y han dejado de lado el análisis de cómo el parroquialismo puede expresarse de manera más sutil a través del lenguaje cotidiano. El lenguaje es una herramienta poderosa para revelar patrones de pertenencia, referencia y exclusión, por lo que analizarlo ofrece una oportunidad única para entender este fenómeno de forma más profunda y a gran escala.

Este proyecto busca precisamente llenar ese vacío. A través de técnicas de Procesamiento de

Lenguaje Natural (PLN) y word embeddings multilingües, se analizará cómo más de 800 ciudades son mencionadas en 14 idiomas diferentes. Se construirán índices específicos para medir el nivel de parroquialismo lingüístico y se explorará cómo se relaciona con factores como la apertura cultural, la confianza entre grupos, el acceso a internet y la desigualdad económica.

La importancia de este trabajo radica en que ofrece una comprensión más profunda del fenómeno del parroquialismo nacional, analizado desde una perspectiva lingüística y cultural en distintos idiomas y regiones del mundo. Al identificar qué tan presente está este sesgo en diferentes lenguas y qué factores socioeconómicos y culturales lo explican, el estudio proporciona una base sólida para entender mejor cómo se manifiesta y qué variables lo amplifican o reducen. Además, al cuantificar de manera comparativa el nivel de parroquialismo entre idiomas, los resultados pueden servir de guía para futuros estudios que busquen analizar su evolución, explorar estrategias para su manejo, o profundizar en las dinámicas de cooperación intercultural en distintos contextos. De esta manera, el proyecto contribuye no solo al conocimiento actual, sino también al desarrollo de investigaciones posteriores que busquen enfrentar los desafíos que este fenómeno plantea en un mundo cada vez más interconectado.

2. Objetivos

Objetivo general: Identificar, cuantificar y analizar el fenómeno del parroquialismo en diferentes idiomas, así como los factores socioeconómicos y culturales que lo determinan.

Objetivos específicos:

- Construir bases de datos de ciudades y países que incluyan información relevante para el análisis.

- Traducir los nombres de ciudades seleccionadas a 14 idiomas, usando la API de Google Translate y validando con back translation.
- Diseñar índices comparativos que midan el parroquialismo en base a la frecuencia y posición de las ciudades en los vectores de palabras.
- Analizar la relación entre el nivel de parroquialismo y variables socioeconómicas y culturales, para identificar qué factores lo explican y en qué medida influyen.
- Generar visualizaciones que representen el nivel de parroquialismo por idioma, destacando patrones regionales y lingüísticos.

3. Marco Teórico

Para comprender a profundidad el fenómeno del parroquialismo y su análisis mediante técnicas de procesamiento de lenguaje natural, es esencial definir los conceptos fundamentales que se abordarán a lo largo de esta investigación. A continuación, se presentan las definiciones de los términos más relevantes para este estudio.

3.1 Parroquialismo y parroquialismo nacional

El parroquialismo se entiende como la tendencia de las personas a centrarse en lo que les resulta más cercano y familiar, mostrando preferencia por lo local frente a lo lejano o extranjero. Este comportamiento puede observarse en distintos ámbitos, desde las relaciones sociales hasta la forma en que hablamos y pensamos sobre los lugares. En esta investigación, el enfoque está en el parroquialismo nacional, es decir, en cómo las comunidades tienden a mencionar y dar más importancia a lugares dentro de su propio país o cultura, dejando de lado o minimizando aquellos que están fuera de su entorno inmediato. Estos patrones en el lenguaje no son casuales, sino que

reflejan estructuras sociales, culturales e incluso cognitivas que están muy arraigadas, y que pueden influir de manera importante en la forma en que cooperamos y nos comunicamos a nivel global [2].

3.1.1 Características del parroquialismo nacional

El parroquialismo se caracteriza por una visión limitada del mundo, donde las personas priorizan a su propio grupo, comunidad o entorno cercano por encima de otros. Esta forma de pensar puede llevar a ignorar perspectivas distintas y a marginar lo diferente, especialmente cuando existe poco contacto con otras culturas, ideas o experiencias [3].

Una característica central del parroquialismo es el etnocentrismo, es decir, la creencia de que la cultura propia es superior a las demás. Esta actitud no solo refuerza la identidad colectiva, sino que también puede fomentar prejuicios hacia otras culturas, limitando la comprensión intercultural y generando barreras invisibles al diálogo y la cooperación [3].

Además, el parroquialismo tiende a fortalecer una mentalidad de "nosotros contra ellos". Desde esta lógica, los grupos externos son percibidos como amenazas potenciales, lo que lleva a desconfiar o rechazar perspectivas diferentes. Esta dinámica puede contribuir a la polarización social y debilitar los esfuerzos por construir sociedades más abiertas y diversas [3].

Otro rasgo importante es la insularidad, que se refiere a un tipo de aislamiento cultural. En contextos parroquialistas, las comunidades no solo privilegian su propia visión, sino que también se aíslan de influencias externas, limitando el intercambio de ideas y reduciendo las oportunidades de enriquecimiento mutuo [3].

Finalmente, en casos más extremos, el parroquialismo puede alimentar actitudes de xenofobia, donde el temor o el rechazo hacia personas de otras culturas se convierte en una barrera explícita para la convivencia. Aunque no siempre llega a este punto, el parroquialismo crea un terreno

fértil donde estas actitudes pueden surgir si no se promueve el contacto intercultural [3].

Estas características muestran que el parroquialismo no es simplemente una preferencia por lo propio, sino un fenómeno complejo que puede afectar la forma en que se percibe, se interactúa y se coopera con el resto del mundo.

3.2 Procesamiento de lenguaje natural (PLN)

Para estudiar el parroquialismo nacional y entender cómo distintas comunidades mencionan y perciben su entorno geográfico, es fundamental contar con herramientas capaces de analizar grandes volúmenes de texto. En este sentido, el Procesamiento de Lenguaje Natural (PLN), rama de la inteligencia artificial, resulta esencial. Esta área combina conocimientos de lingüística computacional, que estudia la estructura del lenguaje con técnicas de aprendizaje automático que permiten identificar patrones dentro de los datos [4].

A diferencia de enfoques más básicos que solo detectan palabras o frases sueltas, el PLN busca captar el contexto en que esas palabras aparecen, así como sus matices emocionales o culturales [4]. En este proyecto, el PLN se utiliza para observar cómo distintos idiomas priorizan la mención de ciertas ciudades en el uso cotidiano del lenguaje, permitiendo identificar y representar de manera visual los patrones de parroquialismo nacional.

3.2.1 Hipótesis distribucional y representación vectorial del lenguaje

Uno de los pilares fundamentales del PLN es la hipótesis distribucional, enunciada por Zellig Harris y reafirmada por John Firth, quien expresó: *"You shall know a word by the company it keeps"* [5]. Según esta hipótesis, el significado de una palabra puede inferirse a partir de los contextos en que aparece, principio que dio origen a los modelos conocidos como word embeddings.

Estos modelos representan palabras como vectores en un espacio numérico multidimensional,

donde la cercanía entre dos vectores refleja su similitud semántica. Para construir estas representaciones, los algoritmos analizan grandes corpus de texto, capturando patrones de coocurrencia dentro de ventanas de contexto (por ejemplo, cinco palabras antes y después de una palabra dada) [5].

Una propiedad fundamental de estos embeddings es que permiten medir la similitud semántica mediante la distancia coseno, definida como:

$$\mathrm{sim}(u, v) = \frac{u \cdot v}{||u||||v||} \quad (1)$$

Donde u y v son los vectores correspondientes a dos palabras, $u \cdot v$ es su producto punto, y $||u||$ $||v||$ representa la norma (magnitud) de cada vector.

En este proyecto, la hipótesis distribucional sirvió como fundamento teórico para inferir la importancia relativa de cada ciudad en diferentes idiomas. Bajo esta perspectiva, se asumió que, si una ciudad aparece frecuentemente y en contextos semánticos relevantes dentro de un idioma, su representación vectorial en el modelo de embeddings reflejará esa prominencia. Así, los vectores de ciudades no fueron tratados como elementos aislados, sino como el resultado de su interacción con el conjunto del discurso en cada lengua.

3.3 Redes neuronales en el PLN

El desarrollo de redes neuronales artificiales ha sido crucial para los avances recientes en procesamiento de lenguaje natural (PLN). Inspiradas en el funcionamiento del cerebro humano, estas redes están formadas por unidades llamadas neuronas artificiales, organizadas en capas sucesivas. Los datos pasan a través de estas capas, transformándose progresivamente hasta generar una salida [6].

Una red neuronal básica incluye tres componentes principales. La capa de entrada recibe los

datos iniciales que alimentan el modelo. A continuación, las capas ocultas aplican una serie de transformaciones matemáticas que permiten a la red aprender patrones complejos presentes en los datos. Finalmente, la capa de salida genera la predicción, clasificación o acción deseada. Cada conexión entre neuronas posee un peso (w), que indica la importancia de esa entrada en el procesamiento. Durante el entrenamiento de la red, estos pesos se ajustan mediante procesos de optimización, con el objetivo de minimizar el error del modelo [6].

Una de las arquitecturas más conocidas basadas en redes neuronales simples es la utilizada en el modelo Word2Vec. Esta red neuronal, que cuenta con una sola capa oculta, funciona de la siguiente manera: cada palabra se representa inicialmente como un vector one-hot, es decir, un vector en el cual solo una posición correspondiente a la palabra tiene valor uno, mientras que las demás posiciones son ceros. Este vector de entrada se multiplica por una matriz de pesos (w_{in}) para generar un embedding, que es una representación de dimensión reducida y continua de la palabra. Posteriormente, este embedding se multiplica por otra matriz de pesos (w_{out}) para producir puntuaciones o logits para cada palabra del vocabulario. Finalmente, se aplica la función softmax para convertir estas puntuaciones en probabilidades que permiten predecir o contextualizar la palabra objetivo dentro del modelo [6].

3.4 Arquitecturas del modelo Word2Vec

Existen dos variantes principales del modelo Word2Vec, que se diferencian por la forma en que definen la tarea de predicción: CBOW (Continuous Bag of Words) y Skip-Gram.

3.4.1 CBOW (Continuous Bag of Words)

En CBOW, el modelo recibe como entrada un conjunto de palabras de contexto y aprende a predecir la palabra central. Es decir, dado un grupo de palabras que aparecen antes y después de una posición determinada, el objetivo es identificar cuál es la palabra faltante.

Por ejemplo, en la frase *"La capital de__es famosa por su arquitectura"*, si el modelo recibe

como contexto ["La", "capital", "de", "es", "famosa"], debería predecir la palabra faltante, como "París".

El nombre "Bag of Words" indica que el orden de las palabras del contexto no es relevante: se consideran como un conjunto no secuencial. El embedding del contexto se obtiene promediando los vectores de las palabras vecinas, y se utiliza para predecir la palabra objetivo mediante una capa de salida y una función softmax [7].

3.4.2 Skip-Gram

Skip-Gram invierte la lógica de CBOW: dado una palabra central, el modelo intenta predecir las palabras que la rodean. Usando el mismo ejemplo, si la palabra central es "París", el modelo buscaría predecir palabras como "capital", "famosa" o "arquitectura".

Este enfoque es particularmente útil en grandes corpus y vocabularios extensos, ya que captura relaciones semánticas más ricas, especialmente con palabras poco frecuentes [7].

3.4.3 Fundamento estadístico del modelo Word2Vec

Más allá de su diseño intuitivo, Word2Vec se apoya en fundamentos estadísticos sólidos. Su objetivo es encontrar representaciones vectoriales de las palabras que maximicen la probabilidad de predecir términos que suelen aparecer juntos en los textos.

En su versión básica, Word2Vec busca maximizar la probabilidad de predecir una palabra de contexto w_o dada una palabra central w_i, utilizando la función softmax [7]:

$$P(w_o \mid w_i) = \frac{\exp(\mathbf{v'}_{w_o}^{\top} \mathbf{v}_{w_i})}{\sum_{w \in V} \exp(\mathbf{v'}_{w}^{\top} \mathbf{v}_{w_i})} \quad (2)$$

donde:

- V_{wi} es el vector de entrada de la palabra central,
- V'_{wo} es el vector de salida de la palabra objetivo (contexto),

- V es el vocabulario completo.

Para entrenar el modelo, se utiliza una función de pérdida basada en la entropía cruzada, que penaliza cuando el modelo asigna baja probabilidad a la palabra correcta. Esta pérdida se minimiza mediante técnicas de descenso de gradiente, ajustando los pesos de la red neuronal.

Sin embargo, calcular la función softmax sobre todo el vocabulario en cada paso es computacionalmente costoso, especialmente cuando el vocabulario es muy grande. Esto plantea un desafío importante para proyectos como este, donde se trabaja con modelos multilingües amplios [7].

3.4.4 Negative Sampling

Para superar las limitaciones de la softmax, Word2Vec incorpora la técnica de Negative Sampling. En lugar de evaluar todas las palabras del vocabulario, este enfoque selecciona una palabra real del contexto (positiva) y un pequeño conjunto de palabras aleatorias que no aparecen en ese contexto (negativas) [7, 8].

El modelo aprende entonces a asignar una alta probabilidad a la palabra positiva y bajas probabilidades a las negativas, utilizando la función sigmoide:

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{3}$$

Para una palabra objetivo w_o (positiva) y un conjunto de K palabras negativas, la función de pérdida se define como:

$$L = -\log \sigma(\mathbf{v'}_{w_o}^{\top} \mathbf{v}_{w_i}) - \sum_{j=1}^{k} \log \sigma(-\mathbf{v'}_{w_j^-}^{\top} \mathbf{v}_{w_i}) \tag{4}$$

donde:

- V_{wi} es el vector de entrada de la palabra central,
- V'_{wo} es el vector de la palabra de contexto positiva (salida),
- V'_{wj-} son los vectores de las palabras negativas.

Este procedimiento reduce de manera significativa el tiempo de entrenamiento sin comprometer la calidad de los vectores aprendidos, permitiendo que modelos como fastText trabajen de forma eficiente incluso con vocabularios extensos [7, 8].

3.4.4.1 Ley de Zipf

El procedimiento de muestreo negativo en Word2Vec no se realiza de manera uniforme. En su lugar, las palabras negativas se seleccionan siguiendo una versión ajustada de la Ley de Zipf, una regularidad empírica propuesta por George K. Zipf en 1932. Esta ley establece que la frecuencia f de una palabra es inversamente proporcional a su rango r en una lista ordenada de frecuencias: la palabra más común aparece aproximadamente el doble de veces que la segunda más común, tres veces más que la tercera, y así sucesivamente [9]:

$$f(r) = \frac{C}{r^s} \tag{5}$$

donde:

- C es una constante de normalización que depende del corpus específico y asegura que las frecuencias sumen al total de palabras,
- S es un parámetro que suele aproximarse a 1 en la mayoría de las lenguas naturales, reflejando la relación empírica observada por Zipf.

En Word2Vec y fastText, esta distribución se ajusta exponencialmente, elevando las frecuencias a la potencia de $\frac{3}{4}$, para evitar que palabras extremadamente comunes (como artículos o preposiciones) dominen el proceso de muestreo negativo:

$$P(w) \propto f(w)^{3/4} \quad (6)$$

Este ajuste mejora la cobertura del vocabulario y permite un entrenamiento más equilibrado.

Más allá de su aplicación en modelos de lenguaje, la Ley de Zipf también ha sido utilizada en estudios de análisis cultural y social, donde permite identificar preferencias lingüísticas y sesgos regionales a partir de patrones inesperados de frecuencia [8, 9].

3.4.5 FastText y la representación mediante subpalabras

Aunque Word2Vec representó un avance importante en la representación vectorial del lenguaje, presentaba una limitación significativa: solo podía generar vectores para palabras que habrían sido vistas explícitamente durante el entrenamiento. Para superar este problema, el equipo de Facebook AI desarrolló fastText, una extensión de Word2Vec que incorpora información subléxica mediante el uso de n-gramas de caracteres [10].

Un n-grama es una secuencia de *n* caracteres consecutivos dentro de una palabra. Por ejemplo, para la palabra "Bogotá" utilizando trigramas (n=3), los n-gramas serían: <Bo, Bog, ogó, got, otá, tá>, donde los signos < y > marcan el inicio y el final de la palabra. En el modelo fastText, el vector de una palabra no se aprende como una sola unidad, sino como la suma de los vectores correspondientes a todos sus n-gramas. De esta manera, el modelo puede construir representaciones para palabras nuevas combinando fragmentos ya conocidos, incluso si esas palabras nunca aparecieron en el corpus de entrenamiento [10].

Esta estrategia ofrece varias ventajas importantes. Permite manejar de manera eficaz palabras desconocidas o compuestas, como nombres de ciudades que contienen guiones, tildes o múltiples componentes. Además, captura patrones morfológicos del idioma, como sufijos recurrentes tales como "-stan", "-ville" o "-abad". También mejora la robustez del modelo frente a errores ortográficos o variaciones diacríticas, logrando, por ejemplo, que "Munich" y "Múnich" puedan

ser representados de forma coherente. El enfoque basado en subpalabras resulta particularmente útil en idiomas de morfología compleja, como el árabe, el finlandés o el hindi, donde la composición y variación de las palabras es rica y altamente productiva [10].

3.5 Modelos de regresión regularizada: Lasso, Ridge y Elastic Net

La regresión regularizada es una técnica que busca ajustar modelos de regresión introduciendo un término adicional de penalización sobre los coeficientes. Esta penalización controla la magnitud de los coeficientes, limitando su tamaño, y permite identificar cuáles variables tienen mayor importancia en relación con el fenómeno que se estudia [11].

En este proyecto, se emplean métodos de regresión regularizada para identificar cuáles factores socioculturales y económicos, como apertura cultural, desigualdad, confianza social, entre otros, están más asociados con el nivel de parroquialismo observado en diferentes idiomas. El objetivo es confirmar qué variables son capturadas con más fuerza por el fenómeno, más no construir un modelo predictivo.

La idea general es minimizar una función que combina dos partes: el error de predicción (por ejemplo, qué tan bien el modelo explica el fenómeno observado) y una penalización que limita cuánto pueden crecer los coeficientes de las variables [11].

Matemáticamente, el objetivo de la regresión regularizada es minimizar una función de la forma:

$$\text{Función de pérdida regularizada} = \text{Error de predicción} + \lambda \times \text{Penalización} \tag{7}$$

donde:

- λ es un parámetro que controla la intensidad de la penalización: valores más altos de λ imponen una penalización más fuerte.
- La penalización depende del tipo de regularización aplicada.

En la regresión Ridge, el término de penalización es la suma de los cuadrados de los coeficientes ($L2\ norm$):

$$\lambda \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

Este método reduce el tamaño de los coeficientes, pero no los lleva a cero. En este proyecto, Ridge ayuda a ver qué variables tienen coeficientes más grandes y, por tanto, más influencia en explicar el parroquialismo, considerando todas las variables juntas [11].

En la regresión Lasso, la penalización es la suma de los valores absolutos de los coeficientes ($L1\ norm$):

$$\lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

Una característica clave de Lasso es que puede forzar a algunos coeficientes exactamente a cero. En este proyecto, Lasso permite identificar directamente qué variables son las más relevantes, ya que las variables que no influyen significativamente en el nivel de parroquialismo son eliminadas automáticamente del modelo (sus coeficientes se vuelven cero) [11].

El Elastic Net combina las penalizaciones de Ridge y Lasso:

$$\lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \tag{10}$$

donde:

- α controla el balance entre la penalización Lasso ($L1$) y Ridge ($L2$).
- Cuando α=1, Elastic Net se comporta como Lasso.
- Cuando α=0, se comporta como Ridge.

En este proyecto, Elastic Net sirve como un método intermedio: selecciona algunas variables como Lasso, pero también considera que otras pueden seguir teniendo un efecto pequeño, como Ridge [11].

El uso de estas técnicas permite identificar con mayor claridad qué factores explicativos están más asociados al nivel de parroquialismo en distintos idiomas, proporcionando así una base sólida para interpretar los resultados obtenidos en el análisis.

3.6 Tecnologías complementarias en el análisis semántico multilingüe

Además de los modelos de word embeddings tradicionales, el análisis semántico multilingüe como el propuesto en este estudio se apoya en otras herramientas esenciales. Tecnologías como los modelos de lenguaje profundo (BERT), los sistemas de traducción automática, y las librerías de visualización geográfica y estadística permiten representar los resultados de manera más accesible e interpretable, superando algunas limitaciones de los métodos clásicos.

3.6.1 API de Google translate

La traducción automática es una rama del Procesamiento de Lenguaje Natural que permite transformar texto de un idioma a otro de forma automática. En este proyecto se utilizó la API de Google Translate, una herramienta basada en modelos de redes neuronales encoder-decoder entrenados sobre grandes volúmenes de texto multilingüe, que permiten realizar traducciones rápidas y contextualmente adecuadas [13].

Estos modelos funcionan en dos etapas principales: primero, el encoder convierte la entrada (por ejemplo, el nombre de una ciudad en español) en una representación numérica continua que resume su significado; luego, el decoder genera, paso a paso, la traducción en el idioma de destino. Estas arquitecturas, basadas en redes neuronales recurrentes o transformadores, permiten capturar tanto relaciones gramaticales como aspectos semánticos [12].

3.6.2 BERT: Representaciones contextuales del lenguaje

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de lenguaje

basado en la arquitectura de transformadores, desarrollado por Google AI y presentado por Devlin en 2018. A diferencia de modelos anteriores como Word2Vec o GloVe, que generan un único vector por palabra independientemente del contexto, BERT produce representaciones contextuales: el vector de una palabra cambia dependiendo de las palabras que la rodean en cada oración [14].

El funcionamiento de BERT introduce dos innovaciones principales. Por un lado, utiliza una codificación bidireccional real: al procesar un texto, BERT considera simultáneamente el contexto hacia la izquierda y hacia la derecha de cada palabra, logrando así una comprensión más profunda del significado. Por otro lado, su entrenamiento se basa en la tarea de Masked Language Modeling (MLM): durante el entrenamiento, algunas palabras de la entrada se ocultan de manera aleatoria (por ejemplo, "París es la capital de [MASK]") y el modelo debe predecir cuál es la palabra faltante. Esta técnica fuerza al modelo a construir representaciones semánticas ricas y precisas, entendiendo no solo las palabras individuales, sino también cómo interactúan entre sí dentro de las frases [14].

Para facilitar el uso de BERT y otros modelos similares, se ha desarrollado la librería Transformers, creada por la empresa Hugging Face, la cual es una organización reconocida por democratizar el acceso a modelos de procesamiento de lenguaje natural, ofreciendo plataformas abiertas y herramientas que permiten a investigadores y desarrolladores aplicar modelos avanzados sin necesidad de entrenarlos desde cero [15].

A través de su librería Transformers, Hugging Face proporciona implementaciones preentrenadas de miles de modelos de lenguaje en diferentes idiomas, listas para ser utilizadas en tareas como clasificación, traducción automática, respuesta a preguntas o generación de texto. Esta facilidad de acceso y uso ha sido clave para impulsar la aplicación práctica de modelos como BERT en proyectos de investigación y análisis de lenguaje a gran escala [15].

En este proyecto, se utilizó el modelo bert-base-multilingual-cased, una variante de BERT diseñada para trabajar con 104 idiomas simultáneamente. El término "base" indica que el modelo tiene un tamaño medio (12 capas de transformadores, 768 dimensiones por vector, y 110 millones de parámetros); "multilingual" refleja su entrenamiento multilingüe; y "cased" significa que distingue entre mayúsculas y minúsculas, lo cual es crucial para respetar la forma correcta de nombres propios como "Bogotá" o "São Paulo".

El modelo fue aplicado para resolver casos en los que, tras traducir nombres de ciudades entre idiomas, surgían fragmentos incompletos o inconsistencias. Utilizando la funcionalidad de Masked Language Modeling disponible en transformers, se predijeron palabras faltantes o se completaron secuencias parcialmente traducidas, mejorando así la correspondencia de nombres en los embeddings analizados.

Gracias a su capacidad para interpretar contextos variados y manejar múltiples lenguas de manera robusta, BERT permitió fortalecer la calidad de los datos lingüísticos del proyecto, asegurando que el análisis de parroquialismo se basara en representaciones textuales más precisas y completas.

3.6.3 Visualización geográfica: Mapas estáticos, interactivos y cartogramas

La representación geográfica fue una parte fundamental en este proyecto para visualizar de manera intuitiva los patrones de parroquialismo detectados en distintos idiomas. Para lograrlo, se utilizaron distintos tipos de mapas: mapas estáticos, mapas interactivos y cartogramas deformados.

Los mapas estáticos permiten representar visualmente cómo se distribuye la importancia asignada a los países según el idioma analizado. Utilizando escalas de colores, es posible destacar las diferencias regionales de una manera clara y ordenada, facilitando la identificación de

tendencias generales [16].

Por otro lado, los mapas interactivos amplían esta funcionalidad al permitir que el usuario explore los datos de forma dinámica. A través de herramientas que habilitan el zoom, el movimiento y la visualización de información adicional al pasar el cursor sobre un país, los mapas interactivos permiten un análisis más detallado y flexible de la información [16].

Un tipo especial de visualización utilizado en este proyecto fueron los cartogramas deformados. En los cartogramas, el tamaño de los países no refleja su área geográfica real, sino una variable cuantitativa de interés, como su nivel de importancia en un idioma. Esta distorsión controlada facilita entender visualmente el peso relativo que ciertos territorios adquieren en el discurso lingüístico, independientemente de su extensión física. Los cartogramas ayudan a resaltar de manera intuitiva los sesgos de atención geográfica presentes en los datos [17].

Para crear estas visualizaciones se utilizaron herramientas específicas. La librería cartogram en R permite generar mapas deformados mediante algoritmos que modifican proporcionalmente las áreas geográficas según una variable numérica. Esto permite representar de forma más fiel el impacto o relevancia de un país en el fenómeno estudiado [18]. Por su parte, la librería leaflet facilita la creación de mapas interactivos, permitiendo agregar funcionalidades como ventanas emergentes, control de capas, zoom y personalización de elementos gráficos, todo ello de forma dinámica y accesible desde un navegador web [19].

Gracias a estas técnicas y herramientas, fue posible representar de manera clara, comparativa y accesible la información geográfica obtenida, enriqueciendo así el análisis y la interpretación de los patrones de parroquialismo en diferentes idiomas.

3.7 Orígenes y explicaciones del parroquialismo

Para entender por qué el parroquialismo persiste en tantas sociedades, es necesario explorar

tanto sus raíces históricas como las teorías modernas que explican su funcionamiento. Desde su surgimiento como estrategia de supervivencia en pequeños grupos humanos, hasta su expresión actual en dinámicas sociales, económicas y psicológicas, el parroquialismo refleja una respuesta adaptativa compleja a contextos cambiantes. Esta sección presenta los principales enfoques y factores que ayudan a comprender cómo y por qué las personas tienden a favorecer a su propio grupo frente a los externos.

3.7.1 Estrategias adaptativas y antecedentes históricos

En los primeros momentos de la historia humana, pertenecer a un grupo no solo ofrecía compañía, sino protección frente a peligros constantes. Las comunidades pequeñas dependían profundamente de la cooperación interna para sobrevivir, ya que el trabajo conjunto permitía cazar, defenderse y compartir recursos de manera más eficiente. Favorecer a los miembros del propio grupo era, por tanto, una estrategia adaptativa lógica: confiar en lo familiar aumentaba las probabilidades de supervivencia, mientras que desconfiar de los forasteros ayudaba a proteger lo que se tenía [20].

Estudios antropológicos han demostrado que esta preferencia no era solo emocional, sino una respuesta funcional a entornos inciertos y escasos en recursos. Esta lógica básica de "nosotros contra ellos" parece haber perdurado, adaptándose a nuevas formas de organización social a medida que surgían alianzas más amplias entre clanes y tribus. A lo largo del tiempo, estas alianzas evolucionaron hacia estructuras sociales más complejas como jefaturas, confederaciones y eventualmente estados, institucionalizando así las lealtades internas [20, 21].

Aunque las sociedades modernas operan en un marco globalizado, las dinámicas de favoritismo hacia el grupo propio siguen presentes en decisiones diplomáticas, económicas y sociales. Incluso frente a desafíos globales como pandemias o crisis ambientales, la lógica de priorizar a los propios ha limitado muchas veces la colaboración internacional, reflejando la

persistencia de este patrón adaptativo en un mundo interconectado [22].

3.7.2 Teorías modernas

Una de las teorías más influyentes para explicar el parroquialismo actual es la Teoría de la identidad social, propuesta por Henri Tajfel y John Turner. Según esta teoría, las personas tienden a clasificarse en grupos sociales ("nosotros" y "ellos"), mostrando mayor lealtad y disposición a cooperar con los miembros de su propio grupo (endogrupo) y desconfianza hacia quienes son percibidos como externos (exogrupo). Esta identificación grupal fortalece la cohesión interna, pero también puede generar actitudes excluyentes como el etnocentrismo, donde la propia cultura es vista como superior [23].

Complementando esta perspectiva, la noción de altruismo parroquial plantea que no solo cooperamos más con los nuestros, sino que en contextos de conflicto podemos justificar la hostilidad hacia los forasteros. Desde un enfoque evolutivo, este comportamiento habría ofrecido ventajas en entornos de alta competencia, aumentando la supervivencia grupal. Estudios recientes muestran que este tipo de altruismo se intensifica en situaciones de peligro o competencia por recursos [24].

La teoría de la evolución de la cooperación condicional aporta otra explicación relevante. A diferencia de estrategias como el "tit-for-tat", donde la cooperación depende de acciones pasadas, la cooperación condicional enfatiza que las personas tienden a ayudar más a quienes reconocen como parte de su grupo, independientemente de su comportamiento previo [25]. En escenarios de incertidumbre o competencia, estas dinámicas se refuerzan, priorizando la cohesión interna sobre la apertura externa, lo que perpetúa actitudes parroquialistas.

Finalmente, factores económicos y psicológicos también juegan un papel crucial. Desde la economía, el parroquialismo tiende a intensificarse en contextos de escasez o inseguridad, donde

proteger al propio grupo se vuelve una estrategia para preservar recursos [26]. Desde la psicología, la percepción del riesgo y la necesidad de seguridad emocional fortalecen la cooperación hacia los cercanos y la desconfianza hacia los externos. Esta dinámica se agrava en momentos de crisis o amenaza, alimentando prejuicios y limitando la cooperación amplia [27].

En conjunto, estas explicaciones muestran que el parroquialismo no es el resultado de un solo factor, sino de la interacción entre presiones evolutivas, estructuras sociales, necesidades emocionales y condiciones económicas. Comprender esta complejidad es clave para diseñar estrategias que promuevan una cooperación más abierta en un mundo cada vez más interdependiente.

3.8 Aplicaciones relevantes y casos de estudio

En los últimos años, herramientas como el Procesamiento de Lenguaje Natural (PLN) y el aprendizaje automático se han consolidado como métodos fundamentales para estudiar el comportamiento humano a partir de grandes volúmenes de texto. Estas tecnologías permiten identificar patrones en opiniones, emociones o sesgos sociales, abriendo nuevas posibilidades para la investigación social a gran escala.

Aunque no todos los estudios han abordado directamente el fenómeno del parroquialismo, muchos de sus hallazgos ofrecen pistas valiosas para investigaciones como la presente. Por ejemplo, un estudio publicado en *Frontiers in Psychology* analizó datos de redes sociales utilizando Apache Spark, una plataforma de procesamiento paralelo. Procesando miles de comentarios de plataformas como Dailymotion, los investigadores lograron identificar actitudes colectivas y reacciones compartidas frente a temas específicos. Este enfoque permite detectar patrones que podrían pasar desapercibidos mediante métodos tradicionales [28].

De manera complementaria, un estudio en la *Journal of Big Data* aplicó técnicas de análisis de

sentimientos, combinando diccionarios léxicos y algoritmos de aprendizaje automático para clasificar mensajes como positivos, negativos o neutrales. Esta técnica ha demostrado ser efectiva para rastrear cambios en las actitudes públicas y evidenciar sesgos culturales, como la polarización en línea [29].

Asimismo, metodologías como la clasificación de textos y el modelado de temas permiten agrupar mensajes por similitud o descubrir ideas recurrentes en diferentes contextos. Estas herramientas resultan especialmente relevantes en estudios multilingües y comparativos entre países, donde los patrones de lenguaje pueden revelar fenómenos culturales profundos como el parroquialismo [30].

En conjunto, estos antecedentes destacan el potencial del PLN y de los enfoques de datos masivos para analizar actitudes sociales en distintos contextos. Su aplicación a investigaciones como esta que busca entender cómo el lenguaje refleja sesgos parroquialistas en diferentes idiomas, refuerza la pertinencia del enfoque metodológico elegido y demuestra la capacidad de estas tecnologías para abordar fenómenos culturales de manera global y comparativa.

4. Metodología

En este trabajo se desarrolló un algoritmo basado en técnicas de Procesamiento de Lenguaje Natural (PLN) con el objetivo de identificar y analizar el fenómeno del parroquialismo nacional en distintos idiomas. El enfoque consistió en capturar y medir menciones geográficas a partir de grandes fuentes de datos textuales, utilizando modelos preentrenados de fastText desarrollados por el equipo de Facebook AI.

Estos modelos fueron entrenados sobre dos grandes corpus lingüísticos: Wikipedia, conocida por su contenido curado y estandarizado en más de 200 idiomas, y Common Crawl, un proyecto de rastreo masivo de la web abierta que provee millones de páginas procesadas y depuradas

mensualmente. Mientras Wikipedia aporta calidad editorial, Common Crawl ofrece volumen y diversidad temática, permitiendo robustecer los modelos especialmente en idiomas con menor presencia digital [31, 32]. Los vectores de palabras obtenidos de estos modelos fueron empleados en esta investigación para evaluar la prominencia relativa de las ciudades en diferentes lenguas.

Una vez obtenidas las menciones geográficas, el análisis se complementó correlacionando la posición relativa de las ciudades con variables socioeconómicas y culturales de los países correspondientes. Estas variables incluyen el nivel de desarrollo humano (HDI), la desigualdad de ingresos (Gini), el acceso a internet, la confianza endogrupal y exogrupal, la percepción de corrupción, la valoración de la competencia, la apertura cultural y la actitud frente a la migración. A través de estas relaciones, se buscó comprender con mayor profundidad los factores que influyen en los niveles de parroquialismo expresados lingüísticamente.

Con el fin de cumplir con los objetivos propuestos, el proceso metodológico se estructuró en diferentes etapas, descritas a continuación:

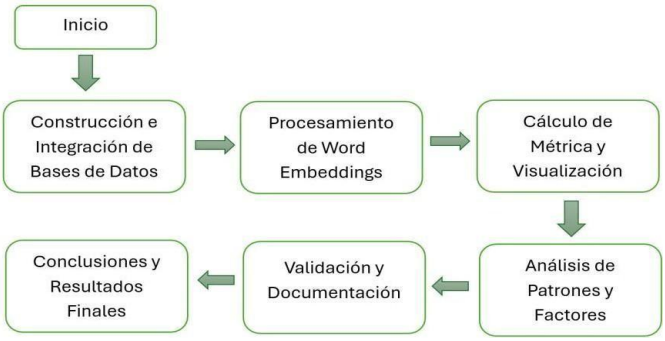


Figura 1. Metodología general del proyecto

4.1 Construcción e integración de bases de datos

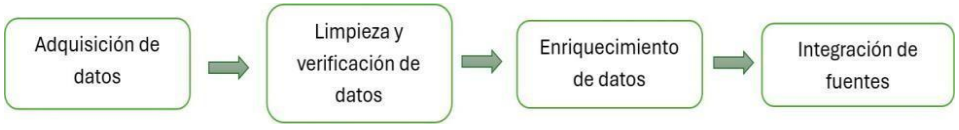


Figura 2. Primera etapa metodología

4.1.1 Recolección de información para la base de datos de las ciudades

La base de datos de ciudades constituye un elemento central en este proyecto, ya que reúne información detallada sobre las 802 ciudades más pobladas del mundo en 2024. Esta información fue obtenida de la fuente World City Listing Table [33], la cual incluye variables como el nombre de la ciudad, el país al que pertenece, el continente y la población actualizada. Inicialmente, la base también contenía datos sobre la población de 2023 y su variación porcentual, pero estas columnas fueron descartadas, enfocándose únicamente en la población más reciente y la ubicación geográfica.

Para enriquecer la información disponible, se integró una segunda fuente: la base de datos Country Annotation, que contiene cerca de 250 registros con aspectos geopolíticos y lingüísticos de cada país. Entre los datos relevantes destacan el código de país ISO 3166-1 alfa-2, el nombre oficial en inglés, el continente de pertenencia y los idiomas principales o cooficiales.

La integración de ambas bases se realizó en R, vinculando los registros mediante el nombre del país. Esto permitió añadir de forma estructurada la información de continente y garantizar una correcta correspondencia geográfica para cada ciudad. Esta etapa fue fundamental para consolidar una base robusta, coherente y lista para los análisis posteriores.

4.1.2 Traducción Multilingüe y Evaluación de Precisión

En la siguiente etapa, el trabajo se trasladó a un entorno de programación en Jupyter Notebook utilizando Python, con el objetivo de enriquecer la base de datos incorporando traducciones de los nombres de las ciudades a distintos idiomas. Para cada ciudad, se añadieron columnas correspondientes a su equivalente en 13 idiomas diferentes, capturando así cómo se nombran los lugares en diversas lenguas representativas de distintas regiones del mundo.

La selección de idiomas se realizó siguiendo un criterio lingüístico, priorizando aquellos que representan familias lingüísticas amplias y diversas. Además del inglés (idioma original de la

base), se incluyeron 13 idiomas adicionales. A continuación, se presenta una tabla que resume los idiomas elegidos y su familia lingüística respectiva:

IDIOMA	FAMILIA LINGÜÍSTICA
Inglés	Indoeuropea
Español	Indoeuropea
Francés	Indoeuropea
Portugués	Indoeuropea
Chino-mandarín	Sino-tibetana
Árabe	Afroasiática
Indonesio	Austronesia
Japonés	Japonés (aislado)
Coreano	Coreana (aislado)
Vietnamita	Austroasiática
Suajili	Níger-Congo
Tamil	Dravídica
Turco	Altaica
Finés	Urálica

Tabla 1. Idiomas seleccionados con sus familias

Para traducir los nombres de ciudades a los distintos idiomas analizados en este proyecto, se utilizó la API de Google Translate mediante la librería *deep_translator* en Python, que permite enviar solicitudes de traducción de manera programada y recibir respuestas rápidas. Se configuró un traductor que detectaba automáticamente el idioma de origen y traducía hacia el idioma objetivo (como español, francés o árabe), a través de una función personalizada que procesaba cada ciudad de forma individual. Anticipando posibles errores, como fallos de conexión, límites de uso o caracteres especiales, la función incorporó un manejo de excepciones: en caso de error, se registraba el problema y se mantenía el nombre original de la ciudad, asegurando así la continuidad del proceso. Finalmente, la traducción se aplicó a toda la columna de nombres de ciudades, generando nuevas columnas por idioma y construyendo una base de datos multilingüe robusta y consistente.

Para ilustrar el resultado del proceso de enriquecimiento, se muestra un ejemplo de la base de datos multilingüe generada:

city_in_Spanish	city_in_Chinese	city_in_Arabic	city_in_Indonesian	city_in_Swahili	city_in_Tamil	city_in_Turkish	city_in_Finnish	city_in_Japanese
Tokio	东京	طوكيو	Tokyo	Tokyo	டோக்கியோ	Tokyo	Tokio	東京
Delhi	德里	دلهي	Delhi	Delhi	டெல்லி	Delhi	Delhi	デリー
Llevar a la fuerza	上海	شنغهاي	Kota Shanghai	Shanghai	ஷாங்காய்	Şanghay	Shanghai	上海
Dacca	达卡	دكا	Kota Dhaka	Dhaka	டாக்கா	Dakka	Dhaka	ダッカ
San Pablo	圣保罗	ساو باولو	Sao Paulo	São Paulo	சாவ் பாலோ	Sao Paulo	São Paulo	サンパウロ
El Cairo	开罗	القاهرة	Kairo	Cairo	கெய்ரோ	Kahire	Kairo	カイロ
Ciudad de México	墨西哥城	مدينة مكسيكو	Kota Meksiko	Mexico City	மெக்சிகோ நகரம்	Meksika Şehri	Mexico City	メキシコシティ
Pekín	北京	بكين	Kota Beijing	Beijing	பெய்ஜிங்	Pekin	Peking	北京
Bombay	孟买	مومباي	Mumbai	Mumbai	மும்பை	Bombay	Mumbai	ムンバイ
Osaka	大阪	أوساكا	Osaka	Osaka	ஒசாகா	Osaka	Osaka	大阪

Figura 3. Representación multilingüe de los nombres de ciudades en la base de datos

Posteriormente, se aplicó una técnica de back translation: se tradujeron nuevamente al inglés los nombres de las ciudades traducidos, con el fin de comparar el resultado con los nombres originales y evaluar la precisión de las traducciones automáticas.

Esta comparación permitió calcular una métrica de accuracy para cada idioma, es decir, el porcentaje de nombres que se mantuvieron coherentes o correctos tras el proceso de ida y vuelta. Los resultados evidenciaron diferencias notables entre idiomas, reflejando los distintos niveles de fidelidad alcanzados por la traducción automática según el sistema lingüístico.

A continuación, se presentan los niveles de precisión obtenidos:

IDIOMA	PRECISIÓN
Español	78%
Chino	83%
Árabe	71%
Indonesio	49%
Suajili	89%
Tamil	72%
Turco	74%
Finés	89%
Japonés	80%
Coreano	70%
Vietnamita	71%
Portugués	85%
Francés	80%

Tabla 2. Niveles de precisión en la traducción de nombres de ciudades

4.1.2.1 Caso del idioma Indonés

Entre todos los idiomas evaluados, el indonesio presentó el menor nivel de precisión en el proceso de traducción inversa, con una tasa de acierto inicial inferior al 50%. Este resultado generó inquietudes y motivó una revisión más detallada de las traducciones obtenidas en este idioma.

Durante el análisis, los casos problemáticos se clasificaron en dos grupos: errores reales y diferencias aceptables. Los errores reales correspondían a traducciones que generaban palabras completamente distintas al nombre original, sin vínculo alguno con su forma o significado. Por ejemplo, la ciudad "Heze" fue traducida como "Wet", un término sin relación aparente, y "Sao Jose dos Campos" se tradujo incorrectamente como "Saint Joseph of Padang", alterando el referente geográfico.

Sin embargo, un examen más profundo reveló que muchas de las supuestas imprecisiones no constituían errores estrictos, sino variaciones lingüísticas razonables. Entre ellas se encontraron adiciones como "City", el uso de palabras locales como "Kota" (que significa "ciudad" en indonesio), o ligeras modificaciones ortográficas, como "Xi-an" en vez de "Xi'an". También se observó que el contexto cultural del idioma influía en la forma de traducir ciertos nombres, generando versiones reconocidas o de uso común en el idioma local.

A continuación, se presenta una tabla con ejemplos representativos de estas diferencias:

	city	city_from_Indonesian_to_English
	<chr>	<chr>
1	Shanghai	Shanghai City
2	Dhaka	Dhaka City
3	Sao Paulo	São Paulo
4	Beijing	Beijing City
5	Chongqing	Kota Chongqing
6	Kinshasa	Kinshasa City
7	Lagos	Lagos City
8	Manila	Manila City
9	Guangzhou	Guangzhou City
10	Tianjin	Tianjin City
11	Bangalore	Bengaluru
12	Chennai	Chennai City
13	Bogota	Bogota, Colombia
14	Hyderabad	Kota Hyderabad
15	Nanjing	Nanjing City
16	Chengdu	Chengdu City
17	Luanda	Luanda City
18	Tehran	Teheran
19	Ho Chi Minh City	Kota Ho Chi Minh
20	Nagoya	Nagoya City
21	Xi-an	Xi'an
22	Ahmedabad	Ahmedabad City
23	Wuhan	Wuhan City
24	Hangzhou	Kota Hangzhou
25	Suzhou	Suzhou District
26	Baghdad	Bagdad
27	Shenyang	Shenyang City
28	Riyadh	Riyadh City
29	Hong Kong	Hongkong
30	Foshan	Foshan City
31	Dongguan	Dongguan City
32	Pune	Kota Pune
33	Santiago	City of Santiago
34	Khartoum	Kota Khartoum

Figura 4. Variaciones lingüísticas en la traducción de nombres de ciudades

Tras este análisis, se concluyó que solo alrededor del 12% (aproximadamente) de las traducciones en indonesio correspondían a errores reales. Esto permitió asumir que, en los demás idiomas, cuyo nivel de precisión inicial superaba el 70%, muchos de los supuestos errores también serían simplemente diferencias aceptables y no fallos verdaderos.

A partir de este hallazgo, se estableció que el margen de error general en el conjunto de traducciones es bajo y manejable, probablemente inferior al 12%. Aunque se reconoce la existencia de cierto nivel de ruido, este fue considerado aceptable dentro del alcance del proyecto, y se asumió que no afectaría de manera significativa las conclusiones generales sobre el fenómeno del parroquialismo.

Adicionalmente, se evaluó la representatividad geográfica y lingüística de la base de datos, verificando la distribución de las ciudades por continente. Este análisis aseguró un balance adecuado en la cobertura global, con un total de 132 países y 86 idiomas representados, incluidos los 14 idiomas principales seleccionados para el estudio:

CONTINENTE	NÚMERO DE CIUDADES
África	109
Europa	78
Norte América	66
Asia	481
Oceanía	6
Sur América	62

Tabla 3. Representatividad de ciudades por continente

Para complementar esta descripción, a continuación, se presenta también una tabla con los metadatos básicos de la base de datos, que resume la información contenida en cada columna:

COLUMNA	TIPO DE DATO	DESCRIPCIÓN
city	Texto	Nombre original de la ciudad (en inglés).
country	Texto	País al que pertenece la ciudad.
continent	Texto	Continente al que pertenece la ciudad.
population	Numérico	Población estimada de la ciudad en el año 2024.
city_in_spanish	Texto	Nombre de la ciudad traducido al español.
city_in_portuguese	Texto	Nombre de la ciudad traducido al portugués.
city_in_french	Texto	Nombre de la ciudad traducido al francés.
city_in_arabic	Texto	Nombre de la ciudad traducido al árabe.
city_in_chinese	Texto	Nombre de la ciudad traducido al chino mandarín.
city_in_korean	Texto	Nombre de la ciudad traducido al coreano.
city_in_japanese	Texto	Nombre de la ciudad traducido al japonés.
city_in_vietnamese	Texto	Nombre de la ciudad traducido al vietnamita.
city_in_finnish	Texto	Nombre de la ciudad traducido al finés.
city_in_tamil	Texto	Nombre de la ciudad traducido al tamil.
city_in_swahili	Texto	Nombre de la ciudad traducido al suajili.
city_in_turkish	Texto	Nombre de la ciudad traducido al turco.
city_in_indonesian	Texto	Nombre de la ciudad traducido al indonés.
city_from_Spanish_to_English	Texto	Resultado de la back translation desde el español al inglés.
city_from_Portuguese_to_English	Texto	Resultado de la back translation desde el portugués al inglés.
city_from_French_to_English	Texto	Resultado de la back translation desde el francés al inglés.
city_from_Arabic_to_English	Texto	Resultado de la back translation desde el árabe al inglés.
city_from_Chinese_to_English	Texto	Resultado de la back translation desde el chino al inglés.
city_from_Korean_to_English	Texto	Resultado de la back translation desde el coreano al inglés.
city_from_Japanese_to_English	Texto	Resultado de la back translation desde el japonés al inglés.
city_from_Vietnamese_to_English	Texto	Resultado de la back translation desde el vietnamita al inglés.
city_from_Finnish_to_English	Texto	Resultado de la back translation desde el finés al inglés.
city_from_Tamil_to_English	Texto	Resultado de la back translation desde el tamil al inglés.
city_from_Swahili_to_English	Texto	Resultado de la back translation desde el suajili al inglés.
city_from_Turkish_to_English	Texto	Resultado de la back translation desde el turco al inglés.
city_from_Indonesian_to_English	Texto	Resultado de la back translation desde el indonés al inglés.

Tabla 4. Metadatos de la base de datos de las ciudades

4.1.3 Creación y enriquecimiento de la base de datos de países

La base de datos de países utilizada en este proyecto se construyó integrando múltiples fuentes confiables, con el objetivo de proporcionar una visión detallada, geográfica, lingüística y socioeconómica de cada país.

El punto de partida fue la base Country Annotation, mencionada y usada anteriormente. Esta estructura inicial fue clave para representar adecuadamente la diversidad lingüística y regional en el análisis.

Para enriquecer la base y adaptarla a los objetivos del estudio, se incorporó información demográfica, específicamente la población estimada para 2024, extraída del reporte *Countries in the World by Population (2025)* de la División de Población de las Naciones Unidas [34]. Aunque inicialmente se consideraron variables adicionales como densidad poblacional y cambio anual, estas fueron descartadas para mantener el enfoque en datos actuales y relevantes.

4.1.3.1 Integración de variables socioeconómicas y culturales

Con el fin de explorar posibles causas estructurales del parroquialismo nacional, la base fue complementada con variables socioeconómicas y culturales provenientes del World Values Survey (WVS) [35], una referencia reconocida en el análisis comparativo internacional. El World Values Survey es un proyecto de investigación global que recopila datos sobre valores, creencias y actitudes de las personas en diferentes países del mundo. A través de encuestas periódicas realizadas desde 1981, el WVS permite analizar cómo factores como la confianza social, la tolerancia cultural, las percepciones de corrupción, las actitudes hacia la competencia o la apertura hacia la migración varían entre sociedades y cómo estos factores influyen en fenómenos sociales, económicos y políticos.

La selección de variables utilizadas en este proyecto respondió a criterios teóricos planteados en el marco conceptual, priorizando indicadores directamente vinculados a dimensiones

relevantes para entender el parroquialismo, como la confianza en otros grupos, la apertura hacia lo externo y la desigualdad percibida.

Dado que cada país cuenta con múltiples respuestas individuales para cada pregunta, fue necesario agrupar los datos por país y calcular el promedio de las respuestas para cada bloque de interés. Este procedimiento permitió obtener un valor representativo por país, facilitando su integración con el análisis de parroquialismo lingüístico posterior.

4.1.3.1.1 Trust Index

Para la construcción del índice de confianza (trust Index), se utilizaron las preguntas Q58 a Q63 del World Value Survey, que abordan el nivel de confianza que las personas tienen en diferentes grupos sociales. Estas preguntas evalúan la confianza en miembros del endogrupo (in-group), como la familia, el vecindario y personas conocidas personalmente, así como en miembros del exogrupo (out- group), como personas que se encuentran por primera vez, personas de otra religión y personas de otra nacionalidad.

Para calcular el Trust Index, las respuestas se agruparon en dos bloques. El primero, denominado *Trust in-group*, corresponde al promedio de las preguntas Q58, Q59 y Q60, que miden el nivel de confianza hacia personas cercanas, amigos y familiares. El segundo, llamado *Trust out-group*, incluye las preguntas Q61, Q62 y Q63, enfocadas en evaluar la confianza hacia personas externas o de grupos diferentes al del encuestado. Esta división permitió distinguir cómo varía la confianza dependiendo del grado de cercanía o pertenencia social del otro.

En el contexto del análisis de parroquialismo, un mayor nivel de confianza hacia el in-group en comparación con el out-group podría interpretarse como un indicador de mayor parroquialismo social: cuanto más acentuada sea la diferencia entre la confianza en los propios y en los externos, mayor podría ser la tendencia a favorecer al propio grupo, reflejando patrones

similares a los observados en el lenguaje y en las dinámicas de mención geográfica analizadas en este estudio.

Cada pregunta ofrece cuatro opciones de respuesta, que medirían el nivel de confianza:

I'd like to ask you how much you trust people from various groups. Could you tell me for each whether you trust people from this group completely, somewhat, not very much or not at all? (*Read out and code one answer for each*):

		Trust completely	Trust somewhat	Do not trust very much	Do not trust at all
Q58	Your family	1	2	3	4
Q59	Your neighborhood	1	2	3	4
Q60	People you know personally	1	2	3	4
Q61	People you meet for the first time	1	2	3	4
Q62	People of another religion	1	2	3	4
Q63	People of another nationality	1	2	3	4

Figura 4. Preguntas índices de confianza WVS

4.1.3.1.2 Corruption Index

El Corruption Index fue construido a partir de la pregunta Q112 del World Value Survey, la cual busca medir la percepción de corrupción en el país del encuestado. La pregunta se formula de la siguiente manera:

"¿Cómo calificaría la corrupción en su país en una escala del 1 al 10, donde 1 significa 'no hay corrupción en mi país' y 10 significa 'hay abundante corrupción en mi país'?"

Las respuestas están organizadas en una escala de 10 puntos:

- **1**: No hay corrupción en el país.
- **10**: Hay abundante corrupción en el país.

Para calcular el Corruption Index, el promedio por país representa la percepción general de corrupción, brindando un valor único para cada uno. Por ejemplo, si la mayoría de los encuestados en un país respondieron entre 8 y 10, se interpretaría una percepción alta de corrupción; mientras que promedios cercanos a 1 o 2 indicarían una percepción baja. Este índice es especialmente relevante para este proyecto, ya que una percepción elevada de corrupción podría asociarse con mayores niveles de parroquialismo, al reflejar una desconfianza más profunda hacia lo externo y lo institucional.

4.1.3.1.3 Competition Index

El Competition Index fue calculado utilizando las respuestas de la pregunta Q109 del World Values Survey (WVS). Esta pregunta mide la percepción de las personas sobre la competencia, presentando dos extremos en una escala de 1 a 10, donde:

- **1** significa "Competition is good" (La competencia es buena)
- **10** significa "Competition is harmful" (La competencia es perjudicial)

Los participantes debían seleccionar un número entre 1 y 10 para indicar su nivel de acuerdo con cada afirmación, lo que permitió cuantificar sus percepciones sobre la competencia. Un promedio cercano a 1 indica una percepción positiva de la competencia, considerándola beneficiosa y un promedio cercano a 10 sugiere que la competencia es vista de manera negativa, como algo perjudicial.

4.1.3.1.4 Migration Index

El Migration Index se calculó a partir de la pregunta Q121 del World Values Survey, en la cual se pedía a los encuestados evaluar el impacto de los inmigrantes en el desarrollo de su país. Las respuestas estaban codificadas en una escala de 1 a 5, donde 5 representaba "muy bueno" y 1 "muy malo". Este formato permitió capturar las percepciones de las personas sobre si los inmigrantes contribuían positiva o negativamente al progreso nacional.

Un promedio cercano a 5 indica que, en general, los inmigrantes son percibidos como una influencia positiva en el desarrollo del país, mientras que un promedio más cercano a 1 refleja una percepción negativa sobre su impacto. De esta manera, cada país quedó asociado a un único valor que resume las percepciones generales de sus ciudadanos sobre el impacto de los inmigrantes.

Este índice es relevante para el análisis del nivel de parroquialismo, ya que una percepción positiva sobre los inmigrantes puede estar relacionada con una mayor apertura cultural y menor

parroquialismo, mientras que una percepción negativa podría correlacionarse con actitudes más cerradas y excluyentes.

4.1.3.1.5 Openness Index

El Openness Index se construyó a partir de las respuestas a las preguntas Q19, Q21, Q22, Q23 y Q26 del World Values Survey. Estas preguntas buscan conocer las actitudes de los encuestados hacia distintos grupos sociales y culturales, pidiéndoles que indiquen si no les gustaría tener como vecinos a personas de diferentes razas (Q19), inmigrantes o trabajadores extranjeros (Q21), personas de una religión diferente (Q22), homosexuales (Q23) y personas que hablan otro idioma (Q26). Las respuestas estaban codificadas como "1" si el grupo era mencionado como no deseado y "2" si no lo mencionaban:

(SHOW CARD 3)
On this list are various groups of people. Could you please mention any that you would not like to have as neighbors?
(*Code an answer for each group*):

		Mentioned	Not mentioned
Q18	Drug addicts	1	2
Q19	People of a different race	1	2
Q20	People who have AIDS	1	2
Q21	Immigrants/foreign workers	1	2
Q22	Homosexuals	1	2
Q23	People of a different religion	1	2
Q24	Heavy drinkers	1	2
Q25	Unmarried couples living together	1	2
Q26	People who speak a different language	1	2

Figura 5. Preguntas para el índice de apertura WVS

Un valor promedio alto indica una mayor apertura cultural y social, reflejando una menor aversión hacia estos grupos, mientras que un valor promedio más bajo sugiere actitudes menos abiertas.

4.1.3.2 Incorporación de indicadores adicionales

Además de las variables obtenidas del World Values Survey, se consideró importante agregar otros indicadores que ayudaran a comprender mejor el contexto estructural de cada país y cómo este podría influir en los niveles de parroquialismo. En particular, se incluyeron tres variables

clave que permiten observar dimensiones como el desarrollo humano, la conectividad digital y la desigualdad económica.

La primera fue el Índice de Desarrollo Humano (HDI), extraído del *Human Development Report* del Programa de las Naciones Unidas para el Desarrollo (PNUD) [36]. Este índice combina aspectos como la esperanza de vida al nacer, el acceso a la educación y el ingreso nacional bruto per cápita, brindando una visión general del bienestar y las condiciones de vida en cada país. Al integrarlo, se busca explorar si existe alguna relación entre el nivel de desarrollo de un país y su apertura hacia otras geografías y culturas.

La segunda variable añadida fue el porcentaje de penetración de internet, tomado de *World Population Review* [37]. Este dato refleja qué porcentaje de la población tiene acceso a internet en cada país, lo que puede influir en la exposición a diferentes contenidos culturales y, por tanto, en la diversidad de referencias geográficas presentes en el lenguaje.

Por último, se incorporó el Índice de Gini, obtenido del Banco Mundial [38], que mide la desigualdad en la distribución del ingreso. Un valor cercano a 0 representa mayor equidad, mientras que uno más alto indica una mayor concentración de la riqueza. La inclusión de esta variable parte de la hipótesis de que en sociedades con mayor desigualdad podría haber menos apertura cultural, lo cual se reflejaría en una mayor tendencia a enfocarse en lo local.

El resultado fue una base de datos robusta, que ofrece una perspectiva más completa sobre los países analizados y permite interpretar el parroquialismo nacional considerando no solo factores culturales, sino también contextos estructurales como el desarrollo, la conectividad y la desigualdad.

Para complementar la descripción de la base de datos de países y brindar mayor claridad sobre su estructura, a continuación, se presenta una tabla con los metadatos correspondientes. Esta tabla resume las principales columnas incluidas en la base final, explicando el contenido y la función

de cada una de ellas dentro del análisis:

COLUMNA	TIPO DE DATO	DESCRIPCIÓN
code	Texto	Código del país en formato ISO 3166-1 alfa-2.
name	Texto	Nombre oficial del país en inglés.
continent	Texto	Continente al que pertenece el país
languages	Texto	Lista de idiomas oficiales o cooficiales hablados en el país.
Population	Entero (int)	Población total estimada del país en 2024.
B_country_alpha	Texto	Código del país en formato ISO alfa-3.
corruption_index	Decimal (float)	Índice promedio de percepción de corrupción.
competition_index	Decimal (float)	Índice promedio de percepción de la competencia.
migration_index	Decimal (float)	Índice promedio sobre la percepción del impacto de los inmigrantes.
openness_index	Decimal (float)	Índice promedio de apertura cultural.
HDI	Decimal (float)	Índice de Desarrollo Humano.
internet.penetration	Decimal (float)	Porcentaje de la población con acceso a internet.
gini_index	Decimal (float)	Índice de Gini que mide la desigualdad.
trust_index_in_group	Decimal (float)	Promedio de confianza en personas del grupo cercano.
trust_index_out_group	Decimal (float)	Promedio de confianza en personas externas.

Tabla 5. Metadatos de la base de datos de los países

4.2 Análisis de frecuencia de ciudades en word embeddings multilingües

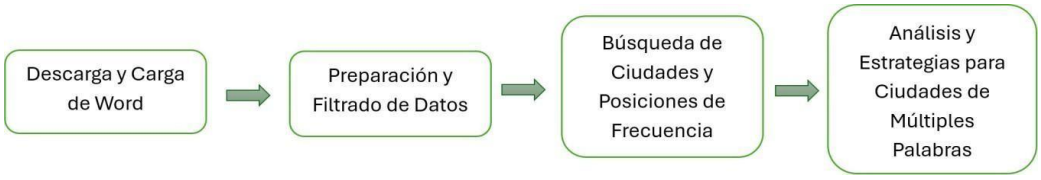


Figura 6. Segunda etapa de la metodología

Una vez finalizada la construcción y enriquecimiento de las bases de datos de ciudades y países, se avanzó hacia la siguiente etapa: la extracción y análisis de la frecuencia relativa de las ciudades en distintos idiomas, utilizando modelos de word embeddings preentrenados. Esta fase fue clave para cuantificar el nivel de parroquialismo lingüístico, tomando como referencia la posición que cada ciudad ocupa en el vocabulario del modelo, ordenado según frecuencia de aparición.

Para este análisis, se utilizaron archivos .bin correspondientes a los modelos preentrenados de fastText para 157 idiomas, disponibles públicamente a través de la plataforma Word Vectors for 157 Languages, desarrollada por Facebook AI Research (FAIR) [39]. Estos modelos no fueron entrenados desde cero en este proyecto, sino que se descargaron ya entrenados y se utilizaron

como insumo para los análisis.

Los modelos fueron entrenados originalmente por el equipo de FAIR utilizando grandes corpus multilingües de texto, principalmente provenientes de Wikipedia y Common Crawl.

- Wikipedia aportó textos bien estructurados y con contenido enciclopédico, lo cual garantiza una cobertura sólida del vocabulario general y de nombres propios relevantes como ciudades y países.
- Common Crawl, por su parte, es una colección masiva de textos extraídos automáticamente de páginas web públicas de todo el mundo, lo que aporta volumen, diversidad y representatividad del lenguaje real utilizado en Internet.

Ambas fuentes combinadas permitieron construir vectores robustos y multilingües, capaces de capturar relaciones semánticas entre palabras, incluso en idiomas con menor disponibilidad de recursos.

Los modelos utilizan la arquitectura fastText, que mejora a Word2Vec al incorporar información de subpalabras o n-gramas. Gracias a esto, el modelo puede construir representaciones para palabras nuevas o con errores ortográficos combinando fragmentos internos de palabras, lo cual es especialmente útil para trabajar con nombres de ciudades poco frecuentes, variantes ortográficas o palabras compuestas.

Aunque el repositorio ofrece una amplia variedad de lenguas, en este proyecto se trabajó únicamente con 14 idiomas seleccionados, priorizando aquellos con diversidad geográfica, lingüística y amplia cobertura poblacional.

4.2.1 Procesamiento de frecuencia y manejo de ciudades de múltiples palabras

En esta etapa, el objetivo fue identificar la posición de cada ciudad dentro de los archivos de los 14 idiomas seleccionados. Estos archivos contienen modelos de word embeddings entrenados

con fastText, donde las palabras están ordenadas por frecuencia: las más comunes aparecen primero y las menos frecuentes al final. Esta posición funciona como indicador del nivel de presencia que una ciudad tiene en el lenguaje cotidiano de cada idioma, siendo clave para calcular el nivel de parroquialismo nacional.

Cuando el nombre de la ciudad aparecía exactamente igual en el archivo, se registraba directamente su posición. Sin embargo, muchos nombres compuestos por dos o más palabras no se encontraban de forma literal en los vocabularios, debido a que los modelos de embeddings tienden a aprender palabras individuales (tokens únicos). Para solucionar este problema, se implementó un enfoque robusto que combinó modelos de lenguaje BERT, búsquedas individuales en el vocabulario y un sistema de penalización semántica.

La dificultad principal era que BERT, al ser multilingüe, necesitaba contexto para reconocer correctamente el idioma de entrada. Dado que los nombres de ciudades son generalmente nombres propios, se introdujo un prefijo contextual traducido al idioma correspondiente, como "En la ciudad de", que ayudaba al modelo a identificar el idioma. Este prefijo no se consideraba en el cálculo de posiciones, sirviendo únicamente para mejorar el reconocimiento:

Idioma	Prefijo usado	Ejemplo completo
Español	En la ciudad de	En la ciudad de San José
Inglés	In the city of	In the city of San José
Árabe	في مدينة	في مدينة القاهرة
Japonés	「都市」	都市 東京
Chino	在城市	在城市 北京

Figura 7. Prefijo usado para los diferentes idiomas

Aplicando la técnica de Masked Language Modeling (MLM), se enmascaraban sucesivamente

las palabras del nombre de la ciudad para predecir su probabilidad en el contexto generado. Por ejemplo, en "San José", se evaluaba la predicción de "José" tras "San [MASK]". A partir de estas probabilidades, se calculaba una penalización inversa: a menor probabilidad, mayor penalización. La penalización total se sumaba a la posición de la palabra base (la primera palabra) en el archivo para estimar la posición completa de la ciudad compuesta. Este procedimiento se aplicaba a nombres de dos o más palabras, como "San José del Cabo", donde se evaluaban múltiples predicciones encadenadas.

La implementación técnica de esta metodología se realizó mediante la función *predict_next_word*, utilizando el modelo *bert-base-multilingual-cased* de HuggingFace. Esta función tokenizaba la frase, obtenía los logits predichos por BERT, aplicaba softmax para convertirlos en probabilidades, y calculaba la penalización correspondiente para cada palabra esperada.

De forma paralela, se utilizó otro método alternativo: buscar todas las palabras componentes de la ciudad en el archivo y quedarse con la posición más alta (menos frecuente). Para cada ciudad compuesta, se conservó la posición más conservadora entre ambos métodos.

Respecto a la tokenización, se diferenciaron dos estrategias según el idioma:

- Para lenguas con separación natural entre palabras (español, inglés, francés, etc.), se utilizó una separación básica por espacios.
- Para idiomas sin espacios naturales (chino, japonés, coreano), se recurrió a tokenizadores específicos del idioma para identificar correctamente las unidades léxicas.

Idioma	¿Espacios naturales?	¿Tokenizador especial usado?	¿Por qué?
Chino	No	Jieba	No hay espacios
Japonés	No	MeCab	No hay espacios
Coreano	Sí (parcial)	KoNLPy (tokenizadores dentro de KoNLPy)	Hay espacios, pero se necesita segmentar mejor
Árabe	Sí	AraBERT Tokenizer	Hay espacios, pero morfología compleja

Figura 8. Idiomas con tokenizador específico

Es importante resaltar que durante la búsqueda en los archivos se utilizaban tokenizadores especializados, mientras que para las predicciones con BERT siempre se empleaba su tokenizador propio, basado en subpalabras (WordPiece), garantizando coherencia en las entradas.

Finalmente, se construyó una matriz de posiciones, donde cada fila corresponde a una ciudad y cada columna a un idioma, registrando la posición relativa de cada nombre en su respectivo vocabulario. A menor posición, mayor frecuencia de aparición. Esta matriz constituye la base del análisis comparativo de parroquialismo lingüístico realizado posteriormente:

city	en_posiciones	sp_posiciones	ch_posiciones	sw_posiciones	ta_posiciones	tr_posiciones	fn_posiciones	jp_posiciones	ko_posiciones	viet_posiciones	ar_posiciones
Tokyo	5872	7624	4662	4852	21427	8728	23280	211	681	4183	7813
Delhi	4853	22845	11384	6482	2558	29739	48244	23766	6829	20922	30765
Shanghai	9021	26129	344	9130	47973	31958	22204	5165	867	7914	26213
Dhaka	22451	217225	82180	19070	22134	117593	247733	59505	71855	55995	64010
Sao Paulo	24092	1136	27780	5076	26236	27514	16688	24286	18481	70471	17991
Cairo	12249	14915	27651	5852	31985	13538	72933	14539	15093	16759	614
Mexico City	2258	810	36649	2565	12063	9028	24562	48184	1968	4650	25034
Beijing	7223	11110	172	5001	21351	11795	38503	5261	879	4145	10633
Mumbai	6388	30581	35115	8969	1545	74960	57816	43246	35831	20714	27677
Osaka	20495	23167	4818	12672	151229	31918	38314	428	1216	13036	31191
Chongqing	45959	76342	1451	84346	177294	350718	153285	31453	29664	32068	182359
Karachi	21056	81600	65375	9797	26069	106496	206055	79522	120255	41026	43458
Kinshasa	67647	99986	163932	3885	163932	191384	165866	103983	204840	64198	95616
Lagos	17041	10952	98664	3457	154897	67357	92270	79325	36704	47668	67383
Istanbul	12656	479219	26448	3851	41901	70	25288	23117	2979	14775	6879
Buenos Aires	15101	854	48322	6101	271710	29893	34561	30170	63096	259812	23145
Kolkata	15268	32627	61341	8407	5449	130744	375263	73906	77960	34104	156095
Manila	12339	22288	30532	13420	28598	61945	76390	17933	16123	10573	40173
Guangzhou	26711	14439	1114	23404	203616	60378	104035	19169	11820	11502	115388
Tianjin	42219	71455	1374	49341	318362	121022	138260	16721	24155	7350	97822
Lahore	21769	80608	122561	35306	17833	87529	187236	85124	130813	43771	48600
Bangalore	8803	64785	87820	16452	4013	138039	161816	66618	89514	26607	78020
Rio de Janeiro	18712	7770	38151	9290	173913	32923	20698	22148	22769	16278	18346
Shenzhen	28652	54890	1008	53782	771333	619623	92087	42537	44143	64459	230120
Moscow	6019	5483	5941	3458	14084	3213	7040	9866	4770	42376	4096

Figura 9. Posiciones de ciudades en los idiomas

4.3 Construcción de la métrica de parroquialismo y visualización de resultados

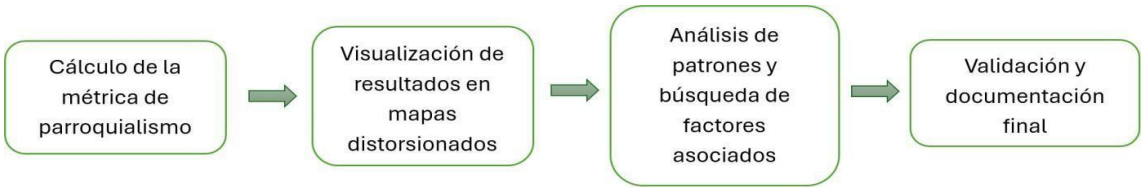


Figura 10. Tercera etapa de la metodología

4.3.1 Creación del primer índice de parroquialismo

Una vez obtenidas las posiciones de las ciudades en los 14 idiomas seleccionados, a partir de los modelos de word embeddings entrenados con fastText, se procedió a construir una métrica que permitiera cuantificar el nivel de parroquialismo lingüístico de cada idioma. Esta métrica busca capturar qué tanto tiende cada lengua a mencionar con mayor frecuencia las ciudades que comparten su mismo idioma, en comparación con ciudades de otros contextos. En otras palabras, se mide si un idioma "mira hacia adentro" o si mantiene referencias amplias y diversas hacia el exterior.

La lógica parte de la Ley de Zipf, que establece que la posición de una palabra en el vocabulario es inversamente proporcional a su frecuencia: cuanto más baja es la posición, más frecuente es su mención. Así, ciudades ubicadas en posiciones iniciales son mencionadas más recurrentemente en el idioma analizado.

Siguiendo esta idea, se diseñó un índice de parroquialismo por idioma basado en las posiciones promedio de las ciudades:

1. Para cada idioma, se identificaron las ciudades cuyo idioma principal coincide con el analizado (por ejemplo, Bogotá o Lima para el español) y se calculó el promedio de sus posiciones en el archivo correspondiente.
2. Se calcularon también los promedios de las posiciones de las ciudades cuyo idioma predominante es diferente al analizado (por ejemplo, Londres, Tokio o Moscú en el caso del español).
3. Se tomó la relación entre ambos promedios (promedio externo dividido entre promedio interno), obteniendo así una medida de cuánto más "lejos" están, en términos de frecuencia, las ciudades extranjeras respecto a las locales.
4. Para facilitar la comparación entre idiomas y suavizar la escala, se aplicó una transformación

logarítmica natural sobre el índice, generando una nueva variable denominada indicelog.

Este procedimiento fue implementado en R mediante una estructura iterativa que recorrió los 14 idiomas, utilizando para cada uno la columna de posiciones correspondiente (por ejemplo, sp_posiciones, en_posiciones). Para cada idioma, se filtraron las ciudades locales y externas, se calcularon los promedios respectivos y se almacenaron tanto el índice crudo como su versión log-transformada.

El resultado fue un DataFrame que consolidó, para cada idioma, la siguiente información: el número de ciudades locales utilizadas, el promedio de posiciones internas, el promedio de posiciones externas, el índice base y el índice logarítmico (indicelog):

idioma	n_ciudades	promedio_interno	promedio_externo	indice	indicelog
English	201	87334.100	229172.99	2.624095	0.9647361
Spanish	78	28289.987	336467.84	11.893531	2.4759946
Chinese	234	220844.331	261912.44	1.185960	0.1705522
Swahili	12	17508.200	72707.07	4.152744	1.4237692
Tamil	1	2084.000	283648.74	136.107840	4.9134475
Turkish	13	10204.846	374145.93	36.663555	3.6017832
Finnish	1	206.000	501414.31	2434.050069	7.7973118
Japanese	9	5043.333	197633.08	39.186996	3.6683449
Korean	15	13345.067	266961.60	20.004516	2.9959581
Vietnamese	6	13855.167	115169.17	8.312362	2.1177439
Arabic	58	43690.293	298578.48	6.833978	1.9219069
Indonesian	20	3336.900	135573.31	40.628520	3.7044703
Portuguese	39	25314.179	326442.19	12.895626	2.5568882
French	27	28691.407	284061.55	9.900579	2.2925932

Figura 11. Dataframe inicial con primer índice

4.3.2 Creación del segundo índice de parroquialismo

Aunque el primer índice de parroquialismo (indicelog) permitió identificar diferencias importantes entre idiomas, su análisis más detallado reveló una limitación: el número de ciudades internas disponibles por idioma parecía estar afectando sus valores. En particular, los idiomas con menos ciudades tendían a mostrar índices artificialmente más altos, probablemente debido a la inestabilidad estadística generada por tamaños de muestra pequeños.

Para evaluar este posible sesgo, se calculó la correlación de Pearson entre indicelog y el número de ciudades internas (n_ciudades) por idioma. El resultado fue una correlación negativa moderada y significativa (r = -0.627, p = 0.016), lo que confirma que, a menor número de ciudades, mayor tiende a ser el índice. En la práctica, esto significaba que el índice inicial favorecía injustamente a idiomas con bases de datos más pequeñas.

Para corregir este efecto, se propuso una versión ajustada del índice. El nuevo índice (new) se construyó multiplicando indicelog por el logaritmo natural del número de ciudades internas, añadiendo 1 para evitar problemas con log(0):

$$New = indicelog * \log(1 + n_{ciudades}) \quad (11)$$

Esta transformación ofrece varias ventajas importantes. Por un lado, modera el impacto que podría tener el hecho de contar con un número muy alto o bajo de ciudades en la base de datos, evitando que los extremos distorsionen los resultados. Además, otorga un mayor peso a los idiomas que cuentan con bases de datos más amplias y promedios más estables, lo que fortalece la consistencia estadística del análisis. Al mismo tiempo, reduce la inflación de valores en idiomas con pocos datos, sin llegar a excluirlos completamente del estudio, lo cual permite mantener una representación amplia y diversa.

De esta manera, el índice *new* conserva la lógica conceptual del *indicelog*, pero incorpora un ajuste estadístico que mejora su robustez y su comparabilidad entre idiomas. Este nuevo índice fue utilizado como base para los análisis posteriores de correlación con variables socioeconómicas y culturales, permitiendo interpretar el fenómeno del parroquialismo de una manera más justa, equilibrada y metodológicamente sólida:

idioma	n_ciudades	promedio_interno	promedio_externo	indicelog	new
English	201	87334.100	229172.99	0.9647361	5.1210776
Spanish	78	28289.987	336467.84	2.4759946	10.8187295
Chinese	234	220844.331	261912.44	0.1705522	0.9311441
Swahili	12	17508.200	72707.07	1.4237692	3.6518960
Tamil	1	2084.000	283648.74	4.9134475	3.4057423
Turkish	13	10204.846	374145.93	3.6017832	9.5053124
Finnish	1	206.000	501414.31	7.7973118	5.4046847
Japanese	9	5043.333	197633.08	3.6683449	8.4466764
Korean	15	13345.067	266961.60	2.9959581	8.3065595
Vietnamese	6	13855.167	115169.17	2.1177439	4.1209393
Arabic	58	43690.293	298578.48	1.9219069	7.8366472
Indonesian	20	3336.900	135573.31	3.7044703	11.2783429
Portuguese	39	25314.179	326442.19	2.5568882	9.4320523
French	27	28691.407	284061.55	2.2925932	7.6393896

Figura 12. Dataframe actualizado con integración del segundo índice

4.3.3 Creación del tercer índice de parroquialismo

Además de los índices anteriores, se propuso una tercera aproximación para medir el nivel de parroquialismo en cada idioma, basada en analizar cómo jerarquizan los países a través de las ciudades que mencionan más frecuentemente, y qué tan alineada está esa jerarquía con un escenario ideal no parroquialista.

La idea central es que, si un idioma prioriza principalmente ciudades de su propio país, su ranking estará sesgado hacia lo local. En cambio, un idioma más abierto tendería a destacar ciudades de países grandes y relevantes globalmente, reflejando un patrón más universal.

Para construir este índice:

- Se agruparon las ciudades por país y se calculó el promedio de sus posiciones en el archivo de cada idioma. Un promedio bajo indica que las ciudades de ese país son mencionadas con mayor frecuencia.
- Con estos promedios se elaboró un ranking de países para cada idioma: el país con el promedio más bajo ocupó la posición 1, el siguiente la posición 2, y así sucesivamente.
- Para evitar sesgos derivados de países con muy pocas ciudades, se estableció un filtro: solo se consideraron países que tuvieran al menos cuatro ciudades en la base de datos.

Este proceso se repitió para cada idioma, generando una matriz de rankings: una fila por país, y una columna por idioma (English_rank, Spanish_rank, etc.):

country	language_selected	English_rank	Spanish_rank	Chinese_rank	Swahili_rank	Tamil_rank	Turkish_rank	Finnish_rank	Japanese_rank	Korean_rank	Vietnamese_rank	Arabic_rank	Indonesian_rank
France	French	12	13	5	12	35	9	6	5	7	5	16	4
Belgium	French	10	10	14	61	15	17	10	12	5	28	18	22
Spain	Spanish	6	2	6	4	40	18	3	3	3	2	6	2
Portugal	Portuguese	9	8	9	60	17	5	8	7	9	11	15	12
Algeria	Arabic	24	23	24	11	26	4	20	23	35	23	5	5
Canada	English	4	16	19	14	28	20	2	14	19	15	33	17
Netherlands	Dutch	8	24	4	6	29	8	7	8	10	8	27	6
Italy	Italian	13	12	12	15	38	13	11	10	12	4	20	14
Poland	Polish	16	14	10	38	12	14	13	11	11	16	22	27
Costa Rica	Spanish	2	1	22	5	6	6	1	16	68	25	12	3
Australia	English	3	17	2	10	16	15	14	6	6	6	26	9
United Kingdom	English	5	28	23	8	33	34	9	20	14	12	25	7
Israel	Arabic	11	15	13	18	27	10	15	17	21	31	3	11
Syria	Arabic	59	20	20	23	30	7	36	18	23	33	9	37
Germany	German	7	21	1	3	24	1	4	2	2	10	14	18
Kenya	English	17	34	31	2	68	29	24	25	34	27	29	35
Japan	Japanese	19	29	11	20	36	21	21	1	1	9	35	10
Bolivia	Spanish	21	4	28	16	3	28	23	58	28	21	34	26
Egypt	Arabic	18	30	3	40	19	22	30	15	16	46	4	21
Sudan	Arabic	44	27	17	43	54	19	26	19	40	14	1	32
Republic of the Congo	French	57	31	36	49	2	41	50	55	54	32	44	30
Kazakhstan	Kazakh	68	38	32	21	41	30	18	45	24	24	52	59
Morocco	Arabic	29	25	29	34	55	32	27	29	29	34	10	46
United Arab Emirates	Arabic	15	33	18	9	7	26	19	31	20	61	7	19
Peru	Spanish	23	5	59	51	20	31	29	40	32	30	46	29

Figura 13. Rankings de paises por idioma

Como referencia neutral, se construyó un ranking ideal basado en la población de los países, bajo la hipótesis de que, en ausencia de sesgos, los países más poblados deberían tener mayor visibilidad.

La distancia entre el ranking real de cada idioma y el ranking ideal se midió utilizando el Error Cuadrático Medio (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{idioma,i} - r_{ideal,i})^2} \quad (12)$$

Donde:

- $r_{idioma,i}$ es el ranking de un país i en un idioma determinado.
- $r_{ideal,i}$ es su posición en el ranking ideal por población.
- n es el número de países comunes entre ambos rankings.

Un RMSE bajo indica que el idioma organiza los países de manera similar a la población mundial, sugiriendo menor parroquialismo. Un RMSE alto revela una desviación significativa, reflejando un sesgo más fuerte hacía lo propio o hacia ciertos contextos culturales.

A diferencia de los primeros índices, que miden directamente la diferencia entre ciudades internas y externas, este enfoque evalúa el patrón general de priorización de países en cada idioma, ofreciendo una perspectiva complementaria y más estructural del fenómeno.

Una vez construido el RMSE para cada idioma, se consolidaron los resultados en una única tabla comparativa junto con los otros dos índices (indicelog y new), lo que permitió tener una visión integrada de los niveles de parroquialismo desde distintas dimensiones de análisis:

RMSE_idioma	RMSE_valor	Indicelog_idioma	Indicelog_valor	New_idioma	New_valor
Vietnamese	27.111	English	0.965	Tamil	3.406
Korean	27.598	Swahili	1.424	Swahili	3.652
English	27.763	Arabic	1.922	Vietnamese	4.121
Indonesian	29.036	Vietnamese	2.118	English	5.121
Japanese	29.112	French	2.293	Finnish	5.405
Turkish	29.453	Spanish	2.476	French	7.639
Portuguese	29.590	Portuguese	2.557	Arabic	7.837
Swahili	29.637	Korean	2.996	Korean	8.307
Chinese	29.724	Turkish	3.602	Japanese	8.447
Finnish	29.836	Japanese	3.668	Portuguese	9.432
Arabic	30.329	Indonesian	3.704	Turkish	9.505
Tamil	31.028	Tamil	4.913	Spanish	10.819
Spanish	31.631	Finnish	7.797	Indonesian	11.278

Figura 14. Orden de idiomas por índice

Los tres índices muestran correlaciones bajas entre sí (todas con p-valores superiores a 0.49), lo cual sugiere empíricamente que están capturando perspectivas diferentes del mismo fenómeno. Estas cifras refuerzan la idea de que cada índice ofrece una lectura distinta del parroquialismo: uno evalúa la cercanía a un patrón universal (RMSE), otro el sesgo entre lo local y lo externo (indicelog), y el tercero ajusta ese sesgo considerando el número de ciudades disponibles en el idioma (new). Esta diversidad analítica fortalece la robustez del enfoque general:

Correlaciones Entre Índices

	Índice 1	Índice 2	Correlación	p-valor
1	RMSE	New	0.137	0.64
2	RMSE	Indicelog	0.147	0.615
3	Indicelog	New	0.201	0.491

Figura 15. Correlaciones entre índices

En las siguientes secciones, estos índices serán explorados con mayor profundidad para determinar cuáles capturan de forma más útil el fenómeno en estudio y cómo pueden ser utilizados de manera efectiva en el análisis correlacional con variables socioeconómicas y culturales por país.

4.3.4 Análisis de índices con variables explicativas

Una vez construidos los tres índices de parroquialismo (indicelog, new y RMSE), el siguiente paso metodológico consistió en evaluar qué tan bien se alinean con las variables socioeconómicas, culturales y de confianza recogidas en la base de datos de países. El objetivo principal era determinar cuál o cuáles de los tres índices resultaban como los más útiles y representativos para explicar el fenómeno del parroquialismo nacional, siguiendo la lógica teórica planteada en el marco teórico.

En concreto, se calcularon las correlaciones de cada índice con las siguientes variables:

- Nivel de percepción de corrupción (corruption_index)
- Actitudes hacia la competencia (competition_index)
- Opinión sobre el impacto de la migración (migration_index)
- Apertura cultural (openness_index)
- Nivel de desarrollo humano (HDI)
- Porcentaje de penetración de internet (internet.penetration)
- Desigualdad económica (gini_index)

- Confianza en el endogrupo (trust_index_in_group)
- Confianza en el exogrupo (trust_index_out_group)

Cada correlación se acompañó de su respectivo valor p y del intervalo de confianza al 95%:

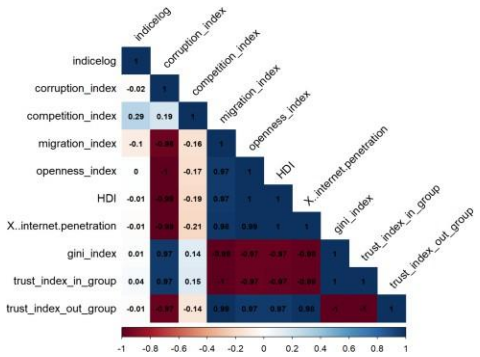


Figura 16. Correlación indicelog vs. variables explicativas

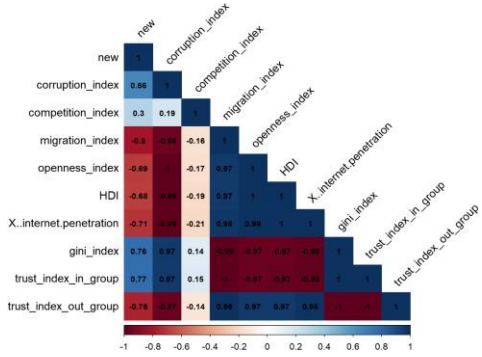


Figura 17. Correlación índice new vs. variables explicativas

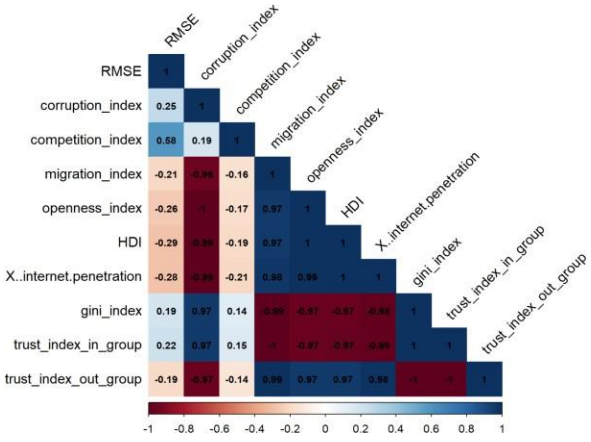


Figura 18. Correlación índice RMSE vs. variables explicativas

El indicelog presentó correlaciones muy bajas y no significativas con todas las variables analizadas, lo que limita su utilidad desde una perspectiva explicativa, ya que, no logró capturar adecuadamente las dinámicas sociales o económicas asociadas al parroquialismo. Además, como ya se había advertido, el indicelog mostró una correlación negativa significativa con el número de ciudades por idioma, evidenciando un sesgo estructural que afecta su estabilidad comparativa.

En contraste, el índice new, que ajusta el indicelog para controlar el número de ciudades, demostró correlaciones altas y estadísticamente significativas con varias variables clave. Estas correlaciones no solo son estadísticamente significativas, sino que también se alinean de manera coherente con las hipótesis teóricas, consolidando al índice new como la métrica más robusta para analizar el parroquialismo nacional.

Por su parte, el índice RMSE, que mide la distancia entre el ranking real de países generado por cada idioma y un ranking ideal basado en población, mostró algunas correlaciones moderadas, aunque menos consistentes. La más relevante fue con la percepción de competencia ($r = 0.58$, $p = 0.028$), lo cual sugiere que este índice podría capturar ciertos patrones estructurales del discurso, relacionados con dinámicas de rivalidad geopolítica, pero no se vincula directamente con dimensiones como la apertura cultural, la confianza social o el desarrollo humano.

El análisis comparativo de los tres índices desarrollados permite destacar sus fortalezas y limitaciones. El indicelog, pese a su valor conceptual como medida de parroquialismo basado en la frecuencia relativa de menciones locales frente a externas, se ve fuertemente afectado por el número de ciudades presentes en cada idioma, limitando su capacidad para realizar comparaciones justas entre lenguas. Asimismo, no presenta ninguna relación significativa con las variables asociadas.

El índice RMSE, ofrece una representación robusta de cómo cada idioma jerarquiza países en relación con una cobertura ideal proporcional a la población. Esta propiedad lo hace especialmente útil para visualizar y comparar la estructura global de visibilidad territorial entre idiomas, aunque no permite explicar directamente las causas sociales o culturales del parroquialismo.

Finalmente, el índice new emerge como la mejor métrica explicativa del fenómeno. Su diseño

ajusta el indicelog mediante una transformación logarítmica que compensa las diferencias en el número de ciudades internas, lo que permite una comparación más equitativa entre idiomas. Además, fue el único índice que presentó correlaciones fuertes y coherentes con variables clave como la apertura cultural, la desigualdad o la confianza social, posicionándolo como el más adecuado para analizar la influencia de factores socioculturales en la intensidad del parroquialismo lingüístico.

Un elemento que refuerza la solidez de los índices new y RMSE es su independencia frente a la cantidad de ciudades analizadas por idioma. En el caso de new, la correlación fue de –0.38 ($p = 0.18$), y para RMSE, de –0.07 ($p = 0.79$), lo cual confirma que ambos pueden utilizarse sin introducir sesgos derivados del tamaño muestral.

En conjunto, estos hallazgos justifican la decisión metodológica de priorizar el uso combinado del índice new, para explicar las causas del parroquialismo con base en variables asociadas, y del índice RMSE, para representar de forma comparativa la distribución global de importancia asignada a los países. Esta estrategia dual ofrece una visión más completa del fenómeno, integrando tanto su dimensión explicativa como su estructura de visibilidad territorial a nivel internacional.

4.3.4.1 Modelos de regresión regularizada para identificación de variables relevantes

En este proyecto no solo se buscó construir índices que midieran el nivel de parroquialismo lingüístico en distintos idiomas, sino también entender cuáles variables socioeconómicas y culturales están más asociadas a dicho fenómeno. Para ello, se implementaron modelos de regresión regularizada como herramientas exploratorias, permitiendo identificar qué factores explicativos aportaban más información sobre el comportamiento del índice new, el cual fue considerado la métrica más robusta y ajustada para explicar este fenómeno en fases anteriores del

estudio.

Aunque el número total de idiomas disponibles como unidades de análisis fue pequeño (n = 14), el uso de estos modelos resulta estadísticamente válido y recomendable cuando el objetivo no es la predicción, sino la exploración de patrones y la identificación del peso relativo de las variables explicativas. En este sentido, la regresión regularizada se utilizó como estrategia analítica para aportar solidez al análisis y detectar relaciones consistentes entre las variables y el fenómeno del parroquialismo.

El modelo principal empleado fue la regresión LASSO (Least Absolute Shrinkage and Selection Operator), caracterizada por la penalización L1, que no solo reduce la magnitud de los coeficientes, sino que también puede llevarlos exactamente a cero. Esto permite realizar, de manera simultánea, selección automática de variables, descartando aquellas que no aportan explicaciones significativas. Esta propiedad resulta especialmente útil en contextos con pocas observaciones, como este estudio, para centrarse en las dimensiones más informativas.

Además del modelo LASSO, se ajustaron también:

- Modelo Ridge: Aplica penalización L2, que atenúa todos los coeficientes, pero no los elimina, manteniendo todas las variables en el modelo.
- Modelo Elastic Net: Combina las penalizaciones L1 y L2, permitiendo tanto reducción como eliminación de coeficientes. En este caso, se utilizó un valor de alpha = 0.5, que representa un balance entre ambas estrategias.

Los modelos fueron ajustados mediante la función cv.glmnet() del paquete glmnet en R, realizando validación cruzada automática para seleccionar el valor óptimo del parámetro de regularización lambda (λ). Este parámetro controla la intensidad de la penalización aplicada: cuanto mayor es λ, más fuertes son las restricciones sobre los coeficientes.

Se optó por utilizar una configuración de 5 folds, dividiendo la muestra en cinco subconjuntos

de tamaño lo más equilibrado posible. Esta decisión respondió a la necesidad de evitar conjuntos de validación demasiado pequeños, lo cual podría haber ocurrido con esquemas de mayor división, como 10 folds, y a la intención de mantener la estabilidad del proceso y mejorar la robustez del análisis. Es importante señalar que, en esta parte del proyecto, no se buscó construir un modelo predictivo, sino explorar de forma explicativa cuáles variables están más asociadas al nivel de parroquialismo observado. Por ello, no se utilizó una base de prueba, y se priorizó una validación cruzada interna. Esta decisión también se justifica por el número limitado de observaciones disponibles (n=14) y por la posibilidad de alta correlación entre variables explicativas.

En cuanto a la preparación de los datos, la matriz de variables explicativas, X, incluyó aquellas seleccionadas según el marco teórico, todas ellas estandarizadas (standardize = TRUE) para garantizar comparabilidad en la penalización aplicada durante el ajuste. La variable de respuesta, y, fue el índice *new*, utilizado para representar el nivel de parroquialismo lingüístico por idioma. La estandarización resultó ser un paso esencial, ya que las penalizaciones L1 (LASSO) y L2 (Ridge) dependen directamente de las magnitudes absolutas de los coeficientes. Sin este ajuste previo, las variables con escalas numéricas mayores habrían dominado artificialmente el proceso de ajuste, sesgando las interpretaciones.

Respecto a la interpretación de los modelos, en el caso de LASSO, los coeficientes exactamente iguales a cero fueron interpretados como indicadores de variables descartadas automáticamente por el modelo, mientras que las variables con coeficientes distintos de cero fueron consideradas como las más relevantes para explicar el fenómeno. En Ridge, todos los coeficientes se reducen, pero permanecen en el modelo, lo cual permite comparar sus magnitudes relativas dado que todas las variables han sido previamente estandarizadas. Por su parte, Elastic Net combina ambos enfoques: descarta algunas variables y mantiene otras con penalizaciones

parciales, ofreciendo una perspectiva híbrida que equilibra entre la selección estricta propia de LASSO y la atenuación propia de Ridge.

Como conclusión de esta etapa, el uso combinado de estos modelos de regresión regularizada permitió identificar de manera sistemática cuáles variables socioeconómicas y culturales están más estrechamente asociadas al nivel de parroquialismo detectado.

4.3.5 Representatividad y cobertura de los datos

Antes de avanzar hacia la visualización y el análisis de resultados, es importante presentar una breve caracterización de la cobertura combinada de las bases de datos utilizadas (la de ciudades y la de países anteriormente explicadas), en términos de su representatividad geográfica, lingüística y demográfica. Esta descripción permite dimensionar el alcance del estudio y reconocer posibles limitaciones en la cobertura.

En conjunto, estas bases integran información de 132 países y 802 ciudades, con una distribución especialmente amplia en regiones como Asia y África principalmente. Además, se consideraron 14 idiomas principales, seleccionados por su diversidad geográfica y lingüística, con especial representación del chino, inglés, español, árabe y portugués.

Desde la perspectiva demográfica, los idiomas con mayor número de hablantes representados son el inglés (2.9 mil millones), el chino (1.4 mil millones), el árabe (502 millones), el español (466 millones) y el francés (442 millones). Estos factores serán relevantes para interpretar los patrones de parroquialismo identificados en las siguientes etapas del análisis.

A continuación, se presenta un resumen de la representatividad general derivada de los conjuntos de datos analizados:

Categoría	Dato resumido
Países representados	132 países
Ciudades totales	802 ciudades
Distribución continental	Asia (481), África (109), Europa (78), América del Norte (65), América del Sur (62), Oceanía (6)
Idiomas principales	14 idiomas seleccionados
Idiomas con más ciudades	Chino (234), Inglés (201), Español (78), Árabe (58), Portugués (39)
Idiomas con más países	Inglés (21), Árabe (21), Español (20), Francés (18), Portugués (4)
Idiomas con más hablantes	Inglés (2.9 mil millones), Chino (1.4 mil millones), Árabe (502M), Español (466M), Francés (442M)

Figura 19. Representatividad de los datos

4.3.6 Visualización de resultados

Con el objetivo de representar de forma clara e intuitiva los patrones de parroquialismo lingüístico detectados, se desarrollaron distintos tipos de visualizaciones geográficas. Estas representaciones se basan en el ranking de países por idioma, calculado previamente a partir del promedio de posiciones de sus ciudades en los modelos de word embeddings.

Para facilitar la interpretación, el ranking fue invertido (1/ranking), de manera que valores más altos representaran mayor importancia o presencia de un país en un idioma dado. A partir de esta métrica, se generaron las siguientes visualizaciones:

4.3.6.1 Mapas coropléticos estáticos

Para clasificar los países según su nivel de importancia en cada idioma, se utilizó un esquema basado en quintiles. Se calculó el valor inverso del ranking (1/ranking) para cada país y luego se distribuyeron estos valores en cinco grupos de tamaño aproximadamente igual. Cada quintil representó un nivel de importancia relativa:

"Muy bajo", "Bajo", "Medio", "Alto" y "Muy alto", facilitando así la interpretación visual de los

mapas.

Se empleó una escala cromática de amarillo (baja importancia) a rojo (alta importancia), a partir de datos combinados entre el shapefile mundial (ne_10m_admin_0_countries.shp) y el ranking lingüístico procesado.

Para asegurar la unión correcta entre datos textuales y geográficos, se utilizó countrycode() para homologar los códigos ISO de los países. La visualización se generó con geom_sf() bajo un diseño minimalista (theme_minimal()), garantizando legibilidad y estética clara.

4.3.6.2 Mapa interactivo con leaflet

Con el fin de ofrecer una exploración más dinámica, se construyeron mapas interactivos utilizando leaflet, permitiendo al usuario consultar información detallada al pasar el cursor sobre cada país: nombre, ranking y grupo de importancia.

La escala cromática se mantuvo consistente con los mapas estáticos. Las ventanas emergentes fueron creadas mediante *htmltools*, y para países sin datos se empleó un color gris claro con la etiqueta "Sin datos de ranking". Todo el proceso fue automatizado mediante una función, que genera mapas para cada idioma.

La librería htmltools en R permite crear y manejar contenido HTML de manera programática dentro de aplicaciones y visualizaciones interactivas. En este proyecto, htmltools se utilizó para construir las ventanas emergentes (popups) en los mapas interactivos, permitiendo mostrar información detallada: como el nombre del país, su ranking y su nivel de importancia, cuando el usuario pasa el cursor sobre cada país.

4.3.6.3 Cartogramas por continente

Finalmente, se desarrollaron cartogramas que deforman geométricamente los países en función de su importancia relativa en cada idioma. Para construirlos: primero se trabajó con el shapefile mundial utilizando la librería sf (simple features), el cual es un paquete de R diseñado para manejar datos espaciales de forma eficiente y sencilla, representando geometrías como objetos especiales que pueden ser manipulados y analizados dentro de un DataFrame. Permite leer, transformar, unir y visualizar datos geográficos de manera estructurada.

El primer paso fue transformar el sistema de coordenadas del shapefile a EPSG:3857 utilizando sf::st_transform(). EPSG:3857 es un sistema de proyección métrica (también conocido como Web Mercator), que convierte las coordenadas de latitud y longitud a metros, permitiendo así realizar operaciones geométricas precisas, como escalado y deformaciones necesarias para crear un cartograma.

Luego, se aplicó cartogram::cartogram_cont(), una función que genera cartogramas continuos. Esta función deforma las geometrías de los países proporcionalmente a un valor numérico especificado, en este caso, el valor de 1/ranking. Esto significa que los países más importantes en el idioma analizado se agrandan, mientras que los menos mencionados se reducen. El parámetro principal es el valor por distorsionar (1/ranking), que define el objetivo de la transformación espacial.

Durante el proceso de deformación, es común que se generen errores geométricos, como polígonos mal cerrados o autointersecciones. Para corregir estos problemas, se utilizó sf::st_make_valid(), que repara automáticamente las geometrías defectuosas, asegurando que los polígonos puedan ser representados correctamente en la visualización final.

Para mejorar la legibilidad de los cartogramas, se decidió etiquetar únicamente los países ubicados en el quintil más alto de importancia dentro de cada continente.

La generación automatizada de estos cartogramas se implementó mediante la función

cartograma_por_idiomatags(), permitiendo crear versiones deformadas y anotadas específicas para cada idioma y continente.

4.3.3 Validación analítica e interpretación de resultados

Tras construir los índices de parroquialismo y desarrollar las visualizaciones geográficas, se realizó una etapa de validación analítica para evaluar la coherencia de los resultados. Esta fase fue clave para contrastar los hallazgos con las expectativas teóricas del proyecto y asegurar que las visualizaciones reflejaran de forma fiel los patrones de referencia geográfica en los distintos idiomas.

La validación combinó tres enfoques complementarios:

1. Análisis visual y cualitativo: Se revisaron los mapas estáticos, interactivos y cartogramas de cada idioma, verificando si los países más destacados (en color rojo) correspondían a territorios donde se habla el idioma en cuestión o a naciones de alto peso geopolítico y cultural. Por ejemplo, en inglés se esperaba una alta prominencia de Estados Unidos, Reino Unido y Australia; en español, de México, España o Colombia. La coherencia entre los patrones observados y las intuiciones culturales sirvió como una primera validación semántica.

También se identificaron países que aparecían consistentemente como poco relevantes en múltiples idiomas, evaluando si estas tendencias concordaban con factores como su tamaño demográfico o su inserción cultural internacional.

2. Interpretación estadística de los índices: Se analizó la relación entre los valores del índice new y las variables socioeconómicas como apertura cultural, confianza intergrupal, desigualdad o desarrollo humano. Tal como preveía el marco conceptual, los idiomas con mayor parroquialismo tendieron a asociarse con menor apertura y mayores niveles de desigualdad y desconfianza, validando la capacidad interpretativa de este índices.

3. Construcción de un ranking consolidado de idiomas: Para integrar las perspectivas de los dos índices más robustos (new y RMSE), se elaboró un ranking combinado. Cada idioma recibió una posición relativa (1 a 14) en cada índice, y luego se promedió su posición para generar un puntaje final. Esto permitió establecer una jerarquía global que sintetiza tanto el sesgo interno hacia las ciudades propias (capturado por el índice new) como la desviación respecto al patrón poblacional mundial (capturado por el RMSE).

Además, se llevó a cabo un análisis específico sobre las ciudades que no lograron ser posicionadas, incluso después de aplicar procesos de traducción, tokenización y estimación mediante BERT. En este análisis se distinguieron tres posibles causas principales: en primer lugar, errores técnicos relacionados con fallos en la tokenización o codificación incorrecta de los nombres de las ciudades; en segundo lugar, la rareza de ciertos topónimos o la falta de equivalentes claros en algunos idiomas, lo que dificultó su identificación adecuada; y, en tercer lugar, la baja visibilidad global de esas ciudades en el corpus lingüístico utilizado, lo que limitó las posibilidades de su detección por los modelos.

Este análisis crítico permitió reconocer los límites inherentes a las herramientas y metodologías empleadas, reforzando la necesidad de interpretar con cautela los vacíos de datos observados. Asimismo, resaltó la importancia de considerar tanto factores técnicos como culturales al analizar la representación geográfica en corpus multilingües.

5. Resultados y análisis

Luego de completar todas las etapas de construcción de datos, desarrollo metodológico y visualización, esta sección presenta los resultados centrales del estudio sobre parroquialismo lingüístico en diferentes idiomas. El objetivo principal es interpretar las métricas obtenidas, evaluar sus implicaciones y contrastarlas con los factores socioeconómicos y culturales

previamente definidos. A lo largo de este apartado, se expondrán los valores obtenidos por cada idioma en los distintos índices de parroquialismo (RMSE, índice *new* y su promedio integrado), se identificarán patrones comunes y casos excepcionales, y se analizarán tanto los países como las ciudades que se destacaron, ya sea por su alta presencia o por su ausencia significativa en los modelos de lenguaje. Asimismo, se interpretarán las visualizaciones generadas y se propondrán hipótesis que contribuyan a explicar los resultados desde una perspectiva íntegra y crítica.

5.1 Resultados de los Índices de Parroquialismo

5.1.1 Resultados del índice RMSE por idioma

Una vez calculado el índice RMSE para cada idioma, se obtuvieron las primeras estimaciones concretas sobre el nivel de parroquialismo lingüístico desde una perspectiva estructural. Como se explicó en la sección metodológica, este índice mide qué tan alineado está el ranking promedio de países por idioma con un escenario teóricamente neutral o "ideal", en el que los países más poblados ocuparían los primeros lugares en las menciones geográficas, sin que mediara un sesgo particular por idioma.

A continuación, se presentan los valores del RMSE obtenidos para cada uno de los catorce idiomas analizados. Los idiomas fueron ordenados de menor a mayor valor de RMSE, lo cual permite identificar cuáles tienden a alinearse mejor con el patrón global por población y cuáles muestran un sesgo más marcado hacia lo local o lo propio.

IDIOMA	RMSE	RANKING
Vietnamita	27.1114	1
Coreano	27.5977	2
Inglés	27.7633	3
Indonesio	29.0357	4
Japonés	29.1121	5
Turco	29.4531	6
Portugués	29.5901	7

Suajili	29.6372	8
Chino	29.7237	9
Finés	29.8364	10
Árabe	30.3285	11
Tamil	31.0278	12
Español	31.6314	13
Francés	31.8709	14

Tabla 6. Resultados del índice RMSE por idioma

De acuerdo con los resultados, los idiomas con menor RMSE fueron el vietnamita, el coreano y el inglés, lo que sugiere que en estos idiomas se tiende a mencionar con mayor frecuencia países que, objetivamente, tienen una mayor población global, sin mostrar una fuerte preferencia por los países con los que se comparte lengua o cercanía cultural.

En contraste, idiomas como el francés, el español y el tamil presentaron los valores más altos de RMSE, lo que indica que sus patrones de menciones geográficas están más alejados del escenario ideal basado en población. Esta desviación puede interpretarse como una mayor inclinación a privilegiar ciudades y países relacionados con su propio idioma o contexto regional, lo que sugiere un mayor nivel de parroquialismo desde esta perspectiva estructural.

El análisis del RMSE permite también observar patrones particulares asociados a las familias lingüísticas y a la dispersión geográfica de los hablantes. Por ejemplo, el idioma inglés, gracias a su distribución global y su estatus de lengua franca internacional, tiende a presentar un RMSE relativamente bajo. Esto puede atribuirse a que sus hablantes se encuentran repartidos en múltiples regiones del mundo, lo que diversifica de manera natural las menciones geográficas. En cambio, idiomas cuya distribución territorial es más concentrada, como el tamil o el finlandés, tienden a registrar valores más altos de RMSE, reflejando una menor diversidad en los contextos lingüísticos donde se generan las menciones.

Para tener en cuenta cómo fue el ranking ideal basado en la población, en el **anexo 1**, se muestran los primeros y últimos 20 países según este escenario ideal.

Además, para facilitar la lectura y la interpretación de los resultados, se elaboró una tabla

resumen que muestra el top 5 de países mejor y peor posicionados por algunos de los idiomas, junto con sus posiciones respectivas en el ranking ideal basado en población. Esta comparación permite observar de manera más concreta cómo algunos idiomas tienden a priorizar sistemáticamente a países lingüísticamente afines o geográficamente próximos, en comparación de otros países más poblados, pero culturalmente más distantes.

A continuación, se presentan cuatro ejemplos seleccionados: dos correspondientes a idiomas que obtuvieron los RMSE más altos (francés y español) y dos asociados a los RMSE más bajos (inglés y vietnamita).

FRANCÉS	RANKINGS
Francia	1
Bélgica	2
España	3
Canadá	4
Italia	5
…	…
Tailandia	62
Libia	63
Bangladesh	64
India	65
China	66

Tabla 7. Rankings más altos y bajos del francés

ESPAÑOL	RANKINGS
España	1
Chile	2
Bolivia	3
Perú	4
Ecuador	5
…	…
Bangladesh	62
India	63
Nigeria	64
Burkina Faso	65
China	66

Tabla 8. Rankings más altos y bajos del español

INGLÉS	RANKINGS
Estados Unidos	1
Australia	2
Canadá	3
Reino Unido	4
España	5
…	…
Taiwán	62
China	63
Somalia	64
Libia	65
Kazajistán	66

Tabla 9. Rankings más altos y bajos del inglés

VIETNAMITA	RANKINGS
Taiwán	1
España	2
Vietnam	3
Francia	4
Italia	5
…	…
Venezuela	62
Nigeria	63
Tailandia	64
Zambia	65
Mozambique	66

Tabla 10. Rankings más altos y bajos del vietnamita

Los resultados de esta comparación detallada mostraron diferencias importantes en la estructura de prioridades entre idiomas. En aquellos con menor RMSE, como el vietnamita y el inglés, los rankings tienden a alinearse más estrechamente con el ideal poblacional, mostrando una distribución de menciones más balanceada. Por el contrario, en idiomas como el francés y el español, los primeros lugares están ocupados predominantemente por países con vínculos culturales o lingüísticos directos, desplazando a un segundo plano a naciones con grandes poblaciones.

Por ejemplo, en el caso del idioma francés, los países mejor posicionados fueron Francia, Bélgica, España, Canadá e Italia, todos cercanos culturalmente. Mientras tanto, países como China, India o Bangladesh, que en el ranking ideal deberían figurar en los primeros puestos por su población, se ubicaron entre los últimos lugares. Un patrón similar se observa en el idioma español, donde predominan en las primeras posiciones España, Chile, Bolivia, Perú y Ecuador, reflejando una concentración marcada en el mundo hispanohablante.

En contraste, el idioma inglés, aunque favorece países como Estados Unidos, Australia, Canadá y Reino Unido, muestra una dispersión más moderada en las menciones, integrando también a España dentro de sus primeras posiciones. El idioma vietnamita, que presentó el RMSE más bajo entre todos los analizados, reveló un patrón aún más equilibrado, con países como Taiwán, España, Vietnam, Francia e Italia liderando su ranking. Este comportamiento sugiere una estructura de menciones más abierta, donde no predomina exclusivamente el criterio lingüístico o regional.

La comparación entre los rankings reales y el ranking ideal revela que la forma en que los idiomas distribuyen simbólicamente la importancia de los países está profundamente condicionada por factores culturales y lingüísticos. Mientras algunos idiomas tienden a construir una visión del mundo más globalizada, otros refuerzan su centralidad cultural al privilegiar

entornos familiares. En este sentido, el RMSE ofrece una medida valiosa para dimensionar el grado de alineación o distancia de cada idioma respecto a un estándar universal basado en población.

5.1.2 Resultados del índice new por idioma

El índice *new* fue diseñado para capturar el parroquialismo lingüístico desde una perspectiva relacional, comparando el trato promedio que reciben las ciudades asociadas a un idioma frente a aquellas que no lo están. A diferencia del RMSE, que mide la distancia respecto a un patrón global basado en población, *new* se enfoca en la diferencia interna entre "lo propio" y "lo ajeno" en cada idioma.

A continuación, se presentan los resultados ordenados de menor a mayor valor del índice *new*, destacando los idiomas más abiertos frente a los más parroquialistas:

idioma	new
Chinese	0.9311441
Tamil	3.4057423
Swahili	3.6518960
Vietnamese	4.1209393
English	5.1210776
Finnish	5.4046847
French	7.6393896
Arabic	7.8366472
Korean	8.3065595
Japanese	8.4466764
Portuguese	9.4320523
Turkish	9.5053124
Spanish	10.8187295
Indonesian	11.2783429

Figura 20. Resultados del índice new por idioma

El idioma chino registró el valor más bajo, reflejando un trato relativamente equitativo entre sus ciudades y las externas. En el otro extremo, el indonesio, el español y el turco mostraron una marcada preferencia interna. En particular, el español evidenció una diferencia promedio de más de 300 posiciones entre ciudades internas y externas:

idioma	promedio_interno	promedio_externo	new
English	87334.100	229172.99	5.1210776
Spanish	28289.987	336467.84	10.8187295
Chinese	220844.331	261912.44	0.9311441
Swahili	17508.200	72707.07	3.6518960
Tamil	2084.000	283648.74	3.4057423
Turkish	10204.846	374145.93	9.5053124
Finnish	206.000	501414.31	5.4046847
Japanese	5043.333	197633.08	8.4466764
Korean	13345.067	266961.60	8.3065595
Vietnamese	13855.167	115169.17	4.1209393
Arabic	43690.293	298578.48	7.8366472
Indonesian	3336.900	135573.31	11.2783429
Portuguese	25314.179	326442.19	9.4320523
French	28691.407	284061.55	7.6393896

Figura 21. Índice new por idioma con promedios

Aunque tanto el RMSE como el índice *new* evalúan el parroquialismo, lo hacen desde ángulos distintos. El vietnamita, por ejemplo, combina un RMSE bajo con un *new* también reducido, reforzando su perfil más abierto. El español, en cambio, se ubica consistentemente entre los idiomas más parroquialistas en ambos índices. Sin embargo, no todos los casos son paralelos: el chino, que no obtuvo el RMSE más bajo, sí emergió como el menos parroquialista según el índice *new*, sugiriendo que, aunque su ranking global se aleja un poco del ideal poblacional, internamente no discrimina de manera acentuada entre ciudades propias y externas.

Estas diferencias resaltan la importancia de emplear múltiples métricas para una comprensión más completa y matizada del fenómeno del parroquialismo lingüístico.

Para comprender los factores que explican las diferencias en el índice *new* entre idiomas, se calcularon las correlaciones entre este índice y diversas variables estructurales, tecnológicas y socioculturales asociadas a los países donde se hablan.

Los resultados mostraron asociaciones estadísticamente significativas:

Variable	Correlación	p-valor	Interpretación
HDI	-0.683	0.0071	A mayor desarrollo humano, menor parroquialismo
Penetración de Internet	-0.711	0.0043	A mayor conectividad, menor parroquialismo
Índice Gini (desigualdad)	0.764	0.0015	A mayor desigualdad, mayor parroquialismo
Confianza In-Group	0.769	0.0013	A mayor confianza hacia el propio grupo, mayor parroquialismo
Confianza Out-Group	-0.764	0.0015	A mayor confianza hacia otros grupos, menor parroquialismo
Índice de Migración	-0.803	0.0005	A mayor migración, menor parroquialismo
Índice de Apertura	-0.694	0.0059	A mayor apertura cultural, menor parroquialismo
Índice de Corrupción	0.657	0.0106	A mayor corrupción, mayor parroquialismo
Índice de Competencia	0.303	0.2916	Sin correlación significativa

Figura 22. Correlaciones del índice new vs. variables

Estos resultados confirman que el parroquialismo lingüístico está estrechamente relacionado con las condiciones sociales y estructurales de los hablantes. En general, los idiomas que presentaron mayor parroquialismo (según el índice new) tienden a estar asociados a contextos marcados por alta desigualdad, menor desarrollo humano, baja conectividad digital, fuerte confianza hacia el endogrupo y limitada apertura hacia grupos externos.

Por el contrario, los idiomas menos parroquialistas se caracterizan por niveles más altos de desarrollo humano, menor desigualdad, mayor penetración de internet, y actitudes más abiertas hacia personas de otras culturas o religiones. Esta relación sugiere que las lenguas no solo reflejan patrones internos de comunicación, sino también la forma en que las sociedades estructuran sus relaciones con lo externo, consolidando o desafiando fronteras simbólicas a través del lenguaje.

5.1.3 Resultados de comparación e integración de índices

Para sintetizar los resultados obtenidos mediante los dos principales índices de parroquialismo (RMSE y *new*), se construyó un ranking combinado basado en el promedio de las posiciones que

cada idioma alcanzó en ambos indicadores. Esta integración permite identificar de forma consistente qué idiomas tienden a mostrar mayores o menores niveles de parroquialismo lingüístico, considerando tanto su alineación con un patrón poblacional ideal como su preferencia interna frente a lo externo:

Idioma	Posición RMSE	Posición new	Promedio de posiciones
Vietnamita	1	4	2.5
Inglés	3	5	4.0
Coreano	2	9	5.5
Chino	9	1	5.0
Indonesio	4	14	9.0
Japonés	5	10	7.5
Turco	6	12	9.0
Portugués	7	11	9.0
Francés	14	7	10.5
Suajili	8	3	5.5
Árabe	11	8	9.5
Finés	10	6	8.0
Tamil	12	2	7.0
Español	13	13	13.0

Figura 23. Promedio entre índices

El ranking combinado sintetiza de manera robusta la posición de los idiomas en un continuo de menor a mayor parroquialismo, integrando dos dimensiones complementarias del fenómeno:

Ranking	Idioma	Promedio de Posiciones	Tipo de patrón observado
1	Vietnamita	2.5	Apertura estructural y relacional alta
2	Inglés	4.0	Apertura estructural y relacional alta
3	Chino	5.0	Apertura estructural, pero con sesgos leves
4	Suajili	5.5	Apertura moderada
5	Coreano	5.5	Apertura moderada con leve preferencia local
6	Finés	8.0	Apertura relacional moderada pero estructuralmente más sesgado
7	Japonés	7.5	Intermedio: leve tendencia al parroquialismo
8	Tamil	7.0	Relacionalmente más abierto, estructuralmente sesgado
9	Turco	9.0	Autocentrismo regional moderado
10	Portugués	9.0	Autocentrismo regional moderado
11	Indonesio	9.0	Autocentrismo fuerte, con baja apertura estructural
12	Árabe	9.5	Moderado-alto parroquialismo
13	Francés	10.5	Sesgo de afinidades culturales externas (no autocentrismo puro)
14	Español	13.0	Autocentrismo extremo (parroquialismo interno muy fuerte)

Figura 24. Ranking final de idiomas

Notas sobre la asignación de patrones:

- Apertura estructural y relacional alta: Ambos índices (RMSE y new) son bajos → el idioma muestra alta apertura tanto en su cobertura global como en su tratamiento de ciudades internas.
- Apertura estructural con sesgos leves: RMSE algo alto pero new bajo → el idioma cubre países variados pero puede priorizar internamente ciertos lugares.
- Apertura moderada: Ambos índices son intermedios → sin sesgos fuertes pero tampoco globalidad perfecta.
- Relacionalmente abierto, estructuralmente sesgado: Índice new bajo, pero RMSE más alto → apertura hacia externos pero estructura interna menos equilibrada.
- Autocentrismo regional moderado: Ambos índices son intermedios-altos → clara preferencia por regiones lingüísticas o culturales cercanas, aunque no de manera extrema.

- Autocentrismo fuerte o extremo: Ambos índices son altos → fuerte sesgo hacia ciudades y países del propio grupo lingüístico/cultural.
- Sesgo de afinidades externas: RMSE alto pero new moderado → el idioma no es tan parroquialista internamente, pero prioriza culturalmente a países afines (ej. francés prioriza Europa francófona y excolonias).

5.1.4 Resultados del LASSO, Ridge y Elastic Net

Para identificar las variables socioeconómicas que mejor explican el comportamiento del índice *new*, se aplicaron tres técnicas de regresión regularizada: LASSO, Ridge y Elastic Net. Estos métodos permiten seleccionar y ponderar variables controlando la complejidad del modelo y evitando sobreajuste en presencia de alta multicolinealidad.

A continuación, se presentan los coeficientes obtenidos en cada modelo:

Variable	LASSO	Ridge	Elastic Net
HDI	**6.03**	-1.66	**22.75**
Internet penetration	0	-0.02	0
Gini index	0	0.07	0
Trust in-group	0	**4.20**	0
Trust out-group	0	-0.73	0
Migration index	**-5.16**	-0.60	**-8.31**
Openness index	0	-0.33	0

Figura 25. Resultados regresión regularizada

Basándose en la magnitud de los coeficientes y su consistencia entre modelos, las variables se agruparon en tres categorías. El Grupo A incluye las variables de alta importancia, relevantes en todos los modelos. El Índice de Desarrollo Humano (HDI) presentó coeficientes destacados en LASSO (6.03) y Elastic Net (22.75), y aunque negativo en Ridge (-1.66), su magnitud fue considerable. También se destaca el índice de migración, con coeficientes negativos consistentes

en LASSO (-5.16), Elastic Net (-8.31) y Ridge (-0.60).

El Grupo B agrupa variables de importancia moderada o intermitente. La confianza in-group mostró un coeficiente positivo en Ridge (4.20), mientras que la confianza out-group presentó un coeficiente negativo (-0.73), aunque ninguno fue seleccionado por LASSO o Elastic Net. De manera similar, el índice de desigualdad (Gini) y el de apertura cultural (Openness) mostraron efectos pequeños en Ridge (0.07 y -0.33, respectivamente), perdiendo fuerza al modelarse junto a otras variables.

Finalmente, el Grupo C comprende las variables descartadas, eliminadas en LASSO y Elastic Net y con coeficientes mínimos en Ridge. Entre ellas se encuentra la penetración de internet, cuyo impacto independiente en el parroquialismo resultó marginal.

En conjunto, los tres modelos coinciden en señalar que el desarrollo humano y la movilidad internacional son los factores más robustamente asociados a menores niveles de parroquialismo lingüístico. Idiomas hablados en contextos de mayor desarrollo y mayor migración tienden a mostrar una visión geográfica más abierta, probablemente gracias a mejor acceso a información global, niveles educativos más altos y mayor apertura cultural. En contraste, variables como la confianza social, la desigualdad y la apertura, aunque relevantes en análisis bivariados, perdieron peso explicativo al ser consideradas conjuntamente, sugiriendo que sus efectos podrían estar mediados por dimensiones estructurales más fundamentales.

5.2 Resultados de las visualizaciones

Con el objetivo de complementar los análisis cuantitativos, esta sección presenta visualizaciones espaciales que ilustran los patrones de parroquialismo lingüístico en los idiomas analizados. La sección se organiza en tres partes: mapas estáticos y dinámicos, cartogramas continentales y análisis de patrones comunes. Estas herramientas buscan articular de forma crítica y accesible los hallazgos cuantitativos del estudio.

5.2.1 Mapas estáticos y dinámicos

Como referencia inicial, se presenta un mapa teórico basado en la distribución poblacional global, donde la importancia de cada país depende exclusivamente de su población.

En la figura a continuación, los países están coloreados según su quintil de importancia teórica, utilizando una escala de colores que va de "Muy bajo" a "Muy alto":

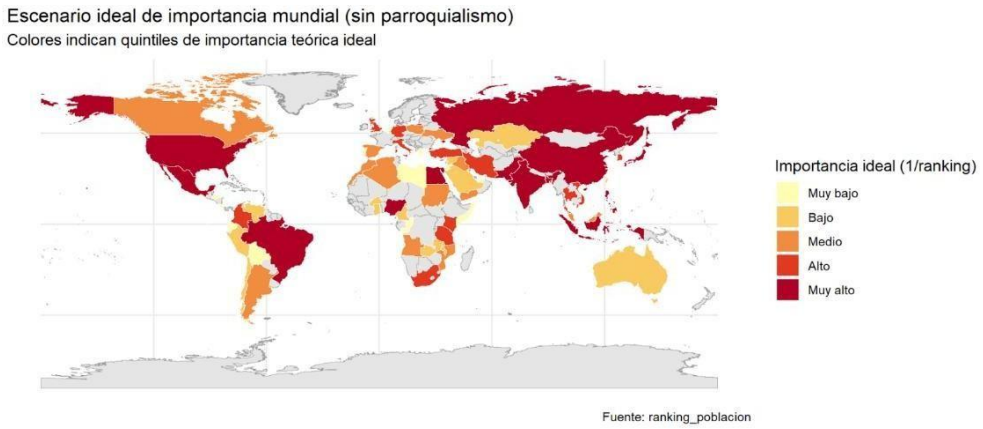


Figura 26. Mapa escenario ideal

En este escenario ideal, países como China, India, Estados Unidos, Brasil, Indonesia y Nigeria destacan como los más relevantes, mientras que naciones menos pobladas presentan una visibilidad menor. Este patrón sirve como base para comparar las distribuciones observadas en cada idioma.

Se agruparon los idiomas en tres niveles: menor, intermedio y mayor parroquialismo lingüístico. A continuación, se presentan tres ejemplos seleccionados por su representatividad de distintos niveles de parroquialismo lingüístico: bajo (vietnamita), intermedio (japonés) y alto (español). Cada mapa refleja la importancia relativa de los países en el ranking de menciones para cada idioma, utilizando una escala de quintiles que va de "Muy bajo" a "Muy alto", como se explicó previamente:

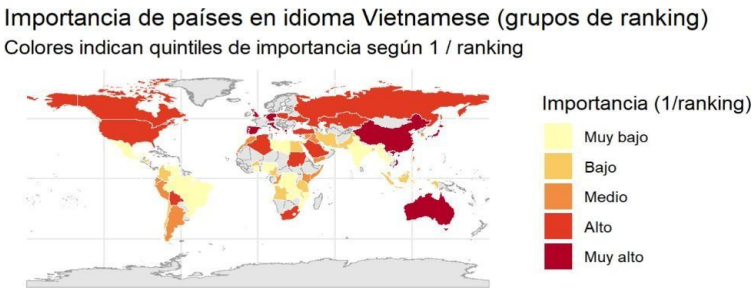


Figura 27. Mapa vietnamita

El primer caso es el vietnamita, considerado uno de los idiomas con menor parroquialismo según los índices calculados. Su mapa muestra una distribución geográfica relativamente abierta, con países de Asia, Europa y América del Norte incluidos en los quintiles superiores. Esto sugiere una mayor diversidad en las menciones y menor centralización en el entorno cultural propio.

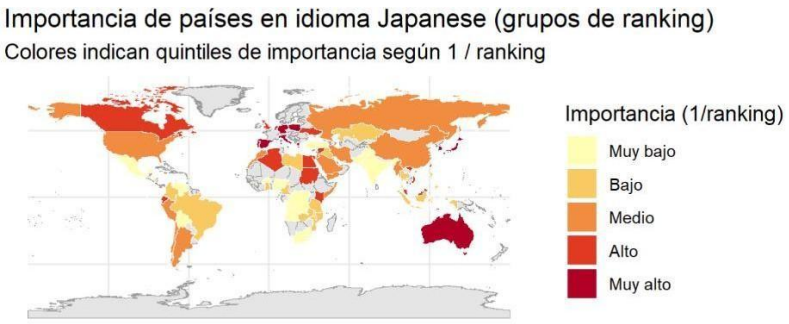


Figura 33. Mapa japonés

En segundo lugar, el japonés representa un nivel intermedio de parroquialismo. El mapa

evidencia una cobertura algo más sesgada, con énfasis en Japón, Europa y algunos países del Pacífico, aunque con menor amplitud en regiones como África o América Latina.



Figura 40. Mapa español

Finalmente, el español destaca como uno de los idiomas con mayor parroquialismo. El mapa revela una fuerte concentración en el mundo hispanohablante, especialmente América Latina y España, con escasa visibilidad otorgada a regiones fuera del espacio cultural propio.

Estos tres ejemplos permiten observar, de forma comparativa, cómo varía el sesgo geográfico entre idiomas y qué tan centrado o abierto es el patrón de mención en cada caso. Para consultar los mapas correspondientes a los demás idiomas analizados, puede dirigirse al **Anexo 2**, donde se presentan las visualizaciones completas de los 14 idiomas incluidos en el estudio.

Con el fin de ofrecer una exploración más dinámica y facilitar el acceso a los resultados del análisis por parte de otras personas interesadas, se construyeron mapas interactivos utilizando la librería leaflet. Estos mapas representan la misma información que los mapas estáticos, pero en un formato navegable que permite consultar datos específicos al pasar el cursor sobre cada país: nombre, ranking y grupo de importancia asignado.

La escala cromática se mantuvo consistente con la empleada en los mapas estáticos, asegurando coherencia visual. Las ventanas emergentes (popups) fueron generadas permitiendo

mostrar información detallada de manera clara e intuitiva. Para los países sin datos disponibles, se asignó un color gris claro acompañado de la etiqueta "Sin datos de ranking".

Este tipo de visualización se diseñó no solo como complemento metodológico, sino también como una herramienta accesible que puede ser de utilidad para docentes, investigadores o cualquier persona interesada en explorar de forma interactiva los patrones de parroquialismo detectados en los distintos idiomas.

Para ver ejemplos visuales de estos mapas interactivos y su funcionalidad, puede consultarse el **Anexo 3**, donde se incluyen capturas representativas de la interfaz y su modo de interacción.

5.2.2 Cartogramas por continente

Para complementar la visualización de los resultados, se elaboraron cartogramas que muestran cómo varía el peso visual de los países dentro de cada continente según el idioma analizado. A diferencia de los mapas estáticos, los cartogramas deforman proporcionalmente el tamaño de los países en función de su importancia relativa en el ranking de menciones.

Debido a que los cartogramas son un recurso visual complementario y no una repetición exhaustiva de los datos, se seleccionaron idiomas representativos para cada continente, optimizando el espacio y facilitando comparaciones claras. Para cada región, se presentan cartogramas de idiomas hablados localmente (como el español, el inglés o el portugués en América) y de idiomas externos (como el francés o el chino).

Este enfoque permite analizar cómo varía la percepción de importancia geográfica entre idiomas locales y externos, y ayuda a identificar patrones de parroquialismo interno, apertura externa y dinámicas de representación diferenciadas en cada región.

5.2.2.1 América

Para analizar el continente americano, se seleccionaron cuatro idiomas representativos que contrastan diferentes niveles de parroquialismo: español, coreano, turco e indonesio. Esta elección permite analizar tanto idiomas locales como externos

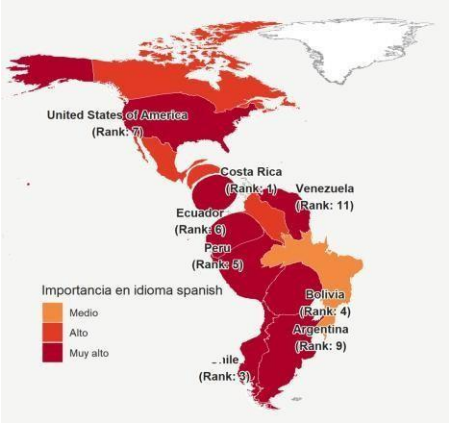


Figura 42. Cartograma América (español)

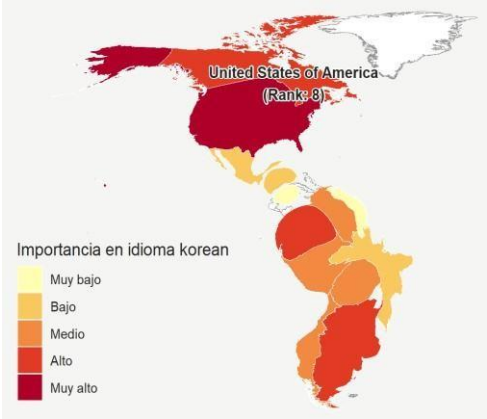


Figura 43. Cartograma América (coreano)



Figura 44. Cartograma América (indonesio)

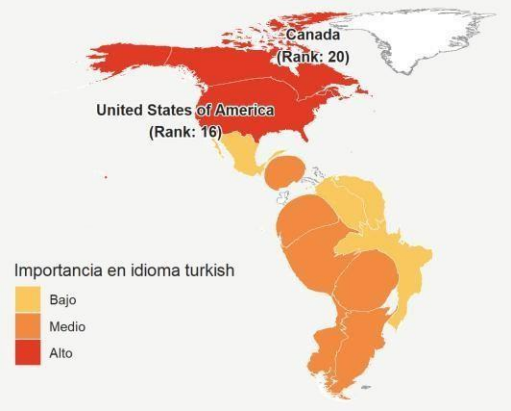


Figura 45. Cartograma América (turco)

El cartograma del español muestra una notable expansión de países latinoamericanos como Costa Rica, Chile, Bolivia, Perú y Ecuador, todos en el grupo de "Muy alto", reforzando el fuerte parroquialismo previamente identificado en este idioma. No hay países clasificados en importancia baja o muy baja, lo que evidencia una concentración clara en la región hispanohablante.

En contraste, el coreano, uno de los idiomas menos parroquialistas, presenta una distribución

mucho más equilibrada. Estados Unidos destaca en el grupo de "Muy alto", mientras que América Latina se reparte entre categorías altas y medias, con deformaciones moderadas. Pese a no ser un idioma local, el coreano otorga relevancia significativa a varios países americanos, lo que reafirma su apertura geográfica.

El turco, con un nivel intermedio de parroquialismo, muestra un patrón moderado. Estados Unidos conserva un tamaño importante, pero América Latina aparece mayoritariamente en categorías medias o bajas, con deformaciones visibles, pero no extremas. El patrón indica un sesgo geográfico presente pero menos marcado que en el español.

Finalmente, el cartograma del indonesio refleja una distribución similar al turco. Estados Unidos y Canadá ocupan posiciones altas, mientras América Latina se ubica principalmente en niveles medios o bajos. No se registran países en el nivel "Muy bajo", lo que confirma que, aunque el indonesio presenta algo de parroquialismo, su tendencia es relativamente baja.

5.2.2.2 África

En África, los cartogramas revelan diferencias claras entre idiomas locales y externos en términos de parroquialismo:

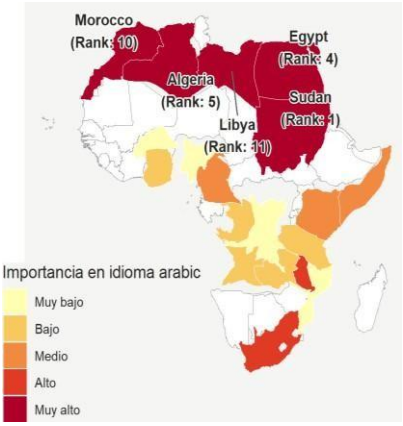


Figura 46. Cartograma África (árabe)

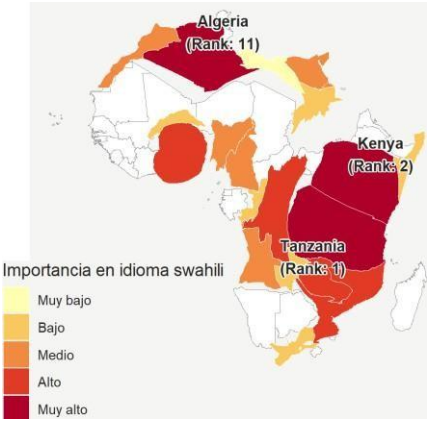


Figura 47. Cartograma África (suajili)

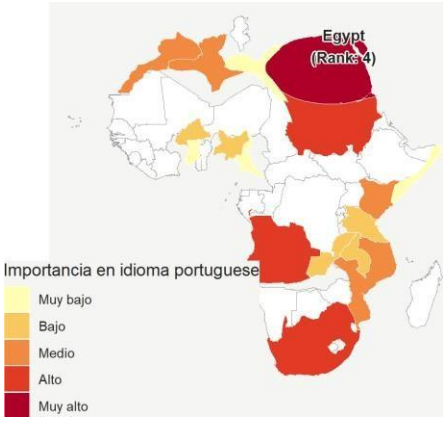


Figura 48. Cartograma África (portugués)

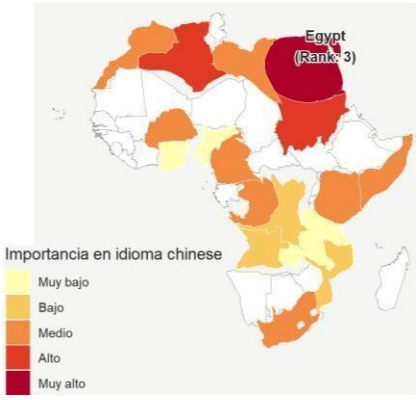


Figura 49. Cartograma África (chino)

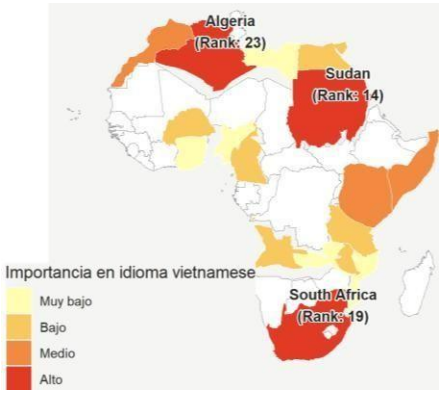


Figura 50. Cartograma África (viertnamita)

El árabe, uno de los idiomas más parroquialistas, muestra una fuerte concentración en el norte de África. Países como Sudán, Egipto, Argelia, Libia y Marruecos aparecen ampliamente agrandados, y gran parte del continente alcanza categorías de importancia media a muy alta, evidenciando un fuerte sesgo regional.

De manera similar, el suajili prioriza fuertemente el este y centro de África. Tanzania y Kenia destacan en "Muy alto", y casi toda la región suajilióFona se mantiene en niveles altos o medios, sin apenas zonas clasificadas como "Muy bajo".

En contraste, el chino presenta una distribución más equilibrada. Aunque Egipto alcanza "Muy alto", varios países africanos ocupan categorías altas o medias, reflejando un interés más global, a pesar de no ser un idioma local en la región.

El vietnamita también muestra baja concentración. Si bien no registra países en "Muy alto", destaca regiones como Sudán, Argelia y Sudáfrica en niveles altos, confirmando su bajo nivel de parroquialismo.

Por último, el portugués exhibe un patrón intermedio. Países como Egipto, Angola y Mozambique alcanzan niveles altos o medios, pero también aparecen regiones de baja importancia, lo que concuerda con su clasificación como idioma de parroquialismo moderado.

5.2.2.3 Europa y Asia

El análisis de los cartogramas de Europa y Asia permite observar cómo varía la concentración geográfica de la importancia en función del nivel de parroquialismo de cada idioma. A continuación, se describe el comportamiento observado en los idiomas seleccionados:



Figura 51. Cartograma Europa y Asia (finés)

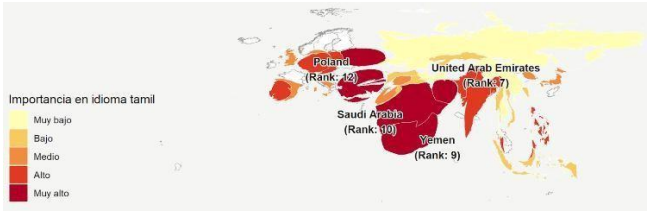


Figura 52. Cartograma Europa y Asia (tamil)

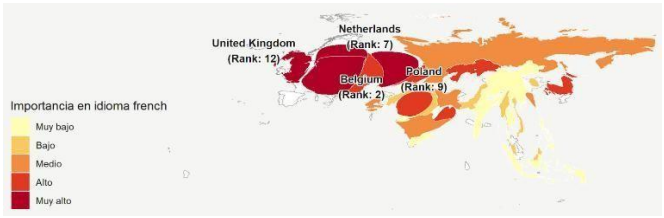


Figura 53. Cartograma Europa y Asia (francés)

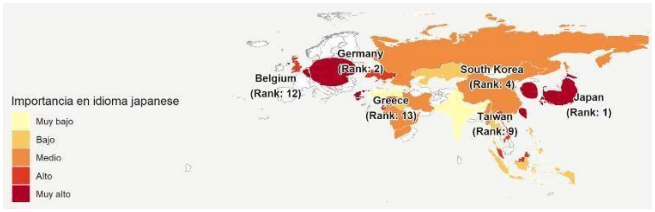


Figura 54. Cartograma Europa y Asia (japonés)

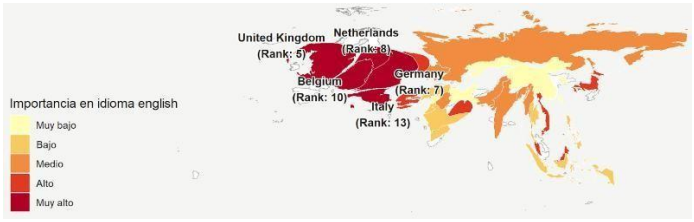


Figura 55. Cartograma Europa y Asia (inglés)

El finlandés refleja un fuerte patrón parroquialista: aunque Finlandia no aparece resaltada por el filtro de ciudades, los países vecinos como Suecia, Noruega, Alemania, Países Bajos y Bélgica ocupan posiciones de “Muy alto”, reforzando la concentración en su entorno cultural y geográfico.

El tamil también muestra un patrón altamente regionalizado. Aunque India no lidera el ranking, zonas asiáticas cercanas como Sri Lanka, Malasia, Singapur y partes de Oriente Medio

destacan, confirmando su fuerte anclaje en su región cultural.

El francés, clasificado como de parroquialismo intermedio-alto, presenta un foco marcado en Europa, especialmente en Bélgica y países francófonos o culturalmente cercanos. Asia, en contraste, aparece con menor relevancia, evidenciando la preferencia europea en su patrón de menciones.

El japonés, con un nivel intermedio, pero más abierto que el francés, muestra una concentración esperable en Japón, pero también otorga alta importancia a países como Alemania, Corea del Sur y Taiwán. Esto sugiere un sesgo hacia lo propio, pero con mayor apertura hacia otras regiones relevantes.

Por último, el inglés reafirma su perfil global. Aunque el Reino Unido destaca, la importancia se distribuye ampliamente tanto en Europa como en Asia, reflejando un alcance equilibrado y baja tendencia al parroquialismo.

Aunque el análisis principal se centró en América, África, Europa y Asia, también se observó un patrón consistente en Oceanía. En todos los cartogramas analizados, Australia y algunos países cercanos aparecen clasificados como de "Alta" o "Muy alta" importancia, a pesar de que, según el ranking ideal basado en población, deberían ubicarse en niveles bajos.

Esta discrepancia sugiere que la relevancia de Australia no responde a criterios demográficos, sino a factores culturales, mediáticos o de afinidad internacional, especialmente en idiomas de amplia circulación como el inglés. El caso de Australia ilustra cómo dinámicas culturales y geopolíticas pueden moldear la representación geográfica, más allá de consideraciones estrictamente poblacionales.

5.2.2.4 Análisis de patrones comunes

Al observar los cartogramas, se identificaron patrones recurrentes de visibilidad geográfica. Algunos países, como Estados Unidos, Alemania, Reino Unido, Australia, Japón y Canadá,

mantienen posiciones de alta o muy alta importancia en múltiples idiomas, independientemente de su nivel de parroquialismo. Esto sugiere que su relevancia no depende solo de su población o cercanía lingüística, sino también de factores globales.

Comparado con el escenario ideal basado en población, se evidencian desviaciones claras. Australia, por ejemplo, aparece sistemáticamente sobrerrepresentada, a pesar de su baja prioridad demográfica. Lo mismo ocurre, en menor medida, con Canadá y algunos países europeos de tamaño medio como Países Bajos y Bélgica. Estas diferencias reflejan cómo factores no demográficos moldean la percepción de importancia geográfica.

La sobre-representación puede explicarse por la influencia mediática global (especialmente de Estados Unidos y Reino Unido), el poder económico (caso de Alemania, Japón y Australia), el rol del inglés como lengua franca y las dinámicas de migración, diplomacia y turismo, que amplifican la visibilidad internacional de ciertos países.

En contraste, se detecta la sub-representación de países muy poblados como India o Nigeria, cuya presencia simbólica es menor de lo esperado. Esto podría atribuirse a menor cobertura mediática, barreras lingüísticas o limitaciones culturales en el alcance global, evidenciando que el volumen poblacional no garantiza visibilidad internacional.

Ejemplos concretos refuerzan estas dinámicas: Australia destaca en numerosos idiomas pese a su baja población, mientras que Estados Unidos mantiene una posición hegemónica constante. India, por su parte, muestra menor relevancia en varios idiomas, especialmente en aquellos más abiertos globalmente. Canadá y Alemania, aunque no lideran siempre, conservan una importancia media-alta sostenida.

En conjunto, estos resultados confirman que la visibilidad geográfica está profundamente moldeada por dinámicas de poder, redes culturales y circulación de información, más que por criterios estrictamente demográficos. El parroquialismo coexiste con una jerarquía global que

favorece a ciertas naciones, dinámica crucial para interpretar cómo los idiomas configuran colectivamente la relevancia geográfica en el discurso mundial.

5.2.2.5 Limitaciones del análisis y sugerencias para futuras investigaciones

Aunque el enfoque adoptado en este proyecto permitió identificar patrones sólidos de parroquialismo y visibilidad geográfica, es importante considerar ciertas limitaciones al interpretar los resultados.

Una primera limitación proviene de los nombres comunes en las ciudades. Palabras de uso frecuente como "Paz", "Ángeles", "Flor" o "Victoria" pudieron elevar artificialmente la posición de algunas ciudades debido a coincidencias lingüísticas o usos figurativos, y no necesariamente a su relevancia geográfica real.

Asimismo, durante el procesamiento de datos se identificaron casos en los que algunas ciudades quedaron con posición final "None", principalmente debido a su ausencia en los vocabularios o a una baja frecuencia de mención en los corpus analizados. Esta situación afectó a menos del 5% de las ciudades por idioma, y se asumió que refleja una baja relevancia cultural, mediática o poblacional.

Otra limitación está asociada a la ambigüedad semántica: algunos nombres de ciudades, como "Jerusalén", pueden aparecer en textos con sentidos simbólicos, religiosos o metafóricos, sin referencia directa a su dimensión geográfica, lo que pudo inflar su importancia de manera no intencionada.

Aunque estas limitaciones no invalidan los resultados obtenidos, sí invitan a interpretarlos con cautela, reconociendo los márgenes de error inherentes al análisis de corpus lingüísticos extensos.

De cara a futuras investigaciones, se plantean varias oportunidades de expansión y mejora. Por un lado, se podría ampliar el número de ciudades, países e idiomas analizados, lo cual permitiría aumentar la representatividad geográfica y cultural del fenómeno. También sería valioso

implementar mecanismos más avanzados de desambiguación semántica para diferenciar claramente entre menciones geográficas literales y usos simbólicos. Asimismo, se sugiere construir un dashboard interactivo que integre todos los resultados —índices, mapas, correlaciones y visualizaciones, en una sola plataforma dinámica, facilitando el análisis comparativo entre idiomas y el acceso público a los hallazgos. Finalmente, se podrían explorar corpus alternativos, como redes sociales, medios de comunicación o literatura contemporánea, para analizar si el patrón de parroquialismo lingüístico se mantiene constante o varía según el tipo de discurso.

5.3 Síntesis integrada

La siguiente tabla resume los principales indicadores obtenidos para cada uno de los idiomas:

Idioma	RMSE (posición)	Índice New (posición)	Promedio posiciones	Variables más asociadas	N° Ciudades	% None
Vietnamita	1	4	2.5	Migración (+), HDI (+), Confianza out-group (+)	6	2.98%
Inglés	3	5	4.0	HDI (+), Migración (+), Internet (+)	201	0.25%
Coreano	2	9	5.5	HDI (+), Internet (+), Apertura (+)	15	4.59%
Chino	9	1	5.0	HDI (+), Internet (+), Migración (+), Confianza out-group (+)	234	4.96%
Indonesio	4	14	9.0	Desigualdad (+), Migración (-), HDI (-), Internet (-)	20	2.23%
Japonés	5	10	7.5	HDI (+), Educación (+), Migración (+)	9	2.48%
Turco	6	12	9.0	Desigualdad (+), Migración (-), Confianza in-group (+)	13	2.73%
Portugués	7	11	9.0	HDI (+), Internet (+), Migración (+)	39	1.61%
Francés	14	7	10.5	HDI (+), Confianza in-group (+)	27	2.60%
Swahili	8	3	5.5	Confianza in-group (+), Desigualdad (+), Baja migración (-)	12	3.72%
Árabe	11	8	9.5	Confianza in-group (+), Desigualdad (+), Baja apertura	58	0.62%
Finés	10	6	8.0	Confianza out-group (+), HDI (+), Apertura (+)	1	4.71%
Tamil	12	2	7.0	Confianza in-group (+), Baja apertura, Desigualdad (+)	1	4.34%
Español	13	13	13.0	Desigualdad (+), Confianza in-group (+), Baja migración (-), Baja internet (-)	78	1.24%

Figura 56. Resumen índices e indicadores por idioma

De acuerdo con los resultados, los idiomas que mostraron menores niveles de parroquialismo (tanto estructural como relacional), fueron el chino, el vietnamita, el inglés y el coreano, caracterizados por bajos valores de RMSE, bajos índices new y una mayor apertura geográfica en los cartogramas.

Los idiomas que mostraron menores niveles de parroquialismo, tanto estructural como relacional, fueron el chino, el vietnamita, el inglés y el coreano, caracterizados por bajos valores de RMSE, bajos índices *new* y patrones de mención geográfica más abiertos en los cartogramas.

Por el contrario, el español, el francés, el portugués y el tamil reflejaron altos niveles de parroquialismo, combinando altos valores de RMSE y *new*, junto con patrones de mención concentrados en regiones lingüísticamente cercanas.

Respecto a las variables socioeconómicas asociadas, se observó que el parroquialismo disminuye con mayor conectividad digital (penetración de Internet) y desarrollo humano (HDI), y aumenta con mayores niveles de confianza endogrupal y desigualdad (índice Gini). Esto sugiere que la apertura social y el acceso a información global favorecen patrones lingüísticos más globales.

Finalmente, el porcentaje de ciudades sin posición final ("None") fue bajo, manteniéndose por debajo del 5% en todos los idiomas, lo que confirma que la pérdida de información fue mínima y que las tendencias observadas son representativas.

6. Conclusiones

Este proyecto permitió abordar de manera sistemática el fenómeno del parroquialismo lingüístico, entendido como la tendencia de los hablantes a priorizar y mencionar más las ciudades o regiones culturalmente cercanas a su idioma. A través de un enfoque empírico basado

en vectores de palabras y análisis cuantitativos, se logró capturar, representar y comparar la importancia geográfica asignada a distintas ciudades en 14 idiomas, abarcando la mayoría de las principales regiones del mundo y proporcionando así un estudio de carácter global y comparativo.

Uno de los principales aportes metodológicos fue la construcción de índices específicos que permitieran cuantificar el nivel de parroquialismo en cada idioma. Estos índices fueron diseñados para capturar diferentes dimensiones del fenómeno: uno reflejando la relación entre menciones locales y externas, y otro evaluando la desviación respecto a un escenario ideal de cobertura geográfica. Esta estrategia permitió no solo medir la intensidad del sesgo, sino también ofrecer una base robusta para su comparación entre idiomas.

Además, el análisis de las asociaciones entre los niveles de parroquialismo y variables socioeconómicas clave, como la desigualdad, el desarrollo humano, la confianza social y la conectividad digital, proporcionó evidencia sólida de que el parroquialismo no es un fenómeno aleatorio, sino que está estructuralmente vinculado a las condiciones sociales y culturales de cada comunidad lingüística.

La metodología de comparación entre rankings reales e ideales, junto con la implementación de cartogramas regionales, facilitó la visualización efectiva de los patrones de apertura o concentración geográfica en cada idioma. Asimismo, se reconocieron limitaciones inherentes al análisis de corpus extensos, como la ambigüedad de algunos nombres de ciudades y la ausencia de ciertos términos en los modelos de vectores.

Finalmente, este trabajo abre la puerta a futuras investigaciones orientadas a comprender cómo los factores sociales, educativos y mediáticos moldean la construcción de visiones geográficas más abiertas o cerradas, y cómo el lenguaje actúa como un espejo de las dinámicas culturales y de identidad en distintas regiones del mundo.

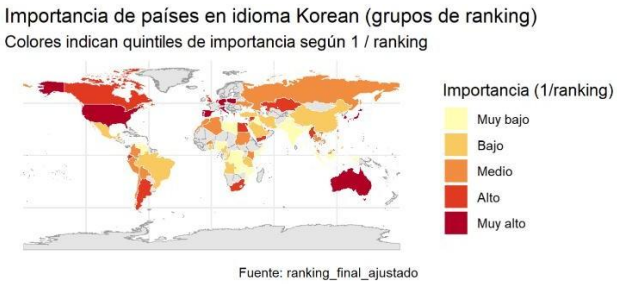
7. Anexos

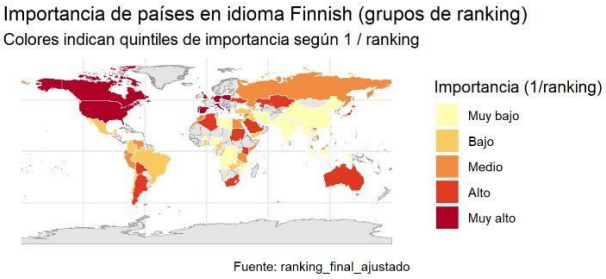
Anexo 1. Primeros y últimos 20 paises (respectivamente) del ranking ideal sin parroquialismo según población

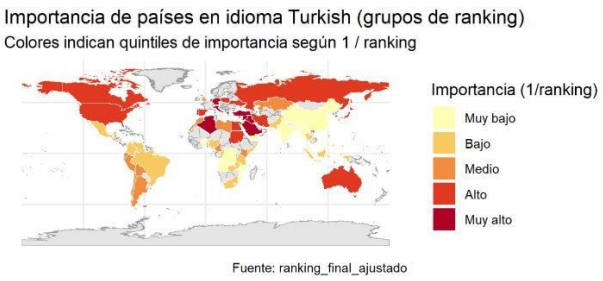
India
China
United States
Indonesia
Pakistan
Nigeria
Brazil
Bangladesh
Russia
Mexico
Japan
Egypt
Philippines
Vietnam
Iran
Turkey
Germany
Thailand
United Kingdom
Tanzania

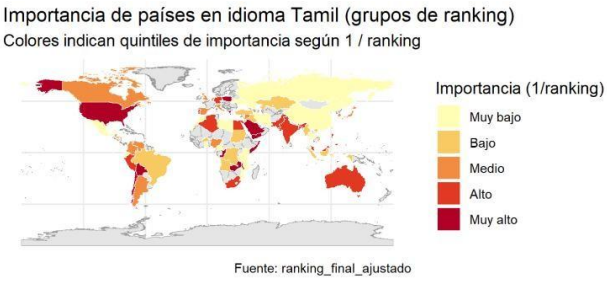
Syria
Burkina Faso
Taiwan
Malawi
Zambia
Kazakhstan
Chile
Somalia
Netherlands
Ecuador
Bolivia
Belgium
Jordan
United Arab Emirates
Honduras
Portugal
Greece
Israel
Libya
Republic of the Congo
Costa Rica

Anexo 2. Mapas estáticos

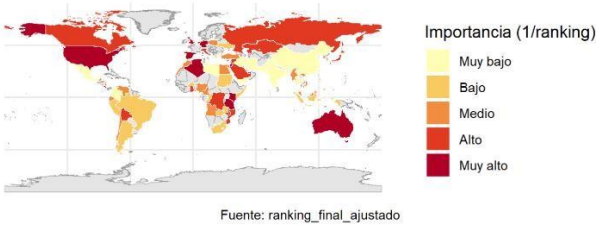






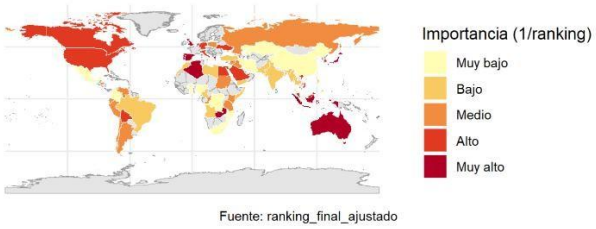


Importancia de países en idioma Swahili (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



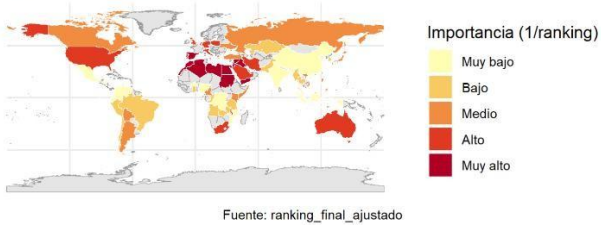
Fuente: ranking_final_ajustado

Importancia de países en idioma Indonesian (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



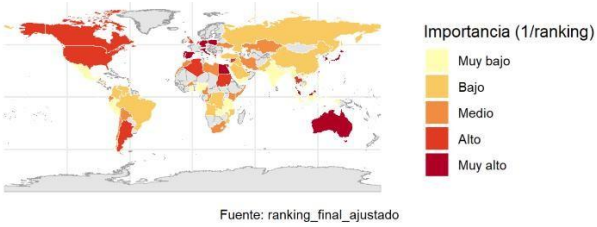
Fuente: ranking_final_ajustado

Importancia de países en idioma Arabic (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



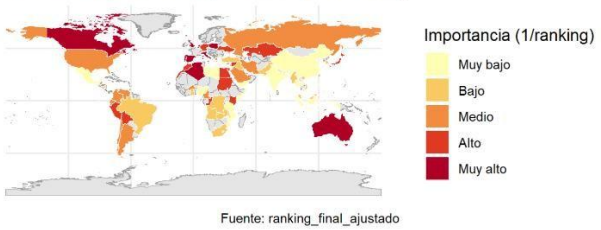
Fuente: ranking_final_ajustado

Importancia de países en idioma Chinese (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



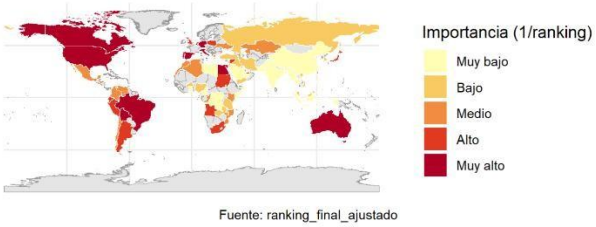
Fuente: ranking_final_ajustado

Importancia de países en idioma French (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



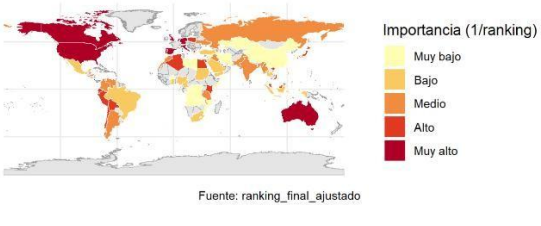
Fuente: ranking_final_ajustado

Importancia de países en idioma Portuguese (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



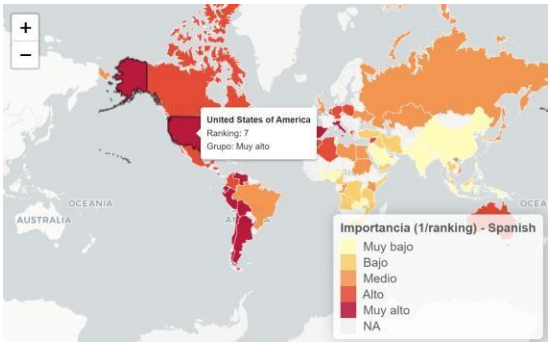
Fuente: ranking_final_ajustado

Importancia de países en idioma English (grupos de ranking)
Colores indican quintiles de importancia según 1 / ranking



Fuente: ranking_final_ajustado

Anexo 3. Muestra de mapas interactivos



Referencias

[1] Romano, A., Sutter, M., Liu, J.H. *et al.* National parochialism is ubiquitous across 42 nations around the world. *Nat Commun* **12**, 4456 (2021). https://doi.org/10.1038/s41467-021-24787-1

[2] Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556–1581. https://doi.org/10.1037/a0037737

[3] English Studies. (n.d.). Parochialism in literature: Literary theory. Recuperado de https://english-studies.net/parochialism-in-literature-literary-theory/

[4] IBM. (2024). Natural language processing. Recuperado de https://www.ibm.com/topics/natural-language-processing

[5] Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

[6] Politi, M. (2020, abril 29). Word2Vec Neural Network from Scratch. *Towards Data Science*. https://medium.com/towards-data-science/skip-gram-neural-network-from-scratch-485f2e688238

[7] Rong, X. (2014). *word2vec parameter learning explained* (arXiv:1411.2738). arXiv. https://arxiv.org/abs/1411.2738

[8] Barla, N. (2024, April 16). *The ultimate guide to word embeddings*. Neptune AI Blog. https://neptune.ai/blog/word-embeddings-guide#approximation

[9] Hosch, W.L. (2025, March 9). Zipf's law. Encyclopedia Britannica. https://www.britannica.com/topic/Zipfs-law

[10] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://aclanthology.org/Q17-1010.pdf

[11] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://www.stat.purdue.edu/~tlzhang/mathstat/ElasticNet.pdf

[12] Hoang, V. C. D., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative back-translation for neural machine translation. En A. Birch, A. Finch, T. Luong, G. Neubig, & Y. Oda (Eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 18–24). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2703

[13] Wikipedia contributors. (2025, April 18). *Google Translate*. Wikipedia. https://en.wikipedia.org/wiki/Google_Translate

[14] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805

[15] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.6.pdf

[16] **Stefanovic, M.** (2023, July 12). *What is the difference between a static and*

dynamic map? HERE360. https://www.here.com/learn/blog/static-map-vs-dynamic-map

[17] Wikipedia contributors. (2025, March 10). *Cartogram*. Wikipedia. https://en.wikipedia.org/wiki/Cartogram

[18] Jeworutzki, S. (2023). *cartogram: Create Cartograms with R* (Versión 0.3.0). CRAN. https://cran.r-project.org/web/packages/cartogram/cartogram.pdf

[19] Cheng, J., Karambelkar, B., & Xie, Y. (2024). *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library* (Versión 2.2.2). CRAN. https://cran.r-project.org/web/packages/leaflet/leaflet.pdf

[20] Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science (New York, N.Y.)*, *318*(5850), 636–640. https://doi.org/10.1126/science.1144237

[21] De Dreu, C. K. W., Balliet, D., & Halevy, N. (2014). Parochial cooperation in humans: Forms and functions of self-sacrifice in intergroup conflict. En A. J. Elliot (Ed.), *Advances in Motivation Science* (Vol. 1, pp. 1–47). Elsevier. https://doi.org/10.1016/bs.adms.2014.08.001

[22] Berg, N., Kim, JY. & Lee, K.M. Why is parochialism prevalent?: an evolutionary approach. *J Econ Interact Coord* **16**, 769–796 (2021). https://doi.org/10.1007/s11403-021-00328-3

[23] McLeod, S. (2023, October 5). *Social Identity Theory in Psychology (Tajfel & Turner, 1979)*. Simply Psychology. https://www.simplypsychology.org/social-identity-theory.html

[24] Pisor, A. C., & Ross, C. T. (2024). Parochial altruism: What it is and why it varies. *Evolution and Human Behavior, 45*(1), 2–12. https://doi.org/10.1016/j.evolhumbehav.2023.06.005

[25] Kiss, H. J., Rosa-Garcia, A., & Zhukova, V. (2020). Conditional cooperation in group contests. *PLOS ONE, 15*(12), Article e0244152. https://doi.org/10.1371/journal.pone.0244152

[26] Civai, C., Elbaek, C. T., & Capraro, V. (2024). Why scarcity can both increase and decrease prosocial behaviour: A review and theoretical framework for the complex relationship between scarcity and prosociality. *Current Opinion in Psychology, 60*, Article 101931. https://doi.org/10.1016/j.copsyc.2024.101931

[27] Zhou, R., Shi, K., Song, X., Li, S., & Zhou, W. (2024). The relationship between risk perceptions and negative emotions in the COVID-19: A meta-analysis. *Frontiers in Psychology, 15*, Article 1453111. https://doi.org/10.3389/fpsyg.2024.1453111

[28] Tariq, M. U., Babar, M., Khattak, A. S., Alshehri, M. D., & Kaleem, S. (2021). Human behavior analysis using intelligent big data analytics. Frontiers in Psychology, 12, Article 686610. https://doi.org/10.3389/fpsyg.2021.686610

[29] El Alaoui, I., Gahi, Y., Messoussi, R., Ounacer, H., & Lyhyaoui, A. (2018). A novel adaptable approach for sentiment analysis on big social data. Journal of Big Data, 5, Article 12. https://doi.org/10.1186/s40537-018-0120-0

[30] Dhaduk, H. (2024). 8 must-know NLP techniques to extract actionable insights from data. Simform. Recuperado de https://www.simform.com/blog/nlp-techniques/

[31] Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Relative Quality and

Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics*, *4*(4), 43. https://doi.org/10.3390/informatics4040043

[32] Kargaran, A. H., Yvon, F., & Schuetze, H. (2024). GlotCC: An open broad-coverage CommonCrawl corpus and pipeline for minority languages. En *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=aJ1yse8GEr

[33] Largest Cities by Population 2024. World Population Review. Retrieved February 23, 2025, from https://worldpopulationreview.com/cities

[34] Worldometer. (2025). Countries in the world by population (2025). Recuperado de https://www.worldometers.info/world-population/population-by-country/

[35] World Values Survey. (2022). World Values Survey Wave 7 (2017-2022). World Values Survey Association. Disponible en https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

[36] United Nations Development Programme. (n.d.). Documentation and downloads. Human Development Reports. Disponible en https://hdr.undp.org/datacenter/documentation-and-downloads

[37] Internet Penetration by Country 2024. World Population Review. Retrieved February 23, 2025, from https://worldpopulationreview.com/country-rankings/internet-62 penetration-by-country

[38] Gini Coefficient by Country 2024. World Population Review. Retrieved February 23, 2025, from https://worldpopulationreview.com/country-rankings/gini-coefficient-bycountry

[39] Facebook AI Research. (n.d.). Word vectors for 157 languages. FastText. Disponible en https://fasttext.cc/docs/en/crawl-vectors.html