Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Molina Salinas, Claudio; Sierra Martínez, Gerardo Eugenio
Hacia una normalización de la frecuencia de los corpus CREA y CORDE
Revista Signos, vol. 48, núm. 89, diciembre, 2015, pp. 307-331
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Disponible en: http://www.redalyc.org/articulo.oa?id=157043352002

Cómo citar el artículo
Número completo
Más información del artículo
Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2015 PUCV, Chile · DOI: 10.4067/S0718-09342015000300002 · 48(89) 307-331

Hacia una normalización de la frecuencia de los corpus CREA y CORDE

Towards a frequency normalization of CREA and CORDE corpora

Claudio Molina Salinas
Universidad Nacional Autónoma de México
Escuela Nacional de Antropología e Historia
México
claudio.molina.salinas@gmail.com

Gerardo Eugenio Sierra Martínez
Universidad Nacional Autónoma de México
México
gsierram@iingen.unam.mx

Recibido: 09-V-2014 / **Aceptado:** 26-I-2015

Resumen

El Corpus Diacrónico del Español (CORDE) y el Corpus de Referencia del Español Actual (CREA) integran uno de los bancos de datos documentales más grande del español y son usados por especialistas en el estudio de la lengua española. Este recurso presenta algunas limitaciones en términos de tamaño, unidad de muestra y representatividad, que condicionan sus resultados y, por tanto, las descripciones de cualquier fenómeno lingüístico estudiado. En el presente trabajo identificamos estas limitaciones y proponemos un método de normalización de frecuencias de documento, por medio del cálculo de medias móviles. Esto permite una interpretación más realista de la lengua española, a través de los datos del corpus, y un aprovechamiento más efectivo del propio recurso.

Palabras Clave: Lingüística de corpus, estadística de corpus, normalización de frecuencias de corpus, lingüística cuantitativa, corpus CORDE y CREA.

Abstract

CORDE (*Corpus Diacrónico del Español)* and CREA (*Corpus de Referencia del Español Actual)* are two of the largest and most frequently used databases in the study of the Spanish language. However, they have some limitations in terms of size, sample unit and representativeness that may influence the results of studies and descriptions of linguistic phenomena. In this paper we identify these limitations and propose a method for the normalization of document frequencies by computing moving averages. We show how this method allows for a more realistic interpretation of corpus data and, thus, a more effective use of these resources.

Key Words: *Corpus* linguistics, corpus statistics, corpus frequency normalization, quantitative linguistics, CORDE and CREA *corpora*.

INTRODUCCIÓN

Aunque existen un sinnúmero de procedimientos para formalizar una investigación lingüística sobre cualquier fenómeno particular (seguir una sola postura metodológica, retomar ideas de distintos trabajos de investigación, valerse de una argumentación muy bien articulada para construir una teoría, entre otros), es muy común que la gran mayoría de los trabajos lingüísticos se basen en evidencias, hechos representativos y suficientemente claros del fenómeno que se pretende describir.

Independientemente del marco teórico que se elija como el más adecuado, en general, la base empírica de una investigación lingüística son hechos reales, ejemplos de uso de la lengua que constituyen un fundamento para el análisis lingüístico y hacen posible establecer una relación entre teoría y datos lingüísticos en el orden propio, además de legitimar hipótesis formuladas respecto al funcionamiento de una lengua natural. Suele llamarse corpus lingüísticos a este conjunto de ejemplos de uso, mismos que no siempre son colecciones documentales de gran extensión.

En la bibliografía lingüística es frecuente encontrar trabajos de investigación basados en pequeñas listas de ejemplos, sin que esto minimice su valor y contribución a la descripción de las lenguas naturales y el desarrollo del conocimiento lingüístico. En realidad, parecería que no importa la cantidad de datos seleccionados para la descripción de un fenómeno lingüístico, tanto como que el fenómeno estudiado esté bien representado y que el corpus recoja, bajo ciertos criterios de ordenamiento, descripciones o notas sobre el hecho lingüístico en cuestión.

Centrándonos en el caso de los corpus de gran magnitud, más específicamente, el caso de la lengua española, la Real Academia Española (RAE) pone a disposición de cualquier interesado en el estudio de nuestra lengua un recurso en línea, de

acceso gratuito y fácil consulta: un banco de datos compuesto por cuatro conjuntos documentales bien diferenciados: el Corpus Diacrónico del Español (CORDE), el Corpus de Referencia del Español Actual (CREA), el Corpus del Español del Siglo XXI (CORPES XXI) y el Corpus del Diccionario Histórico (CDH). Sin embargo, esta base de datos presenta algunas limitaciones relacionadas con su diseño, imputables a su gran extensión, selección de géneros textuales y unidad de muestra que afectan su representatividad. Lo que nos hace suponer que el análisis cuantitativo de la gran mayoría de los fenómenos lingüísticos que se estudien con ejemplos provenientes de este recurso pudieran presentar sesgos.

Con base en algunos ejemplos de análisis del léxico del español señalamos estos aspectos identificados y proponemos una opción de normalización de frecuencias, conceptualmente sencilla y de fácil aplicación, que permitirá a cualquier analista aprovechar más efectivamente el recurso académico y ofrecer una interpretación cuantitativa más precisa del comportamiento histórico y contemporáneo del léxico de la lengua española. Método que pudiera aplicarse al análisis de otros niveles de la lengua e, incluso, dar cuenta de la variación geográfica, temática, entre autores u obras.

1. Marco teórico

1.1. Diseño de grandes corpus

Los trabajos lingüísticos sobre corpus de gran tamaño, útiles para un sinfín de aplicaciones, concuerdan que un corpus lingüístico es una colección de elementos que fueron seleccionados, descritos y ordenados con una finalidad explícita: ser usados como muestras representativas de la lengua o de un fenómeno particular de ella.

Sinclair (1996) define corpus lingüístico como "a collection of pieces of a language that are selected and ordered according to explicit linguistic criteria in order to be used as sample of the language". En esta definición, el autor señala la importancia de los criterios de orden y selección de los documentos que componen un corpus, ya que a partir de ellos se establecen las diferencias existentes entre un corpus lingüístico y cualquier otro tipo de colección textual. Considerando esto, se puede decir que no toda colección de textos es un corpus: el acervo de una biblioteca como tal no lo sería, pues no existen criterios de selección lingüística y representatividad de un estado de la lengua (Sierra, 2008).

Un corpus debería ser una selección de datos lingüísticos reales que constituya una buena aproximación a la realidad de la lengua estudiada, agrupados bajo ciertos criterios que garanticen que todas o algunas variedades lingüísticas estén representadas (dependiendo de la finalidad del corpus) y, además de construirse a partir de lineamientos específicos sobre su propia definición y su estructura, se debe ajustar a un principio fundamental: ser representativo.

En la noción de representatividad está implícito qué tan apegado a la realidad de la población estudiada está un corpus. Esta realidad es dinámica, es decir, se puede modelar considerando aspectos como la localidad geográfica de donde proceden los textos, el autor de ellos o informante, el tópico, el tipo de texto, una fuente explícita, el tiempo en el que fue creado o el momento de su producción, entre muchos más (Sierra, 2008). Criterios que establecen las diferencias entre distintos tipos de corpus y las variedades que estos representan. Entonces, el tipo de variedad de lengua o tipo de lenguaje en particular que se incluyen en un corpus está relacionado y condiciona directamente la oportunidad y representatividad de este.

Biber (1993) afirma que la representatividad de un corpus, entre otras cosas, radica en determinar qué porcentaje de cada tipo textual se incluirá en este y en qué proporción aparecerá un género respecto a otros, noción designada en inglés como '*balance*', que en español suele llamarse: 'equilibrio'. La representatividad tiene relación directa con el equilibrio en la medida de que un corpus es capaz de reproducir las diferencias de las variables que se supone representa, en la proporción en la que ocurren en el estado natural de la lengua.

Además de considerar la variedad y equilibrio implicados en el diseño y caracterización de un corpus, la unidad de muestra resulta indispensable para orientar la pertinencia de un corpus y su funcionalidad, ya que garantiza la pertinencia de este. La unidad de muestra determina la utilidad de un corpus, ya que si se considera el texto completo, el corpus podría servir para estudiar todos los niveles de la lengua en general, mientras que los corpus que consideran fragmentos de un discurso, párrafos u oraciones aisladas de diferentes documentos sirven exclusivamente para estudios del léxico (Biber, 1993; Torruella & Llisterri, 1999).

Considerar una unidad de muestra es conveniente, ya que a partir de ella se recuperan sistemáticamente datos que se incluirán en un corpus y con ellos se forma una colección estable y comparable de las unidades y fenómenos de la lengua que se pretenden describir (Biber, 1993). Sin embargo, al considerar documentos completos resulta virtualmente imposible tener unidades estadísticamente comparables dentro del corpus sin recurrir a un método de normalización de frecuencias. En particular, el diseñador de un corpus debiera procurar alternativas que subsanen esta desproporción, por ejemplo, orientar la selección de la muestra de documentos completos a fragmentos del discurso limitados o proponer un método de normalización de frecuencias que permita comparar una novela de 290 cuartillas de extensión con una nota breve de 500 palabras proveniente de una publicación periódica, por citar un ejemplo frecuente.

La representatividad no solo radica en la diversidad de variables de una lengua que puedan estar incluidas en un corpus, sino en la proporción en la que estas se registran respecto a la totalidad de los datos considerados. Por ello se recomienda seguir un método de recopilación y ordenamiento de los datos en el corpus, ya sea tomar muestras aleatorias no estratificadas o estratificar el muestreo. Una muestra aleatoria

no estratificada, en la mayoría de los casos, pudiera condicionar la aparición de algunos fenómenos o propiciar un desequilibrio en un corpus con fines no lexicográficos. Por tanto, para evitar tendencias indeseables, lo recomendable es utilizar un método de muestreo aleatorio estratificado que, si bien considere la elección aleatoria de la unidad de muestra dentro del texto, también tome en cuenta tipo y género textual, tipo de población y otros factores en las proporciones que ocurren en la lengua o bien establecidas por quien diseña el corpus y sus intereses de investigación; procedimiento que debiera seguirse incluso en el caso de corpus de documentos completos (Biber, 1993).

1.2. CORDE y CREA: Diseño y representatividad[4]

Los corpus CORDE y CREA recopilan en conjunto más de 200 millones de palabras desde el origen del español hasta la época actual. Por un lado, el CREA es un 'corpus monitor' que abarca los últimos 25 años más próximos a nuestro tiempo, en tanto el CORDE recopila documentos desde los orígenes del español hasta los límites temporales del CREA, año 1975. Ambos corpus constituyen 'muestras primordiales' (Kupietz, Belica, Keibe & Witt, 2010) del español en sus distintas etapas históricas, esto debido a que no se ajustan a la población estadística de la lengua que se supone representan.

En un principio, el CREA se describe como un recurso que ofrece muestras del español estándar recogiendo más de 125 millones de formas para los últimos veinticinco años. Hoy en día, sin embargo, se pueden recuperar ejemplos recopilados en el CREA, registrados entre 1974 y 2004, esto significa que el CREA ha crecido cronológicamente de veinticinco a treinta y un años y, obviamente, la cantidad de formas que recopila en la actualidad no coincide con la mencionada en su manual. Además, al ser un corpus periódico de 25 años, toda documentación etiquetada con fecha anterior al primero de enero de 1980, cuando menos en el estado de actualización en el que se encuentra, debiera haber pasado a formar parte de la documentación del CORDE.

Las variedades del español representadas en el CREA están divididas así: el 50% del corpus son documentos provenientes de España y el otro 50% restante de toda América; el 90% son documentos escritos y el restante 10% son documentos orales, aunque no se especifica si esta proporción es para cada división del corpus o para todo el conjunto documental.

Además, se plantean diferentes áreas lingüísticas o zonas dialectales americanas, en las que se ponderan en proporción documental algunas zonas sobre otras. Lamentablemente, esta proporción documental no coincide con el número de hablantes de español como lengua materna para cada una de las zonas geográficas que se describen. Solo por señalar un ejemplo, el número de hablantes de español en la península ibérica no es mayor que el número de hablantes de la variante del español

mexicano residentes en México, según las cifras descritas por el Instituto Cervantes[2]; desigualdad que afecta directamente la representatividad del corpus.

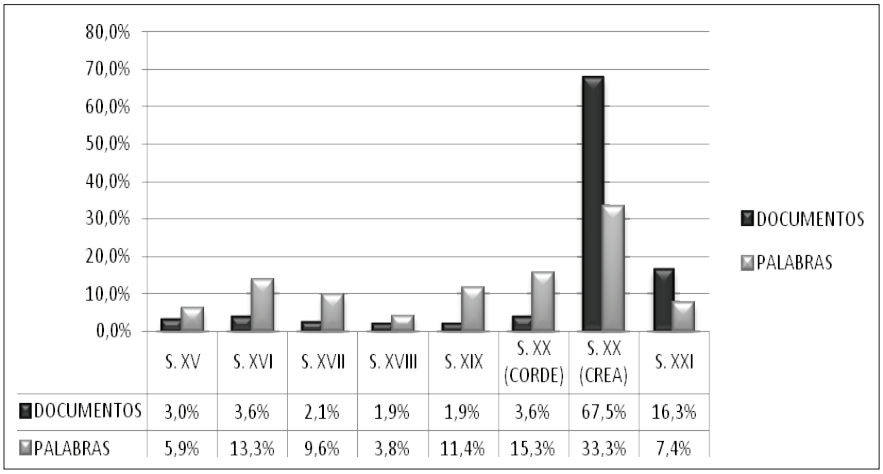
A partir de este hecho se podría decir que, en el CREA, la zona española se encuentra sobrerrepresentada, mientras que el conjunto documental para la zona mexicana está representada por debajo de la realidad. En este caso, no ajustarse a un criterio demográfico para la inclusión de documentos afecta directamente el equilibrio y, por ende, la representación de países dentro del corpus. Sin embargo, este desequilibrio podría justificarse argumentando que la producción escrita y disponibilidad de documentación no es la misma para todas las zonas geográficas representadas dentro del corpus, lo cual significaría desestimar el criterio demográfico.

De igual manera, se describe que, para este corpus, en el periodo comprendido entre 1995 y 1999 se concentra el 30% de los documentos; entre 1990 y 1994, el 25%; y así continúa descendiendo un 5% por periodo de 5 años hasta el lustro de 1975 a 1979 en el que solo se concentran el 10% del total de documentos. Esta distribución documental desigual no se justifica por ningún motivo, pero sí afecta nuevamente la representatividad del corpus, además de que no describe el periodo comprendido entre los años 2000 y 2004, que sí está representado.

El CORDE, por otra parte, es un conjunto muy grande de textos completos que perduraron hasta nuestros días a pesar de las vicisitudes del tiempo, conservación y disponibilidad. Este corpus recoge un 74% de documentos históricos del español peninsular y solo un 26% para el resto de las zonas dialectales propuestas. Esta distribución documental refleja la historia compartida entre España y sus colonias, así como por los problemas que representó en su momento la difusión del español, sin dejar de considerar factores como la producción de manuscritos y su preservación.

Este corpus agrupa la documentación en tres grandes etapas históricas: la Edad Media (de los orígenes del español a 1492) que reúne el 21% del total de los documentos, los Siglos de Oro (de 1493 a 1713) que recogen el 28% y la Época Contemporánea (de 1714 a 1974) que alberga la gran mayoría de los documentos, 51% del total.

Si analizáramos todo este planteamiento desde una perspectiva global, en la que se considere el total de la documentación para ambos corpus, se vería una desproporción entre el siglo XV y el XX que se conserva, incluso, si consideráramos el número de palabras que hay por siglo en la totalidad de la base de datos. El Gráfico 1 ilustra la proporción en que se distribuyen los documentos y el número de palabras dentro de todo el banco de datos.



	S. XV	S. XVI	S. XVII	S. XVIII	S. XIX	S. XX (CORDE)	S. XX (CREA)	S. XXI
DOCUMENTOS	3,0%	3,6%	2,1%	1,9%	1,9%	3,6%	67,5%	16,3%
PALABRAS	5,9%	13,3%	9,6%	3,8%	11,4%	15,3%	33,3%	7,4%

Gráfico 1. Frecuencia de palabras y documentos por siglo en el CORDE y CREA.

Aunque ya señalamos la diferencia entre siglos para la proporción documental de la base de datos e hicimos hincapié en la desproporción de estos, nos interesaba presentar el Gráfico 1 para ilustrar la tendencia que se observa en el CORDE en la que el número de palabras, en términos porcentuales, siempre es mayor que el número de documentos, independientemente de las diferencias que existan entre siglos.

En esta misma gráfica, el caso del CREA resulta interesante, ya que para los siglos XX y XXI el porcentaje de documentos es menor que el porcentaje de número de palabras, lo que nos permite concluir que en los primeros siglos se tenían pocos documentos de mayor extensión respecto a la época reciente, en la que se tiene un mayor número de documentos pero de menor extensión.

Todo el escenario descrito hasta el momento afecta la representatividad de ambos corpus y, adicionalmente, los hace incomparables entre sí. Sin embargo, tomando en cuenta que el CORDE y el CREA cuentan con una interfaz de consulta en la que se puede realizar búsquedas atendiendo a diferentes criterios o filtros (cortes temporales, áreas temáticas, variantes diatópicas, el medio que lo difunde, la obra de un autor o, incluso, hacer consultas en una obra específica), y con ello modelar cuantos 'subcorpus' o 'corpus virtuales' (Kupietz et al., 2010) se desee, el mismo recurso ofrece una alternativa para comparar sus datos entre sí.[3]

Por último, hemos identificado que la base de datos solo puede cuantificar la frecuencia absoluta para las ocurrencias solicitadas en su buscador y, lamentablemente, no considera ofrecer valores normalizados para los resultados ofrecidos, mismos que podrían presentar un sesgo a la luz de los aspectos que ya hemos señalado, situación que podría propiciar análisis cuantitativos e interpretaciones de estos que no están relacionadas con la realidad histórica y contemporánea de la lengua española.

1.3. Análisis cuantitativo y normalización de frecuencias de corpus

Desde una perspectiva cuantitativa, en un estudio lingüístico basado en corpus, se consideran las frecuencias de aparición de un fenómeno estudiado; con base en ello, se suelen sustentar o refutar hipótesis considerando ya la mayor o menor frecuencia de un fenómeno respecto a otros, ya el aumento o decremento de esta. Sin embargo, las frecuencias absolutas obtenidas de particiones con tamaños desiguales no son equiparables, incluso se corre el riesgo de obtener resultados sesgados al comparar un hecho lingüístico en recuentos de países o años con distinto número de palabras.

Para aclarar las debilidades de este método se propone una inspección sobre la distribución de la frecuencia de la palabra del español 'rayuela', en este caso, para señalar la concentración de esta unidad léxica en la obra homónima de Julio Cortázar. Pensemos, por ejemplo, que nos propusiéramos formalizar un estudio diacrónico en el que describiéramos los usos y frecuencias de la palabra 'rayuela' en todo el mundo hispánico. Naturalmente, para hacer este estudio recuperaríamos ejemplos del CORDE, y luego de hacer una búsqueda simple, obtendríamos 103 ocurrencias para la palabra 'rayuela', registradas entre el año 1490 y el año 1974.

Con base en un análisis cualitativo general de los datos observaríamos que 'rayuela' remite a dos realidades: el diminutivo de 'raya', usado entre los años 1490 (primera aparición) y 1625 (aproximadamente), uso que no se registraría después de esta fecha; y el juego infantil ya por todos conocido, que se registra desde el año 1817 hasta 1963, además de un ejemplo aislado de 1974 en el que se usa como diminutivo.

Si centráramos nuestra atención exclusivamente en los datos que remiten al juego infantil veríamos que la frecuencia de la palabra 'rayuela', para estos últimos 200 años considerados en el CORDE, es regular, lo que nos permitiría concluir, desde una perspectiva cuantitativa, que 'rayuela' es una palabra de uso estable en la lengua, información que se corrobora cualitativamente en el Diccionario de la Real Academia Española (en sus diferentes ediciones), al observar que el lema 'rayuela' nunca ha tenido alguna marca cronológica de obsolescencia o neologicidad[4].

No obstante esta generalidad, existe, desde el punto de vista cuantitativo, un momento histórico en el corpus en que la frecuencia de la palabra sale de una tendencia regular: el año 1963, en el que dicha unidad léxica aparece 36 veces o el 41% del total de ocurrencias registradas en el CORDE entre los años 1817 y 1971, 36 ocurrencias que remiten a un autor y a un mismo documento: Julio Cortázar y su conocidísima novela *Rayuela*. Este ejemplo ilustra por qué se desaconseja este método de cuantificación para este corpus y otros en general. El problema que subyace a un análisis como este es que la temática o el estilo de ciertos autores podrían condicionar la aparición de algunas formas lingüísticas o giros y la ausencia de otros. Por tanto, se corre el riesgo de que las conclusiones que se pudieran derivar de un análisis similar describan un comportamiento del fenómeno estudiado ajeno a la realidad de la lengua.

Una alternativa que subsana las limitaciones de un análisis como este, es calcular la 'frecuencia relativa', método descrito por Muller (1973) y generalmente aceptado para el tratamiento de los datos de corpus lingüísticos de gran tamaño. La frecuencia relativa de un fenómeno lingüístico (f_i) se obtiene calculando el cociente de la frecuencia absoluta de este hecho en la muestra (n_i) y el tamaño de esta (N), procedimiento que permite reasignarle a un número de ocurrencias un valor en relación con el tamaño de la partición o del mismo corpus. La fórmula utilizada para calcular la frecuencia relativa es la siguiente:

$$f_i = n_i / N$$

Si retomamos el caso de la palabra 'rayuela' e intentáramos aplicar el método descrito por Muller (1973) para calcular la frecuencia relativa, necesitaríamos conocer el tamaño de la muestra en el CORDE o el CREA, información que se encuentra disponible en el mismo corpus y que se puede recuperar de manera relativamente simple en el apartado "Nómina de autores y obras". Ya que recuperamos la información del tamaño muestral y la frecuencia absoluta de rayuela para cada uno de los periodos de tiempo analizados, hacemos las operaciones pertinentes (cocientes) y así obtenemos la frecuencia relativa de rayuela.

Además, repetimos el procedimiento para otra forma léxica del español, la palabra 'país', y centramos nuestro análisis en un periodo específico (decenio comprendido entre los años 1955 a 1964). La idea de comparar una palabra más en este análisis se justifica, ya que este procedimiento serviría para tener un punto más de comparación de los resultados.

Como se verá, calcular la frecuencia relativa de las palabras, para el caso del CORDE y el CREA, no es un método conveniente, ya que esta alternativa de normalización de frecuencias funciona parcialmente para la palabra 'país', pero no resuelve el caso de 'rayuela', en el que la temática y la extensión de la novela de Julio Cortázar siguen sesgando los resultados en el decenio representado. En el Gráfico 2 se representan ambas frecuencias relativas de aparición de ambas formas léxicas en el CORDE para el decenio analizado y se puede dar cuenta de esto.

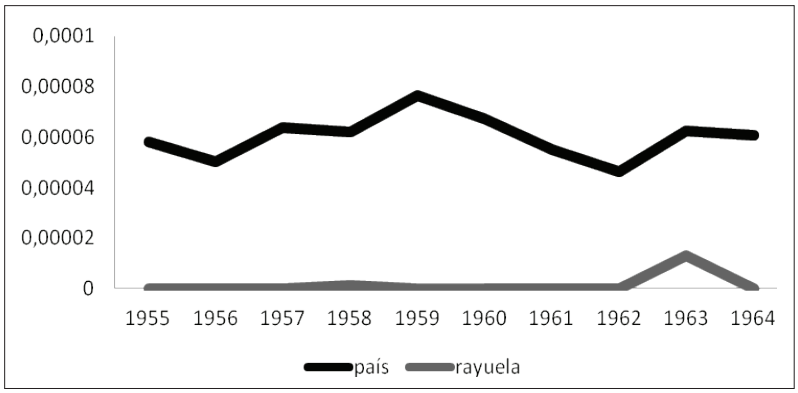


Grafico 2. Frecuencia relativa en el decenio de 1955 a 1964 de las palabras 'país' y 'rayuela'.

Una variante de la 'frecuencia relativa' es la 'normalización de frecuencias por millón de palabras' (Biber, Conrad & Reppen, 1998) que naturalmente se expresa en tantos por millón y, aparentemente, no aporta más al procedimiento originalmente descrito por Muller (1973). Pese a esto, este método es ampliamente usado en lingüística de corpus y con él se han normalizado las frecuencias del Corpus del Español de Mark Davies, *The British National Corpus via the Internet* (BNC)*, Corpus of Historical American English* (COHA)*, Corpus of Contemporary American English* (COCA)*, O Corpus do Português* de Mark Davies, entre otros.

La frecuencia por millón de palabras (*fpmw*, por sus siglas en inglés) se calcula dividiendo la frecuencia absoluta del hecho o fenómeno lingüístico cuantificado (*nt*) entre el número total de palabras en el corpus o recogidas en la partición estudiada de este (*nw*), resultado que después se multiplica por 1 millón. La fórmula descrita por Biber et al. (1998) es la siguiente:

$$fpmw = (nt \,/\, nw) \times 1{,}000{,}000$$

Otra alternativa de normalización a la 'frecuencia relativa' y la 'normalización de frecuencias por millón de palabras' es el 'índice normalizado de dispersión' (Ham, 1979), aplicado a la normalización de frecuencias del Corpus del Español Mexicano Contemporáneo (CEMC). Este ofrece una medida que da cuenta de la frecuencia de un vocablo, su dispersión entre géneros y el tamaño relativo de cada uno de ellos. Se calcula a partir de la 'frecuencia corregida' (*Korrigierte Frequenz (KF)*), propuesta original de Jan Lanke, según reporta Ham (1979), que sirve para subsanar diferencias ocasionadas por ponderaciones de ciertos tipos documentales frente a otros, dentro de un corpus. El 'índice normalizado de dispersión' (C_i) considera el tamaño relativo del género en el que aparece el vocablo (r_i) y un índice de dispersión de este entre géneros (S_i). La fórmula descrita por Ham (1979) es la siguiente:

$$C_i = (100\, S_i - min_i\, r_i) \,/\, (100 - min_i\, r_i)$$

En dónde la dispersión entre géneros (S_i) es el resultado del cociente de la 'frecuencia corregida' (KF_i) y la frecuencia del vocablo (T_i):

$$S_i = KF_i / t_i$$

Lamentablemente para los usuarios del CORDE o el CREA, el diseño de la interfaz no permite recuperar la información necesaria para la normalización mediante procedimientos relativamente sencillos como la 'frecuencia relativa' o la 'frecuencia por millón de palabras' (Muller, 1973; Biber et al., 1998), así como con el 'índice normalizado de dispersión', propuesto por Ham (1979), o cualquier otro método que se pretenda adoptar para ello.

Además de esto, se tiene la fuerte sospecha de que la información recuperada a partir de la 'Nómina de autores y obras' no necesariamente describe la totalidad del conjunto documental, ya que podría ser el caso de que en este apartado se dé cuenta de las obras con autoría reconocida y nunca las obras de autor anónimo; es decir, no existe forma alguna de saber con precisión el número de palabras que tienen los corpus en periodos de tiempo específicos, excepto cuando en estos no haya obras anónimas.

La imposibilidad de conocer con certeza el número de palabras que hay en el CORDE y el CREA o particiones específicas de estos convierte cualquier método de normalización de frecuencias en procedimientos poco confiables, por ello, en el apartado siguiente proponemos una alternativa de normalización de frecuencias que, en un principio, no requiere el conocimiento del número de palabras que hay en los corpus ni en particiones específicas.

2. Normalización de frecuencias del CORDE y CREA[5]

Considerando lo que se ha planteado hasta el momento, a saber: que la unidad de muestra de los corpus académicos son documentos completos con tamaños tan dispares que los hacen estadísticamente incomparables entre sí, aunado al hecho de no poder conocer con certeza el número de palabras recogidas por periodos de tiempo, género, zona geográfica, entre otras; parecería irremediable reconocer que la base de datos documental de la Real Academia Española serviría, únicamente, como base para trabajos con orientación cualitativa, útil, desde luego, para documentar ejemplos de uso.

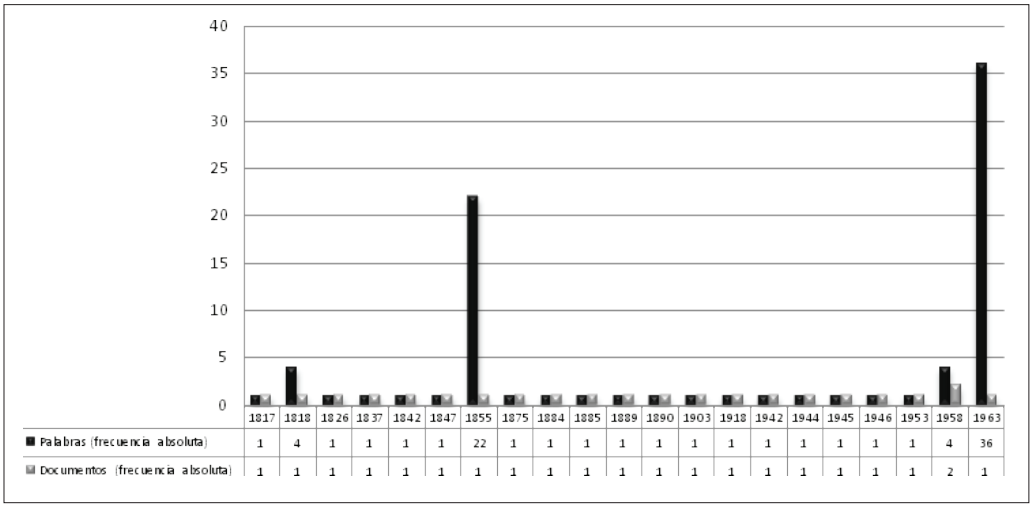
Asumir esta postura sería inadecuada, ya que los datos lingüísticos recogidos y clasificados en el recurso académico, luego de una apropiada normalización de frecuencias, sirven como base confiable para la descripción del fenómeno lingüístico que cualquier analista intentase describir.

El primer punto que debiera subsanar cualquier procedimiento de normalización de frecuencias de los corpus académicos es situar en un plano comparable documentos

de gran extensión, novelas por ejemplo, contra notas informativas de publicaciones periódicas, epístolas breves o cualquier otro documento de extensión reducida; ya que partiendo de la información ofrecida por el motor de búsqueda de ambos corpus es imposible tener certeza sobre el número de palabras contenidas en un documento, periodo de tiempo, zona geográfica, género textual o cualquier otra variable.

Esta situación es la que hace virtualmente imposible poder aplicar, desde la perspectiva de un usuario corriente, cualquier método de normalización de frecuencias descrito en la literatura sobre corpus lingüísticos, por ello, se plantea cuantificar la frecuencia de documento. Una decisión como esta, por arbitraria que parezca, se justifica desde el punto de vista de cuantificar el uso de un informante de una forma o giro lingüístico y no el número de veces que este lo utiliza en un documento.

Regresando al ejemplo de la palabra 'rayuela' referido anteriormente, más específicamente al año 1963: lo que se cuantificaría con este procedimiento es que un hablante utilizaba la forma 'rayuela' y no que la utilizó 63 veces, mismas que quedaron recogidas en una sola obra. Cuantificar el número de documentos en los que aparece 'rayuela' reduce notablemente el condicionamiento señalado, resultado que se puede comparar en el Gráfico 3 que se muestra a continuación:



	1817	1818	1826	1837	1842	1847	1855	1875	1884	1885	1889	1890	1903	1918	1942	1944	1945	1946	1953	1958	1963
Palabras (frecuencia absoluta)	1	4	1	1	1	1	22	1	1	1	1	1	1	1	1	1	1	1	1	4	36
Documentos (frecuencia absoluta)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1

Grafico 3. Frecuencia absoluta de aparición de la palabra rayuela (juego infantil) por año vs. aparición por documento en el CORDE.

Como se puede apreciar en el Gráfico 3, centrar un análisis en la frecuencia de aparición por documentos permite ver el fenómeno desde otra perspectiva: la que explicaría que 'rayuela', el juego infantil, se usa regularmente desde principios del siglo XIX sin aumentos o decrementos en su frecuencia. Hecho que aquí se representa gráficamente, pero que se puede corroborar desde la quinta edición del Diccionario de la lengua castellana (1803) y subsecuentes, en las que esta unidad léxica aparece

con el significado de juego infantil, uso que se mantiene, incluso, hasta la edición más reciente del diccionario (DRAE, 22ª ed.).

Aunque el caso de normalización de frecuencias para 'rayuela' resulta en un principio muy afortunado, ya que al analizar los datos se podría tener la sensación de que este procedimiento solo es suficiente, la normalización de frecuencias todavía requiere algunos pasos adicionales. Revisaremos, por tanto, el otro ejemplo señalado anteriormente: la palabra 'país'.

Si cuantificamos las ocurrencias por documento de 'país', entre el periodo comprendido entre 1900 y 1975, y representamos los resultados, se verá que la frecuencia de uso de la palabra 'país' no es homogénea para este periodo de tiempo. Como se puede ver en el Gráfico 4, la frecuencia de esta en las décadas de los veinte y treinta aumenta, posteriormente decae, para, al final de la representación gráfica, en la década de los setenta, aumentar notablemente.

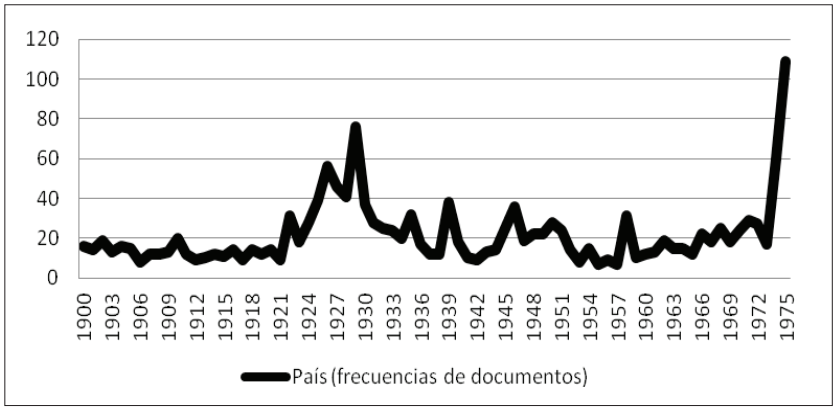


Gráfico 4. Frecuencia absoluta de documentos de 'país' en el siglo XX, en el CORDE.

Para los corpus académicos, cuantificar la aparición de un hecho lingüístico por documentos subsana parcialmente el sesgo de su representatividad asociada con la unidad de muestra, por ello, en la gran mayoría de los casos son necesarios algunos pasos adicionales, no considerados en la literatura, para la normalización de frecuencias de corpus.

La normalización de frecuencias que aquí se propone, basada en cuantificar la frecuencia documental, requiere como siguiente paso inmediato calcular las medias móviles (también conocido como promedios móviles) para cada uno de los cortes temporales estudiados. Este procedimiento, descrito por Murphy (1999) y Suárez (2001), entre otros, que resulta muy usual en el análisis técnico de los mercados financieros, recoge un número de valores que se promedian y, a partir de este cálculo, se le reasigna un nuevo valor a cortes temporales específicos. Con este método se suele representar más efectivamente la dirección y duración de una tendencia suavizando algunos picos accidentales, como los que hemos señalado en el Gráfico 4.

El procedimiento en sí es conceptualmente sencillo de entender. Regresando de nuevo al ejemplo de 'país', el cálculo de las medias móviles requiere, primero, determinar un grupo de valores, podrían ser 3, 5 o más, dependiendo del número de datos totales analizados (en el caso de 'país', debido a que el análisis contempla 76 cortes temporales (1900-1975), se decidió calcular las medias móviles de periodos de 5 años). El paso siguiente es promediar los valores de los elementos incluidos en cada uno de los grupos, por mencionar un simple ejemplo: sumaríamos el valor (*v)* correspondiente a los años comprendidos entre 1920 y 1924, resultado que luego se dividiría entre cinco; posteriormente, el valor de este promedio se le asigna al corte temporal intermedio, correspondiente al valor de la media móvil del año 1922 (*ma* de 1922).[6] El procedimiento descrito hasta aquí sería el siguiente:

ma de 1922 = (*v* de 1920 + *v* de 1921 + *v* de 1922 + *v* de 1923 + *v* de 1924) / 5 = 31

Al final, se debería repetir el mismo procedimiento para todos los años del CORDE en el siglo XX, hasta obtener todas las medias móviles restantes.

Aplicando este método al análisis de las frecuencias de la palabra 'país', en el periodo de tiempo comprendido entre los años 1900 y 1975, se obtiene una línea de tendencia suavizada que podría representar su uso en la lengua escrita recogida en la selección documental del CORDE. En el Gráfico 5, presentado a continuación, se puede ver el resultado obtenido luego de aplicar este procedimiento, además de comparar los resultados con la representación de las frecuencias por documentos, sin normalización de frecuencias.

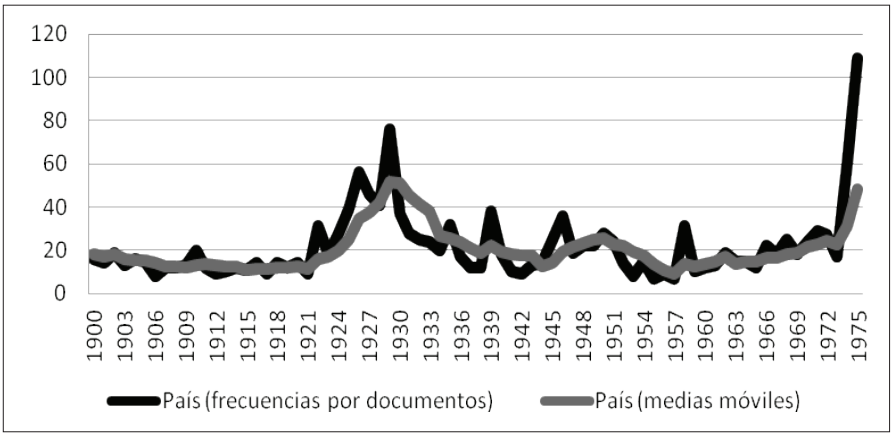


Gráfico 5. Frecuencia de 'país' por documentos vs. valores de promedios móviles.

El cálculo de las medias móviles atenúa las diferencias entre cortes temporales y dibuja una tendencia, pero de ninguna manera normaliza las frecuencias registradas para cada uno de los cortes temporales. Por ello, en este punto es necesario otro paso adicional para este método de normalización de frecuencias: recuperar un valor que refleje el número máximo de documentos posibles que hay en el corpus, para este

mismo periodo de tiempo, tomando en cuenta los mismos cortes temporales que se han considerado para el análisis de la unidad léxica 'país'.

Aunque esta información se pudiera recuperar en la 'Nómina de autores y obras' de los corpus, lo desaconsejamos porque, como ya hemos señalado, al parecer no se reflejan las obras anónimas. Ponemos como ejemplo de ello este caso particular: la 'Nómina de autores y obras', en el año 1993, reporta la existencia de 423 documentos, mientras que en este mismo año, el motor de búsqueda principal señala que la palabra 'de' aparece en 588 documentos. Esta situación imposible es ocasionada porque existen documentos, como los de las agencias noticiosas, que no tienen un autor reconocido y son etiquetados como anónimos o 'sin autor'.

Dicho lo anterior, no es conveniente recuperar el número de documentos que hay en un periodo estudiado desde la 'Nómina de autores y obras', ya que esto pudiera generar imprecisiones basadas en un valor asignado a un fenómeno lingüístico que no refleja la realidad del corpus y, mucho menos, de la lengua que representa, sino que la alternativa propuesta es recuperar, por medio del buscador de ambos corpus, la frecuencia de aparición en documentos de la palabra más frecuente en estos conjuntos textuales.

En este caso, se encuentran disponibles en línea[7] cuatro listas de frecuencias de las unidades léxicas recogidas en el CREA, a saber: '1.000 formas más frecuentes', '5.000 formas más frecuentes', '10.000 formas más frecuentes' y 'Lista total de frecuencias'. La información que aportan estas listas indica que la preposición 'de' es la forma más frecuente del español en el CREA. Entonces, considerando que la palabra más frecuente del CREA es la preposición 'de' y asumiendo que también podría serlo para el CORDE, el siguiente paso sería plantear un método de comparación efectiva entre esta palabra y la unidad léxica estudiada (en este caso: 'país').

Este procedimiento es una alternativa diferente de normalización de frecuencias, respecto a los trabajos existentes, en la que está implícita una comparación entre la aparición de otro hecho lingüístico: otra unidad léxica regular y altamente frecuente en el corpus, y no una normalización de frecuencias basada en el tamaño del léxico recogido en él.

Con base en lo anterior, el siguiente paso para la normalización de frecuencias sería cuantificar la aparición de 'de' para cada uno de los cortes temporales considerados en el análisis de 'país' y, luego, calcular las medias móviles para estos mismos periodos de tiempo; así tendríamos dos tendencias comparables, la de 'país' y la de 'de'. En este punto, no deberíamos perder de vista que las tendencias de las que hablamos, los incrementos y decrementos de estas, pudieran estar condicionadas por el número de documentos que hay registrados en cada año. Razón por la cual es necesario un procedimiento más para la normalización de frecuencias.

Este paso de la normalización de frecuencias considera hacer comparables distintos años del corpus para, así, poder contrastar efectivamente las líneas de tendencia de

'país' y 'de', ejemplo que hemos venido ilustrando. Este procedimiento se justifica ya que, aparentemente, el año 1997, en el CREA, recoge un total de 26.280 documentos en los que aparece la preposición 'de', mientras que en el año 1975, tan solo 296 documentos.

Para poner en una misma escala selecciones documentales tan disímiles es necesario calcular un valor logarítmico para cada uno de los valores anuales registrados correspondientes a la forma léxica 'de' y 'país'. Convertir a valor logarítmico es un procedimiento usual que pone dentro de un rango más manejable los datos analizados que cubren una amplia gama de valores, como en estos casos. En este modelo de normalización de frecuencias se propone usar un 'valor logarítmico natural' o 'logaritmo neperiano' (ln *x*), base logarítmica que utiliza la 'constante matemática *e*' para su cálculo (Maor, 2006).

Posteriormente, es preciso restarle al nuevo valor asignado, 'valor logarítmico natural', a 'país', el respectivo valor recalculado de 'de', para cada periodo de tiempo. La diferencia, que siempre será menor a cero, representará la distancia que existe entre la forma 'país' y el fenómeno más frecuente o máxima posibilidad de aparición de este dentro del corpus, en este caso los documentos en los que aparece 'de', en los cortes temporales señalados.

Grosso modo, este valor ubica dentro de un continuo o ranking, que va de lo más frecuente en el corpus a lo menos, la posición que ocupa 'país', respecto a 'de' en cada uno de los cortes temporales. Con este procedimiento se puede tener una referencia del máximo posible de apariciones de una palabra dentro de un corpus y con esa medida asignar un valor máximo para cada periodo estudiado.

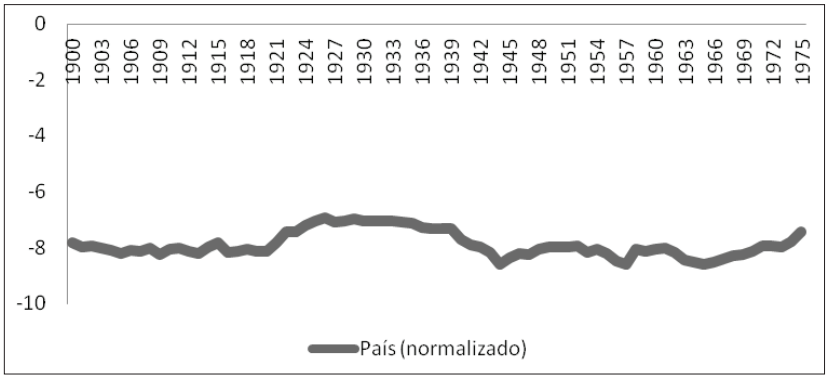
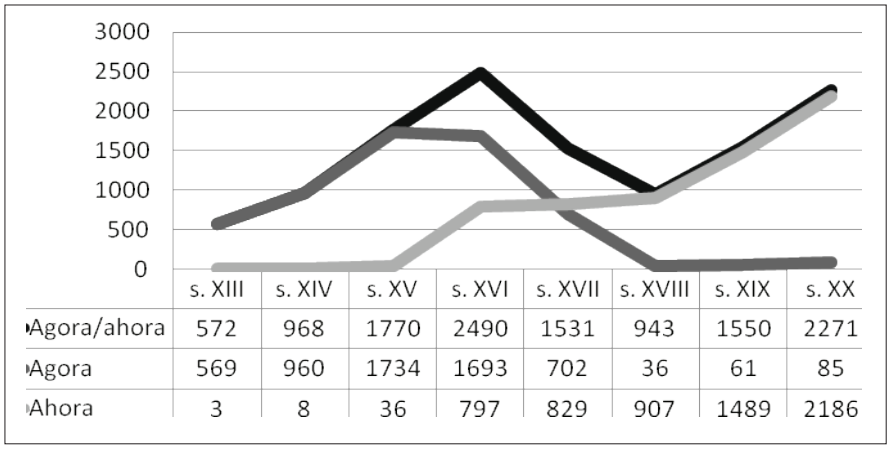


Gráfico 6. Frecuencia de aparición normalizada de 'país' en el CORDE, en el siglo xx.

El Gráfico 6 es un suavizado mediante medias móviles del logaritmo de la frecuencia absoluta de documento que coloca la frecuencia de 'de' como una línea recta equivalente a 'cero' y que representa la posibilidad máxima de aparición de cualquier fenómeno lingüístico en el corpus para cada corte de tiempo analizado.

Esta propuesta de representación nos permite corroborar lo que es por todos bien sabido, que 'país' es una forma de uso estable en la lengua para estos periodos de tiempo; al mismo tiempo que descalificamos la idea de que en las décadas de los veinte, treinta y setenta hubiera un factor que condicionara el uso de esta. Además de probar que el método de normalización de frecuencias propuesto resulta efectivo en los ejemplos, en donde fallan los otros procedimientos señalados.

Existe también la posibilidad de hacer análisis de periodos de tiempo más grandes y utilizar, al mismo tiempo, el CORDE y el CREA. Tal como se muestra en el Gráfico 7, en el que se contrasta el uso de 'ahora' y 'agora' en un periodo comprendido entre los siglos XIII y XX, haciendo cortes temporales para cada siglo y centrando la cuantificación del fenómeno en el recuento de los documentos en los que aparecen ambas formas. En este gráfico se muestran evidencias de uso de la unidad léxica del español 'agora', reconocida generalmente como un arcaísmo léxico, y se destaca el proceso de implantación del cambio fonético que lo transformaría en la forma actual 'ahora' (Sánchez, 1995).



	s. XIII	s. XIV	s. XV	s. XVI	s. XVII	s. XVIII	s. XIX	s. XX
Agora/ahora	572	968	1770	2490	1531	943	1550	2271
Agora	569	960	1734	1693	702	36	61	85
Ahora	3	8	36	797	829	907	1489	2186

Gráfico 7. Frecuencia de aparición, por documentos, del arcaísmo 'agora' y la palabra 'ahora' en el CORDE, por siglos.

Las líneas representadas en este gráfico sirven para constatar el hecho conocido de que 'agora' es un arcaísmo léxico que cayó en desuso entre los siglos XVI y XVII, momento a partir del cual 'ahora' se difundió en el español general y relegó a esta forma léxica a usos exclusivamente literarios.

No obstante este hecho pudiera ser cierto y quizás difícilmente refutable, cuando menos basándonos en este gráfico un análisis más detallado de ella nos orillaría a concluir erróneamente que aunque 'agora' y 'ahora' son formas patrimoniales del español en uso desde el siglo XIII, 'agora' se difunde en el español durante las dos primeras centurias representadas en el gráfico, hasta que en los siglos XV y XVI se observa cierta estabilidad en su uso que después de este tiempo decaería; mientras

que el uso de 'ahora' ha venido creciendo a partir del siglo XVI hasta nuestros días e incluso continúa en crecimiento.

En este caso, la diferencia entre el número de documentos recogidos en los cortes temporales planteados es la que acentúa estas diferencias. La evidencia es que la suma de las frecuencias de estas dos palabras (línea en color negro) refleja aumentos y disminuciones conforme transcurren los años representados, aunque en esta cuantificación nos basamos en la frecuencia de documentos.

En este punto de la normalización de frecuencias es necesario, primero, darle un valor a las ocurrencias cuantificadas respecto al tamaño de las muestras para cada corte temporal y, luego, una alternativa de representación de los datos que subsane estas diferencias. Los pasos seguidos en la normalización de frecuencias son los mismos que en el ejemplo de 'país', a saber, primero, calcular las medias móviles de 'agora' y 'ahora', para los siglos representados. Veamos, en el Gráfico 8, cómo se suavizan las líneas de tendencia después de este procedimiento.

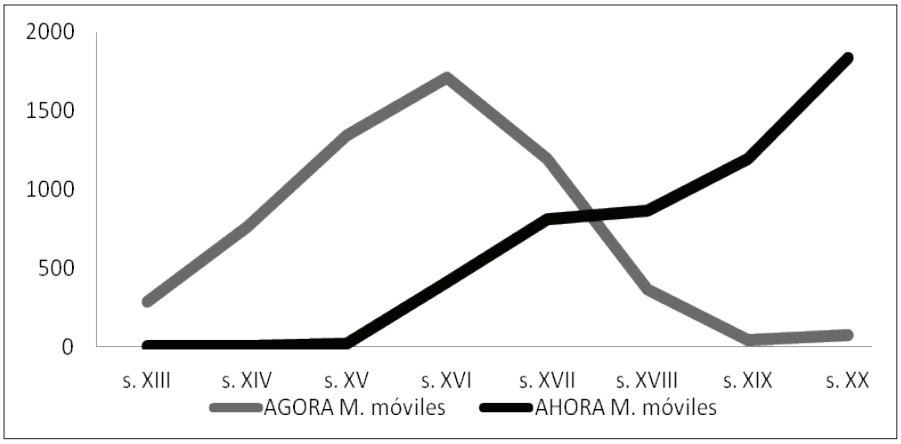


Gráfico 8. Representación de las medias móviles de 'agora' y 'ahora' en el CORDE y CREA, por siglos.

El paso inmediato sería cuantificar la aparición de la palabra de mayor frecuencia en el corpus, en este caso sabemos que es 'de', para los mismos periodos considerados para 'agora' y 'ahora', y con base en ello calcular las medias móviles de las frecuencias absolutas basadas en documentos.

Teniendo los promedios móviles de las frecuencias de documentos 'ahora', 'agora' y 'de', el siguiente paso es convertir estos valores a una escala logarítmica y, al final, es necesario restarle a 'ahora' y 'agora' el valor correspondiente de 'de', para cada periodo de tiempo. Con ello, los resultados deberán ser menores a cero. El resultado de todo este proceso, para el ejemplo de 'ahora' vs. 'agora', se muestra en el Gráfico 9.

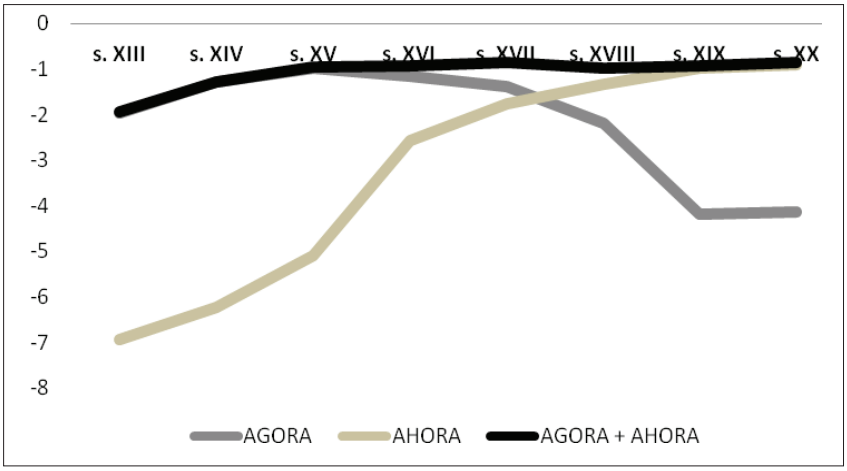


Gráfico 9. Frecuencia de aparición normalizada de 'agora' vs. 'ahora' en el CORDE, por siglos.

A partir de esta representación y del método que la antecede, se puede concluir fácilmente que 'agora' es una forma usual en el español desde el siglo XIII hasta el XVII, época en la que se observa una caída en su uso. En el periodo comprendido entre los siglos XVII y XX se observa un decremento de su frecuencia que no implica el desuso total de la forma, ya que este arcaísmo léxico es, a su vez, un cultismo de uso frecuente en la literatura. Hecho que se concluye a partir de la revisión del artículo lexicográfico correspondiente a esta forma léxica incluida en la 22ª edición del DRAE, en la que se señala que es una forma de uso exclusiva de la poesía y se le asigna la marca estilística: 'poét.'. Por su parte, la frecuencia de 'ahora' se incrementa hasta que, a partir del s. XVII, su frecuencia de uso parece más estable.

En el Gráfico 9, representamos la suma de las frecuencias de estas dos unidades léxicas, en todo caso, complementarias, y en ella se refleja cierta estabilidad en el uso de ambas formas durante las centurias representadas, situación que no ocurría en el Gráfico 8. Lo anterior demuestra que el método de normalización de frecuencias es efectivo aún si se toman ejemplos de ambos corpus y se estudian periodos de tiempo considerables.

Una objeción que se pudiera plantear a todo este procedimiento es el hecho de representar el comportamiento de uso o tendencia dentro de la lengua de ciertos hechos lingüísticos en valores menores a cero. Hecho que en realidad no entraña gran dificultad conceptual, sin embargo en algunos casos puede resultar conveniente aplicar un procedimiento extra para convertir los valores negativos en positivos. Este procedimiento consiste simplemente en sumar a cada uno de los valores representados en la gráfica un valor superior al representado en la escala para finalmente dividir entre 10. El resultado de este procedimiento sería el siguiente:

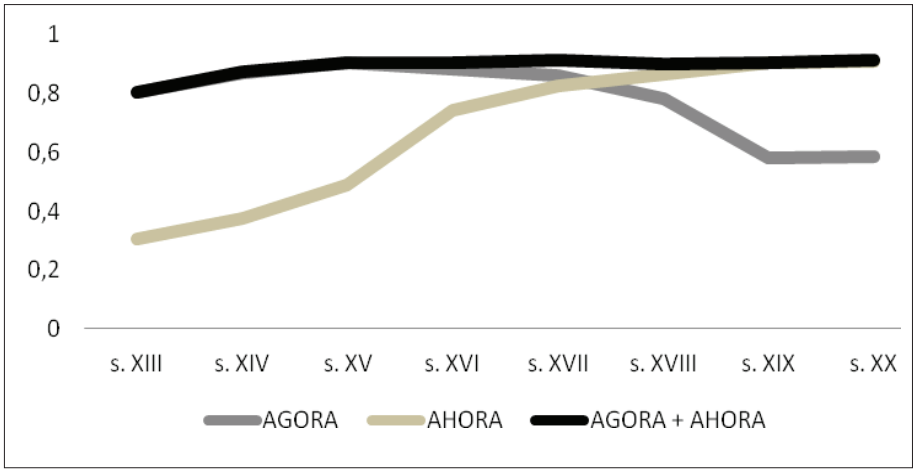


Gráfico 10. Frecuencia de aparición normalizada de 'agora' vs. 'ahora' en el CORDE, por siglos, con valores positivos.

Como se puede ver, el trazado de las líneas del Gráfico 9 y el Gráfico 10 son iguales, la única diferencia es la escala de cada una de ellas y que 'de' ahora sería el número 1 y no el cero.

Por último, este mismo procedimiento se podría aplicar para medir distintos fenómenos de variación de la lengua en los corpus. El método sería exactamente el mismo, la única diferencia radica en la selección de los datos, por ejemplo: si interesara contrastar los distintos procesos de implantación de las formas, en su momento neológicas, 'teléfono celular' y 'teléfono móvil', en México y España durante el mismo periodo de tiempo, tendríamos que fijar un filtro en el corpus para cada zona geográfica y hacer dos análisis.

De la misma forma, se podrían formalizar estudios en los que se contrasten fenómenos para distintos autores, medios de publicación (incluso, en el CREA existe la posibilidad de contrastar la producción oral vs. la escrita), temáticas y, como ya señalamos, diferencias geográficas. Todo ello dependerá de los fines particulares que persiga el analista de la manera en que se plantee recuperar las ocurrencias en ambos corpus.

CONCLUSIONES

En los estudios lingüísticos basados en corpus, sobre todo aquellos en los que se consideran grandes periodos de tiempo, suelen hacerse cortes temporales que permiten comparar momentos históricos de una lengua, además de cuantificar ocurrencias de cierto fenómeno y manejar más fácilmente la información recuperada.

Específicamente para los estudios de la lengua española, el CORDE y CREA son útiles, pero debido a su representatividad, tamaño y unidad de muestra, si se consideraran las ocurrencias recuperadas en cualquiera de los dos conjuntos documentales sin una debida normalización de frecuencias, pudieran formularse interpretaciones equivocadas sobre un hecho lingüístico estudiado.

El proceso de normalización de frecuencias que aquí se propone, tal cual se ha descrito, no pretende manipular los datos para sustituir la realidad lingüística del español, sino que es una alternativa que permite la comparación de hechos lingüísticos que se encuentran representados desproporcionadamente en los corpus académicos.

El método es útil para cualquier estudio de morfología, lexicología, sintaxis y análisis del discurso en el que se consideren, cuantifiquen y contrasten diferencias diacrónicas, diatópicas, diastráticas, diafásicas o diatécnicas. Este procedimiento no debe incidir en las decisiones metodológicas sobre determinar cortes temporales, restringir el fenómeno estudiado a ciertas temáticas, usos geográficos o medios de publicación específicos, sino que en sí mismo plantea otra vía de cuantificación de los datos del CORDE y CREA, considerados de forma aislada o conjunta.

Los pasos propuestos para esta normalización de frecuencias son los siguientes, en este orden: primero, considerar la aparición de cualquier fenómeno lingüístico que se estudie en la base de datos por documento (*D*). Después, es preciso calcular las medias móviles para los periodos de tiempo estudiados (ma_D), con ello lo que se estudiaría es una tendencia de uso. Luego, se calcula el valor logarítmico natural de estas medias móviles ($\ln ma_D$), para darle al fenómeno estudiado un valor comparable independientemente del alto o bajo número de documentos que haya en cada periodo. En este punto es indispensable repetir este mismo procedimiento para las ocurrencias de la forma léxica más frecuente en ambos corpus, a saber, la preposición "de"; con ello obtendremos el número virtual de documentos por periodo de tiempo estudiado (*N*). Del que, luego, calcularemos las medias móviles y su valor logarítmico natural ($\ln ma_N$).

A partir de esto, se resta al valor del fenómeno que interese estudiar el valor de 'de', con la intención de obtener una serie de valores menores a cero que representan la frecuencia normalizada (*fn*). La fórmula que representaría esta serie de procedimientos es la siguiente:

$$fn = \ln ma_D - \ln ma_N$$

Adicionalmente a esto, se podrían representar los valores con cualquier otro método que resulte conveniente para un más fácil entendimiento de los datos, nosotros proponemos esta simple fórmula para la representación de los valores positivos (vp), pero pudiera seguirse cualquier otro procedimiento:

$$vp = (fn+10)/10$$

Por último, conviene señalar que en casos en los que se comparen dos o más hechos lingüísticos, conviene graficar también la suma de ellos, así se tendría una perspectiva global de la difusión o estabilidad de todo el fenómeno estudiado, incluso, podría adoptarse la suma de estos como el número virtual de documentos por periodo de tiempo estudiado *(N)*.

REFERENCIAS BIBLIOGRÁFICAS

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.

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

CLARIN (2010). *Virtual collections* [en línea]. Disponible en: http://www.clarin.eu/sites/default/files/virtual_collections-CLARIN-ShortGuide.pdf.

Ham, R. (1979). Del 1 al 100 en lexicografía. En L. F Lara, R. Ham & I. García (Eds.), *Investigaciones lingüísticas en lexicografía* (pp. 43-83). México: El Colegio de México.

Instituto Cervantes (2012). *El español: Una lengua viva* [en línea]. Disponible en: http://eldiae.es/wp-content/uploads/2012/07/2012_el_espanol_en_el_mundo.pdf

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German reference corpus DeReKo: A primordial sample for linguistic research. En N. Calzolari (Ed.), *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)* (pp. 1848-1854). Malta: European Language Resources Association (ELRA).

Maor, E. (2006). *Historia de un número*. México: Libraria Ediciones.

Murphy, J. (1999). *Technical analysis of the financial markets*. Nueva York: Institute of Finance.

Muller, Ch. (1973). *Estadística lingüística*. Madrid: Gredos.

Real Academia Española (1803). *Diccionario de la lengua castellana compuesto por la Real Academia Española, reducido a un tomo para su más fácil uso* (Quarta edición). Madrid: Viuda de Ibarra.

Real Academia Española (2014). *Banco de datos (CORDE), Corpus diacrónico del español* [en línea]. Disponible en: http://www.rae.es.

Real Academia Española (2014). *Banco de datos (CORPES XXI), Corpus del Español del Siglo XXI* [en línea]. Disponible en: http://www.rae.es/recursos/banco-de-datos/corpes-xxi.

Real Academia Española (2014). *Banco de datos (CREA) [en línea], Corpus de referencia del español actual* [en línea]. Disponible en: http://www.rae.es.

Real Academia Española (2014). *Manual de consulta del banco de datos del español* [en línea]. Disponible en: http://www.rae.es.

Real Academia Española (2014). *Nuevo Tesoro Lexicográfico de la Lengua Española* [en línea]. Disponible en: http://buscon.rae.es/ntlle/SrvltGUILoginNtlle.

Sánchez, M. (1995). Observaciones sobre el arcaísmo lingüístico de los textos aljamiado-moriscos. *Sharq Al-Andalus,* 12, 339-348.

Sierra, G. (2008). Diseño de corpus textuales para fines lingüísticos. En Z. Estrada & A. Munguia (Eds), *Memorias del IX Encuentro de Lingüística en el Noroeste* (tomo II) (pp. 445-462). Hermosillo, México: Universidad de Sonora.

Sinclair, J. (1996). *Preliminary recommendations on corpus t polog y* [en línea]. Disponible en: http://www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpustyp.ps.gz

Suárez, *M. (2001). Interaprendizaje de Estadística Básica.* Ecuador: Editoral Fausto Ibarra.

Torruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. En J. M. Blecua (Ed.), *Filología e informática: Nuevas tecnologías en los estudios filológicos* (pp. 45-77). Barcelona: Milenio-UAB.

NOTAS

1 Este apartado se redactó a partir de la información proporcionada en el *Manual de consulta del banco de datos del español* que se encuentra disponible en <http://corpus.rae.es/ayuda_c.htm>, en el que se explica la arquitectura y cómo se ha modelado este banco de datos. Hacemos esta aclaración, ya que a nuestro modo de ver podría ser incómodo citar repetidamente el mismo documento en sus diferentes apartados (que en realidad resulta muy breve y accesible), sin agregar más información relevante que el sitio en el que se pueden encontrar estas informaciones.

2 El Instituto Cervantes publica en el año de 2012 un reporte sobre el español llamado: *El español: una lengua viva*. En este caso, los datos referidos están basados en este reporte.

3 Conviene aclarar que un 'corpus virtual' no es una 'colección virtual', puesto que se considera que una 'colección virtual' es "an aggregation of various data resources that serves a certain research purpose and that covers resources from various repositories most probably generated by different researchers and teams" (CLARIN, 2010: 1). En este sentido, un 'subcorpus' creado a partir del CREA y el CORDE serían un 'corpus virtual', ya que ambos corpus forman parte del mismo recurso, provienen del mismo repositorio y están creados por el mismo grupo de investigación; además de que comparten la misma codificación (anotación y etiquetado), condición que no ocurre en la documentación que conforma 'colecciones virtuales' (CLARIN, 2010).

4 Véase el *Nuevo tesoro lexicográfico de la lengua española*, disponible en línea en: http://buscon.rae.es/ntlle/SrvltGUILoginNtlle, para esta y otras consultas lexicográficas de este tipo.

5 El presente trabajo se centra exclusivamente en los ejemplos recogidos en el CORDE y el CREA y se descartan otros corpus disponibles como el CDH, el CORPES XXI o el Corpus del Español de Mark Davies, primero, porque el CDH es un corpus con orientación lexicográfica que no sirve como referencia para otros niveles de la lengua, por ello se descarta; el CORPES XXI, tal y como se define, es una versión provisional en la que no figuran todavía transcripciones de textos orales y en el que se registra solo una pequeña parte de los documentos que lo conformarán, mismos que día con día se va actualizando (RAE, 2013), razón por la cual se ha planteado retomar la normalización de frecuencias para este corpus en trabajos posteriores; y por último, se descarta el Corpus del Español de Mark Davies de este estudio, ya que este corpus cuenta con un método de normalización de frecuencias por millón.

6 La abreviatura *ma* es la forma convencional utilizada para referir el cálculo de una media móvil, abreviatura proveniente del término inglés: moving average.

7 Información disponible en: http://corpus.rae.es/lfrecuencias.html