



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Hacia un método computacional para detectar léxico contrastivo

Tesis de Licenciatura en Ciencias de la Computación

Damián Eliel Aleman

Director: Juan Manuel Pérez y Santiago Kalinowski

Codirector: Agustín Gravano

Buenos Aires, 2017

A Edgar Altszyler por su gran ayuda para el desarrollo de la métrica para detectar el léxico contrastivo.

Al equipo de lexicógrafos de la Academia Argentina de Letras que analizaron de forma pormenorizada las palabras candidatas a ser contrastivas, entre ellos Pedro Rodríguez Pagni, Gabriela Pauer, María Sol Portaluppi y Josefina Raffo.

A Federico Plager y a Mariela Sued, por tomarse el tiempo y la dedicación de leer esta tesis y por sus devoluciones.

A mis directores, a Juan Manuel Pérez por todas las reuniones y el acompañamiento constante, a Santiago Kalinowski por acercarme a la lingüística y a Agustín Gravano por sus aportes para entender un poco más el proceso de investigación.

A mi mamá, mi papá y mis hermanos por estar siempre.

A Juli, a Edu y al abuelo.

A los amigos de la facu: Martín, Javier, Luis, Guille, Fixman, Ralo, JP, Diego y Guerson.

A Martín, por ser compañero de viajes, de antidomingos y de toda la carrera.

A mis compañeros de Hacoaj, especialmente a Gabi y Eio.

A mis compañeros del Dari.

Resumen

El crecimiento de la cantidad de datos en la web y los recursos computacionales de las últimas décadas dan la posibilidad de investigar fenómenos lingüísticos a gran escala, tarea casi imposible de realizar manualmente. En el área de la lexicografía se conoce como palabras contrastivas, a aquellas de una misma lengua que tienen una frecuencia de uso significativamente distinta en dos o más regiones. En el presente trabajo desarrollamos una métrica para detectar estas palabras. Esta hace uso de la entropía de la información sobre la cantidad de ocurrencias de cada palabra y la cantidad de usuarios que la usan para dar un indicio de la contrastividad. Para evaluarla se recolectó a partir de Twitter un conjunto de datos con más de 190 millones de palabras escritas y más de 20 mil usuarios de las 23 provincias argentinas. Este trabajo multidisciplinario se hizo en colaboración de la Academia Argentina de Letras, la cual realizó la validación y comprobó diferencias relevantes en la extensión de uso de unas 300 palabras. Varias de ellas formarán parte de las palabras del Diccionario del habla de los argentinos del año próximo.

Palabras claves: Procesamiento del Lenguaje Natural, Lingüística Computacional, Lexicografía Contrastiva

Índice general

1.. Introducción	1
1.1. Trabajo previo en el área	2
1.2. Twitter	2
1.3. Lingüística de Corpus y Lingüística computacional	4
1.4. Objetivo del estudio	5
2.. Datos: Extracción y procesamiento	6
2.1. Extracción de Datos	7
2.1.1. Búsquedas geolocalizadas	7
2.2. Tokenización y normalización	8
2.3. Caracterización de la muestra	9
2.3.1. Distribución temporal de tuits	9
2.3.2. Información de palabra por provincia	11
3.. Métricas para detectar palabras salientes	12
3.1. Búsqueda de contrastes	13
3.1.1. Métricas para medir el contraste en la frecuencia de las palabras	14
3.1.2. Valor de información	14
3.1.3. Robusteciendo la métrica desarrollada	16
3.1.4. Frecuencia de las palabras	17
3.2. Distribución de la entropía	20
3.3. Distribución del valor de contrastividad	20
3.4. Proporción acumulada de ocurrencias	21
4.. Análisis de las palabras contrastivas encontradas	23
4.1. Palabras candidatas	24
4.2. Regiones de palabras	24
4.3. Problemas en el conjunto de datos	24
4.4. Caracterización de las palabras identificadas como contrastivas	26
4.5. Validación estadística	29
4.5.1. Test Hipergeométrico	29
4.5.2. Test t de Welch	31
5.. Conclusiones y trabajo futuro	34
5.1. Conclusiones y trabajo futuro	35
6.. Apéndice	36
6.1. La entropía como medida del desorden	37

6.2. Tablas y Gráficos	37
7.. Bibliografía	41

1. INTRODUCCIÓN

1.1. Trabajo previo en el área

La dialectología es un campo que estudia la variación de las lenguas según la región geográfica. La investigación en estos campos suele utilizar variables lingüísticas: atributos fonéticos, sintácticos y léxicos.

Una palabra es *contrastiva* cuando la frecuencia de uso en dos regiones es muy diferente. Por ejemplo, la palabra “che”¹, o “metegol”² son términos más usados en la Argentina que en España. Un ejemplo dentro de la Argentina, es el de “gurisada” conocido en el litoral y noroeste argentino como un conjunto de chicos [dL08]. Actualmente, las palabras con contraste léxico en distintas regiones se detectan por medio de cuestionarios como el de Almeida [AV95] o en métodos que dependen, en mayor o menor medida, de la intuición y sensibilidad de los lexicógrafos. Las encuestas están integradas por grupos temáticos centrales como la casa, la familia, la enseñanza, el cuerpo humano, etc. Sobre cada grupo temático se les indica a las personas entrevistadas un repertorio de palabras por cada noción, para ver si las conocen y con qué frecuencia se las usa. Con el presente trabajo planteamos cambiar el paradigma y detectar automáticamente las palabras usadas en distintas regiones y sus frecuencias.

Uno de los corpus lingüísticos del español más reconocidos es el *CORPES XXI* [Esp], creado por la Real Academia Española con una distribución de 25 millones de formas por cada uno de los años comprendidos en el periodo 2001 a 2012. Sin embargo, dicho corpus tiene dos desventajas importantes: por un lado, la cantidad de palabras de América Latina está subrepresentada en relación a la demografía ya que el 34,30 % de las palabras del dataset proviene de textos de España y 65,70 % de los demás países hispanoparlantes. Por otro lado, uno no dispone de todo el dataset, sino que solamente se pueden hacer las consultas desde su página web. Estas consultas están limitadas en cuanto a la cantidad de solicitudes y a las funcionalidades que estas proveen, especialmente tener que saber de antemano la forma a buscar.

Una de las virtudes de hacer un corpus con un método de recolección de textos de forma automática se desprende del mayor tamaño del corpus en comparación con métodos manuales como digitalización de textos. A pesar de haber comenzado hace varias décadas la recolección de textos de la web para realizar corpus, no hay muchos en el idioma español. Uno es el de Mark Davies, en el cual se utilizaron las páginas web para recolectar los textos, con dos mil millones de palabras en español, y se dividieron las páginas a partir del país de origen identificado por Google [Dav15]. A pesar de ser un corpus muy grande con anotaciones, no permite diferenciar las frecuencias de las palabras o los textos originados en regiones dentro de cada país. Es por eso que para este trabajo elegimos construir nuestro propio conjunto de datos que contenga una gran cantidad de palabras. Para esta tarea, usamos datos de Twitter. A continuación haremos una descripción de esta plataforma y de sus principales ventajas sobre otras alternativas posibles.

1.2. Twitter

Twitter³ es un servicio de microblogging creado en 2006. Los usuarios son variados, desde personas, instituciones gubernamentales y no gubernamentales hasta bots (i.e pro-

¹ Tratamiento que se usa para llamar, pedir atención o dirigirle a alguien la palabra [dL08].

² Juego mecánico en el que con pequeños muñecos se simula un partido de fútbol (fútbolín) [dL08].

³ www.twitter.com

gramas que corren tareas automáticamente). Cada usuario puede escribir textos llamados tuits, que tienen una longitud máxima de 140 caracteres⁴. Las relaciones en Twitter no necesitan ser recíprocas. Es decir, uno puede seguir a una persona en cuyo caso va a poder leer todos los tuits generados por ella, como también puede ser seguido por una persona. Un tuit puede ser respondido, como también puede ser retuiteado. El retuit es un mecanismo para diseminar por la red tuits generados por otros usuarios. De esta manera si un usuario *A* realiza un retuit generado por el usuario *B*, cualquier seguidor de *A* también va a recibir ese tuit en su panel, al cual llamaremos *timeline*. Si bien los tuits que se ven en el timeline son solo aquellos generados por los usuarios que uno sigue, todos los tuits son públicos, es decir que se pueden acceder a ellos a través de búsquedas en la plataforma. Para dar una noción de la cantidad de usuarios en la Argentina, en el año 2016 había 11,8 millones de usuarios de Twitter. Teniendo en cuenta que en ese momento 15 millones de personas tenían smartphones se deduce que el 70 % de la gente con smartphones poseía una cuenta de Twitter.

Las ventajas de Twitter sobre otras plataformas son varias: provee una interfaz pública para obtener tuits de cualquier persona, independientemente de que haya una relación con el usuario que escribió los tuits en la red social. Es decir, uno puede ver los tuits de otra persona, sin necesidad de ser un *seguidor* de la misma. Además, a diferencia de un portal de noticias donde los comentarios suelen estar relacionados con estas, en Twitter son más amplios los tópicos de los comentarios. Por otro lado, Twitter es una tecnología que permite hacer escalable el trabajo a diferentes países ya que con una misma interfaz se pueden obtener los textos de cualquier región. En cambio, si se elige un portal de noticias para sacar comentarios de usuarios, es necesario conocer la estructura de cada página para obtener esos datos. Otra virtud de Twitter es la identificación de los usuarios sobre la cual se podrían inferir datos como género, edad y ubicación de cada uno. Por último, una ventaja sobre otro método de recolección es que a través de Twitter se pueden recolectar palabras con una granularidad regional muy variable y obtener información de quienes los escriben. Existen otras redes sociales de las que se podrían obtener textos para analizar, pero tienen la desventaja de ser privadas, como *Facebook*, o acotadas en términos de los temas que se hablan, como *LinkedIn*.

En los últimos años se han publicado numerosos trabajos que utilizaron datos de Twitter, desde la detección y monitoreo de terremotos en tiempo real[SOM10], análisis de sentimientos y de la opinión pública [Liu12], predicciones del mercado de valores [PP10] o los resultados de elecciones nacionales [TSSW10]. También se ha utilizado para localizar enfermedades por región [PD11].

En cuanto a trabajos relacionados con la lingüística cabe mencionar el trabajo de Eisenstein et al. [EOSX10] en el que identifica palabras con una gran afinidad regional realizando un modelo probabilístico en el que asumen que las distribuciones léxicas dependen de la región geográfica y de una división de tópicos. En otras palabras, suponen que hay una división de temas sobre todo el dataset y dependiendo de la región del autor, este es más propenso a escribir con una variación dada. El mismo autor realizó un trabajo para identificar variables léxicas y detectar regiones dialectales [Eis14]. El trabajo de Gonçalves et al. [GS14] consistió en analizar las variaciones diatópicas de ciertos conceptos en las grandes ciudades hispanoparlantes. Utilizaron la técnica K-means [Bis06] para obtener regiones dialectales. Por otro lado G. Doyle et al. [Doy14] propone un método bayesiano para estimar la distribución de la frecuencia de una palabra (o frase) condicional a la ubi-

⁴ Recientemente han aumentado el límite a los 280 caracteres.

cación de la persona que la escribe. Hay dos grandes diferencias entre el presente trabajo con los que se mencionaron anteriormente. La primera diferencia es que nuestro análisis está hecho con una métrica basada en la teoría de la información. La segunda diferencia es que este trabajo se realizó con textos en español, mientras que la mayoría de los antes mencionados están hechos sobre corpus en inglés.

1.3. Lingüística de Corpus y Lingüística computacional

La lingüística de corpus es una rama de la lingüística que investiga a través de conjuntos de muestras de uso de la lengua [MH11]. Aunque lo más común es que estas muestras provengan de textos, también se puede extraer datos a partir de grabaciones de voz o videos. Si bien esta rama nació analizando corpus de forma manual, el rápido crecimiento tecnológico de las últimas décadas dio la posibilidad de tener corpus con millones de palabras y hacer algunos estudios de manera más automatizada, usando menos recursos humanos y ahorrando tiempo. A continuación, haremos un breve resumen de algunos hitos en lingüística de corpus.

En el año 1967 se publicó el primer corpus con un millón de palabras denominado Brown Corpus, uno de los pioneros en la lingüística de corpus. Recién en el año 1995 se consiguió generar un corpus del inglés británico con 100 millones de palabras, titulado British National Corpus (BNC). El objetivo era que se convirtiera en una muestra representativa del inglés británico de aquella época. Es importante destacar, por la gran cantidad de recursos que se utilizaron, que este trabajo se hizo con la colaboración de tres grandes editoriales, la Universidad de Oxford, la Universidad de Lancaster y la Biblioteca Británica. El 90 % del corpus era de origen escrito y el 10 % restante surgió de grabaciones de conversaciones transcritas, de voluntarios de distintas edades, clases sociales y regiones. Estas conversaciones fueron producidas a partir de diferentes situaciones, algunas formales, como reuniones de gobierno, y otras más informales como programas de radio. Una de las grandes diferencias entre el BNC y los corpus ya existentes en ese momento, es que además de publicar los datos para investigaciones académicas, también se dio acceso a los datos para uso comercial y educativo.

El crecimiento de la cantidad de datos generados mediante sistemas informáticos en las últimas décadas fue tal que en el año 2003 Kilgariff y Grefenstette [KG03] se preguntaron acerca de la posibilidad de utilizar la Web como fuente para recolectar textos. La Web resulta una gran oportunidad para el estudio de las lenguas ya que provee una cantidad inmensa de datos, accesibles de forma gratuita y con disponibilidad inmediata. Cuando Kilgariff y Grefenstette analizaron las posibilidades de la web, solamente se encontraba la posibilidad de estimar la frecuencia de una palabra a través de la cantidad de resultados que devolvían los motores de búsqueda. En ese sentido, había varias críticas que se le podían hacer al contenido que se encontraba en la web. En primer lugar, estaba la dificultad de que un estudio sea reproducible. En segundo lugar, el contenido solía estar desnormalizado, teniendo errores sintácticos, semánticos, ortográficos y otras variaciones. La primera problemática se resuelve fácilmente en la actualidad recolectando un dataset como proponemos en este trabajo. Por otro lado, aunque la necesidad de la normalización siga vigente, esta se compensa con la inmensa cantidad de datos que es posible recolectar.

En particular, la lengua utilizada en las redes sociales nos brinda la posibilidad de identificar palabras muy asentadas en determinada región del español, que en muchos casos no llega nunca a publicarse en las fuentes tradicionales, como la prensa o la literatura.

Esta dificultad proviene de varios factores. Por un lado, encontrar gran cantidad de autores nativos de diferentes lugares no es una tarea sencilla. Por otro lado, en la literatura se suele utilizar un vocabulario más restringido, normalmente excluyendo (o utilizando con menor frecuencia) términos del habla cotidiana. Un ejemplo de esto son los coloquialismos, cuyo uso es más frecuente en las redes sociales que en la literatura.

La gran importancia de conocer el uso de las palabras en ciertas regiones se puede ver reflejado en las marcas geográficas (o diatópicas) que se encuentran en algunas entradas de los diccionarios. Esta información cobra importancia para saber, por ejemplo, si una palabra tiene un uso general o se la utiliza solamente en algunas regiones. El área de la lingüística que estudia los principios teóricos en que se basa la composición de diccionarios se conoce como lexicología. Históricamente se han hecho diccionarios hispanoamericanos comparando con el español que los diccionarios españoles, generalmente el diccionario de la Real Academia Española (Diccionario de la Lengua Española), consideran general [Zim06]. Esto ocurrió en parte por el gran desarrollo de los diccionarios de la lengua española desde principios del siglo XVIII y por el carácter incipiente de la lexicografía americana. Sin embargo, esta metodología que compara dialectos de países latinoamericanos con España tuvo críticas en las últimas décadas, especialmente porque “una comparación adecuada es la que se puede establecer entre los elementos de entidades equivalentes, como las que forman los países.” [Ávi04, p. 9] Tanto Raúl Ávila como Klaus Zimmermann ponen en discusión el sentido de los diccionarios diferenciales de cada país, principalmente por no ser autosuficientes, ya que en un diccionario diferencial no se encuentran las palabras que se usan en ambas regiones, sino que aparecen únicamente los términos cuyo uso es mayor en la región a estudiar sobre la región con la que se compara. Ambos autores concluyen que es de mayor interés realizar diccionarios integrales, donde se marquen las palabras que se usan de forma contrastiva en una región, pero que también estén las palabras de uso general. Creemos que la metodología propuesta en esta tesis, facilitará el armado de estos diccionarios en particular y el estudio del léxico español hispanoamericano en general.

1.4. Objetivo del estudio

En este trabajo presentamos un método semi-supervisado para la detección de léxico contrastivo a través de un conjunto de textos recolectados de Twitter. Si bien se recolectaron textos de la Argentina, este trabajo puede ser replicado sobre otras regiones. Cabe mencionar que el método detecta palabras con valores significativos de contraste en su uso, y es necesaria la posterior supervisión de investigadores lexicógrafos entrenados para seleccionar los términos con interés lingüístico.

En la sección 2 explicamos la metodología para extraer los datos de Twitter y presentamos la caracterización de la muestra. Luego en la sección 3 se muestran las métricas creadas para medir la contrastividad de una palabra y el análisis de estas.

En la sección 4 mostramos las palabras identificadas como más contrastivas a partir de la métrica elegida y la proporción acumulada de sus ocurrencias en regiones de pocas provincias. También detallamos la validación lingüística realizada por la Academia Argentina de Letras, exhibimos una caracterización de las palabras salientes y hacemos una validación estadística de la métrica a través de tests estadísticos.

Finalmente en la sección 5 sacamos conclusiones a partir de los resultados obtenidos e indicamos trabajos posibles para seguir la investigación.

2. DATOS: EXTRACCIÓN Y PROCESAMIENTO

2.1. Extracción de Datos

Para la recolección de tuits, primero se extrajeron una cantidad de usuarios con información geográfica disponible. Los usuarios se buscaron por provincia de modo tal que haya una cantidad aproximadamente equitativa de cada una. La búsqueda de los usuarios se hizo de la siguiente manera:

Por cada provincia de la Argentina, se extrajeron las coordenadas de cada uno de sus departamentos, de los partidos de la provincia de Buenos Aires y de las comunas de la Ciudad Autónoma de Buenos Aires. El conjunto de estas forman la subdivisión de segundo orden de la República Argentina. La lista de departamentos, partidos y comunas fue extraída de los datos publicados del Censo Argentino del año 2010. Para extraer los tuits se utilizó la librería de *python* llamada *tweepy*.

De esta manera se recolectaron aproximadamente 900 usuarios por provincia, lo que resulta en 46.000 usuarios argentinos. Sobre este conjunto de usuarios se buscaron los tuits. Se decidió no tener en cuenta los retuits, dado que no son escritos por los usuarios sino que son una mera copia de otros tuits.

2.1.1. Búsquedas geolocalizadas

La búsqueda geolocalizada es una herramienta que nos da la posibilidad de obtener tuits generados en un área geográfica particular. Para esto, primero intenta buscar tuits cuyas coordenadas sean las buscadas. En caso de no tener éxito, se busca aquellos tuits creados por usuarios que tienen en el campo *location* de su perfil un lugar cuyo geocódigo coincida con el de sus coordenadas. Es decir, si se hace una búsqueda inversa de las coordenadas, devuelve el lugar de su perfil.

Una vez obtenida la lista de ubicaciones, se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas. Sobre el resultado de esta búsqueda, únicamente se seleccionaron los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia. En el gráfico de la Figura 2.1 se muestran las ubicaciones de los usuarios.

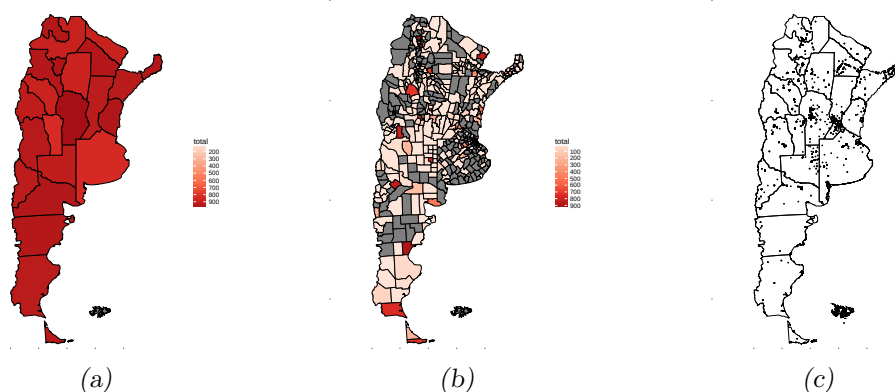


Fig. 2.1: Ubicaciones de los usuarios: (a) Mapa de Argentina con la distribución de los usuarios en las provincias. (b) Distribución de los usuarios en los departamentos de las provincias argentinas. Se muestra con áreas grises los departamentos que no poseen usuarios que hayan definido el campo ubicación de su perfil en ese lugar. (c) Distribución de las coordenadas obtenidas a partir de todos los usuarios. Las coordenadas fueron obtenidas a través de un proceso de geocodificación.

En la Figura 2.1(a) podemos observar que los usuarios que obtuvimos en las búsquedas se distribuyen de manera uniforme a través de las provincias. Si bien en la Figura 2.1(b) hay regiones grises que indican la ausencia de usuarios en ese lugar, cabe destacar que los mapas se realizaron obteniendo las coordenadas geográficas a partir de la ubicación definida en el perfil del usuario. Por lo tanto, si una persona declara que vive en *Tucumán, Argentina*, contabilizamos como que esa persona vive en la capital de esa provincia, lo cual puede no ser cierto. Sin embargo, esto no invalida los resultados puesto que la granularidad del análisis es a nivel provincial. Finalmente para ver la distribución de las coordenadas de los usuarios a lo largo del país mostramos la Figura 2.1(c). Se puede observar que en la mayoría de las grandes ciudades hay usuarios en nuestro conjunto de datos. Podemos observar en la Figura 6.1 del apéndice un mapa donde se contabilizaron todos los tuits con coordenadas geográficas del conjunto de datos. En este gráfico se puede observar que la distribución es mucho más amplia, aunque sigue habiendo mayor concentración de usuarios en aquellos departamentos con mayor densidad poblacional.

2.2. Tokenización y normalización

En cuanto al análisis del texto surge una primer problemática: ¿qué es una palabra? En principio podemos definir a una palabra como cualquier secuencia de caracteres delimitados por espacios blancos. Con esta definición *523456* y *?* serían palabras. Debido a esto podríamos restringir nuestra definición a una secuencia de caracteres alfabéticos. Los ejemplos mencionados anteriormente dejarían de estar dentro de la definición. Sin embargo términos como *asdsdafsdf* también serían palabras. Para evitar este problema podríamos tener un diccionario como filtro para saber si una secuencia de caracteres dada es una palabra. Si bien esto tendría mucha precisión al momento de filtrar los términos, no sería capaz de detectar palabras que existen en una lengua pero que no están incluidos en el diccionario elegido. Además, dada la cantidad de palabras recogidas, es altamente improbable que una secuencia al azar de caracteres alfabéticos reúna las condiciones de frecuencia necesarias para resultar destacada por la métrica que utilizamos. Es por eso que decidimos tomar a una palabra como una secuencia de caracteres alfabéticos.

Es muy posible que tengamos palabras que no sean interesantes a nivel lingüístico, como errores de tipeo (e.g. *computadira*, *escribur*), errores ortográficos o nombres propios. Es importante destacar que Twitter tiene caracteres especiales para mencionar a la gente, como el *@*, o el *#* (hashtag) utilizado para agrupar mensajes. Estos caracteres aparecen mucho, ya que los usuarios suelen responderse en la red, mencionando los mismos temas (aclarando el hashtag), o respondiendo a otros usuarios. Ya que esos caracteres no son alfabéticos, cualquier término que los utilice no va a ser parte del conjunto de palabras, como tampoco lo serán las direcciones de páginas web. Decidimos ignorar estos términos ya que no tienen interés lingüístico y agregarían mucho ruido a los datos.

Además de la tokenización, se realizó una normalización del texto. Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se hizo con la librería *TweetTokenizer* de *NLTK*. Como ya dijimos, se descartó la idea de filtrar las palabras que no estuvieran en un diccionario ya que si bien hubiera eliminado mucho ruido, también nos hubiera filtrado palabras de interés. Este es el caso de los neologismos, o las palabras que, si bien se utilizan hace mucho tiempo, no están en los diccionarios

actuales.

2.3. Caracterización de la muestra

Para tener una noción más completa de la muestra, se presenta la Tabla 2.1, que indica las cantidades de palabras y tuits por provincia. Puede notarse que la muestra está balanceada con respecto a la cantidad de usuarios, tuits por provincia y palabras. Los tuits de la Ciudad Autónoma de Buenos Aires están incluidos en los de la provincia de Buenos Aires. Tomamos esta decisión ya que debido al constante movimiento entre las personas de la ciudad y la provincia, es difícil distinguir de forma fehaciente a los usuarios de ambos territorios.

Provincia	#Palabras Distintas	#Usuarios	#Tuits	#Total Palabras
Buenos Aires	191919	920	1125042	8974372
Catamarca	173104	957	1057019	8161309
Chaco	169476	964	976943	7605991
Chubut	182592	954	1023373	8884745
Córdoba	207307	987	1224266	10075932
Corrientes	183292	939	1044951	8426940
Entre Ríos	188679	969	1193693	9462986
Formosa	169254	903	923352	7184382
Jujuy	171064	971	678004	5951778
La Pampa	186593	935	1085757	8996318
La Rioja	186041	946	704044	6757277
Mendoza	193708	945	1099717	9402399
Misiones	168400	972	984218	7790197
Neuquén	188038	927	1111201	9021449
Río Negro	194383	965	1215361	9991831
Salta	188402	884	830916	7506652
San Juan	183546	926	1002322	8377792
San Luis	164185	896	1006464	8327093
Santa Cruz	174089	935	876621	7432923
Santa Fe	201879	937	1019620	8862328
Santiago del Estero	166540	887	944109	7355729
Tierra del Fuego	197273	964	976426	8559218
Tucumán	195643	962	1093874	9238526

Tab. 2.1: Cantidades del conjunto de datos

También analizamos la cantidad de palabras por tuit promediadas sobre cada usuario. Debido a que los tuits están limitados a 140 caracteres, era de esperar que no hubiera demasiadas palabras promedio por cada tuit. En la Figura 2.2(a) podemos observar que la media para la cantidad de palabras promedio en un tuit está entre 7 y 8.

2.3.1. Distribución temporal de tuits

Los tuits recolectados tienen una particularidad: la cantidad de tuits recolectados crece año a año. Esto se refleja en los gráficos de la Figura 2.3, que muestran la progresión

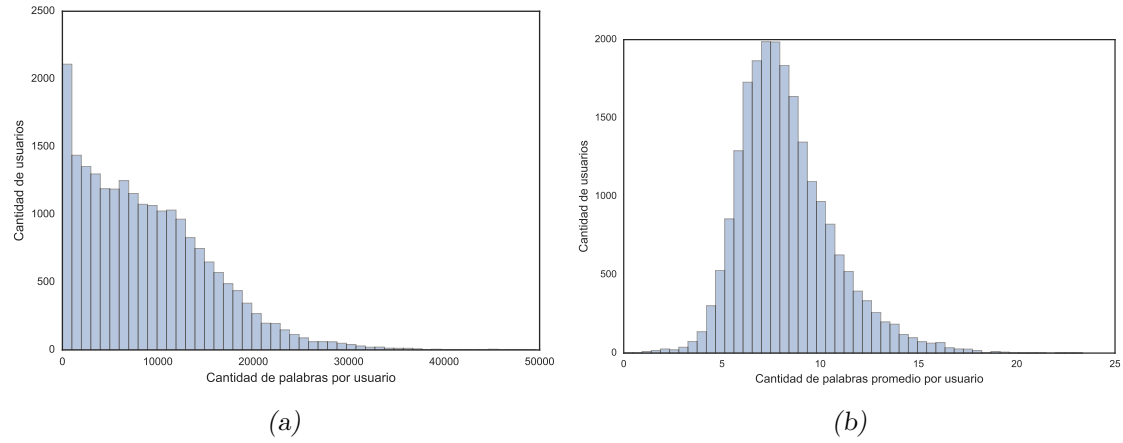


Fig. 2.2: Histogramas: (a) Cantidad de palabras totales por usuario. (b) Cantidad de palabras en promedio por tuit, para todos los usuarios.

temporal de la cantidad de tuits en cuatro regiones representativas: La Pampa, Buenos Aires, Chaco y Neuquén.

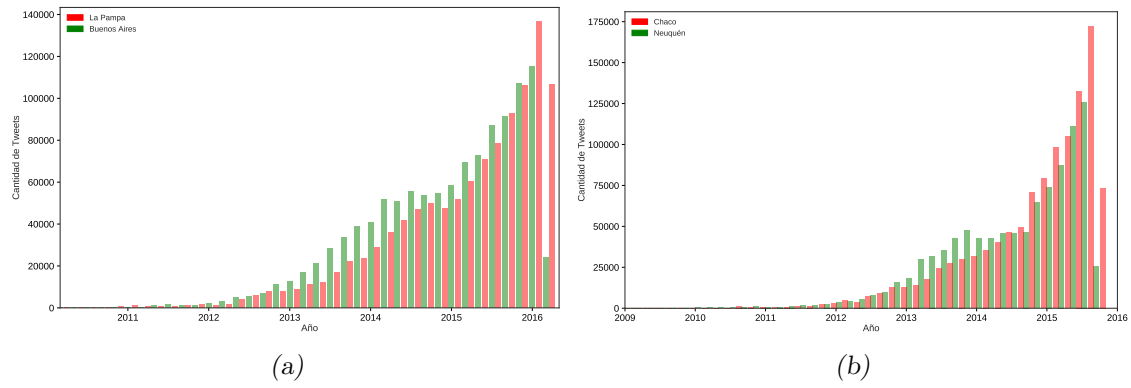


Fig. 2.3: (a) Histograma de la cantidad de tuits que se hicieron por bimestre en las provincias La Pampa y Buenos Aires. (b) Gráfico para Chaco y Neuquén.

2.3.2. Información de palabra por provincia

Una vez recolectados los tuits y luego de procesar los datos, obtuvimos la cantidad de ocurrencias de cada palabra en cada provincia argentina. Esto se refleja en la Tabla 2.2 donde por cada palabra (fila), aparece el número de ocurrencias en cada provincia (columna).

Palabra	Prov.1	Prov.2	...	Prov.23	Totales
w_1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,23}$	$\mathbf{N}(w_1)$
w_2	$N_{2,1}$	$N_{2,2}$	...	$N_{2,23}$	$\mathbf{N}(w_2)$
.	.	.	...	.	.
w_i	$N_{i,1}$	$N_{i,2}$	...	$N_{i,23}$	$\mathbf{N}(w_i)$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.

Tab. 2.2: Cantidad de ocurrencias de cada palabra por provincia.

donde $N_{i,j}$ representa la cantidad de veces que la palabra w_i aparece en la provincia Prov. j , $j = 1, \dots, 23$.

$$\mathbf{N}(w_i) = \sum_{j=1}^{23} N_{i,j} \quad (2.1)$$

3. MÉTRICAS PARA DETECTAR PALABRAS SALIENTES

3.1. Búsqueda de contrastes

Una palabra tiene un contraste cuando tiene un uso con diferencias significativas en distintas regiones. En este trabajo nos propusimos crear un listado con palabras con contrastes que tengan importancia a nivel lingüístico. En este sentido, los nombres de personas, lugares u organizaciones no fueron considerados de interés a pesar de tener contrastes en su uso. Este listado fue ordenado por una métrica que capte en un único valor el nivel contrastivo. De esta manera, se seleccionó un subconjunto de palabras, de acuerdo a la métrica, el cual fue analizado manualmente por los investigadores de la Academia Argentina de Letras (AAL).

El primer acercamiento para ver el contraste de las palabras surgió de comparar las frecuencias de las palabras en cada par de provincias de la Argentina. Para esto calculamos, por cada palabra w , la frecuencia de ocurrencias sobre cada una de las dos provincias p_1 y p_2 . A la mayor frecuencia de ambas, la llamamos frecuencia máxima y a la menor, la frecuencia mínima. Luego el cociente entre la *frecuencia máxima* y la *frecuencia mínima* tiene como resultado lo que llamamos *maxDif*. En caso de que en una de las dos provincias no se hayan recolectado tuits con esa palabra, se toma como frecuencia mínima a la frecuencia mínima distinta de 0 de todas las palabras generadas en esa provincia. Así se evitó la división por cero. Esta métrica se resume en la ecuación 3.1, siendo $frec(w, p)$ igual a la frecuencia de la palabra w en la provincia p .

$$maxDif(w, p_1, p_2) = \frac{f_{max}(w, p_1, p_2)}{f_{min}(w, p_1, p_2)} \quad (3.1)$$

donde

$$f_{max}(w, p_1, p_2) = \max(frec(w, p_1), frec(w, p_2)) \quad (3.2)$$

$$f_{min}(w, p_1, p_2) = \begin{cases} \min(frec(w, p_1), frec(w, p_2)), & \text{si } frec(w, p_1) * frec(w, p_2) > 0 \\ \min_{w \in W(P_{min})} frec(w, P_{min}), & \text{si no.} \end{cases} \quad (3.3)$$

donde P_{min} es la provincia que tiene la menor frecuencia de ambas y $W(P_{min})$ son las palabras mencionadas en esa provincia.

De esa manera se ordenó el listado de cada par de provincias teniendo en cuenta la división de frecuencias. Sin embargo, este método imposibilitaba el trabajo manual para los investigadores de la AAL, que debían mirar estos listados y hacer un análisis más exhaustivo sobre las palabras con mayor diferencia de frecuencias, debido a que había $\binom{23}{2} = 253$ listados (o equivalentemente 253 columnas en un mismo listado) a analizar. Además la métrica solo permitía saber si había un contraste entre dos provincias, pero no se podía tener en cuenta la frecuencia de la palabra en el resto de las provincias. En consecuencia las palabras se encontraban repetidas en los distintos listados y con diferentes valores de *maxDif*, lo cual hacía muy difícil poder identificar en qué regiones había una diferencia significativa de frecuencias. Debido a esto decidimos realizar un nuevo enfoque para encontrar las palabras con alta contrastividad en las distintas regiones, de manera que una métrica pudiera reflejar el nivel de contrastividad de la palabra en un único valor.

En primer lugar intentamos modificar la métrica *maxDif* de la siguiente manera. Sea $f'_{max}(w)$ la frecuencia máxima de la palabra w entre las frecuencias de todas las provincias

y sea $f'_{min}(w)$ la frecuencia mínima distinta de 0 sobre todas las provincias, luego la métrica generalizada $maxDif_g(w)$ se puede definir como en la ecuación 3.4.

$$maxDif_g(w) = \frac{f'_{max}(w)}{f'_{min}(w)} \quad (3.4)$$

Si bien esta métrica logra resumir en un único valor la diferencia de frecuencias, sigue sin considerar la dispersión de las frecuencias en todas las provincias. Por este motivo, nos enfocamos en analizar el contraste de frecuencias de palabras sobre las provincias a través de una métrica superadora.

3.1.1. Métricas para medir el contraste en la frecuencia de las palabras

Dado que se quieren encontrar las palabras con contrastes significativos en distintas regiones se propone generar una métrica basada en la cantidad de información para poder realizar esta tarea.

Una medida que se puede usar para comparar las frecuencias de las palabras en las diferentes regiones del país es la entropía definida por Shannon (ver el Apéndice 6.1), debido a que nos brinda un valor que indica qué tan uniforme es la distribución de las frecuencias de cada palabra. La entropía es máxima cuando los eventos son equiprobables y mínima en el caso que la probabilidad de un evento es 1.

Sin embargo, la entropía como única medida tiene sus desventajas. En particular, una palabra con una sola ocurrencia en una provincia y ninguna en las demás, tiene la entropía mínima, lo cual la pondría en igualdad de condiciones con una palabra con 1000 ocurrencias en una única provincia, algo no deseable.

A pesar de que nos interesan las palabras con un contraste significativo entre regiones, dentro de ellas elegiremos las que tienen mayor cantidad de ocurrencias. Es por esto que elaboramos otra métrica que tenga en cuenta la entropía, entre otras variables.

3.1.2. Valor de información

La métrica que utilizamos para ordenar los listados de palabras y detectar cuáles son las que tienen altos contrastes en su uso en distintas regiones fue inspirada por el trabajo de Zanette y Montemurro [MZ10]. Ellos, a diferencia de Shannon, estudiaron una relación entre una medida de la información y su función semántica en el lenguaje. A continuación detallamos el procedimiento para calcular lo que los autores llamaron el *valor de la información*:

Dado un texto dividido en V partes iguales llamadas *ventanas*, se calcula la entropía $H(w)$ sobre el vector de probabilidad de ocurrencia de w en cada una de las V ventanas.

Es decir, $H(w) = - \sum_{i=1}^V \frac{w_i}{N_w} * \log(\frac{w_i}{N_w})$ con N_w igual a la cantidad total de ocurrencias de la palabra w en toda el corpus y w_i la cantidad que se obtuvo en la ventana i en el conjunto de datos recolectado. Notar que $\frac{w_i}{N_w}$ es una estimación de la probabilidad de ocurrencia de la palabra w en la ventana i . Por otro lado, $\hat{H}(w)$ es un promedio de $H(w)$ sobre todas las posibles permutaciones de las palabras del texto. Formalmente, sea $\tilde{W} = (w_1, \dots, w_V)$ un vector aleatorio con w_i igual a la cantidad de ocurrencias de la palabra w en la ventana i . Luego $\hat{H}(w) = \mathbb{E}[H(\tilde{W})]$. Si bien Zanette y Montemurro dividen un texto en V ventanas que pueden variar dependiendo del tamaño de la ventana fijado, en este trabajo se tomo como ventana todos los textos provenientes de una provincia.

Es de esperar que en la mayoría de los casos la entropía sobre todas las permutaciones del texto sea mayor que la medida en el cálculo original. Esto se debe a que en una gran cantidad de las permutaciones las palabras se distribuyen de forma más uniforme en las distintas partes.

Zanette y Montemurro definieron al *valor de la información de una palabra* como

$$\Delta I(w) = p(w) (\hat{H}(w) - H(w)) = p(w) \Delta H(w) \quad (3.5)$$

siendo $p(w)$ la frecuencia total de la palabra en el texto. De esta manera se les da más importancia a las palabras que son más frecuentes y a las palabras que tienen una baja entropía, ya que en estas el término de la diferencia es más grande.

Los autores analizaron el valor de la información de las palabras sobre tres textos, *Análisis de la mente* de Bertrand Russell, *Moby Dick* de Herman Melville y *El origen de las especies* de Charles Darwin. En los tres libros las palabras con mayor valor de la información están altamente relacionadas con los temas principales.

Zanette y Montemurro utilizaron el *valor de la información de una palabra* de manera tal que sumando el valor para todas las palabras $W = \{w_1, \dots, w_K\}$ se obtiene el *valor de información de la distribución de las palabras* $\Delta I(s)$ definido en la ecuación 3.6, siendo $s = N/V$, N la cantidad de palabras en el texto, V la cantidad de ventanas en la que se dividió y $p(w)$ la frecuencia total de la palabra en el texto.

$$\Delta I(s) = \sum_{w \in W} p(w) (\hat{H}(w) - H(w)) = \sum_{w \in W} p(w) \Delta H(w) \quad (3.6)$$

Zanette y Montemurro definieron esta medida para “cuantificar la relación entre las heterogeneidades de la distribución de palabras debido a su función lingüística y la partición de texto” [MZ10, p. 136]. De esta manera, para cada texto buscaban cual era el valor de s tal que la información $\Delta I(s)$ sea máxima.

Si bien el *valor de la información de una palabra* definida por Zanette y Montemurro es muy bueno a la hora de calcular el tamaño óptimo de ventana que maximice el valor de la información de un texto $\Delta I(s)$, también refleja un valor que no es conveniente al hacer la comparación entre dos palabras cuando una de ellas es demasiado frecuente (como la palabra *que*). Esto sucede ya que las palabras, al tener valores de frecuencias muy distintos (ver más detalle de esto en la sección 3.1.4), generan un valor de ΔI muy alto en las palabras más frecuentes en comparación con el resto. Este fenómeno se pone en evidencia en la Tabla 3.1 donde se observa que palabras como *que* y *me* tienen los valores máximos de la métrica ΔI .

Es por esto que, entre otras cosas, usamos al logaritmo sobre las cantidades de ocurrencias para generar una mejor dispersión de los datos. Así agrupamos a las palabras con un valor de $\Delta H(w)$ parecido y le restamos relevancia a las palabras sumamente frecuentes. Otra dificultad que surge de la métrica propuesta por Zanette y Montemurro es la imposibilidad de realizar la media de todas las posibles permutaciones del texto por la limitación computacional ya que tenemos una cantidad muy grande de datos. Es por eso que diseñamos una métrica similar basada en una aproximación de la esperanza de la entropía de una palabra.

Sea $\#w$ la cantidad de ocurrencias de la palabra w en toda la Argentina y W todas las palabras posibles en todo el país, definimos

$$m = \min_{w \in W} \#w \quad (3.7) \quad M = \max_{w \in W} \#w \quad (3.8)$$

Palabra	ΔI
que	109932.23
me	101754.11
rioja	59244.90
a	54872.75
re	52207.93
no	51470.18
la	47693.43
de	44035.71
ushuaia	42972.62
jujuy	39087.57
salta	33310.16
el	32628.74
q	32596.95
y	29905.83
comodoro	25902.79

Tab. 3.1: 3.1 Las 15 palabras con mayor valor de I_w

Luego, realizamos una normalización lineal de $\log(\#w)$

$$norm_w(w) = \frac{\log(\#w) - \log(m)}{\log(M) - \log(m)} \quad (3.9)$$

Así $norm_w$ tiene su imagen en el rango $[0, 1]$, tomando el valor 1 sobre la palabra que tiene la cantidad de ocurrencias máxima y el valor 0 cuando se aplica a la palabra con menor cantidad de ocurrencias. Vale la pena aclarar que tomamos 40 como umbral mínimo de la cantidad de ocurrencias de las palabras para ser estudiadas, por lo tanto $m_w = 40$. A partir de la función $norm_w$ definimos el *valor contrastivo sobre las palabras* I_w como:

$$I_w(w) = norm_w(w) \cdot (\widehat{H}_w(w) - H_w(w)) \quad (3.10)$$

siendo $H_w(w)$ la función de entropía calculada sobre las cantidades de ocurrencias de la palabra w sobre las 23 provincias argentinas. Como $\widehat{H}_w(w)$ es un promedio de $H(w)$ sobre todas las posibles permutaciones del texto, al ser este de gran tamaño resulta muy difícil de computar. Es por esto que simulamos la cantidad de ocurrencias de la palabra w en todas las provincias con una distribución multinomial $M(N_w, \frac{1}{23}, \dots, \frac{1}{23})$ y realizamos un promedio de la entropía sobre las ocurrencias simuladas para aproximar $\widehat{H}_w(w)$. Elegimos esta distribución con igual probabilidad de ocurrencia en todas las provincias, para simular el hecho de que todas las palabras ocurren de manera uniforme a lo largo del país.

3.1.3. Robusteciendo la métrica desarrollada

Ahora bien, una determinada provincia o región puede tener muchas ocurrencias de una palabra formuladas por algunos pocos usuarios que utilizan constantemente el término. Un ejemplo de esto podrían ser bots que escriben automáticamente textos iguales (o similares) en grandes cantidades. Otra posible causa de este fenómeno podría ser la de usuarios que hablan de personas, lugares o marcas de forma constante. Es por esto que realizamos

una métrica similar que tenga en cuenta la diferencia de la entropía sobre la cantidad de personas que mencionaron la palabra w , que definimos como $\#^u w$. Así como normalizamos la diferencia de entropías por la cantidad de ocurrencias, también desarrollamos la variable normalizadora $norm_p$ de la cantidad de usuarios que mencionan la palabra w . Para eso, sea

$$u = \min_{w \in W} \#^u w \quad (3.11) \quad U = \max_{w \in W} \#^u w \quad (3.12)$$

Luego, definimos

$$norm_p(w) = \frac{\log(\#^u w) - \log(u)}{\log(U) - \log(u)} \quad (3.13)$$

Con esta variable normalizadora definimos el *valor contrastivo sobre las personas* I_p ,

$$I_p(w) = norm_p(w) \cdot (\hat{H}_p(w) - H_p(w)) \quad (3.14)$$

donde H_p es la entropía calculada sobre el vector de probabilidad de la cantidad de usuarios que mencionan una palabra. Es decir, si p_{ui} es la probabilidad de que un usuario que menciona una palabra sea de la provincia i , $H_p(w) = -\sum_{i=1}^{23} p_{ui} \cdot \log(p_{ui})$. A su vez $\hat{H}_p(w)$ es el promedio de H_p sobre todas las posibles permutaciones de la cantidad de usuarios que mencionan a w . En otras palabras, sabiendo la cantidad total de usuarios que mencionan a w en toda la Argentina y calculando el promedio de H_p sobre todas las permutaciones posibles de esa cantidad en las 23 provincias argentinas obtenemos $\hat{H}_p(w)$.

Debido a que queremos tener en cuenta tanto a la variación de la cantidad de ocurrencias de la palabra como a la variación de la cantidad de usuarios, elegimos como métrica la multiplicación de ambas métricas definidas, es decir

$$I(w) = I_w(w) \cdot I_p(w) \quad (3.15)$$

A esta métrica la denominaremos el *valor de contrastividad*. Es importante aclarar que tanto $norm_w$ como $norm_p$ realizan una normalización del logaritmo de esas variables. Esto se debe a que el logaritmo genera una dispersión tal que agrupa los valores altos que se encontraban muy dispersos mientras que separa los valores pequeños que estaban concentrados. Esto se puede ver comparando los histogramas de la Figura 3.2 con los de la Figura 3.1.

Para eliminar los valores atípicos se procedió a remover tanto las palabras que no superaran las 40 ocurrencias, como también aquellas que eran dichas por menos de 6 usuarios. La métrica se evaluó en este conjunto filtrado de palabras. En la Tabla 3.2 se muestran las 20 palabras más contrastivas de acuerdo a nuestra métrica.

3.1.4. Frecuencia de las palabras

En la Figura 3.1(a) graficamos la distribución de la cantidad de ocurrencias de las palabras. Podemos observar que la mayoría de las palabras ocurren poco. En particular el 50 % de las palabras tienen una frecuencia relativa menor a $7,22e - 07$. Por otro lado hay pocas palabras que ocurren mucho, por ejemplo la palabra *que* o la preposición *de*. En la Tabla 6.1 del apéndice se encuentran las 20 palabras más frecuentes.

Si comparamos la posición de la palabra en un listado ordenado podemos ver que las cantidades de ocurrencias parecieran seguir una distribución zipfiana. La ley de Zipf es

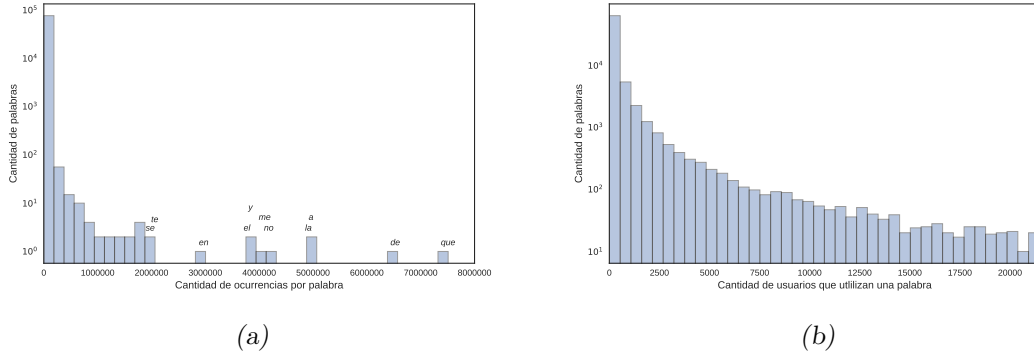


Fig. 3.1: (a) Histograma de la cantidad de ocurrencias de las palabras. (b) Histograma de la cantidad de usuarios que utilizan una determinada cantidad de palabras.

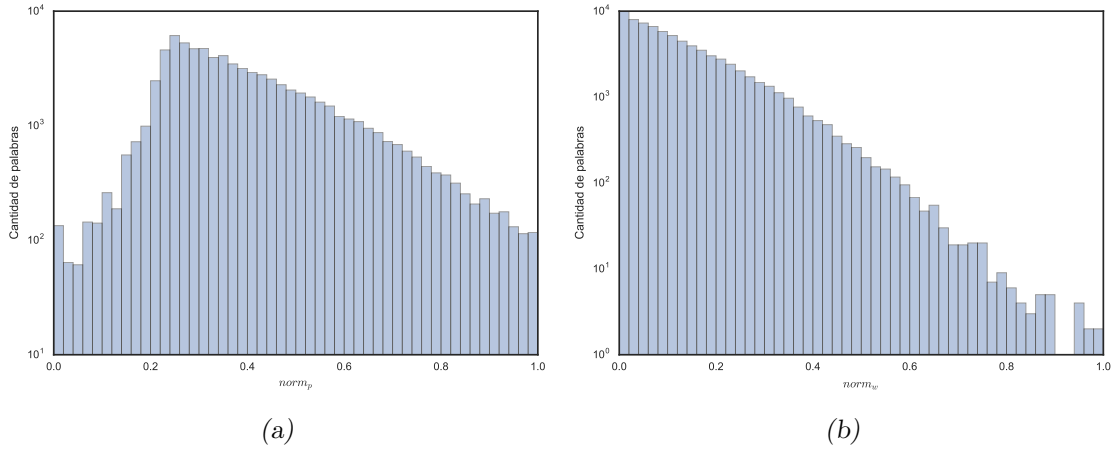


Fig. 3.2: (a) Histograma de la normalización sobre la cantidad de usuarios que utilizan una determinada cantidad de palabras definida como $norm_p$. (b) Histograma de la normalización de la cantidad de ocurrencias de las palabras definida como $norm_w$.

una ley empírica formulada por George Zipf en el año 1932 en la cual se establece una relación entre la frecuencia de una palabra con su posición dentro del listado de palabras ordenadas por frecuencia decreciente [Mon01, Zip16]. En particular, sea n la posición de la palabra en el listado ordenado y sea $f(n)$ la cantidad de ocurrencias de la n -ésima palabra, se puede hacer la siguiente aproximación:

$$f(n) \approx \frac{A}{n^\alpha}$$

donde α toma un valor levemente mayor a 1 y A es una constante normalizadora. Entonces, bajo la ley de Zipf uno puede saber que la frecuencia de la segunda palabra más dicha en un corpus es aproximadamente la mitad que la primera. La palabra con posición 3 en el listado ordenado por frecuencias, va a tener aproximadamente la tercera parte de la cantidad de ocurrencias que la primera y así sucesivamente. De esta manera hay una relación lineal entre el logaritmo de la posición del listado ordenado por frecuencias y el logaritmo de la cantidad de ocurrencias de cada palabra. Realizando una regresión lineal pesada según las cantidades de ocurrencias de las palabras obtuvimos un valor de $\alpha = 1,089$.

Palabra	$I(w)$
chivilcoy	1.533974
oberá	1.491781
ushuaia	1.461703
ush	1.434766
obera	1.244550
breñas	1.227546
viedma	1.213424
bragado	1.202022
logroño	1.188416
nqn	1.143036
tdf	1.126729
riojanos	1.121459
charata	1.071503
chivil	1.043653
cldo	1.036830
blv	1.005322
rioja	1.002387
choele	0.999114
tolhuin	0.985641
rada	0.965216

Tab. 3.2: Las 20 palabras más contrastivas de acuerdo a nuestra métrica $I(w)$.

Otra forma de utilizar esta ley empírica es la siguiente: sabiendo la posición de una palabra w en el listado ordenado por frecuencias de un corpus **A** y sabiendo la cantidad de palabras totales de un corpus **B**, puede estimarse la cantidad de ocurrencias de w en el corpus **B**.

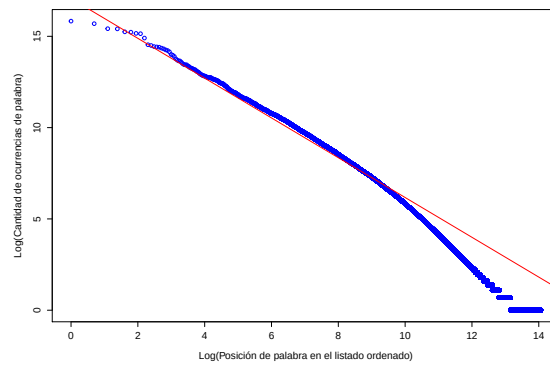


Fig. 3.3: Cantidad de ocurrencias de palabra vs posición en listado ordenado. Se aplicó el logaritmo natural a las cantidades de ocurrencias, como también a los valores de las posiciones para mostrar la proporcionalidad entre $f(n)$ y $\frac{1}{n^\alpha}$, obteniendo un valor de $\alpha = 1,089$.

3.2. Distribución de la entropía

Teniendo el listado de palabras, hicimos un cálculo de entropía tomando en cada provincia la cantidad de ocurrencias de cada palabra. En la Figura 3.4 podemos observar la distribución del valor de la entropía sobre todas las cantidades de ocurrencias de las palabras con más de 40 apariciones y dichas por más de 5 usuarios.

Podemos ver que la mayor parte de las palabras tienen un valor de entropía entre 2.5 y 3. Esto quiere decir que hay un gran conjunto de palabras que tiene una cantidad de ocurrencias relativamente uniforme a lo largo de todas las provincias.

Sin embargo, hay otro conjunto de palabras que tienen una entropía menor a 2, la cual podemos considerar como baja. Estas últimas palabras serán las que tienen mayor interés debido a que tienen una variación marcada en cuanto a su utilización en las distintas regiones. El máximo valor alcanzado de la entropía es de 3,1350 con la palabra *el*. Como aclaración la entropía calculada se realizó con logaritmos naturales, por lo tanto el máximo valor posible es de $\ln(23) = 3,1355$ donde habría una distribución uniforme en la cantidad de ocurrencias sobre las 23 provincias argentinas.

Tener en cuenta únicamente a la entropía de las palabras nos puede generar la detección de palabras que no son de interés, ya sea porque no ocurren una cantidad significativa de veces o porque la variación de las ocurrencias en las distintas provincias se debe solamente a pocos usuarios que la utilizan mucho. Es por esto que también se calculó la entropía teniendo como variable la cantidad de personas que usaron cierto término en una determinada provincia.

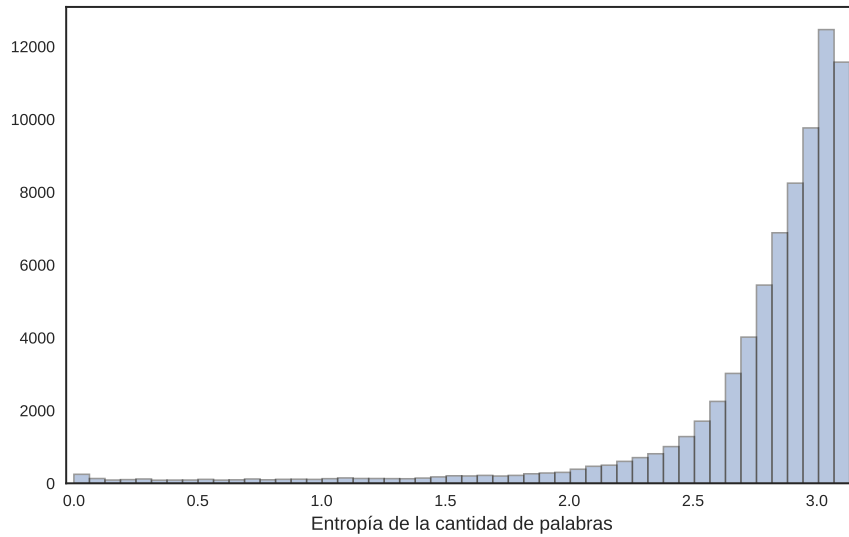


Fig. 3.4: Histograma del valor de la entropía de las palabras (H_w).

3.3. Distribución del valor de contrastividad

En la Figura 3.5(a) se muestra una clara relación entre la cantidad de ocurrencias que tiene una palabra y su valor de contrastividad definido en la ecuación 3.15, indicado por el

color: cuanto más oscuro, más alto el valor. A su vez, se nota que el *valor de contrastividad* suele ser mayor a medida que el valor de la entropía es menor. Esto no sucede siempre debido a que hay palabras que tienen una entropía de palabras H_w baja, pero sin embargo la entropía de personas H_p es alta, lo cual hace que el valor de contrastividad sea bajo.

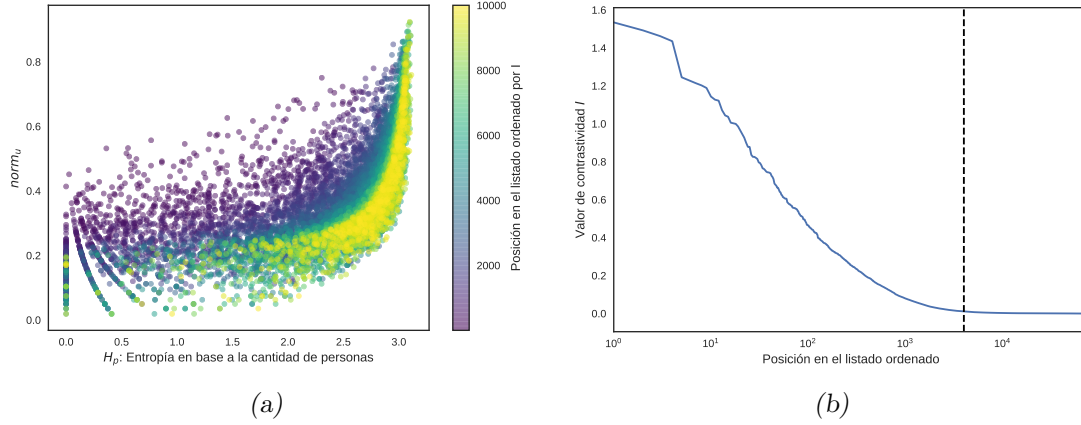


Fig. 3.5: (a) Gráfico de dispersión que muestra la posición en el listado ordenado según el valor de contrastividad reflejado por la escala cromática. A su vez se muestra para cada palabra el valor de la entropía de las personas (H_p) y la cantidad normalizada de personas que utiliza dicho término ($norm_p$). (b) Valor de contrastividad según la posición de la palabra en el listado de palabras. El gráfico se realizó sobre el conjunto de palabras cuya cantidad de ocurrencias era mayor a 40 y la cantidad de usuarios que utilizaron cada término era mayor a 5. Desde la palabra cuya posición es 4000, el valor I se estabiliza alrededor de 0.

En la Figura 3.5(b) se puede ver el valor de contrastividad según la posición en la que se encuentra en el listado ordenado por la métrica. Notamos que el valor se estabiliza aproximadamente a partir de la palabra en la posición número 4000, a partir de donde tiende a acercarse a 0.

3.4. Proporción acumulada de ocurrencias

Además de la detección de las palabras contrastivas en su uso, nos interesa saber en qué regiones se utilizan más. Para esto ordenamos, por cada palabra, las provincias de acuerdo a la cantidad de menciones, formando una lista de provincias: $ps = p_1, p_2, \dots, p_{23}$, donde $\#w(p_1) \geq \#w(p_2) \dots \#w(p_{23})$. Luego elegimos al conjunto de provincias que superen un cierto porcentaje de todas las ocurrencias. Esto lo realizamos en distintas muestras de palabras, las 1000-2000-5000-10000 más contrastivas según nuestra métrica y sobre el conjunto total de las palabras (sin incluir a las palabras que son dichas por menos de 5 usuarios o con menos de 40 ocurrencias en todo el conjunto de datos). Reflejamos este análisis en la Figura 3.6 donde cada curva representa la proporción acumulada media según la cantidad de provincias, eligiendo por cada palabra y una cantidad de provincias determinada aquel conjunto de provincias que maximice la proporción. Es notable la diferencia de proporciones acumuladas según la muestra de palabras. Solamente con una provincia para cada palabra ya se puede cubrir, en promedio, el 76 % del total de ocurrencias sobre las mil palabras con mayor valor de nuestra métrica.

En el gráfico de la Figura 3.7 se observa que la variación del cubrimiento de ocurrencias

disminuye a medida que se aumenta la cantidad de provincias tomando la muestra con las primeras 5000 palabras con mayor valor de contrastividad.

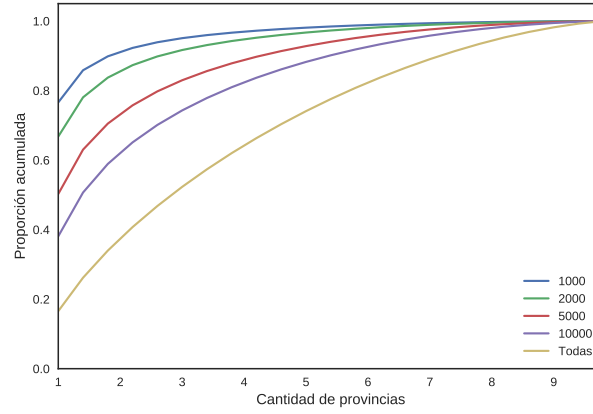


Fig. 3.6: Proporción de ocurrencias acumulada media según la muestra de palabras. El número de la leyenda indica la cantidad de palabras contrastivas elegidas para la muestra respectiva, siempre seleccionando las más contrastivas según la métrica. En cada curva se refleja el promedio de la proporción acumulada de una muestra variando la cantidad de provincias que forman las regiones maximales.

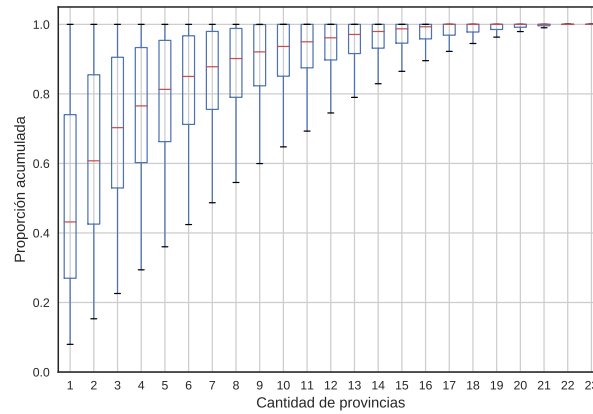


Fig. 3.7: Variación de la proporción de ocurrencias acumulada a partir de la muestra con las primeras 5000 palabras con mayor valor de contrastividad.

4. ANÁLISIS DE LAS PALABRAS CONTRASTIVAS ENCONTRADAS

4.1. Palabras candidatas

Para buscar las palabras candidatas a tener contrastes significativos en cuanto a la cantidad de ocurrencias en distintas provincias, elegimos el conjunto de las primeras cinco mil (5000) palabras con mayor valor de nuestra métrica. El número 5000 surgió de ver la distribución de los valores de la información graficado en la Figura 3.5(b), donde hay una caída pronunciada de la métrica y a partir de la palabra cuya posición es 4000 se observa que empieza a estabilizarse con valores muy cercanos a 0. Es por esto que nos pareció razonable dar un margen de 5000 palabras para evaluar manualmente y, entre estas, seleccionar las palabras con contrastes significativos que tienen interés a nivel lingüístico.

Como era de esperar, los topónimos, como los nombres de ciudades y provincias, son palabras que ocurren mayormente en sus respectivas regiones. Esto causa que haya una gran variación en la cantidad de ocurrencias sobre las distintas provincias, lo que genera un valor alto en la métrica de valor de la información. Para facilitar la detección de palabras contrastivas con mayor interés lingüístico buscamos un conjunto de datos con los nombres de las localidades y departamentos de la República Argentina de modo tal que podamos resaltarlas para que el equipo de filólogos tenga una primera alerta sobre posible toponimia.

4.2. Regiones de palabras

En la sección 3.4 analizamos la variación de la proporción acumulada de las palabras en las regiones formadas por un conjunto de provincias. Estas regiones tienen la particularidad que son establecidas por cada palabra y el conjunto de provincias que la forman son aquellas que tienen más ocurrencias en el conjunto de datos. Establecimos un umbral del 80% de ocurrencias para definir a una región, es decir, estas eran formadas por el conjunto de provincias que contenían al 80% de las ocurrencias de la palabra. Una vez definidas las regiones para cada palabra, nos propusimos analizar cuales son las más frecuentes. Para eso generamos una lista de regiones de las 5000 palabras con mayores contrastes de acuerdo al valor de contrastividad y la ordenamos según su frecuencia. De estas regiones quitamos aquellas que eran formadas por más de 7 provincias o por una sola. En la Tabla 4.1 se muestran las 30 regiones más frecuentes. Analizando esta tabla pudimos notar que la mayoría de las regiones están compuestas por provincias contiguas, es decir que para cada provincia dentro de una región hay otra provincia limítrofe. Mostramos algunos de los conjuntos de palabras de cada región en la Tabla 6.2 (ver apéndice).

4.3. Problemas en el conjunto de datos

Cuando vimos las palabras con mayor valor de la información, observamos que algunas palabras de la provincia de La Rioja eran provenientes de España. Analizando la causa de este problema, notamos que la API de Twitter no realiza las búsquedas localizadas como uno esperaría. En particular, no solo se fija en los tuits geolocalizados, sino que también hace una búsqueda inversa a través de los nombres de las ciudades que tienen esa coordenada. Específicamente, La Rioja es una provincia Argentina, como así también una provincia de España. Es por eso que al hacer búsquedas con las coordenadas de ciudades de La Rioja en Argentina, tuvimos resultados de tuits de España. Lo mismo sucedió con San Juan (capital de Puerto Rico), Santiago Del Estero (Santiago de Chile) y Córdoba (ciudad de Andalucía, España). A pesar de que los tuits no fueron escritos en Argentina,

Conjunto de provincias	Cantidad de Palabras
Jujuy - Salta	24
Mendoza - San Juan	19
Neuquén - Río Negro	18
Corrientes - Misiones	16
Chaco - Corrientes - Formosa	16
Chaco - Corrientes	16
Chubut - Santa Cruz	13
Catamarca - La Rioja	12
Santa Cruz - Tierra del Fuego	12
Corrientes - Entre Ríos - Formosa - La Rioja - Misiones	12
Formosa - Misiones	12
Corrientes - Formosa - Misiones	12
Córdoba - La Rioja	11
Catamarca - Salta - Santiago del Estero - Tucumán	11
Catamarca - Jujuy - La Rioja - Salta - Santiago del Estero - Tucumán	10
Chaco - Corrientes - Misiones	10
Chaco - Corrientes - Formosa - Misiones	9
Catamarca - Santiago del Estero - Tucumán	9
Catamarca - Tucumán	9
Salta - Tucumán	8
Catamarca - Jujuy - Salta - Santiago del Estero - Tucumán	8
Neuquén - San Juan	7
Chubut - Santa Cruz - Tierra del Fuego	7
Buenos Aires - La Pampa	7
Salta - Santiago del Estero - Tucumán	7
Buenos Aires - La Pampa - Río Negro	6
Corrientes - Formosa	6
Catamarca - Jujuy - Salta - Tucumán	6
Chaco - Corrientes - Entre Ríos - Formosa - Misiones	6
Catamarca - Santiago del Estero	6

Tab. 4.1: Indica cuantas palabras tienen un cubrimiento del 80 % de sus ocurrencias en cada conjunto de provincias a partir de las 5000 palabras con mayores contrastes (de acuerdo al valor de la información).

consideramos que su cantidad no es lo suficientemente grande como para afectar de manera significativa nuestros resultados.

4.4. Caracterización de las palabras identificadas como contrastivas

Dentro de las palabras contrastivas identificadas a través de la métrica, podemos hacer una caracterización de ellas según el fenómeno lingüístico que representan.

En base al listado de palabras identificadas como contrastivas a partir de la métrica, lexicógrafos de la Academia Argentina de Letras hicieron la validación lingüística. Consistió en un estudio pormenorizado, palabra por palabra, para determinar si la palabra en cuestión forma parte del repertorio léxico de una comunidad de hablantes. Esto excluyó, como es tradicional en lexicografía, nombres propios y topónimos locales, que la métrica sube a los puestos altos de las listas porque efectivamente su uso es abundante y contrastivo.

En la lista 4.4 presentamos una caracterización de las palabras. La lista apenas incluye algunos ejemplos en cada categoría. Sin embargo sirve para ilustrar ejemplos de uso de las palabras contrastivas identificadas. La lista completa arroja un resultado dentro del rango de las 300 palabras dignas de estudio por cada 5000 palabras, es decir, 1 palabra cada 17 aproximadamente. A pesar de que no existen otros proyectos que provean un término de comparación para evaluar el grado de éxito implicado en esta relación, no cabe ninguna duda de que, al menos en la detección de coloquialismos locales actualmente en uso, la herramienta plantea un verdadero punto de inflexión para la lexicografía contrastiva. Esta área del léxico es justamente la más elusiva, puesto que su impacto en cualquier medio impreso llega notablemente más tarde y, todavía más importante, en la mayoría de los casos no llega nunca. Se incluyeron como relevantes varias palabras que ya están incluidas en el Diccionario del habla de los argentinos [dL08], dado que ese hecho es una confirmación adicional de la pertinencia de la ubicación que asignó la métrica.

Las formas cuyo uso se ejemplifica son las que están en negrita y solo en ellas se normalizó la tildación.

■ Coloquialismos o vulgarismos

“Perdon pero tenes que ser muy **culiado/a** para ir a mc y pedirte una ensalada” (Córdoba)

“Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda” (Mendoza)

“Que **carnasas** poniendole rosas rojas a toda la ropa, para mi queda horrible sorry” (Neuquén)

“Teres, **pororós** y pelis con Carlita y Flor” (Chaco)

“Ver un negro **chuño** con musculosa y gorro.. se ve que el tipo no quería pasar ni frío ni calor.” (San Juan)

“Tenía la re expectativa para este sábado y al final **trancó** todo ” (Formosa)

■ Indigenismos

“Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo (Del guaraní mitaí “pequeño”)” (Formosa)

“**Angá** mi negrito, esta triste (Del guaraní angá aprox. “pobre”) (Corrientes)” (Corrientes)

“Gracias tormenta **ura** por sonar como una pochoclera de chasquibums a las 3 de la mañana en mi ventana durante 50 minutos. (Valor despectivo. Del quechua ura “vulva, vagina”) ” (Tucumán)

■ Gentilicios

Casildense (de Casilda), **concordiense** (de Concordia) y **obereño** (de Oberá).

■ Voces no marcadas en registro, que aluden a una realidad local

“Quiero a alguien que me diga vamos a comer **piadinas**, un pancho, un chori, una hamburguesa lo que sea y soy feliz” (San Juan)

“**Tareferos** que reclamaban asistencia interzafra en Posadas estarían preparando una protesta para hoy en la Fiesta del Inmigrante en Oberá.” (Misiones)

“Me encantan los bohemios anti sistema que usan vans. Es como que seas ecologista y uses un cuaderno hecho con media **yunga**.” (Jujuy)

■ Leísmo

“No te olvides de **saludarle** a tu suegro hoy” (Misiones)

“Vine a **visitarle** a mis primas y estan re colgadas, para eso me quedaba en mi casa no maaa ” (Misiones)

“A **esperarle** a nahuel, que traiga los teresss ” (Formosa)

■ Fusiones y acrónimos que pueden señalar pronunciación o alta frecuencia de uso

“Los sueños de la siesta me dejan **patra** ” (Buenos Aires)

“Si mañana me dice q no, voy sola, necesito ver esa pelicula en el cine siosi” (Córdoba)

■ Voces sospechadas generales pero con acepción local diferente

“Mañana que alguien **atine** con parque y porrones” (Mendoza)

“**Mansas** ganas de sentarme a tomar un te con semitas” (San Juan)

“**Habilítenme** una nueva espaldaa” (Tierra del Fuego)

“sigo **asada** por cosas que han pasado hace como dos dias, que falla (Mendoza) / Que **asada** estoy, tengo la cabeza echa un lío” (San Juan)

■ Voces con una morfología propia de una región

Ejemplo: terminación azo/aza con base adjetiva.

“Creo que va a estar **malazo** lo de esta noche ” (San Juan)

“Esta **locaza** esa mina para hacer eso” (Córdoba)

■ Variantes ortográficas

Ejemplos: culiado (adj. despect. o fórmula de tratamiento de confianza) y tereré.

“Menos mal que soy de los chetos de la carne y mañana tengo **asao** todo el dia jajajajaj” (Tucumán)

“Un lunes con buen humor ta **pasao** ” (Catamarca)

“Ahora a la mañana tengo q ir hacerme la tarjebus jajajajj **mavale** q me estoy por levantarr jajajaj” (Corrientes)

“Q paja volver al colegio **culiaa**” (Córdoba)

“Que pajero el **qliao** este.” (Córdoba)

“Quiero recitaaal **qliaaaa**” (Córdoba)

“**Tereresss** y pile con todos mis primiss” (Entre Ríos)

“No se si hacerme un **tere** o un mate para pasar la siesta” (Corrientes)

“Es lo mas lindo no ir al colegio y quedarme a tomar **teresss**” (Chaco)

■ **Formas verbales coloquiales con sustantivos o adjetivos como base**

“Me calma mucho **mimosear** a mi perro ” (Neuquén)

“Me vine a acostar y ya me dicen que parezco de 80 años ME CHUPA UN HUEVO LO QUE PIENSEN, DEJENME **ABUELEAR** ” (Buenos Aires)

“Estaría bueno que ari venga aunque sea a saludarme y que no se quede todo el tiempo **pollereando**.” (Tierra del Fuego)

■ **Vesres:** Creación de palabras por inversión de sílabas que se usa jergalmente o con fines humorísticos.

“Estoy en lo de villa mateando con él y jimmy. Pinta **sogui** abundante más tarde dijeron ” (Corrientes)

“Uhhh me acuerdo si no habré saltado el muro del aguapey par colarme a los **cequin**. (cequín “fiesta de quince”)” (Chaco)

■ **Intejercciones**

“**Aijué**, encima me decís vieja, re que no pinta esto facundo jaja ya te dije como es la onda, fin ” (Formosa)

“**Ains**, una mujer hablando de fútbol.” (Formosa)

“Al fin una buena: hora libreeee! **Yirr** ” (Corrientes)

■ **Guaranismos** Cabe destacar que la detección de términos en guaraní coincide exactamente con la región guaranítica¹. Un ejemplo de esto fueron las palabras *angá*, *angaú* y *mitaí*. Como se puede ver en la Tabla 4.2 el contraste entre las frecuencia normalizadas ² de la región guaranítica y la del litoral da una noción de la importancia que tienen estos términos en el noreste argentino.

Estos términos serán agregados al Diccionario del habla de los argentinos [dL08].

¹ Según las regiones dialectales establecidas por Vidal de Battini[VdB64]

² La frecuencia normalizada es una medida de estandarización que indica la cantidad de veces que aparece una determinada forma por cada millón de palabras.

	Región Guaranítica		Región Litoral	
	#Ocurencias	Frecuencia Normalizada	#Ocurencias	Frecuencia Normalizada
Angá	1017	32,80	27	0,337
Angaú	261	8,42	2	0,025
Mitai	467	15,06	3	0,037

Tab. 4.2: Cantidad de ocurrencias y frecuencias normalizadas de las palabras en la región guaraní-tica y la del litoral. La cantidad total de palabras en la región guaraní-tica es de 31.007.510, mientras que la cantidad de términos en la región litoral es 80.186.170

4.5. Validación estadística

4.5.1. Test Hipergeométrico

Luego de realizar el listado de palabras ordenado por el valor de la información, se aplicó un test estadístico para validar que efectivamente el uso en la región seleccionada es significativamente mayor que en el resto del país. El test se aplicó sobre las 5000 palabras consideradas más contrastivas por nuestra métrica.

Decidimos usar el test hipergeométrico ya que queremos ver que la palabra sobre la que se hace el test no estuvo sobrerrepresentada en comparación con la población. Asumimos que la cantidad de ocurrencias de una palabra se puede modelar con una distribución hipergeométrica ya que se puede pensar como un experimento donde se obtuvieron k palabras exitosas en una región con n palabras y un total de N palabras en la Argentina. Las regiones que utilizamos para cada palabra son el conjunto de provincias que cubren el 80 % de las ocurrencias de dicho término. Luego, queremos calcular la significancia estadística de haber obtenido esas k palabras exitosas.

Luego, por cada palabra seleccionada como contrastiva le aplicamos el test estadístico con la siguiente hipótesis nula: la palabra tienen un uso homogéneo en las distintas regiones de la Argentina, es decir que la frecuencia de ocurrencias de cada palabra debería ser similar independientemente de la región. Por lo tanto, en caso de que la palabra sea contrastiva deberíamos obtener una baja probabilidad de haber obtenido diferencias entre las frecuencias de la palabra en una región con el resto del país. Para aplicar el test hipergeométrico representamos los datos sobre la palabra en una tabla de 2x2 como la de la Tabla 4.3.

	#Palabras Sobre Región	#Palabras en el resto de Argentina	Total
# Palabras w	k	$K - k$	K
# Palabras $\neq$ w	$n - k$	$N + k - n - K$	$N - K$
Total	n	$N - n$	N

Tab. 4.3: Tabla de contingencia

En primer lugar hicimos el test estadístico sobre las 5000 palabras candidatas a ser contrastivas identificadas a través de nuestra métrica. Con este test obtuvimos los p-valores de la Figura 4.1. Debido a que realizamos múltiples tests tuvimos que aplicarle una corrección para evitar falsos positivos. Decidimos usar la corrección de Bonferroni con $\alpha = 0,5$.

Ante los p-valores tan bajos, decidimos hacer el test estadístico sobre palabras que consideramos que no deberían tener una frecuencia muy variada en las distintas regiones,

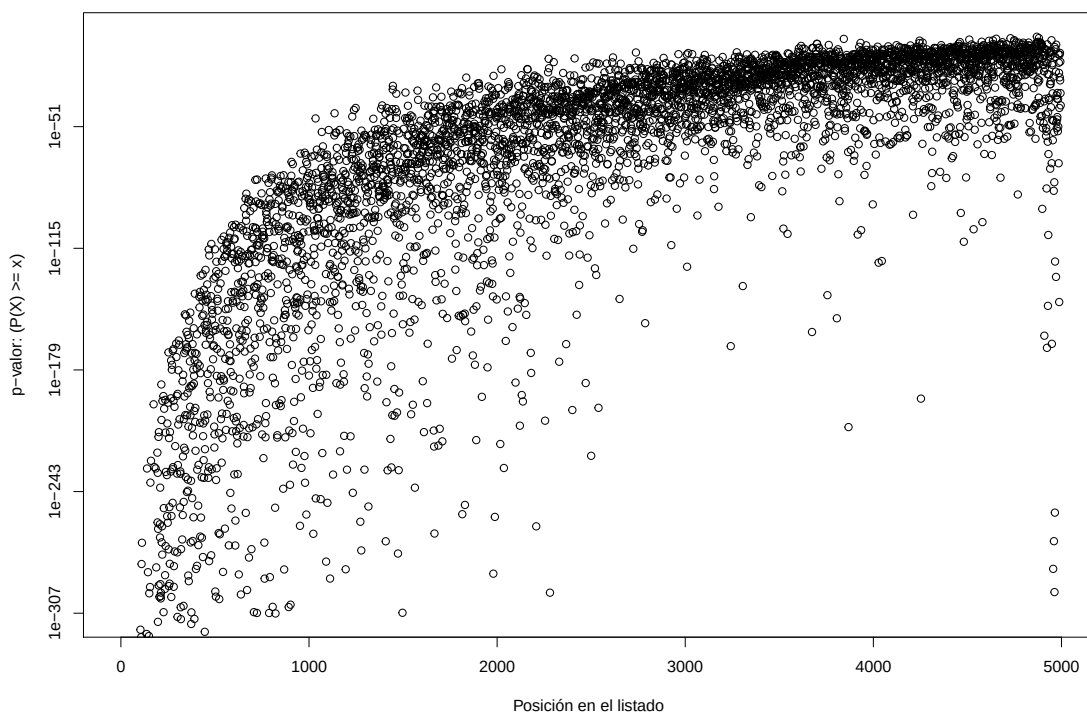


Fig. 4.1: Gráfico de dispersión de los p-valores del test hipergeométrico sobre las primeras 5000 palabras del listado ordenado según la métrica del valor de la información.

es decir, para las palabras *que*, *cuando*, *hola*, y también dieron p-valores menores a 0.001. Frente a esta situación investigamos las posibles causas de este fenómeno.

Un estudio de Adam Kilgarriff titulado “*Language is never, ever, ever, random*” [Kil05] sostiene que el uso de χ^2 y log-likelihood test son problemáticos, ya que se basan en la suposición de que todas las palabras son estadísticamente independientes. El autor afirma que, debido a que el lenguaje no es aleatorio (ya que hablamos y escribimos con una intencionalidad) y que la hipótesis nula de los tests estadísticos suponen aleatoriedad, cuando los datos provienen de fenómenos lingüísticos sobre corpus, la hipótesis nula nunca es cierta. Kilgarriff agrega que cuando hay suficientes datos (casi) siempre podemos ser capaces de rechazar la hipótesis nula. Con esto, el autor no quiere decir que los tests estadísticos no sirven, incluso menciona que en estadística se realizan tests que suponen la aleatoriedad cuando el fenómeno a estudiar no presenta esa propiedad. El problema surge cuando el fenómeno es tan arbitrario como el lenguaje y de esta manera se simplifica en demasía la hipótesis de aleatoriedad. Kilgarriff también hace un análisis sobre trabajos previos y comenta el caso en el que un estudio que quería encontrar palabras con frecuencias significativas entre el inglés británico y el americano, representados por el Corpus Brown (inglés americano) y el Lancaster-Oslo-Bergen Corpus (inglés británico). Para cada palabra testearon la hipótesis nula que afirmaba que la diferencia entre las frecuencias sobre los dos corpus se debía a una fluctuación aleatoria. El test lo hicieron a partir de muestras obtenidas aleatoriamente de los corpus mencionados. En este estudio se marcaron las palabras donde la hipótesis nula se rechazaba con distintos niveles de confianza. Las listas

sugieren que la mayor parte de las palabras *comunes* fueron marcadas, es decir que el test estadístico sugería que todas estas palabras tenían diferencias significativas en su uso. Esto que comenta Kilgariff es lo que tuvimos como resultado a partir de nuestro test hipergeométrico. Lo interesante es que el autor atribuye ese fenómeno a la esencia no aleatoria del lenguaje. Si bien sabíamos que al realizar el test hipergeométrico suponíamos que todas las palabras son estadísticamente independientes, pensamos que esta suposición no iba a afectar tanto los resultados como lo hizo.

Para entender este fenómeno es necesario notar el modelo que se usó al representar los datos para el test hipergeométrico. Este modelo es conocido como bolsa de palabras o *bag-of-words* y consiste en calcular las frecuencias de las palabras sobre todos los textos, sin tener en cuenta la distribución de las frecuencias en cada uno. Por lo tanto, se representa al corpus como un gran texto en el que no importa el orden de las palabras. Tanto Kilgariff [Kil01] como Paquot y Bestgen [PB09] afirman que es posible representar los datos de manera diferente y hacer uso de otros tests suponiendo la independencia entre los textos y no entre las palabras. De esta manera se puede observar la distribución de las palabras dentro de un corpus. En un estudio más reciente, Lijffijt et al. [LNS⁺16] analizan algunas de estas alternativas para hacer un test de hipótesis en los cuales no se supone el modelo de *bag-of-words*. En ese trabajo se analizaron el test t de Welch [Wel47], el test de los rangos con signo de Wilcoxon (Wilcoxon rank sum), el de Bootstrap y el de tiempo entre llegadas (inter-arrival time). Lijffijt et al. explican que la diferencia entre estos tests con los que suponen el modelo de *bag-of-words* (como el hipergeométrico, el χ^2 y el log-likelihood test) reside en la representación de los datos, ergo la unidad de observación.

Para los tests que suponen los modelos de bolsa de palabras, los datos se representan en una tabla de contingencia de 2x2 y el número de muestras equivale al número de palabras en el corpus, mientras que en los otros cuatro tests, los datos son representados en una lista de frecuencias o una lista de *tiempos de llegada*. En estos casos, el número de muestras es mucho menor que la cantidad de palabras en el corpus. Se destaca el número de muestras de los modelos ya que este número generalmente determina nuestro nivel de seguridad en relación a los valores estimados. Es por esto que los resultados experimentales muestran que los tests de modelos de *bag-of-words* tienen una confianza excesivamente alta en los valores estimados de la frecuencia media de las palabras, en el contexto de comparación estadística entre dos corpus [LNS⁺16].

4.5.2. Test t de Welch

Basándonos en las propuestas de Lijffijt, decidimos utilizar el test de Welch. Este nos provee un valor de probabilidad para rechazar la hipótesis nula que afirma que las medias de las dos distribuciones son iguales. Sean S y T dos corpus y sea q la palabra sobre la cual se va a hacer el test, sea x_1 la media de la frecuencia de la palabra w sobre los textos de S , y sea s_1 la desviación estándar. Análogamente, sea x_2 la media de la frecuencia de q en los textos T y s_2 la desviación estándar. El estadístico t se calcula con la siguiente ecuación:

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{|S|} + \frac{s_2^2}{|T|}}} \quad (4.1)$$

Las suposiciones del test consisten en que todos los textos son estadísticamente independientes y que la media de las frecuencias proviene de una distribución normal. En

nuestro caso, agrupamos todos los tuits de cada usuario representando un texto. De esta manera, cada provincia tiene alrededor de 900 textos formados por distintos usuarios.³ Luego, el test es aplicado a cada palabra con las frecuencias entre dos corpus: uno está formado por todos los textos de los usuarios que provienen de las provincias en donde se cubre el 80 % de las ocurrencias, el otro consiste en los textos creados por usuarios del resto de las provincias. Notar que suponer la independencia entre todos los textos de un usuario de una provincia, es una suposición más débil que la de suponer independencia en todos los textos generados en una provincia como se hacía en el modelo de bolsa de palabras.

Puede observarse la distribución de los p-valores para distintos grupos de palabras en la Figura 4.2. En éste, se pone de manifiesto que la distribución de estos no es uniforme, sino que tienden a ser p-valores bajos, y no queda del todo claro cómo se comparan estos resultados con los obtenidos por Lijffijt et al. Cabe destacar que el análisis de uniformidad realizado en dicho trabajo es distinto en varios aspectos: por empezar, nosotros definimos un “corpus” para cada una de las palabras debido a las regiones de uso. Por otro lado, las particularidades de los textos de Twitter podrían estar generando un artefacto en el cálculo de los p-valores.

Sin embargo, a partir de esta validación estadística obtuvimos algunos resultados que dan un indicio de las virtudes de la métrica desarrollada, los cuales exponemos a continuación. Uno de ellos es la comparación de las *tasas de rechazo de la hipótesis nula* sobre las tres métricas desarrolladas, I_p , I_w y I mencionadas en 3.14, 3.10 y 3.15 (páginas 16 y 17) respectivamente. Definimos a la *tasa de rechazo* como la proporción de tests que tienen un p-valor menor a 0.05, es decir:

$$\text{Tasa de rechazo}(\text{tests}) = \frac{\#\{t : \text{tests} \mid p\text{-valor}(t) < 0,05\}}{\#\text{tests}} \quad (4.2)$$

Calculamos la tasa de rechazo para las palabras en distintos intervalos del listado ordenado según las tres métricas elegidas: la métrica que tiene en cuenta la entropía de palabras, la que valora la entropía de personas, y la que contiene a los dos factores. En la Figura 4.3 se muestran los resultados, donde la tasa de rechazo se calcula después de haber aplicado la corrección de Bonferroni con $\alpha = 0,5$ para reducir la probabilidad de falsos positivos. La métrica elegida, que tiene a ambos factores en consideración tiene una mejor tasa de rechazo de la hipótesis nula en las palabras consideradas contrastivas y una menor tasa de rechazo para el resto. Es importante notar que el test de Welch que realizamos tiene como muestras las distintas frecuencias de todos los usuarios en cada región. Por lo tanto, es razonable obtener un resultado como este, en el cual las métricas que consideran la dispersión sobre los usuarios y las palabras tienen un mejor comportamiento sobre la que tiene en cuenta solamente a la dispersión de las palabras.

También es importante destacar que a medida que uno se aleja de las palabras más contrastivas de acuerdo a nuestra métrica, la tasa de rechazo es menor. Esto refleja el buen comportamiento de la métrica. Este resultado se puede observar también en la Figura 4.2 donde se detalla la distribución de los p-valores en el conjunto de las primeras 5000 palabras y el resto de los términos del listado.

³ La cantidad de usuarios recolectados por cada provincia se encuentra detallada en la Tabla 2.1

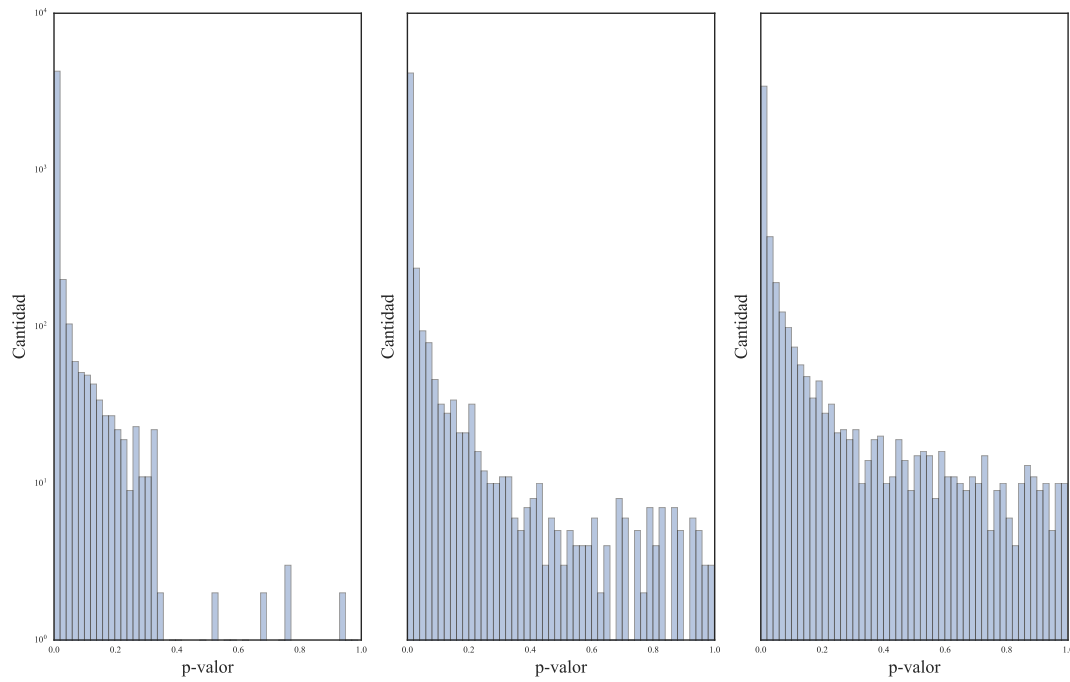


Fig. 4.2: Distribución de p-valores (sin corrección de tests múltiples) en los distintos conjuntos de palabras. De izquierda a derecha se muestran los gráficos dependiendo de los índices de las palabras en el listado ordenado por la métrica del valor de la información: $[0,5000]$, $[20000,25000]$, $[50000,55000]$

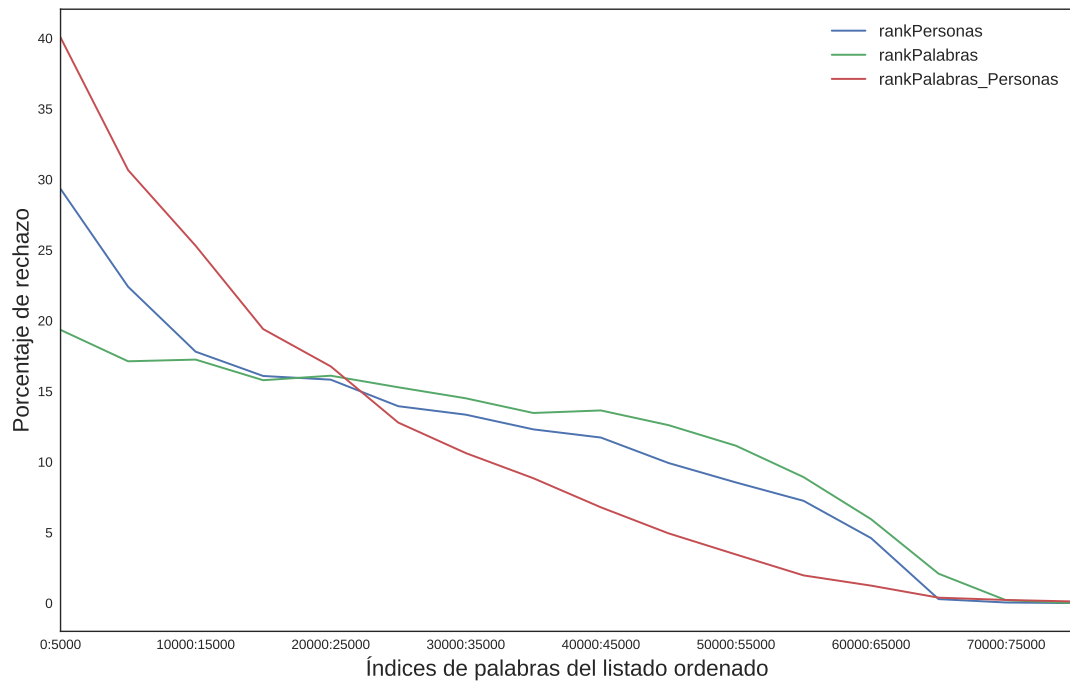


Fig. 4.3: La tasa de rechazo de la hipótesis nula en los distintos conjuntos de palabras, los cuales varían según el índice de estas en el listado ordenado según la métrica elegida.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1. Conclusiones y trabajo futuro

En el presente trabajo multidisciplinario desarrollamos una métrica de la contrastividad de uso de una palabra en distintas regiones. Para probar esta métrica recolectamos un conjunto de datos de textos de la Argentina a través de la API de Twitter.

La métrica que creamos usa la entropía para medir la variación de la cantidad de ocurrencias y de la cantidad de usuarios que la utilizaron en las diferentes provincias del país. Se seleccionaron las 5000 palabras con mayor valor de contrastividad para realizar una validación lingüística por parte de la Academia Argentina de Letras. La validación arrojó un resultado con alrededor de 300 palabras dignas de estudio de esas 5000 palabras, es decir, 1 palabra cada 17. A pesar de que no existen otros proyectos que provean un término de comparación para evaluar el grado de éxito implicado en esta relación, no caben dudas de que, al menos en la detección de coloquialismos locales actualmente en uso, la herramienta plantea un verdadero punto de inflexión para la lexicografía contrastiva. Varias de las palabras detectadas a partir de la métrica desarrollada serán agregadas al Diccionario del Habla de los Argentinos.

En cuanto a la validación estadística, dejamos como trabajo futuro el cálculo de un análisis estadístico aplicable a nuestra métrica, ya que está por fuera del alcance de esta tesis. Sin embargo, en base al análisis hecho a través del test t de Welch también tenemos indicios de las virtudes de la métrica desarrollada.

En este trabajo se analizan las regiones formadas con una provincia como unidad regional, pero esta se puede cambiar para replicar el análisis con distinta granularidad. De esta manera se podrían ver las palabras contrastivas en los distintos países hispanoparlantes y comparar las variaciones entre regiones más grandes o bien, replicar el trabajo en el interior de una sola provincia o ciudad.

Uno de los desafíos que dispara este trabajo es el de poder identificar regiones/clusters con usos dialectales diferentes. A su vez, permitiría validar la vigencia de las regiones propuestas por Vidal de Battini en 1964 [VdB64].

También, el proceso de normalización se podría mejorar para tener una mayor precisión en las palabras utilizadas. A partir de una mejor normalización y de la lematización con información metalingüística del corpus, podríamos trascender el ámbito de léxico para estudiar los fenómenos sintácticos del español, como su variación en distintas regiones. Continuando con la línea de investigación se podría analizar la contrastividad léxica comparando la distribución de n-gramas. Por otro lado, sería útil agregar un sistema de reconocimiento de nombres de entidades para destacar ciertos nombres propios, de manera tal que el listado de palabras tenga más alertas sobre términos sin interés lingüístico.

Es importante señalar las ventajas de *Twitter* ya que nos permitió recolectar un volumen grande de datos de texto, escritos por distintas personas con información de su localización. En cuanto a las desventajas de esta plataforma podemos destacar los errores ortográficos de los textos, la modificación intencionada de las palabras para generar énfasis o con motivo de una escritura más rápida. Todo esto conlleva a un aumento de la dificultad para normalizar el texto. Creemos, a pesar de todo esto, que el volumen de datos prevalece a la hora de decidir una plataforma para recolectarlos.

6. APÉNDICE

6.1. La entropía como medida del desorden

Para ver la cantidad de información que nos aporta cada palabra se hará una introducción a la teoría de la información, específicamente los conceptos que introdujo Claude Shannon [Sha01, Abr63]. Para entender estos conceptos es útil tener una descripción matemática del mecanismo que genera la información. Para eso se define a la *fente* que emite señales de un alfabeto $S = \{s_1, s_2, \dots, s_n\}$ de acuerdo a una función de probabilidad fija. Si la fuente emite señales estadísticamente independientes decimos que es una *fente de memoria nula* y un símbolo s está completamente determinado por el alfabeto S y las probabilidades $p_1, p_2, \dots, p_n$.

Luego, sea $\mathbf{p} = (p_1, \dots, p_n)$ un vector de probabilidad puntual. Es decir, $p_i \geq 0$ y $\sum_{i=1}^n p_i = 1$. Definimos la entropía de $\mathbf{p}$ siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i) p_i. \quad (6.1)$$

De esta forma, la función H satisface las siguientes propiedades:

- (i) $H(\mathbf{p}) = 0$ si y solo si $\mathbf{p}$ esta concentrada en un unico punto: existe i tal que $p_i = 1$ y $p_j = 0$, para todo $j \neq i$.
- (ii) La función H se maximiza tomando $\mathbf{p}$ equiprobable: $p_i = 1/n$, para todo i .

Tenemos entonces que la máxima entropía se realiza cuando todos los elementos tienen igual probabilidad de ocurrir. Ninguno de ellos es más probable que otro. No hay como predecir un valor en particular. En tal caso, decimos que el desorden es máximo. Por el contrario, cuando la probabilidad se concentra en un único punto, estamos en un sistema desteterminístico: orden total. En este sentido es que H es una medida de *orden* del sistema. Mayor entropía dice de más desorden; entendiendo desorden como *aleatoriedad completamente impredecible - distribución uniforme*.

6.2. Tablas y Gráficos

Palabra	Cantidad de Ocurrencias
que	7509160
de	6527014
a	4962492
la	4913854
no	4177810
me	4101998
y	3838370
el	3773455
en	2969783
te	2060662
se	1976027
un	1863075
es	1825892
con	1799979
lo	1712189
mi	1643777
por	1553382
los	1498941
para	1398757
las	1212452

Tab. 6.1: Cantidad de apariciones de las 20 palabras más frecuentes.

Conjunto de Provincias			
Salta-Jujuy	Mendoza-San Juan	Chubut-Santa Cruz-T. Del Fuego	Chaco-Corrientes-Formosa
tribuno	mza	austral	anga
salteño	zonda	chilote	teresss
orán	secamente	calafa	músicas
tartagal	sanjuaninos	chilota	argela
salteña	sanjuanino	palmaso	angaaa
martearena	asar	vueltones	olo
oran	tomba	riviera	cuchale
yuto	queras		corrientesss
purmamarca	traica		angá
yutos	sopaipillas		cts
gorriti	ardente		correntinas
quijano	secamentes		iburrr
tabacal	jáchal		cheraa
desentierro	virreina		cheraá
huaico	tombaaa		bofill
pichanal	mansooo		receppp
diablos	tombino		
bandy	parisi		
aramayo	asadaaa		
ñaño			
colque			
urkupiña			
juj			
guachipas			

Tab. 6.2: Palabras cubiertas sobre el 80 % de las ocurrencias totales por el conjunto de provincias a partir de las 5000 palabras más contrastivas (de acuerdo al valor de la información).

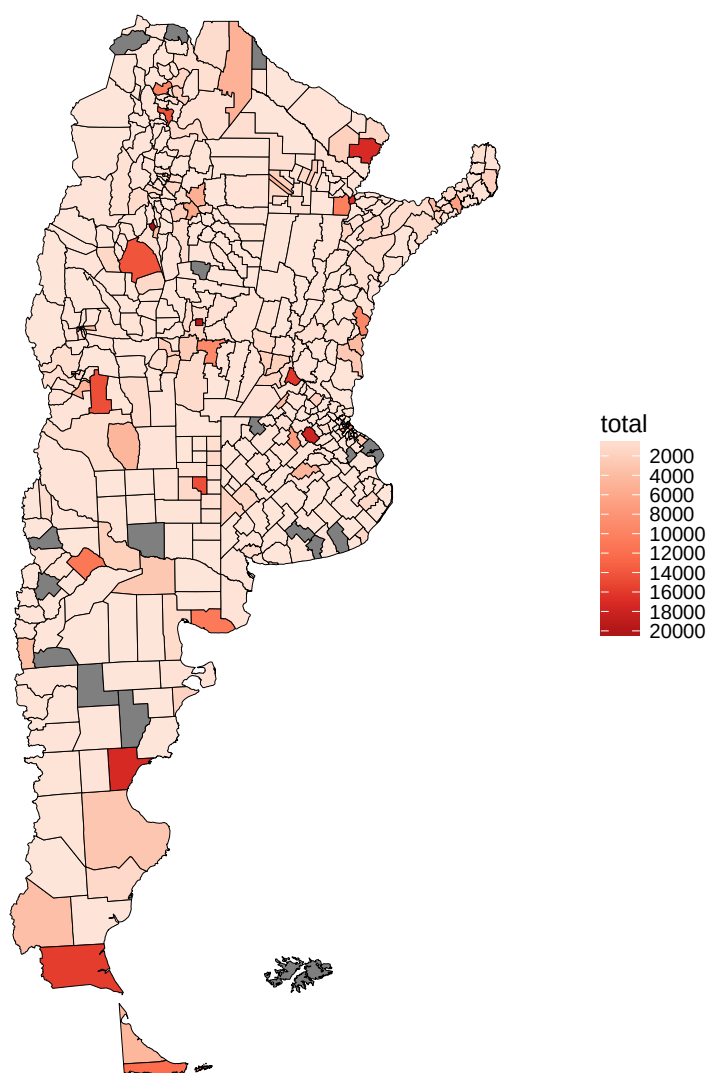


Fig. 6.1: Mapa con la distribución de los tuits que incluyeron sus coordenadas geográficas.

7. BIBLIOGRAFÍA

Bibliografía

- [Abr63] Norman Abramson. Information theory and coding. 1963.
- [AV95] Manuel Almeida and Carmelo Vidal. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):Pág–50, 1995.
- [Ávi04] Raúl Ávila. ¿El fin de los diccionarios diferenciales? ¿El principio de los diccionarios integrales? 2004.
- [Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [Dav15] Mark Davies. Corpus del Español. <http://www.corpusdelespanol.org/xs.asp?c=3>, 2015.
- [dL08] Academia Argentina de Letras. *Diccionario del habla de los argentinos*. Emecé Editores, 2008.
- [Doy14] Gabriel Doyle. Mapping dialectal variation by querying social media. In *EACL*, pages 98–106, 2014.
- [Eis14] Jacob Eisenstein. Identifying regional dialects in online social media. Georgia Institute of Technology, 2014.
- [EOSX10] Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.
- [Esp] Real Academia Española. Banco de datos (corpes xxi)[en línea]. *Corpus del español del siglo XXI (CORPES)*.
- [GS14] Bruno Gonçalves and David Sánchez. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074, 2014.
- [KG03] Adam Kilgarrieff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.
- [Kil01] Adam Kilgarrieff. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133, 2001.
- [Kil05] Adam Kilgarrieff. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276, 2005.
- [Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.
- [LNS⁺16] Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2):374–397, 2016.

-
- [MH11] Tony McEnery and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.
- [Mon01] Marcelo A Montemurro. Beyond the zipf-mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.
- [MZ10] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.
- [PB09] Magali Paquot and Yves Bestgen. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and computers studies in practical linguistics*, 68:247, 2009.
- [PD11] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsn*, 20:265–272, 2011.
- [PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.
- [Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- [TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsn*, 10(1):178–185, 2010.
- [VdB64] Berta Elena Vidal de Battini. El español en la argentina. Technical report, Argentina., 1964.
- [Wel47] Bernard L Welch. The generalization of student’s problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.
- [Zim06] Klaus Zimmermann. El fin de los diccionarios de mexicanismos, colombianismos, argentinismos, cubanismos etc. la situación de la lexicografía del español de américa después de la publicación de los diccionarios contrastivos del español de américa: Español de amér. *Estudios de lingüística del español*, 23, 2006.
- [Zip16] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.