La localización de refranes en corpus lingüísticos digitales: métodos y aplicaciones

Vito Antonio MICCOLI
Universidad Complutense de Madrid (España)
vmiccoli@ucm.es || https://orcid.org/0009-0008-0661-4325

Recibido: 26/03/2025 || Aceptado: 20/04/2025

Resumen: El presente estudio pretende avanzar una propuesta metodológica para la búsqueda de refranes en corpus lingüísticos digitales en español. Para ello, se han diseñado cuatro métodos de búsqueda, destinados a localizar dichos enunciados sentenciosos en gigacorpus que contienen mil millones de tokens, como los de la familia TenTen. Estas técnicas se han elaborado a partir de criterios internos y externos, considerando tanto la estructura bimembre como la extensión variable de los refranes. En este sentido, se ha realizado un análisis empírico y cuantitativo que ha permitido la elaboración de un corpus anotado del que proceden los ejemplos utilizados en este trabajo. La metodología aplicada permite investigar sobre cómo actúan estas paremias en la actualidad y observar los varios procesos de variación y cambio a los que están sometidas. Asimismo, esta metodología puede resultar de gran utilidad en ámbitos del saber cómo la traducción, en la medida en que ofrece un método de consulta fiable para localizar refranes y determinar elecciones traductológicas.
Palabras clave: Paremiología. Refrán. Corpus lingüístico. Español.

Titre : **« Méthodes pour localiser des proverbes dans des corpus linguistiques ».**
Résumé : Cette étude vise à avancer une proposition méthodologique pour la recherche de proverbes dans des corpus linguistiques numériques en espagnol. À cet effet, quatre méthodes de recherche ont été élaborées, visant à localiser ces énoncés sentencieux dans des gigacorpus contenant un milliard de tokens, tels que ceux de la famille TenTen. Ces techniques ont été développées à partir de critères internes et externes, en prenant en compte à la fois la structure bimembre et la longueur variable des proverbes. Dans ce sens, une analyse empirique et quantitative a été réalisée, permettant la création d'un corpus annoté dont proviennent les exemples utilisés dans ce travail. La méthodologie appliquée permet d'étudier ces paremies et d'observer les différents processus de variation et de changement auxquels elles sont soumises. De même, cette méthodologie peut s'avérer très utile dans des domaines du savoir tels que la traduction, dans la mesure où elle offre un moyen de consultation fiable pour localiser des proverbes et déterminer des choix traductologiques.
Mots-clés : Parémiologie. Proverbe. Corpus linguistique. Espagnol.

Title: **"Methods for the localization of proverbs in linguistic corpora."**
Abstract: This study aims to advance a methodological proposal for locating proverbs in digital linguistic corpora in Spanish. To this end, four search methods have been developed, focused on localizing these paremias in gigacorpora containing billions of tokens, such as those in the TenTen family. These techniques were created based on both internal and external criteria, considering the bimember structure and the variable length of proverbs. In this regard, an empirical and quantitative analysis has been conducted, leading to the creation of an annotated corpus from which the examples used in this work are derived. The applied methodology allows for the investigation of how these paremias function today and observing the various processes of variation and change they undergo. Similarly, this methodology may prove very useful in fields of knowledge such as translation, insofar as it offers a reliable consultation method for identifying proverbs and determining translation choices.
Keywords: Paremiology. Proverb. Linguistic corpus. Spanish.

INTRODUCCIÓN

En tiempos recientes, hemos sido testigos del creciente interés y de la necesidad, por parte de paremiólogos, de localizar e investigar con mayor precisión las estructuras polilexemáticas de la

lengua mediante el empleo de análisis y métodos empíricos y cuantitativos. Por ello, el presente trabajo pretende avanzar una propuesta metodológica para la búsqueda de paremias en corpus lingüísticos digitales. En particular, el estudio se circunscribe a una única categoría parémica: el refrán. La localización eficaz de estos enunciados sentenciosos facilita su análisis desde múltiples enfoques de interés, que solo un entorno digital permite. Mediante esta propuesta, se busca promover una comprensión actualizada y reconsiderada del refrán, acorde con las posibilidades que ofrece el tratamiento digital del dato lingüístico.

Indudablemente, la lingüística computacional ha resultado ser un enfoque vanguardista para la paremiología y la fraseología, reflejando una convergencia entre estas disciplinas cada vez más frecuente (Mellado, 2016: 70). Asensio y Sciutto (2024: 190) respaldan esta perspectiva, considerando el uso de corpus como un soporte digital fundamental para el estudio y análisis de las paremias desde diversas perspectivas. Asimismo, Zurdo y Sevilla (2016: 18)[1] subrayan las ventajas que se atesoran al tener fácil acceso a bancos de datos en soporte electrónico cuyas posibilidades de documentación pueden superar cuantitativa y cualitativamente los resultados obtenidos mediante procesos de trabajos de campo tradicionales. En concreto, la contemporánea aparición de los *gigacorpus,* junto con las eficientes herramientas de procesamiento en las que se apoyan, evidencian el verdadero avance cuantitativo del análisis paremiológico frente a la limitada disponibilidad de resultados, muchos de los cuales, derivados de intuiciones personales y originados a raíz de consultas en corpus tradicionales, elaborados de manera manual y fundamentados en las teorías de la década de 1990 (Corpas y Gutiérrez, 2023: 2).

Para fortalecer la postura sobre el uso provechoso que brinda el análisis de las paremias en recursos computarizados intervienen los rasgos irrefutables que definen el concepto mismo de corpus. La definición de corpus a la que nos acogemos es la de Rojo (2014: 371), según la cual:

> Un corpus es un conjunto de (fragmentos de) textos, orales o escritos, producidos en condiciones naturales, conjuntamente representativos de una lengua o una variedad lingüística, en su totalidad o en alguno(s) de sus componentes, que se almacenan en formato electrónico y se codifican con la intención de que puedan ser analizados científicamente.

Los textos integrados en un corpus tienen que ser manifestaciones reales de la comunicación en contexto; por lo tanto, no han sido concebidos con el objetivo de destacar un determinado fenómeno lingüístico. Además, son representativos de una lengua o de sus variedades, es decir, resulta imperativo que un corpus exhiba una visión adecuada de la lengua que pretende representar. Como consecuencia de estos dos rasgos definitorios, existe otra característica inmediata y consecuente: el equilibrio. Según lo plantea Rojo (2021: 1-2, 49), un corpus supondrá cierto grado de equilibrio de los textos que lo integran, esto es, un número suficiente de fragmentos adscribibles a diferentes tipos y géneros textuales, sean estos escritos u orales, de ficción, de corte periodístico o teatral, de textos mexicanos o procedentes de variantes peninsulares, etc. Estos tres rasgos resultan esenciales para el análisis de estas unidades fijas y sentenciosas, en la medida en que contribuyen a que los corpus reflejen la auténtica manifestación del uso por parte del hablante. Adicionalmente, estudios en lingüística computacional han evidenciado cómo los corpus web proporcionan una visión interesante de la lengua general y, dependiendo del tipo de investigación y de las herramientas empleadas,

[1] Cabe destacar que, según las autoras, cuyas consideraciones seguimos, los métodos clásicos de recopilación paremiográfica continúan siendo no solo pertinentes sino también indispensables y complementarios al tipo de documentación proveída por los corpus digitales. Esto se debe a la capacidad única de las encuestas tradicionales «para establecer el término medio del conocimiento y uso de paremias entre los hablantes y por su utilidad para descubrir la incidencia de los factores diatópicos, diastráticos, diafásicos o estilísticos en las respuestas» (Sevilla y Zurdo, 2016: 18).

incluso una visión de los usos lingüísticos tan válida y detallada como la proporcionada por un corpus de referencia tradicional (Kilgarriff y Renau, 2013: 14).

Establecidas las premisas teóricas sobre las cuales se rige el presente trabajo, se han delineado cuatro métodos de búsqueda que hemos perfilado y destinado a la localización de refranes. Se ha llevado a cabo un estudio empírico y cuantitativo que nos ha permitido elaborar un corpus anotado en el cual se detallan métodos de búsquedas, resultados cuantitativos y, en su caso, cualitativos, observaciones preliminares sobre cada paremia que ha formado parte de nuestra investigación, etc. De acuerdo con la metodología empleada, hemos formulado tres hipótesis: a) de los cuatro métodos de búsqueda propuestos, se plantea que uno de ellos se perfila como el más eficiente para proporcionar resultados fiables en términos cuantitativos; b) otros métodos, por su parte, presentan la ventaja de generar abundantes datos relevantes desde un enfoque cualitativo; y, c) considerados en conjunto, los distintos métodos son complementarios y proporcionan una visión global y detallada de cómo actúan los refranes en los corpus digitales, reflejo de su uso en la sociedad actual. Las evidencias que avalan nuestras hipótesis, así como los resultados obtenidos, se presentarán con mayor detalle a lo largo del presente estudio, respaldados por ejemplos extraídos del corpus anotado.

1. PRESUPUESTOS TEÓRICO Y METODOLÓGICOS

Atendiendo a los objetivos planteados, proponemos una metodología para la búsqueda de refranes en corpus lingüísticos digitales. La finalidad es la localización eficaz de estas estructuras sentenciosas de la lengua y, por ende, permitir su investigación desde varios enfoques de interés.

1.1. Criterios de selección del corpus

Para este trabajo, hemos optado por dedicarnos exclusivamente a desarrollar métodos de búsqueda para los refranes, excluyendo por el momento las demás clases parémicas[2]. El motivo principal que nos ha llevado a tomar esta decisión radica en las características constitutivas propias de los refranes, de las cuales suelen carecer las otras clases de enunciados sentenciosos. Concretamente, para la elaboración del cuarto método de búsqueda nos beneficiamos de la estructura bimembre que generalmente presentan los refranes. En lo que concierne a su definición, el refrán se entiende como una paremia de origen por lo general anónimo y uso predominantemente popular, cuya estructura manifiesta una fuerte tendencia al bimembrismo (Sevilla y Crida, 2013: 111; Cantera, Sevilla y Sevilla, 2005: 12; Crida y Sevilla, 2015: 70; Anscombre, 1999: 25-26). Sumado a ello, el alto grado de frecuencia de uso que se otorga a los refranes, *a priori,* permitió pronosticar su presencia en el corpus consultado.

A la luz de estos argumentos, conocer y haber delimitado el objeto de estudio no exime de las dificultades que supone generar un análisis riguroso. En primer lugar, elegir qué refranes investigar y cuáles no, supuso una serie de limitaciones iniciales. Adicionalmente, nos enfrentamos a otro interrogante, esto es, el número satisfactorio y adecuado de refranes que deben analizarse para poder comprobar de manera cuantitativa un fenómeno. Con el fin de prevenir estas dos primeras restricciones, hemos acudido al *Mínimo paremiológico español* (Sevilla y Barbadillo, 2021). Nuestra decisión de extraer y analizar los refranes de este corpus radica en las características propias de esta obra de referencia y su aplicabilidad práctica a los más amplios ámbitos de la lingüística como la didáctica, la traducción y la interpretación, entre

[2] «La tipología de las paremias españolas queda reducida así a seis categorías (proverbios, aforismos, refranes, frases proverbiales, locuciones proverbiales y dialogismos), las cuales a su vez se pueden dividir en subtipos, según diferentes criterios (temáticos, formales, genéticos)» (Sevilla y Crida, 2013: 113)

otros (Zurdo y Sevilla, 2016). Dado su carácter transversal, hemos considerado válido extender su uso también al ámbito computacional, convirtiéndose de tal modo en nuestro punto de partida.

Ahora bien, el *Mínimo paremiológico español* recoge enunciados sentenciosos pertenecientes a las seis categorías parémicas citadas en la nota 2. Por ello, hemos seleccionado solo las formas que las dos autoras identifican como refranes (morales, meteorológicos, supersticiosos, dialogados, etc.). De esta manera, las paremias objeto de nuestra investigación han pasado de 298 a 206. Consideramos el volumen de datos extraídos adecuado para ejecutar un análisis investigativo empírico y cuantitativo veraz. Otro argumento a favor de nuestra decisión de analizar todos los refranes del *Mínimo paremiológico español* sería el hecho de que interviene la representabilidad que le corresponde a esta colección de paremias. De hecho, estos refranes no solo son representativos de cómo una comunidad lingüístico-cultural ha entendido y clasificado el mundo real y sus modelos tradicionales y sociales, sino también del hecho de ser reflejo de las temáticas y de «las tendencias más relevantes que se observan en la actividad comunicativa en el uso de los enunciados fraseológicos por parte de los hablantes» (Zurdo y Sevilla, 2016: 28-41).

1.2. La delineación de métodos de búsqueda de refranes en corpus digitales

Para el presente análisis, hemos trabajado con el corpus esTenTen23, uno de los últimos que integra la familia de los *gigacorpus* TenTen; está disponible mediante la herramienta de procesamiento web Sketch Engine. Para realizar el análisis de las paremias seleccionadas, no se ha restringido la búsqueda a ningún dominio web específico ni a un subcorpus determinado. Esto ha permitido preservar la representabilidad del español en todas sus variedades peninsulares y americanas (Kilgarriff y Renau, 2013: 13-14), en un corpus que cuenta con aproximadamente 33,1 mil millones de *tokens.* El motor de búsqueda usado es el *avanzado* mediante *Concordancia*, ajustando el tipo de consulta (*query type*) a *phrase.* Según lo plantea Gutiérrez (2024; 2022), esta configuración del motor de búsqueda permite localizar combinaciones de palabras idénticas a las introducidas en el buscador. En este sentido, hemos reiterado el mismo proceso para cada uno de los cuatro métodos de búsqueda diseñados que siempre consisten en el uso de una secuencia de palabras, más o menos variable, y que no consta de signos gráficos, como una coma o un punto y coma, que a menudo suelen usarse para enfatizar y dividir los miembros de una paremia.

Los cuatro métodos de búsqueda se diferencian en la extensión de las secuencias de palabras utilizadas, las cuales varían según la longitud del refrán o en función de la estructura que este posee. Al mismo tiempo, también se distinguen porque, por un lado, se inspiran en las investigaciones previas, y por otro, en la estructura frecuentemente bimembre del refrán.

El primer método de búsqueda (m1) es el más intuitivo considerando que representa la forma prototípica de localizar signos lingüísticos en un corpus digital, es decir, introducir la paremia completa en el motor de búsqueda y esperar a que el sistema proporcione los resultados. Por otro lado, para la delineación del segundo y el tercer método de búsqueda y recuperación de refranes en corpus digitales, seguimos las indicaciones de Mieder (en Sevilla, 2024: 24-25). Según lo plantea el autor, es imprescindible que las nuevas investigaciones experimentales no pasen por alto la extensa y excelente bibliografía existente. En este sentido, nos hacemos eco de las técnicas de recopilación paremiográfica en entornos naturales elaboradas por Julia Sevilla (2016). La paremióloga aplica diversas técnicas para cumplir con este objetivo, entre las cuales se encuentra crear encuestas en las que «la forma de los refranes puede estar completa, incompleta con solo la primera parte para aportar la segunda o con los refranes fragmentados en dos partes y en columnas distintas para que el informante las enlace» (Sevilla, 2016: 236). Para nuestro análisis, el segundo método de búsqueda (m2) coincidirá con la secuencia de palabras que corresponde a la primera parte del refrán y el tercero (m3) con la secuencia que corresponde

a la segunda parte. Estos dos diseños pueden interpretarse como el proceso inverso de las encuestas paremiográficas directas a hablantes, en las que se interroga al corpus para recuperar respuestas previamente registradas en él.

En último término, el cuarto método de búsqueda (m4) es el que para nosotros representa un aporte novedoso en este ámbito. Gutiérrez Rubio (2024: 47-48) nos anticipa que una localización eficaz de las paremias en los corpus digitales se obtiene buscando «combinaciones de tres palabras consecutivas invariables e impropias *a priori* del discurso libre» y usa como ejemplo «su san Martín» para localizar la famosa paremia *A cada cerdo le llega su san Martín*. Otro ejemplo que el autor presenta es «se pica ajos» para *Quien se pica ajos come* (Gutiérrez, 2022: 30). Siguiendo esta idea, nuestro cuarto método de búsqueda corresponde a una secuencia de palabras extraídas del discurso repetido, la cual deriva del alto grado de fijación de estos enunciados sentenciosos. Para su elaboración, establecemos dos criterios esenciales: en primer lugar, la secuencia diseñada se apoyará explícitamente en la estructura bimembre del refrán; en segundo lugar, el criterio de longitud de la secuencia será flexible. Dicha flexibilidad es consecuencia directa de la extensión del refrán u otro factor que interviene en el proceso. En otras palabras, la localización de los registros se realiza mediante una secuencia de palabras, que hemos decidido nombrar *secuencia central parémica*: esta siempre respeta la estructura bimembre del refrán integrando elementos consecutivos de sus dos mitades. A modo de ejemplo, para la búsqueda del refrán *Camarón que se duerme / se lo lleva la corriente* se ha usado la secuencia «duerme se lo lleva» o «casado casa» para *El casado / casa quiere*. Como muestran los ejemplos, la secuencia varía en número de palabras según la longitud del refrán y/o de los fenómenos que se quieren investigar. En concreto, conviene distinguir entre las secuencias diseñadas para refranes cortos y aquellos de mayor extensión. Por ejemplo, para paremias cortas como *A rey muerto / rey puesto*, hemos empleado una secuencia más reducida, resultado de la brevedad de la estructura original: «muerto rey». En cambio, en el caso de refranes largos, como *Ante el vicio de pedir / está la virtud de no dar* ha sido necesario perfilar secuencias más largas de manera que la localización fuera igualmente eficaz: «de pedir está la virtud». Se trata de un método que adapta longitud y estructura de la secuencia de búsqueda según la naturaleza del refrán, asegurando que se respeten sus elementos centrales.

1.3. Estrategias de filtrado y validación de datos

Cada método de búsqueda presenta sus ventajas, así como sus límites dependiendo de varios factores, internos y externos. Entre ellos, destacan el refrán mismo y su estructura, su frecuencia de uso en el corpus, el fenómeno lingüístico que se quiere investigar. No obstante, surge otra problemática ya señalada por varios autores, la copiosa presencia de variabilidad de las formas. Según lo plantea Corpas (2014: 134), los corpus reflejan empíricamente una baja frecuencia de uso de estas unidades fijas, mientras que destaca la alta frecuencia de estructuras inestables derivadas de ellas. En este contexto de inestabilidad, proliferan las variantes (ocasionales y no), modificaciones o desviaciones cuya finalidad es el efecto generalmente lúdico, llamativo, expresivo. Otro principal desafío radica, pues, en determinar el tipo de dato que ha de considerarse fiable y en cómo reducir la presencia de dichos resultados inapropiados. Al respecto, algunos autores como Rubio (2024: 49) aluden a la noción de *ruido*, esto es, la aparición de datos irrelevantes que interfieren en la precisión y en la fiabilidad del análisis. Otros, como Pamies y Pazos (2023: 141), alertan sobre la presencia en los corpus de formas duplicadas, resultado de fuentes que se plagian entre sí; pues, son criterios de selección y exclusión de los datos no siempre homogéneos. En el marco de nuestra investigación, se emplea la terminología *fragmentos incidentales* (FI). Nos referimos a aquellos registros que, si bien emergen a raíz de nuestra búsqueda, no son pertinentes para nuestro análisis. Al ser resultados contextuales, tal vez por otro tipo de acercamiento sean relevantes.

Para poder delimitar nuestros FI, partimos del *Refranero multilingüe* (Sevilla y Zurdo, 2009). Empleamos como forma de referencia[3] el refrán que inicia la ficha en cuestión. No se consideran FI ni las variantes documentadas en el proprio refranero, sean estas diatópicas o menos, ni las variantes estructurales (Corpas, 1996: 28) que presentan cambio de número o género de sus constituyentes, o distinta conjugación verbal respecto a la forma canónica[4], siempre que tales modificaciones no alteren la organización interna del refrán ni su sentido original. De ahí que tampoco se clasifiquen FI las variantes diafásicas ni aquellas en las que se ha detectado un descuido ortográfico producido por el hablante[5]. Asimismo, no se incluyen como FI las formas que presentan la ruptura de la estructura del refrán por la inserción de un presentador paremiológico (Corpas, 1996: 137) como en *A Dios rogando, dice el refrán, y con el mazo dando.*

Por otro lado, sí se consideran FI: i) las hipotéticas variantes diatópicas o diastráticas no registradas en el refranero, como *A falta de pan buenas son tortillas* o *A falta de pan galletas* (19 ocurrencias localizadas mediante m2); ii) las formas que presentan procesos de desautomatización (Mena Martínez, 2003) o manipulación creativa (Corpas, 1996: 233-234) de la forma canónica, como *A caballo regalado no le toques los cojones*; iii) las formas que presentan la inserción de comentarios propios del hablante, como *A río revuelto es fácil que haya ganancia de pescadores*; iv) refranes con alteración en el orden de sus miembros, como *pecho a lo hecho* o *buena cara a mal tiempo*; v) las formas truncadas[6], como *A buen entendedor…*; y, finalmente, vi) los fragmentos en los que el refrán pasa a otra categoría fraseológica: *Damián es afortunado en el juego y desafortunado en el amor* (*Afortunado en el juego, desgraciado en amores*).

Para reducir la cantidad de FI, hemos optado por revisiones manuales en aquellos casos en que el número de resultados era inferior a quinientos. Cuando los resultados superaban este umbral, se recurrió a la aplicación de filtros. Una de las grandes ventajas de Sketch Engine ha sido la posibilidad de manejar filtros que redujeran drásticamente la presencia de FI. Uno de los filtros que hemos usado permite conservar únicamente aquellos fragmentos textuales en los que, además de la secuencia buscada, aparece un ítem específico de interés. De forma ejemplificativa, presentamos datos parciales del segundo tipo de búsqueda del refrán *Cuando el río suena agua lleva*. Se ha usado la secuencia «cuando el río suena», que corresponde a la primera mitad de la estructura bimembre, obteniendo un total de 3095 fragmentos. Para eliminar

[3] «El criterio pragmático que rige el *Refranero multilingüe* lleva a buscar la forma más usual de la paremia hoy día o la más utilizada en los textos […] El resto de las formas ocupan su lugar entre las variantes» (Sevilla, 2012).

[4] Entendemos por forma canónica del refrán aquella que registra un uso predominante según los diccionarios actuales o, en el caso de formas caídas en desuso, la realización que ha evidenciado un desarrollo diacrónico más significativo (Klein, 2006: 163; 2017: 38).

[5] A este respecto, Rojo (2021: 265) señala que el problema está casi siempre en las vocales con tilde o sin ella y que es necesario tener en cuenta que en los corpus también puede haber faltas de ortografía en los fragmentos que proceden de actos de habla naturales. A esto se añade otra cuestión, esto es, la incorporación en estas bases de datos de nuevos géneros textuales vinculados a internet como blogs, tuits, wasap, etc. (Rojo, 2021: 184), lo que aumentaría la presencia de variación diafásica y descuidos ortográficos.

[6] Generalmente, los usuarios dan por sobreentendida la segunda parte o, en su competencia paremiológica, solo conocen la primera parte (Barbadillo y Sevilla, 2021: 237; Zuluaga, 1980: 10). Pese a esto, hemos decidido excluir estos fragmentos a causa del alto porcentaje de fenómenos observados de modificación de la forma de referencia. Por ello, como es imposible determinar cómo el hablante habría modulado la segunda parte o, incluso, si la conocía, hemos decidido excluir estas instancias de nuestro análisis.

los FI, empleamos el filtro avanzado *keep lines containing*, con *query type: lemma*[7] y *part of the speech: verb* configurado para el verbo en infinitivo *llevar* en el motor de búsqueda. Adicionalmente, se estableció el *range token* y una *distancia desde el kwic (incluido)* de +3. El resultado es la diminución drástica de 3095 fragmentos a 1090 (se conserva aproximadamente el 35% de las instancias). La aplicación de estos filtros no solo optimiza la precisión del análisis, sino que permite y facilita una revisión personal más eficiente, evitando la confusión que puede producir la revisión manual de miles de formas. Evidentemente los filtros que se usan, cómo se configuran y aplican dependen siempre de factores de interés, externos a la búsqueda.

Cabe subrayar que el número y el tipo de FI cambian de acuerdo con el método de búsqueda que se emplea. A modo explicativo, si se emplea el tercer tipo de búsqueda (segunda parte del refrán), no se recuperarán los FI en los que el elemento que genera irrelevancia se sitúa precisamente en esta parte. Los elementos que generan irrelevancia, si los hay, se encontrarán en la primera parte del refrán. Por ello, los resultados que se producen en cada tipo de método nunca serán homogéneo, tanto en términos cuantitativos como cualitativos.

Con el fin de demostrar la funcionalidad de la metodología descrita, aportamos ejemplos cualitativos de cinco refranes. Para los datos cuantitativos se ofrece la información en la Tabla 1. En esta se detalla el número de ocurrencias y el porcentaje de FI para cada método. Los refranes son: a) *A caballo regalado no le mires el diente*; b) *Camarón que se duerme se lo lleva la corriente*; c) *Donde hay patrón, no manda marinero*; d) *El que roba a un ladrón tiene cien años de perdón*[8]; y e) *Más vale pájaro en mano que ciento volando*[9].

Refrán	m1	m2	m3	m4	FI2%	FI3%	FI4%
a[10]	134	1238	225	1125	43,38%[11]	12,89%	43,56%
b	179	572	576	286	40,04%	39,80%	23,81%
c	167	315	666	227	30,79%	8,11%	5,73%
d	36	97	345	147	35,05%	13,91%	5,44%
e	384	974	774	604	35,42%[12]	24,94%	9,11%

Tabla 1: Ocurrencias y relativos porcentajes de FI para cada método de búsqueda.
Fuente: esTenTen23, The Sketch Engine. Elaboración propia.

[7] Para evitar la exclusión de variantes estructurales en las que aparece el verbo conjugado de manera distinta a la forma de referencia.

[8] Se ha observado que una gran parte de los refranes que incluyen términos numéricos suelen presentar variantes donde la cifra se expresa en números arábigos. Se consideran estas formas válidas, pues, constituyen variedad diafásica.

[9] Para este refrán, se han incluido dentro de las variantes estructurales la forma *Mejor pájaro en mano que ciento volando,* aunque no esté recogida en el *Refranero multilingüe*.

[10] En relación con este refrán, nuestra hipótesis, según la cual el m4 es el que menos FI proporciona, no se cumple. Esto se debe a que la secuencia central parémica utilizada consta únicamente de dos elementos «regalado no». Al pronosticar un resultado negativo, decidimos no abarcar otros elementos en la secuencia, ya que esto habría excluido refranes con variación en el complemento personal. Consideramos más interesante mantener el mayor número posible de variantes diafásicas. De hecho, al ampliar la secuencia a «regalado no le», los resultados obtenidos son los siguientes: ocurrencias 299 y FI 54 (el porcentaje ha bajado al 18,06% respecto al valor anterior).

[11] Filtro empleado: *keep lines containg, query type* lemma, *part of the speech* verb (mirar); *range* token, *distance from de kwic (incluido)* +5; el mismo filtro se ha usado por aFI4.

[12] Filtro empleado: *keep line containing, query type* lemma, *part of the speech* verb (volar); *range* token, *distance from de kwic (incluido)* +5.

2.LOS MÉTODOS DE BÚSQUEDA DE REFRANES EN CORPUS DIGITALES

2.1. Método de búsqueda 1

El primer método (m1) es el más intuitivo, representando la manera usual de buscar secuencias de signos lingüísticos en los corpus digitales. En este caso, se emplea la secuencia que corresponde al refrán entero, lo que conlleva una serie de desventajas. En primer lugar, al usar el refrán completo, se eliminan por completo los FI, así como toda variante, incluso aquellas que representan un interesante ítem de investigación. Se produce una depuración de todas las formas que varían de la forma original. Como consecuencia el número de ocurrencias que arroja el corpus será siempre menor. Otra cuestión que se añade deriva de la metodología adoptada: las búsquedas se realizan siempre sin signos ortográficos, los cuales suelen situarse entre las dos partes del refrán. Por ello, este tipo de búsqueda no detectará resultados como *Donde hay patrón, no manda marinero* por la sola presencia de una coma. Estos datos aparecerán con los demás métodos. En términos generales, es el método menos productivo, aunque podría resultar útil para estudiar el contexto y las finalidades con que se usa actualmente el refrán en su forma canónica.

2.2. Método de búsqueda 2

El segundo método (m2) es el que corresponde a la primera parte del refrán y brinda resultados interesantes. En primer lugar, es útil para localizar la paremia de referencia, así como sus variantes acertadas: *A caballo regalado no se le ve el colmillo*. Pese a esto, excluye aquellos refranes que presentan variabilidad en la primera parte del refrán, como *donde manda capitán no manda marinero* (recuperado con m3). Con la m2 solo aparecerán las formas encabezadas por *Donde manda patrón…*

Adicionalmente, hay una presencia constante de FI. En concreto, desviaciones o fenómenos de desautomatización, como *A caballo regalado no le toques los cojones*, *Camarón que se duerme amanece de coctel* o *Más vale pájaro en mano que no muerde que cien perros volando que ladran*; un alto número de paremias truncadas siendo la segunda parte del refrán sobreentendida o desconocida; hipotéticas variantes diatópicas o diastráticas: *el que roba a un ladrón tiene mil años de perdón, a caballo regalado no le mires la dentadura, a caballo regalado no le miras las muelas* (estas variantes con *mil, dentadura* y *muelas* no se registran en el refranero de referencia, aunque producto de relaciones sinonímicas).

Por lo general, es una metodología eficaz para la localización tanto de resultados relevantes como de FI, cuando estos sean de interés. Para nuestro análisis, no representa la mejor opción de búsqueda, debido a la alta frecuencia de resultados marginales. Además, un dato que conviene señalar consiste en que la búsqueda sistemática de un refrán mediante su primera parte puede llevar a la aparición de millones de formas y, por ende, dificultar la mitigación de FI. Si examinamos los casos del refrán meteorológico *Abril, aguas mil* y del refrán moral *Amigos, hasta el infierno*, los m2 en este caso serían respectivamente *Abril* y *Amigos*. Las secuencias, en estos casos, compuestas por un solo sustantivo[13], producen más de 4 millones de ocurrencias por ambas formas, sin contar que, en nuestro caso, aceptaríamos las variantes estructurales de número y genero de *Amigos*. Asimismo, esta problemática se produce con la m3. Por ejemplo, «pecho» de *A lo hecho, pecho* (743 681 ocurrencias mediante m3).

2.3. Método de búsqueda 3

[13] Lo mismo ocurre con sintagmas que pertenecen tanto al discurso repetido como a la técnica libre del discurso: «año nuevo» y «vida nueva» (270 826 y 25 240 ocurrencias respectivamente) del refrán moral *Año nuevo, vida nueva*. La secuencia central parémica «nuevo vida» produce 708 ocurrencias reduciendo drásticamente el volumen de FI.

Las observaciones que se han hecho sobre este método son similares a las del m2. Esto significa que es posible localizar tanto la forma de referencia como sus variantes acertadas cuando la variación ocurre en la primera parte del refrán. Asimismo, en el caso de los FI, se encuentran procesos de manipulación creativa, como *a suelo recalificado no le mires los dientes*, *Gobernador que se duerme se lo lleva la corriente*, *más vale un par de semanas de felicidad en mano, que ciento volando, donde hay madre no manda marinero, el que mata a un fascista tiene cien años de perdón*; hipotéticas variantes no registradas como *a burro regalado no le mires el diente*; inversión de los constituyentes, como *donde no manda marinero, manda capitán*; y, formas con comentarios intercalados como *Quien le roba a un ladrón (los empresarios) tiene cien años de perdón*, *más vale pájaro (o billete en mano), que ciento volando*. En términos generales, no aparecen formas truncadas con este método.

2.4. Método de búsqueda 4

El cuarto método de búsqueda de refranes en corpus digitales es, según hemos podido verificar, el más novedoso y fructífero en términos cuantitativos para la localización de resultados más fiables. Como se describe en el apartado 1.2, consiste en el uso de una secuencia consecutiva de palabra, que hemos denominado *secuencia central parémica*, y corresponde a la parte central del refrán, acogiendo elementos de ambas partes de su estructura bimembre.

Se trata de un método que presenta tanto ventajas como desventajas. Aunque se ha demostrado (tabla 1) que el número de ocurrencias que produce el corpus es de lo más fiable en la medida en que produce menos FI, parece oportuno señalar las deficiencias de este acercamiento. En primer lugar, se verán excluidas todas aquellas formas que presentan variabilidad en los elementos centrales. A modo de ejemplo, para el refrán *Más vale pájaro en mano que ciento volando* se ha utilizado la secuencia «en mano que ciento». Esto excluirá variantes como *Más vale pájaro en mano que buitre volando* o *Más vale pájaro en mano, que dos en el corral*, que presentan variación léxica precisamente en correspondencia de la secuencia central. Lo mismo ocurriría con la paremia *Donde hay patrón, no manda marinero* cuya secuencia central parémica sería «patrón no manda»; pues, se verían excluidas las variedades que albergan en su estructura el término *capitán* en posición de *patrón*.

En cuanto a la depuración de los registros incidentales, observamos que para el refrán *(e)* las ocurrencias son 604 con solo 55 FI[14] (9,11%). En términos cualitativos, es posible observar fenómenos distintos: i) hipotéticas variantes, como *Más vale paloma en mano que ciento volando* (3 ocurrencias); ii) procesos de desautomatización o desviaciones, como *más vale un Chardonnay en mano que ciento volando* o *más vale mundial en mano que ciento volando*; iii) formas adscribibles a otra categoría fraseológica *preferir pájaro en mano que ciento volando* (8 ocurrencias). Para el refrán *(c)* se pueden sacar las mismas conclusiones. Observamos 227 ocurrencias, de las cuales solo se excluyen 13 FI (5,73%). Además, se observan fenómenos como: i) hipotéticas variantes, como *Donde hay patrón no manda criado*; ii) procesos de desautomatización de la forma canónica, como *donde hay patrón no manda designer* o *donde hay patrón no manda ornitólogo*. Ulteriores ejemplos podrían ser los resultados del refrán *(a)*. Se ha elaborado la secuencia central parémica «regalado no», adaptándola a nuestras exigencias. Como se ha indicado anteriormente, la secuencia se puede adaptar al dato de interés que se quiere resaltar. En este caso, la finalidad era mantener aquellas formas que veían variabilidad en el pronombre personal. Por ello, tuvimos que limitar la secuencia a dos elementos léxicos. Se observa 1125 ocurrencias y se han identificado 490 FI (43,56%). Se trata de: i) fenómenos de modificación como *A vino regalado no le mires el embotellado, a camello regalado no se le mira la joroba*; y, ii) hipotéticas variantes resultado de las relaciones sinonímicas de los

[14] Excluyendo las variantes estructurales *Mejor pájaro en mano que ciento volando* se subiría a 127 FI (21,9%). Se verifica igualmente la hipótesis.

términos usados, como *a burro regalado no le mires el diente* (7 ocurrencias con burro), *a caballo regalado no le mires el bocado/el rabo/no se le mira la boca,* etc.

Pese a las ventajas mencionadas hasta ahora, este tipo de acercamiento puede generar algunas dudas. Efectivamente, para este método carecemos de criterios sobre la extensión en la creación de la secuencia central. No ha sido posible, por el momento, ni determinar un número específico fijo de palabras que deberían componerla, ni fijar un criterio de simetría de los términos abarcables de ambas mitades del refrán. En el primer caso, no sería lo mismo elaborar la secuencia con un número fijo de palabras, especialmente cuando se trabaja con refranes de distinta extensión. En el segundo caso, aplicar como criterio la simetría de los elementos podría excluir aquellos resultados interesantes, por el simple hecho de abarcar una palabra más en la secuencia. Reiteremos, como ejemplos, la paremia *A caballo regalado / no le mires el dentado.* Si se pone como hipotético criterio la simetría de los elementos centrales, la secuencia constaría de «caballo regalado no le», esto es, dos elementos de ambas mitades. El resultado sería la exclusión de las formas donde aparecen sinonimias actanciales como *burro* en lugar de *caballo*. Asimismo, se perderían las variantes diafásicas producto de la variación del pronombre personal. Pues, la utilidad de este método es justamente la posibilidad de adaptarse a factores internos y externos a la búsqueda como podría ser el ítem que se quiera destacar o priorizar en la localización de las formas.

Finalmente, se puede afirmar que también este método permite sacar conclusiones cualitativas, aunque en menor medida si se compara con m2 y m3. Por su parte, nos parece el método más apropiado para la localización de formas fiables, puesto que es el tipo de búsqueda que más FI excluye. Para obviar estas posibles limitaciones, se podría iniciar una búsqueda del refrán, complementando los diversos métodos o, para el último caso, cambiar la secuencia central parémica en función de las variantes que despiertan interés y las necesidades de estudio.

CONCLUSIONES.

El presente artículo es el resultado de un estudio inicial sobre cómo actúan algunos refranes en los corpus digitales del español, aunque para el presente estudio solo hemos trabajado con esTenTen23, alojado en Sketch Engine. A raíz de estas observaciones y después de haber manejado insistentemente este corpus, se han perfilado cuatro métodos para la búsqueda de refranes. Estas técnicas se han inspirado en distintos estudios científicos previos o en la estructura bimembre de esta clase parémica. Los resultados preliminares han permitido plantear una serie de hipótesis: a) el m4 podría ser el método que menos FI produce; b) los m2 y m3 favorecerían la aparición de abundantes datos irrelevantes y alejados de la forma canónica del refrán, pero de interés cualitativo; c) los métodos podrían complementarse de manera armoniosa y coordinada para proporcionar una visión panorámica e integral de los refranes en los corpus digitales. Con el objetivo de poder verificar estas hipótesis, hemos decidido extender esta metodología a una cantidad copiosa de refranes representativos del español y gracias a los datos positivos obtenidos, se ha conseguido desarrollar esta metodología. Asimismo, ha sido posible elaborar un método cuya finalidad fuera delimitar el número de fragmentos incidentales en favor de datos más fiables.

Cada tipo de búsqueda tiene sus ventajas y muestra sus propias limitaciones. En primer lugar, el m1 es el acercamiento más básico y, en consecuencia, produce los resultados menos interesantes y escasos en términos cualitativos y cuantitativos. Se trata de una búsqueda que no genera resultados relevantes en la medida en que solo proporciona una cifra muy pobre de instancias. Las funcionalidades mismas de los corpus digitales se reducen a la mera observación del contexto de uso de la forma de referencia de un refrán. Por su parte, los métodos 2 y 3 suelen presentar problemáticas comunes. Desde un punto de vista cuantitativo, son los métodos

que más datos producen y, por ende, aquellos que más FI presentarán, dificultando tanto la localización del elemento de interés como la depuración de las formas marginales. Esto nos ha obligado a la aplicación de filtros rigurosos cuando el control manual se veía imposibilitado por el inabarcable número de registros. Pese a esto, desde una perspectiva más bien cualitativa, tanto las formas aceptables como los FI resultan de gran interés para investigaciones sobre los fenómenos que se desarrollan a partir de un refrán original. Por último, el m4 no solo destaca por su aporte novedoso en este tipo de acercamiento a los refranes en corpus digitales, sino que también representa la búsqueda que menos FI produce en proporción simétrica con el número de ocurrencias que resulta. Este último método, está limitado por una serie de cuestiones como la imposibilidad de agrupar las formas que varían en la parte central. Sin embargo, esta es una imperfección que muestra cada método. Otra limitación es la falta de criterios que regulen la longitud de la secuencia a un número definido de signos gráficos consecutivos. La elaboración de esta secuencia puede depender de elementos externos a la búsqueda, esto es, del elemento que se quiere resaltar. Lo práctico y útil de este método está precisamente en la posibilidad de elaborar la secuencia en función del elemento que se quiere investigar. Esta adaptabilidad a las más variadas situaciones es el límite que plantean los otros métodos.

Adicionalmente, cabe destacar el gran potencial que tienen estas técnicas al ser usadas de forma complementaria. En efecto, los resultados fiables, así como los FI que surgen a raíz de los diferentes métodos de búsqueda no son homogéneos. Cada búsqueda produce datos distintos, lo que permite acceder a una visión más detallada e, incluso, panorámica del refrán y cómo este actúa en los corpus digitales. En este sentido, la metodología que se describe en este estudio y la consecuente posibilidad de acceder a una información tan abundante abren un abanico de aplicaciones en varios ámbitos del saber. En particular, su relevancia se hace evidente en disciplinas como la traducción donde la búsqueda de correspondencias de un enunciado sentencioso plantea numerosas dificultades. Contar con un número significativo de registros, obtenidos mediante una metodología encaminada hacia la localización de refranes, posibilitaría determinar elecciones traductológicas más próximas a la paremia de partida.

En última instancia, parece oportuno subrayar las razones que nos han llevado a adoptar un enfoque computacional. La importancia de un acercamiento digital no solo reside en su capacidad para abordar el refrán desde nuevas perspectivas, sino también en el potencial de generar un punto de partida para futuras investigaciones. A través de él, se favorece el almacenamiento de abundantes volúmenes de datos a los que, de otro modo, sería imposible acceder. Esta aproximación proporciona una comprensión nueva y enriquecida del refrán.

REFERENCIAS BIBLIOGRÁFICAS

Anscombre, J. (1999). Estructura métrica y función semántica de los refranes. *Paremia*, 8, 25-36. <https://cvc.cervantes.es/lengua/paremia/pdf/008/004_anscombre.pdf>

Asensio Ferreiro, M. D., Sciutto, V. (2024). Estudio teórico-metodológico y contrastivo de la competencia paremiológica en lengua materna y lengua extranjera en el ámbito universitario. *Romanica Olomucensia,* 36 (2), 187-205. <https://romanica.upol.cz/pdfs/rom/2024/02/01.pdf>

Cantera Ortiz de Urbina, J.; Sevilla Muñoz, J.; Sevilla Muñoz, M. (2005). *Refranes, otras paremias y fraseologismos en Don Quijote de la Mancha*. University of Vermont. <https://cvc.cervantes.es/lengua/biblioteca_fraseologica/documentos/cantera-sevilla-sevilla_refranes-otras-paremias-y-fraseologismos-en-don-quijote-de-la-mancha.pdf>

Corpas Pastor, G. (1996). *Manual de fraseología española*. Gredos.

Corpas Pastor, G. (2014). El fraseólogo internauta: cómo pasarlo pipa en la red. V. Durante (ed.), *Fraseología y paremiología: enfoques y aplicaciones*, 133-152. Instituto Cervantes.

Corpas Pastor, G., Gutiérrez Rubio, E. (2023). Computational and corpus phraseology applied to Spanish. *Romanica Olomucensia,* 35 (1), 1-8. <https://romanica.upol.cz/pdfs/rom/2023/01/02.pdf>

Crida Álvarez, C. A.; Sevilla Muñoz, J. (2015). La problemática terminológica en los estudios paremiológicos. *Anuari de Filología. Estudis de Lingüística,* 5, 67-77.

Gutiérrez Rubio, E. (2022). Frecuencia de uso de locuciones y paremias en el corpus Spanish Web 2018 (esTenTen18): implicaciones didácticas y lexicográficas. Corpas Pastor G. *et al*. (Eds.), *Proceedings of the International Conference EUROPHRAS 2022, Malaga, Spain, 28-30 September* (pp. 26-33). INCOMA. <https://www.acl-bg.org/proceedings/2022/EUROPHRAS%202022/pdf/2022.europhras-1.4.pdf>

Gutiérrez Rubio, E. (2024). Estudio de corpus de los presentadores paremiológicos en español. *Paremia,* 34, 45-54. https://cvc.cervantes.es/lengua/paremia/pdf/034/003_gutierrez.pdf

Kilgarriff, A.; Renau, I. (2013). EsTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences*, 95, 12-19. <https://www.sciencedirect.com/science/article/pii/S1877042813041372>

Klein, J. R. (2006). Problemas relacionados coa determinación da forma 'canónica' nunha base de datos de refráns franceses (DicAuPro). *Cadernos de fraseoloxía galega*, 8, 147-163. <https://www.cirp.gal/pub/docs/cfg/cfg08_07.pdf>

Klein, J. R. (2017). Portrait d'un dictionnaire automatique et philologique des proverbes français. *Paremia*, 26, 35-40. <https://cvc.cervantes.es/lengua/paremia/pdf/026/003_klein.pdf>

Mellado Blanco, C. (2016). La equivalencia textual español-alemán en corpus paralelos a través de la fraseología de la película «Todo sobre mi madre» de P. Almodóvar. E. Dal Maso y C. Navarro (Eds.), *Gutta cavat lapidem. Indagini fraseologiche e paremiologiche*. Universitas Studiorum.

Mena Martínez, F. (2003). En torno al concepto de desautomatización fraseológica: aspectos básicos. *Tonos Digital, 5*.

Pamies, A.; Pazos, J. M. (2023). Frasemas combinados «a pedir de boca». *Romanica Olomucensia*, 35 (1), 139-156.

Rojo, G. (2014). Hispanic Corpus Linguistics. En M. Lacorte (Ed.), *The Routledge Handbook of Hispanic Applied Linguistics* (pp. 371-387). Routledge.

Rojo, G. (2021). *Introducción a la lingüística de corpus en español* (1.ª ed.). Routledge.

Sevilla Muñoz, J. (2012). Aplicación de las TIC a la paremiología y la paremiografía: el Refranero multilingüe. González Rey, M. I. (ed.), *Unidades fraseológicas y TIC*, 51-62. Instituto Cervantes. <https://cvc.cervantes.es/lengua/biblioteca_fraseologica/n2_gonzalez/j_sevilla.htm>

Sevilla Muñoz, J. (2016). El refranero hoy. Técnicas paremiográficas. *Paremia,* 22, 235-242. <https://cvc.cervantes.es/lengua/paremia/pdf/025/020_sevilla.pdf>

Sevilla Muñoz, J.; Barbadillo, M.ª T. (2021). *El mínimo paremiológico español*. Centro Virtual Cervantes (Instituto Cervantes). Biblioteca fraseológica y paremiológica, serie «El mínimo paremiológico», n.º 2. <https://cvc.cervantes.es/lengua/biblioteca_fraseologica/m2_sevilla/el_minimo_paremiologico_espanol.pdf>

Sevilla Muñoz, J.; Crida Álvarez, C. A. (2013). Las paremias y su clasificación. *Paremia, 22*, 105-114. <https://cvc.cervantes.es/lengua/paremia/pdf/022/009_sevilla-crida.pdf>

Sevilla Muñoz, J.; Zurdo Ruiz-Ayúcar, M.ª I. T. [Dirs.] (2009). *Refranero multilingüe*. Instituto Cervantes (Centro Virtual Cervantes). <http://cvc.cervantes.es/lengua/refranero/>

Sevilla Muñoz, M. (2024). Wolfgang Mieder, precursor de la paremiología moderna. *Paremia*, 34, 11-28. <https://cvc.cervantes.es/lengua/paremia/pdf/034/001_sevilla.pdf>

Zurdo Ruiz-Ayúcar, M.ª I. T.; Sevilla Muñoz, J. (2016). *El mínimo paremiológico: aspectos teóricos y metodológicos*. Centro Virtual Cervantes (Instituto Cervantes). Biblioteca fraseológica y paremiológica, serie «El mínimo paremiológico», n.º 1. <https://cvc.cervantes.es/lengua/biblioteca_fraseologica/m1_zurdo/el_minimo_paremiologico.pdf>

Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas*. Peter Lang.

Corpus empleado

esTenTen23, The Sketch Engine. <http://www.sketchengine.co.uk>