Análisis empírico de la dispersión del español mexicano

Orlando Ramos[1], David Pinto[1], Belem Priego[1,2], Iván Olmos[1], Beatriz Beltrán[1]

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, México

[2] Laboratoire de Lexiques, Dictionnaires et Informatique, Université Paris 13, Francia

{orlandxrf, belemps, ivanoprkl}@gmail.com, {dpinto, bbeltran}@cs.buap.mx

Resumen. En este artículo se presenta un sistema que pretende facilitar el análisis de la dispersión del español mexicano. Se presentan gráficas resultantes, así como los modelos del sistema. El objetivo es mostrar el avance del sistema y su posible aplicación en el cálculo de la dispersión del lenguaje para otros idiomas. Los experimentos fueron realizados sobre dos tipos de corpora: noticias y tweets.

Palabras clave: Dispersión del idioma, noticias, tweets.

1. Introducción

En el idioma español existen rasgos y características que distinguen a un pueblo de otro de una manera muy particular, ya sea el español de España en donde su pronunciación y significado varian en comparación al de Latinoamérica, especialmente al español mexicano. México es un pueblo rico en cultura y tradiciones, y esto precisamente es lo que hace, que en cada una de sus entidades federativas se encuentren frases o palabras que los distingan de una manera en particular, estas pequeñas diferencias son las que dan pie a esta investigación, para lograr identificar las regiónes de nuestro país en donde se dan estas variaciones.

Como objetivo general nos hemos planteado analizar el uso del idioma español en la República Mexicana y su posible dispersión de acuerdo a la ubicación geográfica. Así, nuestros objetivos específicos son los siguientes:

1. Construir un corpus etiquetado geográficamente del español usado en la República Mexicana.
2. Estudiar diversos métodos para la identificación automática de la dispersión en el uso del lenguaje natural.
3. Construir un mecanismo de visualización para el uso del idioma, de acuerdo a la ubicación geográfica.
4. Evaluar los resultados obtenidos en base a métricas estándar tales como precisión y recall.

5. Comprobar si existe o no una marcada dispersión en el uso del español mexicano de acuerdo a la ubicación geográfica.

En particular, hemos definido la siguiente hipótesi que debe ser evaluada: Consideramos que el español mexicano ha sufrido un fenómeno de dispersión en cuanto a la agregación o modificación del vocabulario en función de la ubicación geográfica. Por tanto, se formulan la siguiente hipótesis: -existe una variabilidad significativa y consistente con la ubicación geográfica en el uso del español mexicano.

Para observar si la hipótesis es correcta, hemos diseñado un sistema basado en el web que permite observar la dispersión del idioma a través de toda la República Mexicana.

2. Trabajo relacionado

En esta sección se introducen algunos trabajos que se encuentran de alguna manera relacionados con el análisis del lenguaje. Si bien, algunos son análisis sintácticos, morfológicos o puramente léxicos, sirven como punto de partida para definir estrategias en el análisis de la dispersión del español en la República Mexicana.

En [7] se hace uso de los datos sobre diversidad de fonemas para estimar la fecha de origen del lenguaje. Esta diversidad fonética denota el número de unidades perceptualmente distintas de sonido (consonantes, vocales y tonos) en un lenguaje, dado que la diversidad de fonemas varía considerablemente entre las lenguas, y varios idiomas funcionan con un número limitado de fonemas. La diversidad de fonemas también se correlaciona positivamente con el número de idiomas que rodean, lo que sugiere que los fonemas, como otros rasgos culturales, se pueden pedir prestados. La diversidad de fonemas de un lenguaje depende del tamaño de la población de los hablantes, el área geográfica sobre la cual se habla el idioma, y la diversidad lingüística local.

En [6] se estudia la competencia entre los lenguajes o los rasgos culturales de dos poblaciones que se difunden en la misma zona geográfica combinando el modelo de competencia del lenguaje de Abrams-Strogatz (AS), y un modelo de dispersión humana en un sustrato no homogéneo. Por "competencia" se entiende que en cualquier momento hablantes pueden cambiar a otro lenguaje, como consecuencia de la interacción entre los hablantes de la lengua 1 y 2.

En [1] se explica cómo la diversidad lingüística y las bases biológicas del lenguaje han sido tradicionalmente tratados por separado. Algunos debates han generado una propuesta que argumenta a favor de un sistema biológico para usos específicos, por analogía como el sistema visual. Otra propuesta es que el lenguaje en cambio, confía en los mecanismos neuronales de dominio general evolucionando para otros fines. Sin embargo, existe un mayor acuerdo sobre el origen de la diversidad lingüística, que normalmente se atribuye a la evolución cultural divergente siguiendo la migración humana. En este trabajo se muestra un modelo teórico de la relación entre la diversidad lingüística y la base biológica para el lenguaje, que consiste en analizar el número de agentes hablantes de un

lenguaje y de cuantos principios cuenta, el conjunto de genes con que cada individuo está dotado y los tres alelos con los que cuenta cada gen.

En [2] se analiza cómo los estudios de lingüística, cultura y sociolingüística, geografía y dialectología es lo que permite un estudio espacial del lenguaje en su contexto geográfico, además de social y cultural. El estrecho vínculo de la dialectología con la necesidad del uso de mapas se ha centrado en la dimensión diatópica (fenómenos que se producen en una lengua en virtud de su extensión geográfica) de la lengua. La dialectología estudia las variaciones de una lengua según los lugares, y la geografía lingüística es uno de los métodos para espacializar y reconocer estas variaciones en cartografías y mapas. Según esto, en la lingüística, y en la elaboración de lo que se conoce como atlas lingüísticos, la dimensión espacial ha estado presente y ha sido reconocida para visualizar distribuciones espaciales. Sin embargo, la manera de enfocar los estudios espaciales en la lingüística antes de los sesenta fue la de representar aspectos lingüísticos en escenarios regionales o desarrollar técnicas cuantitativas y estadísticas de recolección de información en mapas, lo cual ha dado lugar a lo que en conjunto se conoce como geografía lingüística. Así, los estudios regionales clásicos desarrollan una recolección de información en campo presentando esto en cartas geográficas que permiten ver la distribución espacial de los hablares regionales, que se integran en el hablar común de una nación, facilitando la delimitación de zonas dialectales.

Aunque existen varios estudios que realizan análisis morfológico del español, tal vez uno de los más completos ha sido realizado por Gelbukh y Sidorov [5,3,4]. Ellos plantean un sistema computacional que analiza el español usando un modelo denominado "de generación". Se plantea en dicho artículo que este modelo es capaz de obtener mejores resultados para a lenguajes con alternaciones irregulares de raíz, como es el caso del espa´ nol. El modelo consiste en un conjunto de reglas para obtener todas las raíces de una forma de palabra para cada lexema, su almacenamiento en el diccionario, la producción de todas las hipótesis posibles durante el análisis y su comprobación a través de la generación morfológica.

Quizás el trabajo más relacionado es el presentado en [8], en donde se estudian diferentes variedades del español. En particular para los siguientes países: México, Argentina, Perú y España. Se hace uso de textos noticieros también. Desde nuestro particular punto de vista, este trabajo puede servir como base para analizar la dispersión del español en México, teniendo en cuenta que de alguna manera está última tarea (la nuestra) presenta una mayor grado de complejidad que aquella presentada en [8].

3. Análisis y Diseño del Sistema

El sistema desarrollado considera el cálculo de ngramas como un medio de determinación del grado de dispersión del idioma. Se usan n-gramas de letras y n-gramas de palabras. Estos n-gramas son calculados por regiones geográficas (estados de la República Mexicana) y comparados con respecto a la media na-

cional. En general, se calcula la frecuencia de cada n-grama (unigrama, bigrama o trigrama) y se ordenan los n-gramas en forma descendente. Se usa una porción de los n-gramas más frecuentes y se calcula el grado de intersección entre los n-gramas calculados a nivel nacional y aquellos calculados a nivel estatal. El grado de traslape indica la cercanía del idioma en el estado con respecto a la media nacional.

Un primer Corpus que se uso para este análisis esta conformado por notas periodísticas publicadas digitalmente por la Organización Editorial Mexicana (OEM) para cada una de las entidades federativas de la República Mexicana, con más de medio millón de noticias. Si bien, este corpus muestra información perteneciente a cada estado, es muy probable que el modelado basado en n-gramas represente el estilo de escritura de un reportero, más que el estilo de escritura de la población en general. Por esta razón, hemos recopilado un segundo corpus basado en Tweets.

Este segundo Corpus se extrajo de Twitter, con aproximadamente medio millón de tweets. La forma de extracción fue a tráves de geo-coordenadas de las capitales de todas las entidades federativas de la República Mexicana. Se buscaron todos aquellos tweets que estuviesen localizados en un radio de 10 Kilómetros alrededor de cada capital. Consideramos que este corpus es mucho más representativo del lenguaje que el corpus de noticias, aunque tiene las particularidades asociadas a los Tweets: textos cortos y con un vocabulario ligeramente modificado (eliminación de vocales y errores ortográficos).

En la Figura 1 se muestran los casos de uso del sistema propuesto. Los diagramas de procesos y modelos de procesos se muestran en la Figuras 2 y 3, respectivamente.

En general, en el sistema se ha buscado poder mostrar la dispersión del idioma (español mexicano) usando diversos mecanismos de visualización. Se usan tablas de resultados, gráficas y mapas. De momento, los dos corpora usados son estáticos, pero como trabajo a mediano plazo consideramos implementar un módulo que permita subir documentos al sistema para que sean procesados automáticamente.

4. Resultados

El sistema desarrollado permite visualizar en base a tablas, gráficas y mapas el grado de dispersión del español en la República Mexicana. Hemos denominado a este sistema AEDEM por ser las siglas de Análisis Empírico de la Dispersión del español mexicano (ver Figura 4).

En la Figura 5 se muestra una tabla que indica la cantidad de trigramas de letras que se comparten entre cada estado y la media nacional. El cálculo es realizado usando el corpus de Tweets como datos de entrada. La tercera columna muestra la cantidad de Tweets por estado, y las columnas siguientes muestran la cantidad de n-gramas que comparten con la media nacional. Si el usuario coloca el cursor sobre un valor se mostrará el porcentaje de cobertura.

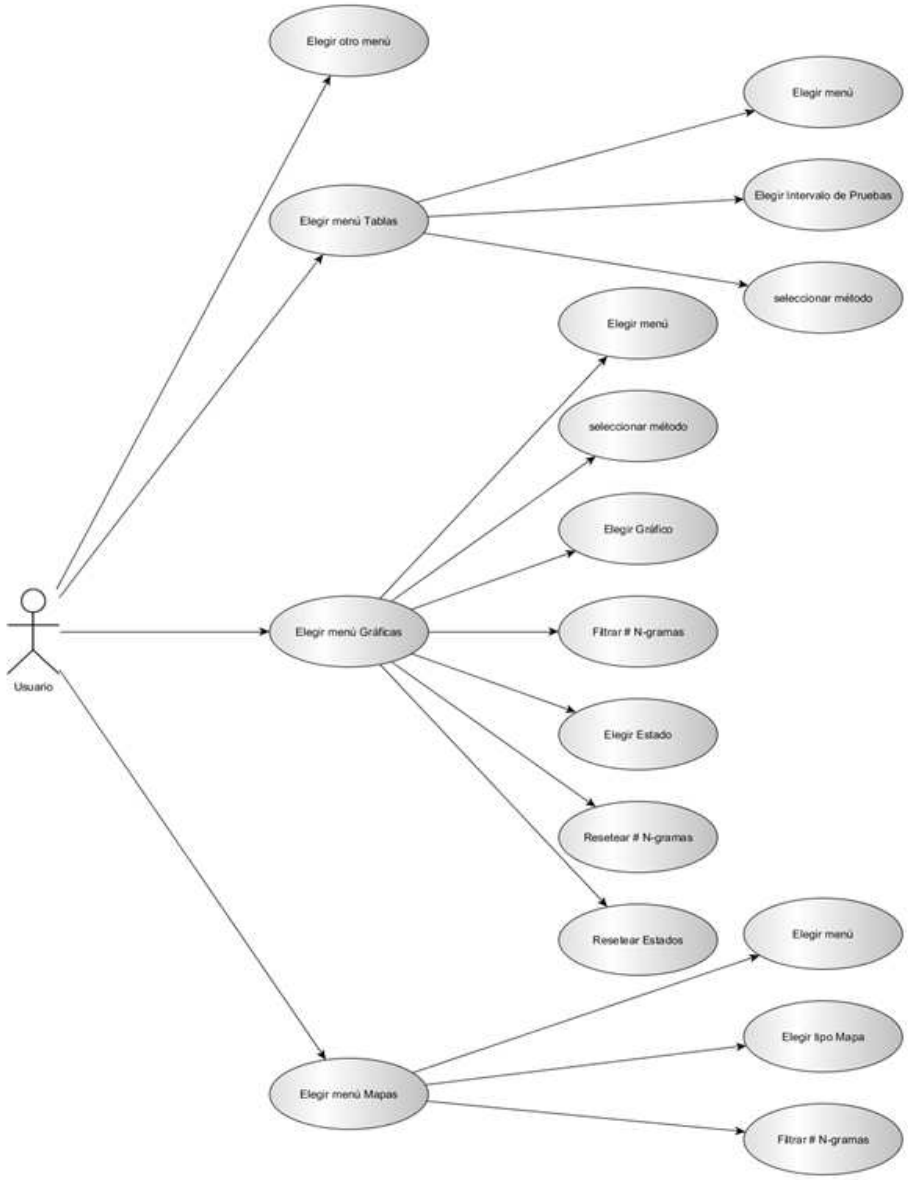


Fig. 1. Casos de uso del sistema AEDEM

En la Figura 6 se presenta una gráfica comparativa entre tres estados de la República Mexicana y su grado de cobertura con respecto a la media nacional. El sistema permite hacer este tipo de comparaciones entre cualquier número de estados.

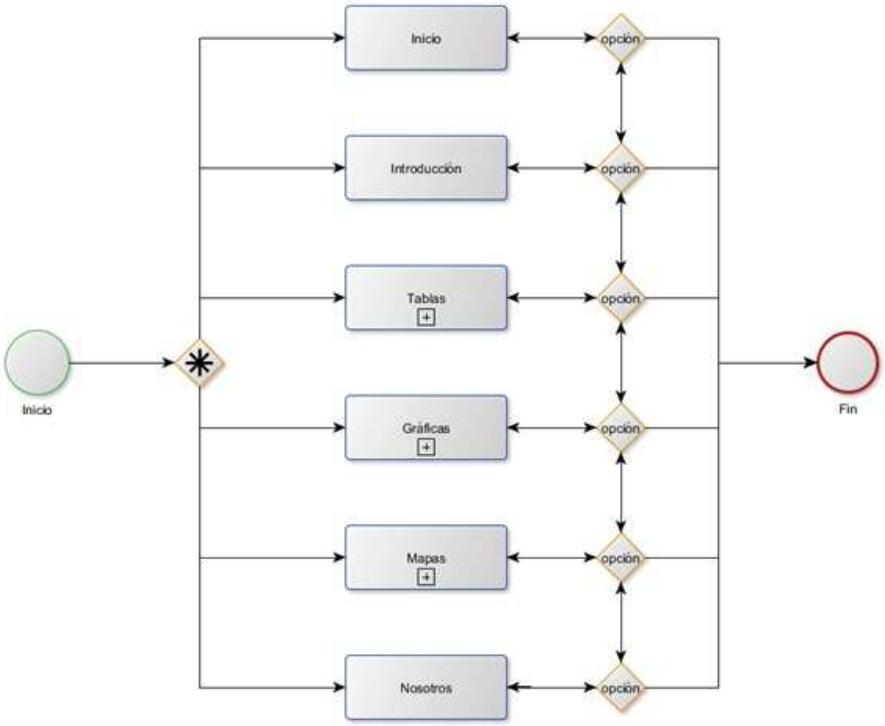


Fig. 2. Diagrama de procesos

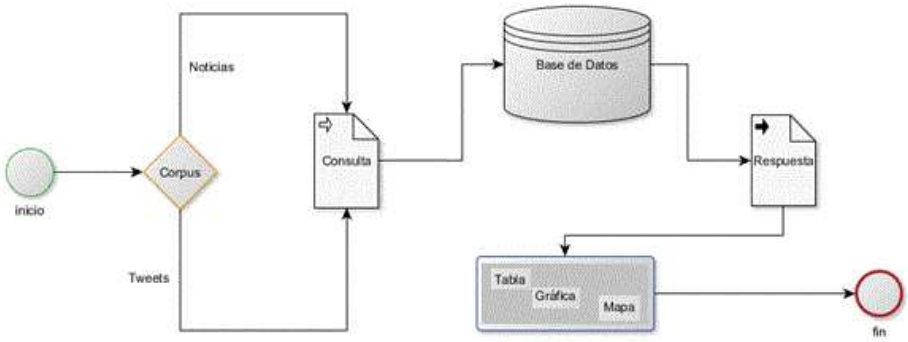
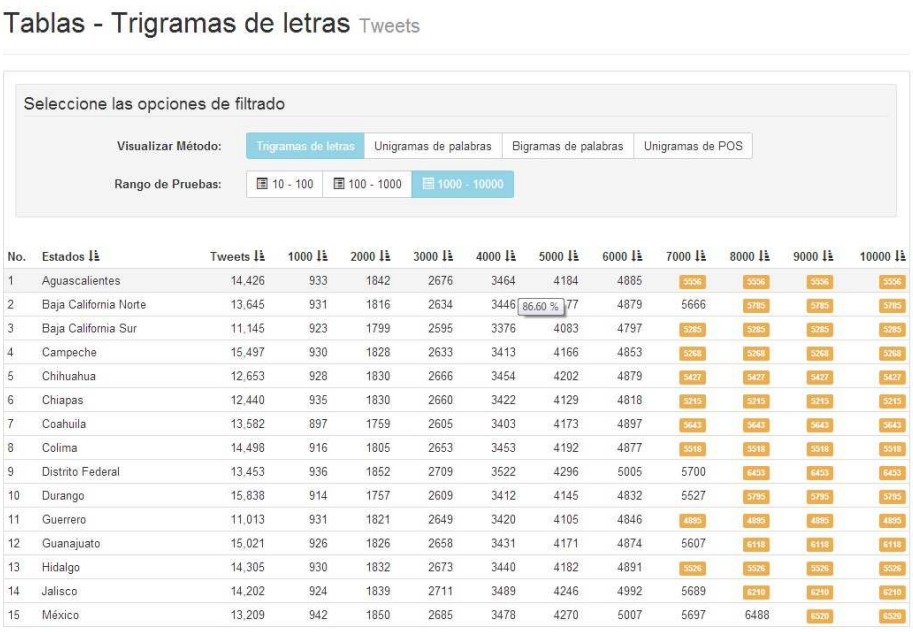


Fig. 3. Modelo de procesos

Observe en las Figuras 7 y 8, cómo ciertos estados tienen un color mucho más obscuro, lo que significa que se acercan mucho más a la media nacional. Mientras que otros estados muestran una coloración mucho más clara, lo cual significa que los modelos del lenguaje están más alejados de la media nacional.

Fig. 4. Pantalla principal del sistema AEDEM



No.	Estados	Tweets	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
1	Aguascalientes	14,426	933	1842	2676	3464	4184	4885	5556	5556	5556	5556
2	Baja California Norte	13,645	931	1816	2634	3446	86.60 %	4879	5666	5785	5785	5785
3	Baja California Sur	11,145	923	1799	2595	3376	4083	4797	5285	5285	5285	5285
4	Campeche	15,497	930	1828	2633	3413	4166	4853	5268	5268	5268	5268
5	Chihuahua	12,653	928	1830	2666	3454	4202	4879	5427	5427	5427	5427
6	Chiapas	12,440	935	1830	2660	3422	4129	4818	5215	5215	5215	5215
7	Coahuila	13,582	897	1759	2605	3403	4173	4897	5643	5643	5643	5643
8	Colima	14,498	916	1805	2653	3453	4192	4877	5518	5518	5518	5518
9	Distrito Federal	13,453	936	1852	2709	3522	4296	5005	5700	6453	6453	6453
10	Durango	15,838	914	1757	2609	3412	4145	4832	5527	5795	5795	5795
11	Guerrero	11,013	931	1821	2649	3420	4105	4846	4895	4895	4895	4895
12	Guanajuato	15,021	926	1826	2658	3431	4171	4874	5607	6118	6118	6118
13	Hidalgo	14,305	930	1832	2673	3440	4182	4891	5526	5526	5526	5526
14	Jalisco	14,202	924	1839	2711	3489	4246	4992	5689	6210	6210	6210
15	México	13,209	942	1850	2685	3478	4270	5007	5697	6488	6520	6520

Fig. 5. Tablas de análisis de n-gramas

En particular, para el caso del corpora de noticias, se puede observar que el uso del lenguaje en el centro de la República tiene un impacto importante sobre el resto de los estados. Es posible que esto se deba al hecho de que la OEM tiene

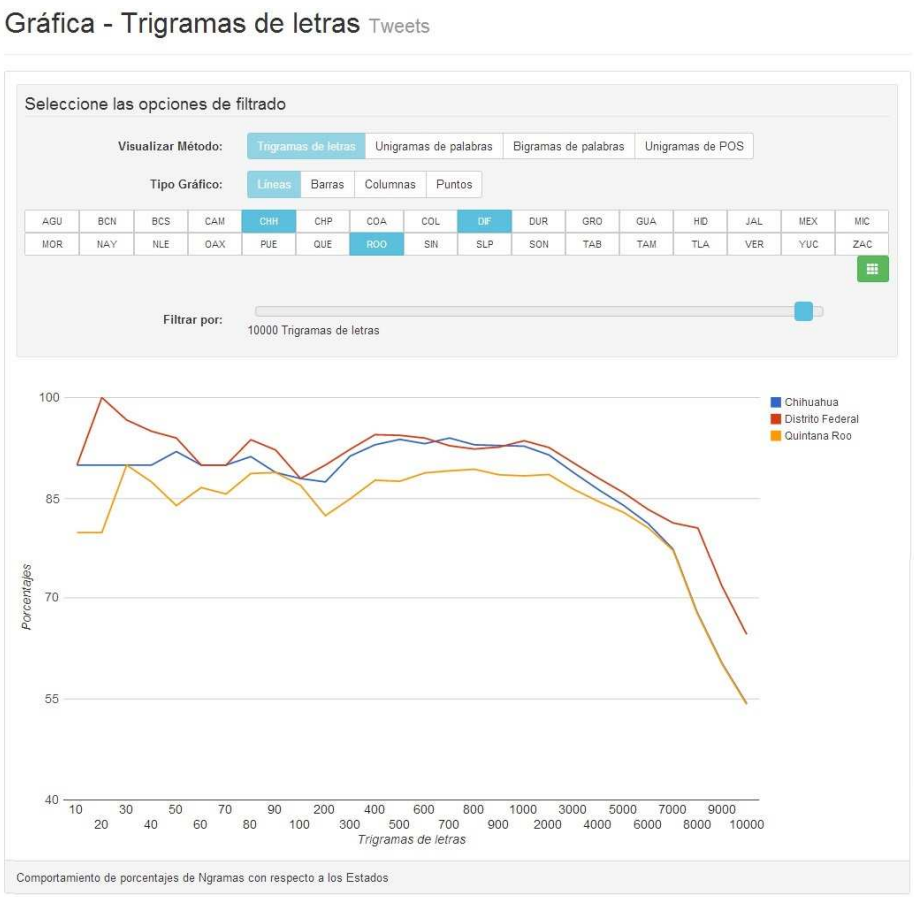


Fig. 6. Gráfica comparativa de n-gramas para tres estados

presencia en todo el país pero existe una gran cantidad de noticias que provienen precisamente del centro de México.

Cuando se toman en cuenta los textos escritos mediante Tweets, la historia es bastante diferente. Se observa una clara dispersión del idioma en estados como Oaxaca y Colima. Algunas zonas de la República como el sureste también muestran una dispersión con respecto al centro de México. Tal y como lo mencionamos anteriormente, consideramos que este mapa refleja mucho mejor el grado de dispersión del español mexicano de cada estado con respecto a la media nacional. Sin embargo, este hecho habrá que analizarlo con mayor detalle.

Con la finalidad de garantizar que la muestra es representativa, en la Tabla 1 se muestra el número de usuarios diferentes por estado considerados en este experimento. Se observa que el total de usuarios es de 196,288, sin embargo, dado que ciertos usuarios fueron localizados en más de un estado, es necesario calcular el número de usuarios diferentes en toda la colección, y este número es

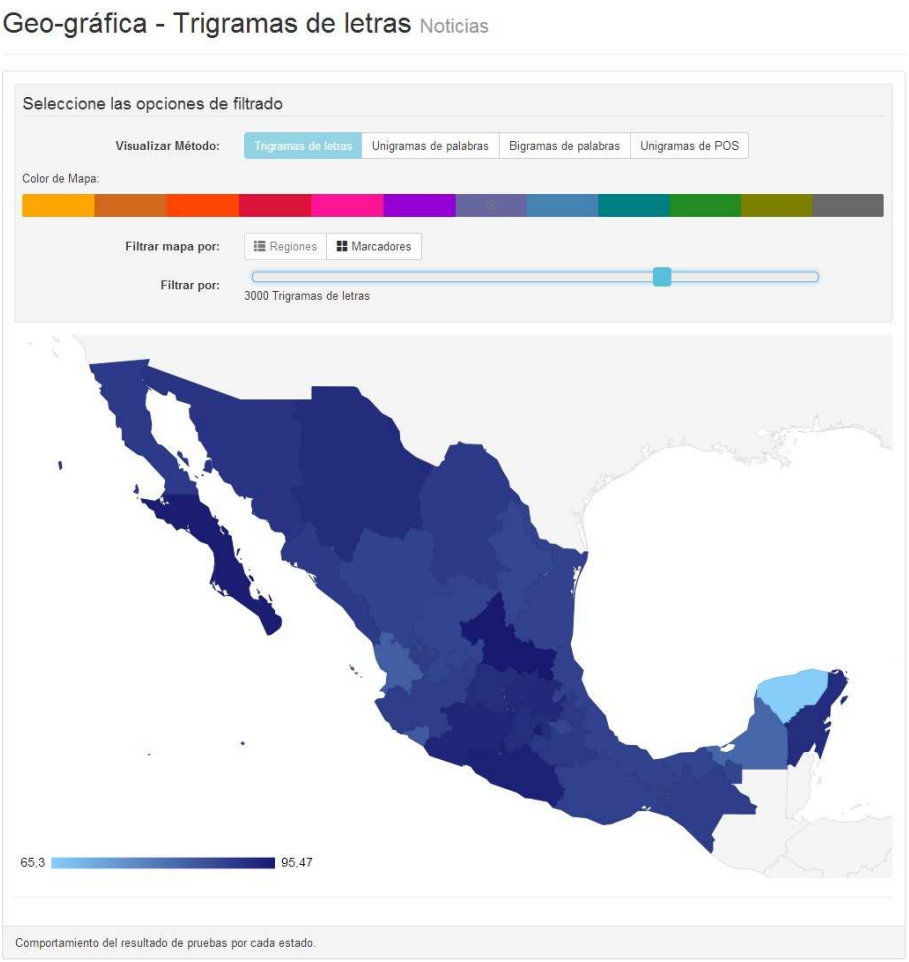


Fig. 7. Visualización de la dispersión del idioma en el mapa (n-gramas de letras) usando el corpus de noticias

177,753. De cualquier manera, consideramos que esta cantidad de usuarios puede ser suficiente para estimar el grado de dispersión del español en la República Mexicana.

5. Conclusiones

Se ha presentado un sistema para la visualización del grado de dispersión del español mexicano en México. El sistema es flexible en el sentido que permite visualizar los datos en formas de tablas, gráficas y mapas. De esta manera, el usuario puede analizar desde diversas perspectivas el grado de cobertura del

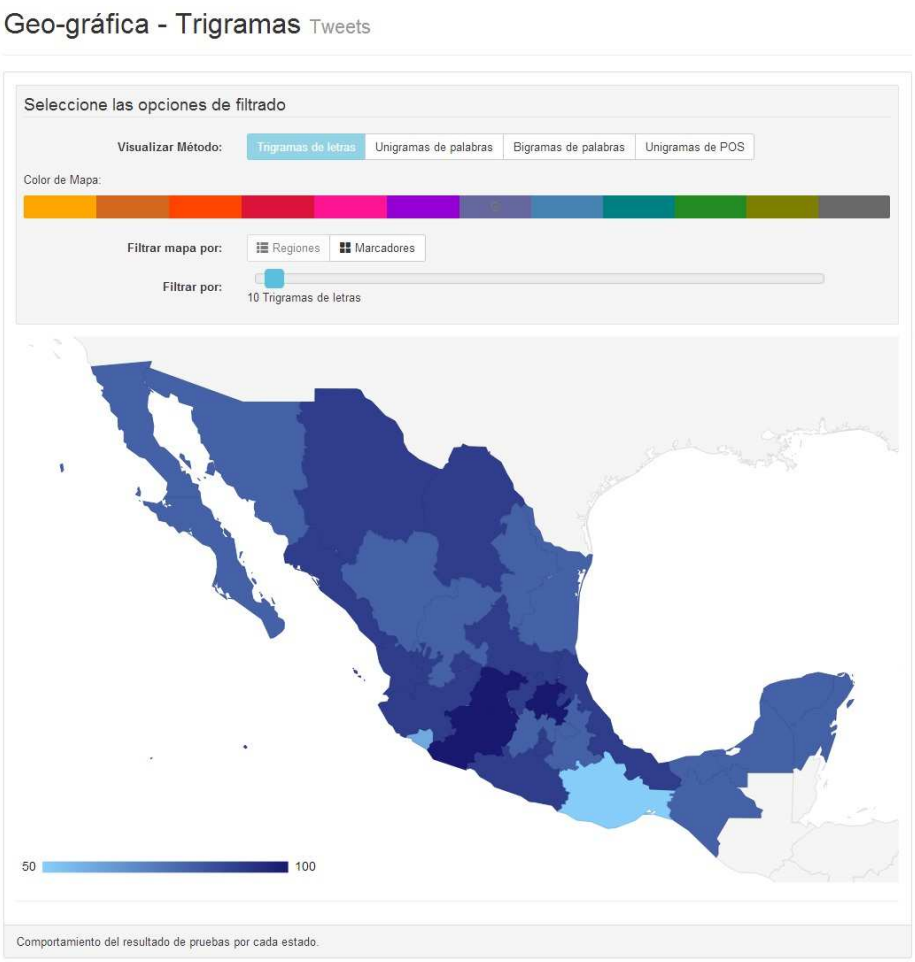


Fig. 8. Visualización de la dispersión del idioma en el mapa (n-gramas de letras) usando el corpus de tweets

idioma en los diferentes estados de la República Mexicana. Para el caso actual, se usan dos corpora, uno del dominio de noticias y el otro extraído desde Twitter. Como trabajo a futuro se implementarán métricas mucho más formales para medir el grado de entropía entre el modelo del lenguaje de un estado y la media nacional. También se pondrá a disposición el sistema con una opción para subir datos a discreción, lo cual permitirá una total flexibilidad en el usuario final.

Referencias

1. Baronchelli, A., Chater, N., Pastor-Satorras, R., Christiansen, M.H.: The biological origin of linguistic diversity. PloS one 7(10), e48029 (2012)

Tabla 1. Cantidad de usuarios diferentes por estado considerados en el corpus.

Estado	**Usuarios**	**Estado**	**Usuarios**
Aguascalientes	4,536	Morelos	7,882
Baja California Norte	5,743	Nayarit	2,351
Baja California Sur	2,158	Nuevo León	8,857
Campeche	7,684	Oaxaca	6,891
Chihuahua	6,178	Puebla	8,623
Chiapas	5,074	Querétaro	5,497
Coahuila	5,939	Quintana Roo	10,189
Colima	3,302	Sinaloa	5,491
Distrito Federal	11,367	San Luis Potosí	4,144
Durango	3,569	Sonora	6,710
Guerrero	2,226	Tabasco	5,993
Guanajuato	6,224	Tamaulipas	3,267
Hidalgo	5,557	Tlaxcala	9,111
Jalisco	8,373	Veracruz	4,863
México	11,426	Yucatán	8,668
Michoacán	4,824	Zacatecas	3,571

2. Córdoba, H., Gloria, A.: Geografía, lingüística y geolingüística. una propuesta para comprender el contacto dialectal. Forma y Función 24, 47–60 (2011)
3. Gelbukh, A., Sidorov, G.: Approach to Construction of Automatic Morphological Analysis Systems for Inflective Languages with Little Effort. In: Computational Linguistics and Intelligent Text Processing. pp. 157–162. Springer (2003)
4. Gelbukh, A., Sidorov, G.: Analizador morfológico disponible: un recurso importante para pln en español. In: Memorias de talleres del congreso internacional Iberamia-2004. pp. 209–216 (2004)
5. Gelbukh, A., Sidorov, G., Velázquez, F.: Análisis morfológico automático en español a través de generación. Revista Escritos 28, 9–26 (2003)
6. Patriarca, M., Heinsalu, E.: Influence of geography on language competition. Physica A: Statistical Mechanics and its Applications 388(2), 174–186 (2009)
7. Perreault, C., Mathew, S.: Dating the origin of language using phonemic diversity. PloS one 7(4), e35289 (2012)
8. Zampieri, M., Gebre, B.G., Diwersy, S.: N-gram language models and pos distribution for the identification of spanish varieties. In: Proceedings of TALN2013. pp. 580–587 (2013)