

<https://doi.org/10.29162/pv.43.1.1176>

UDC 811.131.1'373.7:81'322=134.2

Original scientific paper

Received on 19 January 2026

Accepted for publication on 12 April 2026

VITO ANTONIO MICCOLI

ESTUDIO SOBRE LA FRECUENCIA DE USO DE LAS PAREMIAS: REFLEXIONES DESDE LA LINGÜÍSTICA DE CORPUS

Resumen: El presente trabajo analiza la aplicabilidad de los corpus como herramienta computacional capaz de medir la frecuencia de uso de las paremias. La noción de frecuencia de uso, de aparente simplicidad intuitiva, representa una variable compleja de sistematizar debido a las dificultades que su proceso de catalogación supone. Para ello, en este estudio se opta por combinar metodologías de corte tradicional con otras basadas en corpus. Ambas aproximaciones presentan aspectos metodológicos heterogéneos; no obstante, pese a su distinta naturaleza, los resultados muestran una notable convergencia en los valores frecuenciales proporcionados. Salvo un número limitado de excepciones, la triangulación de los datos revela una coherencia proporcional en la frecuencia de uso de las paremias en las dos formas de cribado. El acercamiento digital se convierte, así, en un método complementario para la extracción de información no solo cuantitativa, sino también cualitativa y contextual, lo que se traduce en una comprensión más amplia y no convencional del fenómeno proverbial.

Palabras clave: encuesta paremiográfica, frecuencia de uso, Lingüística de Corpus, paremias italianas, paremiografía digital

1. Introducción

El auge de los estudios paremiológicos en las últimas décadas ha estimulado una abundante y multidisciplinaria producción

científica internacional. Aun disponiendo de esa prolificidad, al colocar bajo un escrutinio crítico la vasta producción a disposición se advierte una serie de interrogantes que, de forma recurrente, comprometen la plena realización de las investigaciones en determinadas áreas de la paremiología contemporánea. Un eje de reflexión sistemáticamente palpable atañe a la noción de *frecuencia de uso* con la que se verbalizan las paremias. Este tema de fricción puede condensarse en tres preguntas nodales: i) ¿cuál es la frecuencia de uso actual de las paremias?, ii) ¿cómo precisar patrones fiables relativos a su índice de uso? y, iii) ¿qué herramientas resultan pertinentes para evaluar la significatividad y el grado de recurrencia de las formas parémicas en sincronías contemporáneas?

En el marco de las evidencias disponibles, solo para un número finito de paremias y en un conjunto reducido de lenguas ha sido posible formalizar respuestas a las cuestiones planteadas. Así, a título ilustrativo, tras la aplicación de técnicas paremiográficas se ha logrado establecer un índice de aparición para las paremias anexadas en los mínimos paremiológicos¹ actuales disponibles (Sevilla y Zurdo 2016; Sevilla y Barbadillo 2021; Crida 2021; Díaz y Sabio 2024). A pesar de la meticulosidad científica que caracteriza estas obras, persiste cierta resistencia hacia métodos que garanticen el acceso a bancos de datos exhaustivos y capaces de facilitar resultados representativos y estadísticamente relevantes.

Por ello, este estudio pretende indagar sobre la aplicabilidad de los corpus lingüísticos para conseguir una estimación de la frecuencia de uso de las paremias. Con tal fin, se ha seleccionado una muestra de cincuenta enunciados sentenciosos italianos, confeccionados en una lista *ad hoc* considerando criterios lingüísticos y científicos. Para la medición de la frecuencia de uso se ha optado por combinar métodos canónicos de la labor de campo —las encuestas dirigidas a informantes— con recursos computacionales. En concreto, se reflexiona sobre el papel cla-

¹ Hasta la fecha, se dispone de mínimos paremiológicos en tres lenguas, consultables en la *Biblioteca fraseológica y paremiológica* del Centro Virtual Cervantes: https://cvc.cervantes.es/lengua/biblioteca_fraseologica/minimo_paremiologico.htm [última consulta: 15/11/2025].

ve que la Lingüística de Corpus (LC) puede desempeñar como estrategia estadística para el estudio de las unidades parémicas.

La naturaleza gnómica de dichas formas justifica la aproximación mixta. En efecto, que las paremias sean piezas heurísticas vinculadas a la interacción oral queda ampliamente demostrado (Franceschi 1994; Sevilla y Crida 2013: 107; Anscombe 2022: 436; Sevilla y Barbadillo 2021: 30). Según lo observado, es precisamente su naturaleza predominantemente oral otro factor relevante de cara a la aplicación lineal de técnicas estadísticas digitales; por ende, la yuxtaposición de métodos protocolarios y computacionales garantiza una comparabilidad más fiable.

Por ello, los propósitos de este artículo se despliegan en tres direcciones principales: a) deliberar sobre la aplicabilidad de los corpus como método capaz de extraer datos de corte cuantitativo también en ámbito paremiológico; b) facilitar una serie de reflexiones metodológicas cuyo objetivo sea, mediante el empleo de los corpus, asistir al paremiólogo en su labor paremiográfica digital; y, c) demostrar que el corpus es una herramienta idónea para desestimar las incertidumbres que el paremiólogo experimenta al analizar una determinada paremia. Con relación a este último punto, el uso de corpus procesados electrónicamente y las posibilidades de extraer información no solo cuantitativa, sino también cualitativa y contextual, resultan de gran utilidad a fin de comprender ampliamente las particularidades de las paremias. A modo de ejemplo, el corpus puede auxiliar en la compilación de las fichas virtuales de proyectos paremiográficos, como el *Refranero multilingüe* (Sevilla y Zurdo 2009). En el sistema de consulta del refranero, de los numerosos elementos descriptivos que concurren a la presentación de una paremia, los coautores disponen de la opción de asignar marcadores de frecuencia de uso relativos a la distribución de la paremia en la actualidad (Sevilla 2012). Si para el español y otras lenguas existen obras de referencia que facilitan dicha tarea, el italiano no cuenta con tanta fortuna². La carencia de obras actualizadas que registren la

² A estos efectos, es imperativo mencionar las labores —todavía inéditas— que, con meticuloso rigor, se están realizando para la confección del *Mínimo Paremiológico Italiano* (Sardelli 2009; 2011).

difusión de las paremias italianas impone la necesidad de acudir a nuevos enfoques para perseguir tales fines.

2. Revisión de la noción de frecuencia de uso de las paremias

Conforme a lo expuesto en el apartado introductorio, la incógnita del estudio son las marcas de frecuencia de los enunciados sentenciosos. Para abordar el tema, conviene introducir consideraciones generales y, posteriormente, profundizar en las implicaciones de la investigación basada en corpus.

Sevilla y Crida (2013: 107), en su cometido de documentar las características formales y comunes a toda paremia, aluden a la frecuencia de uso, «la cual es gradual y va de un uso alto a estar en desuso». En este sentido, varios autores han pretendido escudriñar el estado actual y la vitalidad de las paremias: algunas investigaciones expresan inquietudes sobre la vitalidad del refranero que, según los hallazgos, se encontraría en proceso de declive por motivo de la progresiva disminución de uso tanto en la población adulta como, de manera más acentuada, entre los jóvenes (Sevilla y Cantera 2002: 255; Crida 2021:15; Asensio 2022: 157; Sevilla 2023: 78-79). En dirección contraria, otros asumen posturas menos pesimistas. Para botón de muestra, Combet (1996: 22) sostiene que las paremias no solo siguen vigentes, sino que manifiestan una continua evolución para adaptarse a las nuevas demandas de la contemporaneidad. Rigurosa es la postura de Shirley Arora (en Sevilla 2000: 14) quien observa como la frecuencia de uso de las paremias se subordina a las variables sociales, con factor determinante de la edad. Efectivamente, la paremióloga argumenta que, aun cuando los jóvenes las reconocen, su utilización práctica resulta mínima. Más allá del ámbito hispanófono, observaciones análogas se registran en la italo fonía. Temistocle Franceschi, pionero de la geoparemiología italiana, alude a un deterioro paulatino de la competencia paremiológica, tendencia todavía más alarmante entre los jóvenes. Según el autor, la causa radicaría en el nexa que los *proverbi*³ mantienen con la oralidad dialectal que las instituciones

³ En italiano, la ausencia de estudios taxonómicos sobre las paremias perpetúa un evidente desfase terminológico (Sardelli 2016: 85). Aunque las diversas peculiaridades

y los medios de comunicación de forma metódica han intentado segregar (en Sardelli 2010: 18).

A la luz de estos datos, resulta lógico sostener que la frecuencia de uso de las paremias está supeditada a múltiples variables, ya sean internas a la lengua o extralingüísticas. Conjuntamente, la opacidad metodológica que rodea el objeto de estudio constituye un ulterior factor de tensión. En este sentido, si la bibliografía citada hasta ahora se fundamenta en importantes investigaciones científicas, otras carecen de tanta solidez. La comunidad investigadora señala la existencia de aproximaciones científicas que indagan el fenómeno paremiológico y fraseológico desde enfoques intuitivos y personales (Gutiérrez 2022: 2). Tal es el caso de obras fraseográficas⁴ en las que se recalca con preocupación «la no existencia de índices de frecuencia de las unidades fraseológicas» (Penadés 1999: 26). El uso de criterios intuitivos expone el riesgo de establecer equivalencias injustificadas entre unidades de escaso uso y otras de mayor relevancia o constancia. Dicha desalineación obedece a las orientaciones teóricas predominantes durante las últimas décadas del siglo XX, por las que el estudio de las unidades multiverbales se ha sustentado en la observación de datos numéricamente limitados y, con frecuencia, orquestados por la intuición personal del investigador (Corpas y Gutiérrez 2023: 2).

El cambio de paradigma en la lingüística contemporánea sugiere prudencia hacia la intuición y la introspección. Según lo plantea Chafe (1991: 90), si bien estas dos habilidades contribuyen a reconocer la relevancia de un fenómeno, lo más satisfactorio y adecuado es cuantificar su ocurrencia mediante la integración de herramientas estadísticas. Efectivamente, los

de las formas parémicas se hallan reconocidas, sin justificación, sigue aplicándose el término *proverbio* como archilexema o hiperónimo que aglutina las categorías más heterogéneas de «todos estos dichos que llegan a hacerse populares» (Sidoti, 2020: 709).

⁴ Este trabajo no parte de una postura ancha que fusione fraseología y paremiología, motivo de discrepancia entre los autores. Aun así, corresponde señalar que el estudio de las UF desde la LC ha sido más fecundo que el de las paremias. Por ello, dadas las características comunes que ambas esferas del discurso repetido comparten, la vía más recomendable es atender a exploraciones de corpus aplicadas a las UF con el objetivo de integrarlas en investigaciones de corte paremiológico.

nuevos corpus a gran escala equivalen a las fuentes más fiables para la obtención de evidencias cuantitativas (Wallis 2021: 7). En términos similares, Pérez-Ávila (2007: 1) identifica en los corpus digitalizados una fuente de recursos lingüísticos que permite prescindir de la intuición; también, destaca su carácter intrínseco como cuantificadores de ocurrencias. Consolidando esta posición, Mellado (2009: 138) sostiene que, de los varios avances en la LC, el más relevante para el lingüista radica en la opción de validar empíricamente sus intuiciones acerca del uso y funcionamiento del ítem lingüístico. Igualmente, conviene recordar cómo dentro de la LC se ha trabajado para el «desarrollo de sistemas y aplicaciones basados en corpus para la detección y la extracción automatizada» de las estructuras multiverbales (Corpas 2013: 351; Corpas 2008: 75). Las paremias son un claro ejemplo de ello.

Ahora bien, habida cuenta de la naturaleza prevalentemente oral de las paremias y la insuficiencia de instrumentos metodológicos estadísticos para medir su dispersión, se revela palmaria la inadecuación para expresar —salvo en aquellos casos en los que la frecuencia de aparición sea inequívocamente manifiesta— un juicio verdaderamente objetivo. En este marco, Corpas (1998: 368) identifica en los corpus el punto de partida del estudio paremiológico. La exposición a una fuente inagotable de datos facilita una comprensión amplia y sistemática del segmento lingüístico. En esta misma línea se colocan las observaciones de Tarnovska (2005: 206), quien los considera un indicador esencial para la medición de la distribución frecuencial de las unidades parémicas. Más recientemente, Gutiérrez (2022) acredita que, de las múltiples funciones del corpus, particularmente fructífera resulta su aplicación estadística en el cómputo de las UF. En el estudio referido, el autor incorpora también una serie de paremias.

Pese a los entusiasmos que genera esta aproximación, conviene precisar los retos metodológicos que presenta el cribado digital de las paremias. Tomando prestadas las palabras de Penadés (2015: 260), indicar la frecuencia de una UF concreta es una tarea altamente intrincada, incluso cuando se limita su búsqueda a un corpus determinado. Según los hallazgos, la misma afirmación puede extenderse al estudio de las paremias basado en corpus, cuyas dificultades se resumen en el siguiente párrafo.

3. Reflexiones sobre la extracción paremiológica en los corpus digitales

La localización de las paremias en corpus digitales exige recorrer un itinerario metodológico específico de esta aproximación.⁵ Se identifican tres nodos críticos de decisión que influyen en el análisis paremiológico corpus-informado: i) evaluar la extensión y los rasgos estructurales de la paremia en cuanto unidades pluriverbales, fijas (Zuluaga 1975: 226) e idiomáticas (Pamies 2007: 173); ii) intervenir en los resultados mediante control manual o estrategias de filtrado electrónicas; y iii) precisar el tipo de resultado que ha de considerarse pertinente en la investigación y, en coherencia con ello, aquello que debe quedar excluido.

El primer punto crítico concierne a los rasgos estructurales de estos enunciados plurimembres. Sevilla y Crida (2017) caracterizan las paremias como «unidades fraseológicas (UF) constituidas por un enunciado breve y sentencioso, el cual corresponde a una oración simple o compuesta». Por ello, queda claro que el rasgo de plurilexicalidad (Corpas 1996: 18; Mellado 2004: 17), en el caso de las paremias, se intensifica hasta el punto de que su extensión máxima adopta la forma de oración compuesta. El caso prototípico en paremiología lo constituyen los dialogismos que, según la forma, llegan a presentar una estructura trimembre adornada por esquemas dialogados (Med 2019: 62; Ugarte 2022: 66). Dicha sobreabundancia de elementos léxicos es el primer factor que condiciona la ejecución de una búsqueda en el corpus tan sencilla y prototípica como la que se configura para la localización de una unidad léxica simple. En palabras de Corpas (2013: 342), la identificación de las UF en los corpus es la dificultad por antonomasia que complica su tratamiento digital.

Con el objetivo de paliar este obstáculo inicial, se acude a ajustes de búsquedas diseñados exclusivamente para las unidades polilexemáticas. En el caso de las locuciones, Corpas (2014: 145) aconseja optimizar los diseños de búsqueda mediante el uso de secuencias truncadas o la extracción de expresiones regula-

⁵ Atendiendo a la postura de Rojo (2021: 48), la LC se considera una aproximación al estudio del fenómeno lingüístico y coadyuva a la observación empírica de grandes cantidades de datos.

res. Refinando la búsqueda al esquema paremiológico, Gutiérrez (2024: 47) explica que, con el intento de reducir la inclusión de combinaciones libres de palabras, la búsqueda de una paremia debería configurarse a partir de «combinaciones de tres palabras consecutivas invariables e impropias *a priori* del discurso libre». ⁶ Efectivamente, aunque la búsqueda de la paremia completa podría constituir el método más intuitivo y usual de buscar signos lingüísticos, la potencialidad del análisis asistido por corpus queda reducida a la mera constatación de los contextos de uso en un número de ocurrencias exiguo (Miccoli 2025: 78). Así, la segmentación se convierte en el procedimiento más conveniente.

Para implementar la segmentación debe considerarse asimismo otro aspecto estructural que atañe a la longitud de la paremia: su variabilidad. Según se aprecia, la extensión de estas unidades, que suele superar con creces el bigrama, representa un rasgo problemático (Corpas 2013: 344). También Tarnovska (2005: 207) comparte esta posición y explica que un aspecto crucial es, justamente, la variable longitud de la unidad proverbial, factor determinante para medir la frecuencia de uso que propicia el corpus: según ella, cuanto más corta es la paremia, más ocurrencias serán proporcionadas. Miccoli (2025: 73) reconoce en la flexibilidad el principio orientador que debe aplicarse a la hora de segmentar la paremia en secuencias. El autor subraya dos razones principales: a) no es lo mismo extraer una secuencia consecutiva y fija de palabras para paremias cortas y aquella de mayor extensión ⁷ y b) sería conveniente ajustar la secuencia al elemento que se quiere resaltar o investigar. Resulta evidente que las decisiones sobre qué secuencia elegir sean el resultado de numerosas pruebas para verificar cuál es el segmento más productivo, cuantitativa y cualitativamente.

En cuanto al segundo punto, el uso deliberado de controles manuales o electrónicos es una repercusión directa de la acción de segmentar la paremia. Si bien las paremias —al igual que las demás UF— se caracterizan por ser fijas, dicha estabilidad se

⁶ El autor pone como ejemplo la secuencia de búsqueda «su San Martín» para la localización de la paremia *A cada cerdo le llega su San Martín*.

⁷ Se comparan *A rey muerto, rey puesto* y *Antes el vicio de pedir, está la virtud de no dar*. La diversa extensión de ambos refranes resulta ostensible.

reconoce como un rasgo gradual que cede el paso a la variación (Penadés 1996; Mellado 2004; García-Page 2008; Sevilla y Crida 2013). Variabilidad y fijación vienen configurándose como las dos caras de la misma moneda (Penadés 2022: 22-23). En los corpus, el resultado de esta dualidad se traduce en que todo elemento que no forma parte del segmento formulado, y debidamente tecleado en el motor de búsqueda, está potencialmente expuesto a procesos de variación y cambio. Esta limitación —o ventaja, dependiendo de los casos— es consecuencia directa de la aproximación basada en corpus. En este ámbito, las UF canónicas y con cierto grado de fijación aparecen con un índice de ocurrencia numéricamente inferior respecto a toda variante derivada a partir de ellas (Corpas 2014: 134). Se constata el mismo escenario con las paremias, lo que comporta intervenir con revisiones para poder limitar el número de los resultados espurios.

La revisión personal de los registros que suministra el corpus puede resultar una práctica tediosa (Corpas 2013: 353). Según lo plantea Svartvik (1992: 10), en la LC, existen fenómenos lingüísticos por los que una atenta revisión manual es imprescindible. Sustituirla con procesos rápidos y automatizados sería, por tanto, lo más cómodo, pero provocaría la pérdida de la representatividad que, en realidad, interesa recuperar. Stubbs (1995: 27) justifica el uso de revisiones personales argumentando que no siempre los procesos computarizados son los más indicados especialmente cuando es el lingüista, y no la máquina, quien tiene la formación adecuada para tipificar tendencias lingüísticas inusitadas. Son las que McEnery y Hardie (2012: 125-126) denominan *hand and eye techniques*, las únicas, según los autores, capaces de permitir al investigador examinar, o en sus palabras *escanear*, los múltiples registros suministrados. De tal modo, se optimiza la identificación de ejemplos y patrones lingüísticos relevantes que no se recuperan mediante algoritmo u otro proceso de filtrado computarizado.

Es innegable que las técnicas de observación personal son aplicables dentro de límites establecidos. Si bien son recomendables a la hora de administrar un número razonable de fragmentos de textos, en otras circunstancias, la enorme presencia de datos vuelve la revisión manual una tarea extremadamente complicada. De ahí que resulte difícil garantizar la ausencia de márgenes

de error (McEnery y Hardie 2012: 2). Ciertamente, la reducción controlada del número de ocurrencias viene siendo una solución frecuente en los estudios basados en corpus, especialmente, cuando se opera con formas propias del lenguaje vivo y oral (Mellado 2009: 142). Una posibilidad es condensar el número de instancias mediante el empleo de técnicas de filtrado⁸ cuyo objetivo es facilitar una revisión manual, optimizar la precisión del análisis y minimizar la confusión generada por el monitoreo personal de miles de formas (Miccoli 2025: 75). Además, la aplicación o no de filtros dependerá del umbral establecido de ocurrencias que el investigador está dispuesto a gestionar manualmente, asegurando la validez analítica y la replicabilidad de los datos.

El tercer punto crítico remite a la delimitación del objeto de estudio. La posibilidad de elegir qué material parémico incluir en el análisis es el beneficio vinculado a la revisión manual de los registros. Como se ha mencionado anteriormente, en la búsqueda de paremias, muchas de las instancias que recupera el corpus están sometidas a alteraciones o no son las previstas en el marco del trabajo. A tal propósito, Corpas (2014: 144) habla de *ruido documental* acerca de aquella información encontrada por el buscador, pero no relevante en cuanto demasiado genérica. También Gutiérrez (2024: 49) emplea el término *ruido* y lo describe como la aparición de datos que interfieren en la precisa realización del estudio. Pamies y Pazos (2023: 141), por su parte, estiman problemáticas las duplicaciones⁹ producto de fuentes que se plagian entre sí. El presente estudio emplea la terminología *fragmentos incidentales* (Miccoli 2025: 73) (en adelante, FI). La decisión de adoptar esta denominación responde a varios motivos: en primer lugar, más que la paremia, el corpus provee al investigador de fragmentos textuales en los que se ubica el elemento de interés.

⁸ En particular, el autor hace hincapié en las ventajas del filtro *keep lines containing* (en la herramienta de gestión de corpus *Sketch Engine*). Mediante este filtro es posible conservar únicamente aquellos fragmentos textuales en los que, además de la secuencia buscada, aparece un ítem específico de interés.

⁹ Esclarecedoras, a tal propósito, son las palabras de Hunston (2022: 41), quien alerta que el problema de la duplicación de los textos en formato electrónico ha sido un obstáculo constante en la LC. Esto ocurre con fragmentos recuperados por el corpus de periódicos, especialmente los que se encuentran en línea, que suelen publicar y referenciar artículos en múltiples ediciones. El resultado son duplicaciones y plagios.

En segundo lugar, la aparición de estos registros se origina a raíz de la configuración de búsqueda formulada. En esencia, se trata de una serie de datos contextualizados. Pues, este catálogo de fragmentos contextuales puede guardar información interesante o, en cierta proporción, recuperar datos carentes de utilidad para la investigación. Estos últimos podrían resultar pertinentes en otras líneas investigativa, circunstancia que justifica su carácter incidental.

Por último, aunque no esté directamente vinculado a estas etapas metodológicas, resulta pertinente introducir otro condicionante de la consulta paremiológica en corpus: la gestión de los signos ortográficos de puntuación. En lo que atañe a las paremias, es frecuente observar una coma o un punto y coma que enfatizan su estructura polimembre, como en el refrán italiano *Tale padre, tale figlio*.

Por lo general, procesar las marcas de puntuación no presenta dificultades significativas. En el motor de búsqueda sería suficiente distanciar los signos de puntuación digitando espacios finos, tanto delante como detrás de las palabras que integran la secuencia formulada.

A pesar de su sencilla disposición, el uso de secuencias comprensivas de signos de puntuación carece de la eficacia deseada, provocando una reducción drástica en el número de ocurrencias. Compárese, por ejemplo, los resultados obtenidos por el corpus *itTenTen20* (en *Sketch Engine*) al buscar la paremia italiana *Il lupo perde il pelo, ma non il vizio*. Las secuencias utilizadas son «il pelo ma» vs «il pelo, ma». La primera excluye la coma, mientras que la segunda la incorpora distanciándola con espacios finos. El corpus proporciona respectivamente 1721 y 493 registros. En la medida de lo posible, convendría emprender los dos tipos de búsqueda para alcanzar una comprensión completa del fenómeno.

Para concluir, es importante aclarar que dichas limitaciones, articuladas en una serie de puntos críticos, son el resultado de observaciones derivadas del uso extensivo de los corpus. Ciertamente, dichas complicaciones varían en función del tipo de *software* de tratamiento de corpus que se emplea, de su tamaño y del elemento en análisis. Asumiendo un escenario favorable, tales

complejidades tenderán a mitigarse con el tiempo, en virtud de las mejoras constantes que se observan en la LC.

4. Procesamiento del dato vía formulario

La exigencia de combinar métodos convencionales y computacionales radica en la resistencia que el objeto de estudio muestra a ser medido. A tal propósito, Sevilla y Zurdo (2016) reconocen el potencial y las ventajas tanto a nivel cuantitativo como cualitativo que ofrece el fácil acceso a bancos de datos procesables mediante herramientas automáticas. Pese a ello, y coincidiendo con las consideraciones de las dos autoras, tales avances no deben ser motivo de supresión de los métodos tradicionales de recopilación paremiográfica, los cuales son irremplazables y complementarios al tipo de documentación facilitada por la aproximación digital (Sevilla y Zurdo 2016: 18). A la luz de estas reflexiones, se ha diseñado un cuestionario comprensivo de cincuenta paremias italianas. La selección de la muestra responde a criterios científicos y lingüísticos, garantizando así la pertinencia del repertorio recopilado.

4.1. Criterios de sistematización del inventario

En lo que respecta a los criterios científicos, la lista constituye un repertorio paremiográfico organizado y seleccionado *ad hoc*. Dicho catálogo ha sido empleado en investigaciones previas a la presente (Asensio y Sciutto 2024: 192). Asimismo, el material parémico ha sido sometido a supervisiones: cada paremia del inventario ha sido localizada en obras paremiográficas de referencia, como el *Refranero multilingüe* (Sevilla y Zurdo 2009) y el *Dizionario dei proverbi Italiani* (Lapucci 2006). Esto ha favorecido la realización de los ajustes necesarios para que la transcripción refleje fielmente la grafía de la paremia tal como figura en las obras de referencia o, cuando resulte pertinente, su adaptación al italiano de uso actual. A modo de ejemplo, el enunciado *Il bove dice cornuto all'asino* (Lapucci 2006: 187 n. 828) contiene el sustantivo *bove*, más arcaizante. Se trata de una variante léxica en desuso en la actualidad. El ítem léxico ha sido sustituido por el sinónimo vigente *bue*, en la formulación *Il bue*

dice cornuto all'asino.¹⁰ Las sistematizaciones se han aplicado a todo el inventario cuando se ha considerado necesario. La inexistencia de obras de referencia exhaustivas para el italiano, los resultados satisfactorios obtenidos en la investigación exploratoria (Asensio y Sciutto 2024) y la validación cruzada de las formas en las obras de referencia aseguran la solidez científica de la lista *ad hoc*.

Si se atiende a los criterios lingüísticos, se ha procurado abarcar solo formas proverbiales que se consideran *proverbi*¹¹ en italiano. Dada la ausencia de una clasificación taxonómica de las paremias italianas, tal como se ha indicado brevemente en la nota 2, se entiende aquí por *proverbio* aquella estructura gnómica de la tradición popular que adopta la forma de enunciado breve y sentencioso (Sevilla y Crida 2015) y que presenta autonomía enunciativa y textual (Zuluaga 1980: 192; Corpas 1996: 132; Alvarado 2007: 6-7; Anscombe 2018: 576-577). Estas precisiones —bien afianzadas en la tradición paremiológica española, pero todavía en fase embrionaria en la tipificación fraseológica y paremiológica italiana— excluyen otras formulaciones de naturaleza fraseológica, como las locuciones. Con estas últimas existen fronteras a veces poco definidas, pero su dependencia enunciativa de otro elemento oracional determina una neta distinción de las paremias (Barbadillo 2024: 211). Por ello, en los cuestionarios se incluyen formulaciones bimembres (*Donna barbata / sempre piaciuta* o *A mali estremi / estremi rimedi*) cuyas propiedades formales permiten asimilarlas a lo que, en ámbito hispanófono, se denomina *refrán*. También se contemplan estructuras unimembres (*Il fine giustifica i mezzi* o *Tutti i nodi vengono al pettine*) en las que la preservación del orden canónico de los constituyentes de la frase evoca a la *frase proverbial*, otra categoría paremial muy presente en la tradición paremiológica española.

¹⁰ La paremia de uso actual con actante bue se recoge también en el *Refranero multilingüe* <https://cvc.cervantes.es/lengua/refranero/Ficha.aspx?Par=58517&Lng=8> [última consulta: 24/11/2025]

¹¹ Si bien en español *proverbio* se clasifica como una subcategoría parémica (Sevilla y Crida 2013: 109), en italiano el término se emplea como archilexema, equivalente de *paremia* en español.

4.2. Disposición del formulario

Para el acopio de los datos orales se ha usado un cuestionario accesible mediante el sistema de gestión de formularios *Google Form*. En virtud de las múltiples opciones de configuración que brinda la plataforma, se ha estructurado el formulario en tres fases secuenciales: a) un bloque inicial que documenta las variables sociodemográficas de los entrevistados; b) el eje central de nuestra investigación correspondiente a las paremias cuyo grado de difusión se pretende evaluar; y, c) una tercera sección, de carácter opcional, en la que se invita al entrevistado a comentar variantes de las paremias empleadas en la sección precedente u otra que emerja espontáneamente, tanto en la lengua estándar como en los distintos dialectos vigentes en el territorio italiano. Si bien la última sección es voluntaria, las fases anteriores son obligatorias. El cuestionario no permite avanzar si no se contesta a cada pregunta marcada visualmente con un asterisco rojo, lo cual indica su obligatoriedad.

En la primera ronda de preguntas, se solicita: 1) la región de procedencia, de entre las veinte que conforman la unidad administrativo-política italiana; 2) la provincia, indicada íntegramente o mediante su abreviatura, por ejemplo, *Roma* o *RM*; 3) la edad, agrupada en cinco rangos: 18-30, 31-40, 41-50, 51-60 y 61+;¹² y, 4) el nivel de educación formal adquirido, fraccionado según el sistema educativo italiano en: *Licenzia media*, *Diploma di Scuola Superiore*, *Laurea Triennale*, *Laurea Magistrale*, *Dottorato* y *Altro/Non Specificato*. Las demás variables socioidentitarias, basadas en supuestos esencialistas o binarios, no se consideran vinculantes para los fines de la investigación. La representatividad sociolingüística exigida queda adecuadamente cubierta mediante (1) la información de corte diatópico, dada la fuerte heterogeneidad lingüística italiana, (2) por la variable edad, puesto que el uso de las paremias se vincula ampliamente a esta dimensión y (3) por el nivel educativo, dado que provee con información de interés diastrático.

¹² Para este grupo, las reducidas capacidades digitales han obligado al entrevistador a garantizar la provisión de soporte personal, o delegado, para la correcta cumplimentación del formulario.

La segunda parte del formulario constituye el núcleo del estudio. En esta sección, se listan cincuenta paremias y para cada una se solicita indicar la frecuencia de uso conforme a la experiencia personal del informante. En este sentido, conviene señalar que la evaluación del grado de uso de una paremia puede generar ciertos niveles de confusión. En particular, este escepticismo surge si se considera, siguiendo a Mellado (2009: 148), que los estudios centrados únicamente en la medición del uso activo de las UF deberían complementarse con aquellos que indagan el grado de conocimiento y percepción de una determinada unidad en análisis, lo que remite a la competencia fraseológica pasiva. Si dentro las UF se consideran exclusivamente las paremias, «pasaremos de la competencia fraseológica a la competencia paremiológica» (Huéscar y Sevilla 2024: 171). Efectivamente, en el caso de las paremias, la competencia paremiológica activa se defiende como la habilidad que tiene el hablante de citar a tiempo y en la circunstancia adecuada la paremia más apropiada (Sevilla y Díaz 1997: 367; Martí 2018: 57). Esta competencia, a su vez, se vincula a la disponibilidad en el repertorio lingüístico del individuo de toda una variedad de informaciones que permite reconocer y comprender la paremia, su competencia paremiológica pasiva (Sevilla y Barbadillo 2005: 5). Siguiendo esta lógica, Tarnovska (2005: 205) subraya la conveniencia de emplear experimentos que midan tanto el reconocimiento pasivo como la disponibilidad activa de las paremias en los informantes. En este marco, Sevilla y Barbadillo (2021: 21) consideran fundamental, en investigaciones de corte empírico, tener en cuenta la competencia pasiva y activa de los hablantes, refiriéndose a ellas en términos de «repertorio individual y colectivo».

Del mismo modo, este estudio pretende recopilar datos integrando tanto la percepción de uso (competencia pasiva) como el uso efectivo (competencia activa) que los hablantes evalúan poseer respecto al material parémico proporcionado. Para ello, las preguntas se han configurado en una cuadrícula de respuesta múltiple (Imagen 1).

Tanto va la gatta al lardo che ci lascia lo zampino *

	Mai	Raramente	A volte	Frequentemente	Molto spesso
Lo uso	☐	☐	☐	☐	☐
Lo sento o leggo	☐	☐	☐	☐	☐

Imagen 1. Cuadrícula de varias opciones.

Tal como se observa, se aplica una escala Likert impar de cinco puntos en orden progresivo mediante el uso de adverbios de frecuencia, de menor a mayor intensidad, que en español se traducen como: nunca, raramente, a veces, frecuentemente y a menudo. El indicador de frecuencia debe señalarse tanto para la competencia activa (entendido aquí como el uso propio del hablante) como para la competencia pasiva (considerado aquí su reconocimiento), respectivamente a través de las fórmulas *Lo uso* (Lo uso) y *Lo sento o lo leggo* (Lo escucho o lo leo). Este enfoque abarca tanto la disponibilidad oral de la paremia (mediante la acción de escuchar), como su provisión escrita (mediante la acción de leer). De tal modo, es posible abarcar toda dimensión diamésica por la que se transmite el dato.

A este respecto, si bien el número de paremias sometidas a encuesta asciende a cincuenta, esta sesión incluye tres preguntas más: ítems de control incorporados con una finalidad metodológica específica. Considerando que el sondeo se ha realizado por medios virtuales y que la cumplimentación del formulario se ejecuta sin la supervisión del entrevistador, existe el riesgo de obtener datos sesgados, derivados de un nivel de concentración insuficiente, de fatiga o de la tendencia del informante a agilizar su cometido tachando las opciones de manera mecánica. Con el fin de mitigar estos riesgos, se ha optado por introducir consultas de control destinadas, por un lado, a mantener elevado el grado de concentración del participante y, por el otro, a desestimar aquellos cuestionarios en los que la consigna no ha sido atendida correctamente. A modo de ejemplo, estos controles se formulan como instrucciones explicativas: *Si estás siguiendo con atención el cuestionario, selecciona la respuesta “raramente”*. Las otras

preguntas de control reproducen la misma proposición, aunque variando el marcador indicado. Asimismo, a fin de mantener la atención del entrevistado, los tres controles se han diseminado de forma aleatoria a lo largo del cuestionario, después de las preguntas número 10, 27 y 44. El propósito es evitar la identificación de patrones de dispersión de los controles. Por ello, aquellos cuestionarios que presentan errores en los ítems de control serán considerados inconsistentes y eliminados.

La tercera y última sección habilita un espacio abierto para aportar, de manera no dirigida, variantes de las premias previamente listadas y/o proponer otras vigentes en el repertorio individual o comunitario. Esta sección busca complementar el enfoque cuantitativo con aportaciones cualitativas, favoreciendo la recopilación de datos espontáneos que enriquecen el corpus y ofrecen futuras líneas de investigación.

Ahora bien, es oportuno señalar que previa su difusión, el cuestionario ha sido sometido a una prueba piloto en la que participaron diez informantes. A estos se les ha solicitado —además de contestar a las preguntas— observar, experimentar y evaluar la herramienta empleada. En otras palabras, su papel ha sido proporcionar retroalimentaciones sobre la aplicabilidad de la encuesta, su facilidad en ser cumplimentada y comprendida, así como sobre posibles mejoras antes de su lanzamiento. Tras recibir comentarios positivos, el cuestionario se ha difundido mediante canales mediáticos. Finalmente, para cumplir con el requisito de sincronía, siguiendo las pautas de Navarro-Brotons (2020: 116), la encuesta ha permanecido abierta durante tres meses, desde el 16 de julio de 2025 hasta el 16 de octubre del mismo año.

5. Triangulación y discusión de los resultados

Expuestas las premisas teóricas y metodológicas que rigen la presente contribución, este párrafo describe el proceso de investigación realizado.

5.1. El estudio de corpus

En cuanto a los aspectos del cribado digital, se ha acudido a la herramienta de gestión de corpus *Sketch Engine* y se ha selec-

cionado un corpus representativo del italiano estándar actual: el *Italian Web Corpus*, conocido mediante la sigla *itTenTen20*. Es el último gigacorporus de la familia *TenTen* para el italiano y cuenta con una amplitud registrada de más de 14 mil millones de tokens. Este corpus de referencia cubre una gran variedad de géneros, tipologías textuales y recursos web, con una descarga de datos realizada entre octubre y diciembre de 2019 y diciembre de 2020. De tal modo, el enfoque sincrónico del estudio queda justificado. El corpus se ha interrogado en su integridad evitando limitar las búsquedas a subcorpus menores. Se ha empleado la configuración de búsqueda concordancia (*concordance*) que ordena al corpus de proporcionar todos los fragmentos textuales en los que aparece la secuencia parémica tal cual se ha digitado en el motor de búsqueda. Siguiendo esta configuración, se ha seleccionado el tipo de consulta (*query type*) frase (*phrase*) que técnicamente se emplea para la identificación de secuencias de palabras.

Se ha procedido a segmentar las paremias en secuencias distintas para verificar cuál equivale a la más prolífica en términos de resultados cuantitativos, considerando las premisas expuestas en el párrafo 3 sobre flexibilidad y puntuación. De manera puntual, para enunciados breves como *Uomo avvisato, mezzo salvato*, el segmento extraído consiste en un número reducido de palabras consecutivas, como «uomo avvisato». En el caso de paremias más extensas, pues, se sigue la lógica de que, a mayor extensión de la paremia, mayor será la extensión del segmento extraído. A modo de ejemplo, «occhio del padrone» para *L'occhio del padrone ingrassa il cavallo*.

El paso sucesivo ha sido la revisión. Estas han sido manuales cuando el corpus ha proporcionado un número de segmentos que no superaba más de mil registros. Cuando las ocurrencias excedían el umbral establecido, se ha optado por operaciones automáticas orientadas a su acotación. Es el caso del filtro *keep lines containing* (nota 8). Limitar el número de ocurrencias a una cantidad operativamente gestionable ha facilitado la observación de los resultados y, por ende, ha permitido compartimentar los FI.

Con respecto a los FI, se establecen criterios de selección que incluyen todo tipo de variación parémica surgida a raíz de la búsqueda. Esto excluye las combinaciones libres de palabras.

Si se piensa en la secuencia «amore è» para la frase proverbial *L'amore è cieco*, es indudable la naturaleza libre, y menos repetida, del sintagma. A través de las estrategias de filtrado por reducción numérica se ha acotado el número de fragmentos textuales que incluyen combinaciones libres, que para este estudio se considerarán los únicos FI. La decisión de incluir todas las modificaciones realizadas a partir de la forma de referencia¹³ radica en el supuesto de competencia paremiológica que se ha adoptado a lo largo del estudio. En términos generales, la capacidad de los hablantes de modificar las estructuras fraseológicas y paremiológicas, según sus propias exigencias, es *sine qua non* un indicador de competencia activa (Martí 2019: 12).

Tal como se puede intuir, toda variación, voluntaria o involuntaria, implica cierta disponibilidad de competencia paremiológica tanto individual, pero primariamente compartida y comunitaria. Por ello en este estudio, aceptamos los registros donde aparezcan: a) desautomatizaciones o manipulaciones creativas, como *Chi va con lo zoppo impara il twist*, *Donald Trump batte dove il dente della Cina duole* o *I lupi di Wall Street perdono il pelo ma non il vizio*; b) hipotéticas variantes marcadas o no, como *La volpe perde il pelo ma non il vizio*, *A buon intenditore basta mezza parola*, *Tra i due contendenti il terzo gode*, *Il cervo che dice cornuto al bue*; c) estructuras en las que se verifica una inversión de sus constituyentes, como *Poche parole al buen intenditore*; d) variaciones estructurales, como *Dal dire al fare*, *c'è di mezzo il mare*, *Via la gatta, i topi ballano*; e) paremias aludidas o diluidas, como *Con la botte piccola, il vinaiolo cura di più la qualità che la quantità*; f) estructuras parémicas fraccionadas por palabras intercaladas, comentarios propios del hablante o un presentador paremiológico, como *A caval donato, dice il detto, non si guarda in bocca*; g) formulas en las que se superponen varios procesos de variación, como *I topi ballano malgrado (e forse grazie a) la presenza del gatto*. Y otras que, por obvias cuestiones de espacio, no se listan.

¹³ En este trabajo es la paremia empleada en el formulario. Para iniciar la labor de segmentación, por lo general se acude a la forma canónica o de referencia de la paremia. Para la definición de forma canónica de una paremia consúltese Klein (2006: 163).

5.2. Resumen de los resultados del formulario

En cuanto a los datos ofrecidos por los informantes, la encuesta ha recibido un alcance fortuito, totalizando 485 cuestionarios remitidos. Las respuestas han sido luego trasladadas a una hoja de cálculo *Excel* para facilitar su gestión y conteo gracias a una visualización más esquemática. Del total de los formularios recibidos, se han excluido 17 que no respetan las órdenes impartidas por los puntos de control. Se ha encuestado a un total de 468 sujetos.

En cuanto a la dispersión geográfica, el cuestionario ha sido cumplimentado en casi todas las regiones italianas, en más de 70 provincias diseminadas a lo largo de la península. En cuanto a las variables edad y nivel de formación formal, se visualizan respectivamente en los Gráficos 1 y 2.

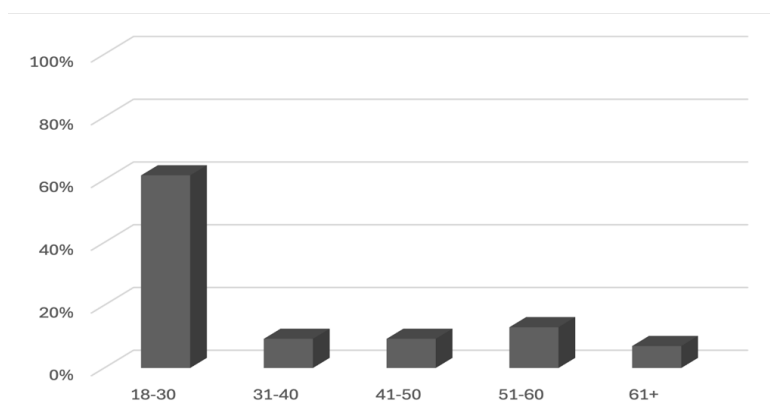


Gráfico 1. Edad de los informantes.

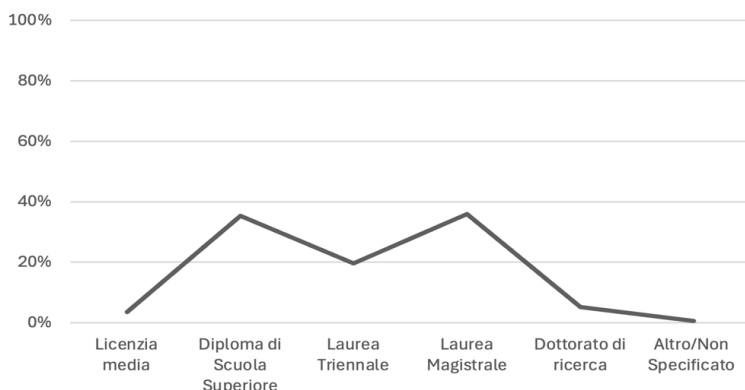


Gráfico 2. Nivel de formación formal de los informantes.

5.3. Validación cruzada de los resultados: corpus vs encuesta

La etapa más espinosa del análisis corresponde a la triangulación de los resultados obtenidos por dos medios de recopilación muy heterogéneos. Por un lado, ha sido necesario contabilizar los registros proporcionados por el corpus; por el otro, aunque el formulario facilita cifras numéricas claras, en términos estadísticos, queda pendiente la creación de un método de comparación apto para triangular los datos. Por ello, se ha decidido considerar solo la vertiente perceptiva de los informantes. De esta manera, se valora la distribución de la unidad lingüística conforme a la experiencia colectiva y no individual.

A modo de ejemplo, presentamos los datos de cinco paremias. Los resultados muestran patrones distribucionales similares tanto en los formularios como mediante el cribado digital. Las cifras numéricas quedan resumidas en la Tabla 1. En la columna *Encuesta* se indica el cálculo de la media.¹⁴

¹⁴ Con el fin de obtener un valor frecuencial, hemos decidido optar por cálculos usuales de la estadística descriptiva: se establece la media de las respuestas proporcionadas por los informantes. En este caso, el proceso es de sencilla resolución. A cada categoría de la escala Likert se le asigna un valor numérico, desde 0 hasta 1, con incrementos de 0,25: Mai: 0 / Raramente: 0,25 / A volte: 0,5 / Frequentemente: 0,75 / Spesso: 1. Asignados los valores, se multiplica el número de respuestas de cada categoría por el valor correspondiente. Los resultados se suman y se dividen entre el número total de participantes (468). El dato resultante es un índice de frecuencia que se coloca entre 0

Tabla 1. Triangulación de los resultados.

	Paremia	Encuesta	itTenTen20 - FA ¹⁵
<i>a</i>	Chi va con lo zoppo, impara a zoppicare	646	641
<i>b</i>	Morto un papa se ne fa un altro	731	1024
<i>c</i>	Sbagliando, s'impára	772	1700+
<i>d</i>	L'occhio del padrone ingrassa il cavallo	136	207
<i>e</i>	Gallina vecchia fa buon brodo	480	461

Las cifras en la tabla evidencian que a un mayor número de registros propiciados por el corpus responde un índice de frecuencia más alto señalado por los entrevistados. Se puede apreciar la misma proporcionalidad en los resultados de menor rendimiento. La paremia *c* es la que registra más ocurrencias en el corpus y, al igual, su media en los cuestionarios responde a un índice de frecuencia superior a 0,75: según los informantes se usa *frecuentemente*. Por otro lado, si se considera la paremia *d*, el *itTenTen20* indica el nivel de ocurrencias más bajo. En términos proporcionales, conclusiones análogas se sacan mediante formulario. Aquí, el índice de frecuencia se sitúa entre *nunca* y *raramente*. Para las paremias *a* y *e* se observan valores numéricos que se acercan al punto medio de la escala Likert. Los indicadores del corpus respaldan la encuesta. Cabe observar que los cambios mínimos en la puntuación decimal de la encuesta responden a cambios exponenciales en el corpus. Efectivamente, comparando *b* y *c*, al cambio mínimo decimal (de 0,73 a 0,77), responde en el corpus un cambio exponencial (de 1024 a 1700 ocurrencias): casi 700 ocurrencias más por un 0,04 decimal.

Los ejemplos en la tabla muestran equilibrio entre los datos recopilados mediante el método convencional y el digital. Sin embargo, es imperativo indicar que, aunque la mayoría de las

y 1. Si el resultado es 0,5 (o cercano) la paremia tiene un grado de frecuencia medio. De manera paralela, cuanto más se acerca a 1 más se utiliza la paremia y, a la inversa, cuando la cifra se aproxima a 0 la paremia se etiqueta como en desuso. Para proyectar el decimal a un número entero, se multiplica por 1000. A modo de ejemplo, para la paremia *a* se observa un valor posicional decimal de 0,6464. Multiplicando por 1000, el número entero comparable será 646,4, que se redondea a 646.

¹⁵ En LC, acrónimo de Frecuencia Absoluta.

paremias analizadas muestra patrones proporcionales bastante armoniosos, otros casos son discordantes. De manera puntual, la célebre paremia *Una rondine non fa primavera*, presente en la mayoría de los repertorios paremiológicos indoeuropeos (Anscombre 2022: 428), presenta una frecuencia considerable en el corpus, con un total de formas registradas que supera las 1500 ocurrencias. En cambio, no se percibe con la misma frecuencia en el cuestionario, donde solo veinte informantes señalan que la escuchan *a menudo* y casi 300 individuos la colocan entre *nunca* y *raramente*. Al contrario, otras formas que en corpus producen resultados escasos, en los cuestionarios registran una frecuencia intensa. Es el caso de *Nella botte piccola c'è il vino buono* que, según los entrevistados se posiciona cerca de *frecuentemente*, aunque en el corpus solo se recogen unas 500 ocurrencias. Otro ejemplo análogo, y más drástico todavía, es relativo a *Parli del diavolo e spuntano le corna* que en el corpus registra poco más de 90 formas; en cambio los hablantes la perciben con destacada frecuencia: 185 votos *a menudo* y 135 *frecuentemente*. Solo 8 personas han tachado la casilla *nunca*.

Dicha asimetría en los resultados constituye una evidencia adicional sobre lo arduo que es realizar un análisis de este tipo, en el que nociones como la frecuencia de uso son muy inconstantes y de compleja clasificación. Acudir a varios métodos que complementen a la vez la escritura y la oralidad es la aproximación más indicada. Por lo general, es posible asumir que algunas formas conviven en ambos canales diamésicos, mientras que otras muestran cierta reticencia en formar parte de la escritura privilegiando el canal donde fueron originadas, la oralidad.

6. Consideraciones conclusivas

Este trabajo pretende demostrar la aplicabilidad de los corpus para el estudio de un fenómeno tan controversial como la frecuencia de uso de las paremias que, hasta la fecha, sigue generando divergencias analíticas en la comunidad científica. La incógnita de partida del trabajo versa sobre qué metodología resulta la más idónea para registrar la intensidad de la unidad lingüístico-paremial, un rasgo de inmediata intuición, pero problemático a la hora de ser catalogado o demostrado. En el caso

específico de las paremias, su naturaleza esencialmente oral dificulta dicha empresa ulteriormente, instando a la coordinación de varios métodos de consulta para medir, a veces de manera exitosa y otras no, la distribución del evento proverbial.

Siguiendo esta lógica, se introduce una metodología mixta para la estimación de la frecuencia de uso de 50 paremias italianas. Combinar más métodos tiene como objetivo el cotejo de múltiples perspectivas y la triangulación de los datos. La falta de obras de referencia que clasifiquen las paremias italianas según su frecuencia, la ausencia de métodos estadísticos digitales y la incertidumbre que sufre el paremiólogo para comprender plenamente dicho fenómeno justifican este planteamiento. Siguiendo esta postura, se ha recurrido al uso de herramientas tradicionales de la labor de campo, como los formularios dirigidos a informantes, y por el otro, el uso de recursos estadísticos computacionales. Estos últimos, los corpus, son imprescindibles cuando el tema es medir la ocurrencia de un fenómeno. Disponer de una fuente ilimitada de datos digitalizados es algo que ya no se puede desestimar en los estudios de corte paremiológico. Por ello, se ha propuesto una metodología de búsqueda articulada en varios puntos críticos que el investigador debería considerar en su cometido de localizar, acotar y tratar estas estructuras tan complejas a nivel no solo estructural, sino también variacional.

Una vez obtenidos los resultados, se ha pasado a la triangulación de los registros. En términos generales, ambos métodos de consulta, censal y cribado digital, permiten establecer patrones simétricos en la difusión de las paremias. Las indicaciones aportadas por los entrevistados y las cuantificaciones establecidas en el corpus coinciden, aunque se presentan excepciones. De hecho, pese a las apreciaciones favorables iniciales, la hipótesis planteada no siempre se verifica. En concreto, por una cifra reducida de paremias, la simetría mencionada se altera hacia asimetría y, dependiendo del caso, con fuerte desproporción. Algunas de las paremias analizadas son muy recurrentes en la oralidad, pero apenas se aprecian en los fragmentos escritos recogidos por el corpus. Contrariamente, otras presentan una notable presencia en el corpus, pero revelan una reducida frecuencia de uso según la encuesta.

Esto permite concluir que, si bien algunos enunciados sentenciosos han entrado a formar parte del repertorio escrito del hablante, otros muestran cierta resistencia a vivir en ambas diamesías, quedándose principalmente relegadas a la oralidad donde nacieron y donde los hablantes las perciben con mayor frecuencia.

Cabe destacar que los resultados que se discuten en este trabajo han de ser interpretados en su contexto: tal vez, acudir a corpus de diferente tipología, como los orales o los mixtos, puede ser un punto de partida igual de satisfactorio. Asimismo, la posibilidad de establecer criterios de proporcionalidad matemáticamente más afinados puede generar resultados más próximos a la realidad. El uso coordinado de más corpus de referencia para el italiano constituye otra posibilidad, dado que lo aconsejable siempre es trabajar ampliando los datos.

Se puede concluir con que el uso de los corpus como herramienta estadística orientada a la medición de la frecuencia de uso representa una aproximación hoy indispensable en las investigaciones paremiológicas. Su uso no se limita al análisis cuantitativo, sino que representa una rica fuente de muestras cualitativas. Si, por un lado, su uso integrado con otras metodologías proporciona una visión ampliada y nueva de cómo actúan estas frases sentenciosas en la actualidad, parece apropiado indicar que para el estudio de la frecuencia de uso de las paremias no es empleable como método exclusivo o principal. Por lo menos, no todavía. El corpus se configura, así, como herramienta de soporte adicional a las técnicas de cribado ya consolidadas. Su función principal es acertar, justificar o desmentir las hipótesis planteadas y las intuiciones que madura el paremiólogo, respetando las funcionalidades propias del estudio basado en corpus.

Referencias bibliográficas

- Alonso Pérez-Ávila, Elena. “El corpus lingüístico en la didáctica del léxico del español como LE”. *Boletín de la Asociación para la Enseñanza del Español como Lengua Extranjera*, no. 37, 2007, pp. 11–32.
- Alvarado Ortega, M. Belén. “Las fórmulas rutinarias como unidades fraseológicas”. *ELUA*, no. 21, 2007, pp. 9–20, <https://rua.ua.es/server/api/core/bitstreams/099be5b7-b398-4a77-b09e-f19408ac4859/content>, última consulta el 23/11/2025.

- Anscombre, Jean-Claude. “La gnomicidad/genericidad de las paremias desde el punto de vista del tiempo y del aspecto”. *RILCE*, vol. 34, no. 2, 2018, pp. 573–604, <https://revistas.unav.edu/index.php/rilce/article/view/8408/22256>, última consulta el 25/11/2025.
- Anscombre, Jean-Claude. “Origen y evolución de las paremias”. *RILCE*, vol. 38, no. 2, 2022, pp. 426–446, <https://revistas.unav.edu/index.php/rilce/article/view/42051/36313>, última consulta el 25/11/2025.
- Asensio Ferreiro, M. Dolores y Virginia Sciutto. “Estudio teórico-metodológico y contrastivo de la competencia paremiológica en lengua materna y lengua extranjera en el ámbito universitario”. *Romanica Olomucensia*, vol. 36, no. 2, 2024, pp. 187–205, <https://romanica.upol.cz/pdfs/rom/2024/02/01.pdf>, última consulta el 11/11/2025.
- Barbadillo de la Fuente, María Teresa. “Los Diarios de Lorenzo: caracterización fraseológica y paremiológica del protagonista”. *Romanica Olomucensia*, vol. 36, no. 2, 2024, pp. 207–224, <https://romanica.upol.cz/pdfs/rom/2024/02/02.pdf>, última consulta el 11/11/2025.
- Cantera Ortiz de Urbina, Jesús, Julia Sevilla Muñoz y Manuel Sevilla Muñoz. *Refranes, otras paremias y fraseologismos en Don Quijote de la Mancha*. Burlington, Vermont: University of Vermont, 2005.
- Chafe, Wallace. “The Importance of Corpus Linguistics to Understanding the Nature of Language”. *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, edited by Jan Svartvik, Berlin/New York: Mouton de Gruyter, 1992, pp. 105–122.
- Combet, Louis. “Los refranes: origen, función y futuro”. *Paremia*, no. 5, 1996, pp. 11–22.
- Corpas Pastor, Gloria y Enrique Gutiérrez Rubio. “Computational and corpus phraseology applied to Spanish”. *Romanica Olomucensia*, vol. 35, no. 1, 2023, pp. 1–8, <https://romanica.upol.cz/pdfs/rom/2023/01/02.pdf>, última consulta el 09/12/2025.
- Corpas Pastor, Gloria. “Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas”. *Fraseopramática*, edited by Inés Olza Moreno y Elvira Manero Richard, Berlin: Frank & Timme, 2013, pp. 335–373.
- Corpas Pastor, Gloria. “El fraseólogo internauta: cómo pasarlo pipa en la red”. *Fraseología y paremiología: enfoques y aplicaciones*, edited by Vanda Durante, Madrid: Instituto Cervantes, 2014, pp. 133–152.
- Corpas Pastor, Gloria. “El uso de paremias en un corpus del español peninsular actual”. *Estudios de fraseología y fraseografía del español actual*, edited by Gerd Wotjak, Madrid: Iberoamericana Vervuert, 1998, pp. 365–390.

- Corpas Pastor, Gloria. *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang, 2008.
- Corpas Pastor, Gloria. *Manual de fraseología española*. Madrid: Gredos, 1996.
- Crida Álvarez, Carlos Alberto, and Julia Sevilla Muñoz. “La problemática terminológica en los estudios paremiológicos”. *Anuari de Filologia. Estudis de Lingüística*, vol. 5, 2015, pp. 67–77, <https://revistes.ub.edu/index.php/AFEL/article/view/AFEL.2015.5.5/18052>, última consulta el 09/12/2025.
- Crida Álvarez, Carlos Alberto. *El mínimo paremiológico griego con correspondencias en español*. Centro Virtual Cervantes (Instituto Cervantes), 2021. *Biblioteca fraseológica y paremiológica*, serie “El mínimo paremiológico”, no. 3, https://cvc.cervantes.es/lengua/biblioteca_fraseologica/m3_crida/el_minimo_paremiologico_griego_con_correspondencias_en_espanol.pdf, última consulta el 02/12/2025.
- Díaz Ferrero, Ana M. José Antonio Sabio Pinilla. *El mínimo paremiológico portugués: el portugués europeo*. Centro Virtual Cervantes (Instituto Cervantes), 2021. *Biblioteca fraseológica y paremiológica*, serie “El mínimo paremiológico”, no. 4, https://cvc.cervantes.es/lengua/biblioteca_fraseologica/m4_diaz/el_minimo_paremiologico_portugues_el_portugues_europeo.pdf, última consulta el 09/12/2025.
- Franceschi, Temistocle. “El Atlas Paremiológico Italiano (API) y el Centro Interuniversitario de Geoparemiología de la Universidad de Florencia”. *Paremia*, no. 6, 1997, pp. 223–234, https://cvc.cervantes.es/lengua/paremia/pdf/006/033_franceschi.pdf, última consulta el 02/12/2025.
- García-Page, Mario. *Introducción a la fraseología española. Estudio de locuciones*. Barcelona: Anthropos, 2008.
- Gutiérrez Rubio, Enrique. “Estudio de corpus de los presentadores paremiológicos en español”. *Paremia*, no. 34, 2024, pp. 45–54, https://cvc.cervantes.es/lengua/paremia/pdf/034/003_gutierrez.pdf, última consulta el 28/11/2025.
- Gutiérrez Rubio, Enrique. “Frecuencia de uso de locuciones y paremias en el corpus Spanish Web 2018 (esTenTen18): implicaciones didácticas y lexicográficas”, *Proceedings of the International Conference EURO-PHRAS 2022, Malaga, Spain, 28–30 September*, edited by Gloria Corpas Pastor et al., Shoumen: INCOMA, 2022, pp. 26–33.
- Huéscar Villa, Myriam y Manuel Sevilla Muñoz. “La competencia paremiológica pasiva de estudiantes universitarios con un nivel avanzado de lengua materna y extranjera: un estudio de caso”. *Onomázein. Revista de Lingüística, Filología y Traducción*, no. especial XIV, 2024.

- Hunston, Susan. *Corpora in Applied Linguistics*. 2nd ed., Cambridge: Cambridge University Press, 2022.
- Klein, Jean René. “Problemas relacionados coa determinación da forma ‘canónica’ nunha base de datos de refráns franceses (DicAuPro)”. *Cadernos de Fraseoloxía Galega*, no. 8, 2006, pp. 147–163, https://www.cirp.gal/pub/docs/cfg/cfg08_07.pdf, última consulta el 17/11/2025.
- Lapucci, Carlo. *Dizionario dei proverbi italiani*. Milano: Le Monnier, 2006.
- Martí Sánchez, Manuel. “De las unidades fraseológicas al conocimiento fraseológico”. *Universidad de La Habana*, no. 285, 2018, pp. 50–68, <https://revistas.uh.cu/revuh/article/view/2681/2339>, última consulta el 13/11/2025.
- Martí Sánchez, Manuel. “La competencia fraseológica como objeto problemático de la fraseodidáctica”. *Innovación en fraseodidáctica. Tendencias, enfoques y perspectivas*, edited by Carlos Alberto Crida Álvarez and Arianna Alessandro, Frankfurt am Main: Peter Lang, 2019, pp. 11–23.
- McEnery, Tony y Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.
- Med, Natalia. “Los dialogismos y wellerismos en las lenguas romances”. *Paremia*, no. 28, 2019, pp. 69–66, https://cvc.cervantes.es/lengua/paremia/pdf/028/006_med.pdf, última consulta el 15/12/2025.
- Mellado Blanco, Carmen. “Utilidad y limitaciones de los corpora informáticos en la elaboración de un tesoro fraseológico (alemán-español)”. *A Survey on Corpus-based Research. Panorama de investigaciones basadas en corpus*, edited by Pascual Cantos Gómez y Aquilino Sánchez Pérez, Murcia: Asociación Española de Lingüística de Corpus, 2009, pp. 138–151, <https://www.um.es/lacell/aelinco/contenido/pdf/10.pdf>, última consulta el 09/12/2025.
- Mellado Blanco, Carmen. *Fraseologismos somáticos del alemán*. Frankfurt am Main/Berlin/Bern/Bruxelles/New York/Oxford/Wien: Peter Lang, 2004.
- Miccoli, Vito Antonio. “La localización de refranes en corpus lingüísticos digitales: métodos y aplicaciones”. *Paremia*, vol. 35, 2025, pp. 69–80, https://cvc.cervantes.es/lengua/paremia/pdf/035/006_miccoli.pdf, última consulta el 05/12/2025.
- Navarro-Brotons Lucía. “Aproximación mediante encuestas a las paremias usuales comunes entre España e Hispanoamérica aplicables a las clases de traducción general”. *Fraseología y variaciones (socio)lingüísticas y diatópicas*, edited by Pedro Mogorrón Huerta y Ana Cuadrado Rey, *ELUA Anexo VII*, 2020, pp. 113–128, <https://rua.ua.es/server/api/core/>

- bitstreams/1e931553-613c-49f0-95dd-b5a3d191a437/content última consulta el 17/12/2025.
- Pamies Bertrán, Antonio y José Manuel Pazos. “Frasemas combinados ‘a pedir de boca’”. *Romanica Olomucensia*, vol. 35, no. 1, 2023, pp. 139–156, https://romanica.upol.cz/artkey/rom-202301-0010_frasemas-combinados-a-pedir-de-boca.php, última consulta el 10/12/2025.
- Pamies Bertrán, Antonio. “De la idiomaticidad y sus paradojas”. *Nouveaux apports à l'étude des expressions figées*, edited by G. Conde Tarrío, Fernelmont: InterCommunications & E.M.E., 2007, pp. 173–204.
- Penadés Martínez, Inmaculada. “La variación en las unidades fraseológicas”. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, vol. 16, no. 32, 2022, pp. 21–39, <https://revistas.nebrija.com/revista-linguistica/article/view/472/426>, última consulta el 12/12/2025.
- Penadés Martínez, Inmaculada. “Implicaciones de la frecuencia de uso de las locuciones en la elaboración de un diccionario”. *ELAU*, no. 29, 2015, pp. 253–277, <https://rua.ua.es/server/api/core/bitstreams/de4c12cbb382-4a0d-b1af-48fdc0f36c4a/content>, última consulta el 15/12/2025.
- Rojo, Guillermo. *Introducción a la lingüística de corpus en español*. London: Routledge, 2021.
- Sardelli, Maria Antonella. “De los estudios paremiológicos en Italia. Estado de la cuestión”. *Paremia*, no. 25, 2016, pp. 79–97, https://cvc.cervantes.es/lengua/paremia/pdf/025/007_sardelli.pdf, última consulta el 02/12/2025.
- Sardelli, Maria Antonella. “Aproximación á elaboración do ‘mínimo paremiológico’ italiano”. *Cadernos de Fraseoloxía Galega*, no. 11, 2009, pp. 191–210.
- Sardelli, Maria Antonella. “Le ricerche geoparemiologiche in Italia”. *Paremia*, no. 19, 2010, pp. 11–18, https://cvc.cervantes.es/lengua/paremia/pdf/019/001_sardelli.pdf, última consulta el 02/12/2025.
- Sardelli, Maria Antonella. “Recopilación y sistematización del mínimo paremiológico italiano”. *Paremiología y herencia cultural*, edited by Antonio Pamies Bertrán, José de Dios Luque y Patricia Fernández Martín, Granada: Granada Lingüística, 2011, pp. 87–95.
- Sevilla Muñoz, Julia y Carlos Alberto Crida Álvarez. “Las paremias y su clasificación”. *Paremia*, no. 22, 2013, pp. 105–114, https://cvc.cervantes.es/lengua/paremia/pdf/022/009_sevilla-crida.pdf, última consulta el 10/10/2025.
- Sevilla Muñoz, Julia y Carlos Alberto Crida Álvarez. “Taxonomía de las paremias en lengua española”. *Phrasis*, no. 1, 2017, pp. 117–129, <https://>

- www.phrasis.it/rivista/index.php/rp/article/view/18/14, última consulta el 13/12/2025.
- Sevilla Muñoz, Julia y Juan Carlos Díaz. “La competencia paremiológica: Los refranes”. *Proverbium*, vol. 14, 1997, pp. 367–381.
- Sevilla Muñoz, Julia y M.^a del Carmen Ugarte García. “La variación en los refranes y frases proverbiales en español: un caso práctico”. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, vol. 16, no. 32, 2022, pp. 92–108, <https://revistas.nebrija.com/revista-linguistica/article/view/477/430>, última consulta el 13/12/2025.
- Sevilla Muñoz, Julia y M.I. T. Zurdo Ruiz-Ayúcar (dirs.). *Refranero multilingüe*. Instituto Cervantes (Centro Virtual Cervantes), 2009, <https://cvc.cervantes.es/lengua/refranero/>, última consulta el 15/12/2025.
- Sevilla Muñoz, Julia y María Teresa Barbadillo de la Fuente. “El máximo y el mínimo paremiológicos y sus aplicaciones didácticas”. *La creatividad en el lenguaje: colocaciones idiomáticas y fraseología*, edited by Antonio Pamies y Juan de Dios Luque Durán, Granada: Método, 2005, pp. 1–13.
- Sevilla Muñoz, Julia y María Teresa Barbadillo de la Fuente. *El mínimo paremiológico español*. Centro Virtual Cervantes (Instituto Cervantes), 2021. *Biblioteca fraseológica y paremiológica*, serie “El mínimo paremiológico”, no. 2, https://cvc.cervantes.es/lengua/biblioteca_fraseologica/m2_sevilla/el_minimo_paremiologico_espanol.pdf, última consulta el 07/11/2025.
- Sevilla Muñoz, Julia. “Aplicación de las TIC a la paremiología y la paremiografía: el Refranero multilingüe”. *Unidades fraseológicas y TIC*, edited by M. Isabel González Rey, Madrid: Instituto Cervantes, 2012, pp. 51–62.
- Sevilla Muñoz, Julia. “Shirley L. Arora y la paremiología hispanoamericana”. *Paremia*, no. 9, 2000, pp. 7–14, https://cvc.cervantes.es/lengua/paremia/pdf/009/001_sevilla.pdf, última consulta el 02/12/2025.
- Sidoti, Rossana. “Las paremias y su clasificación en el ámbito de la Fraseología y de la Paremiología”. *Orillas*, no. 9, 2020, pp. 685–714, <https://www.orillas.net/orillas/index.php/orillas/article/view/94/90>, última consulta el 14/10/2025.
- Stubbs, Michael. “Collocations and semantic profiles: on the cause of the trouble with quantitative studies”. *Functions of Language*, vol. 2, no. 1, 1995, pp. 23–55.
- Svartvik, Jan. “Corpus Linguistics Comes of Age”. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4–8 August*

- 1991), edited by Jan Svartvik, Berlin/New York: Mouton de Gruyter, 1992, pp. 7–13.
- Tarnovska, Olga. “El mínimo paremiológico en la lengua española”. *La creatividad en el lenguaje: colocaciones idiomáticas y fraseología*, edited by Antonio Pamies Bertrán y Juan de Dios Luque Durán, Granada: Méto-do, 2005, pp. 197–217.
- Ugarte García, María del Carmen. “Eso lo sabe mi burra, dijo Corique. Dialogismos en el habla de Roa de Duero (Burgos)”. *Paremia*, no. 32, 2022, pp. 63–73, https://cvc.cervantes.es/lengua/paremia/pdf/032/006_ugarte.pdf, última consulta el 13/12/2025.
- Wallis, Sean. *Statistics in Corpus Linguistics Research: A New Approach*. Routledge, 2021.
- Zuluaga Ospina, Alberto. “La fijación fraseológica”. *Thesaurus*, vol. 30, no. 2, 1975, pp. 225–248, https://cvc.cervantes.es/lengua/thesaurus/pdf/30/TH_30_002_017_0.pdf, última consulta el 09/12/2025.
- Zuluaga Ospina, Alberto. *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter Lang, 1980.
- Zurdo Ruiz-Ayúcar, M.I. T. y Julia Sevilla Muñoz. *El mínimo paremiológico: aspectos teóricos y metodológicos*. Centro Virtual Cervantes (Instituto Cervantes), 2016. *Biblioteca fraseológica y paremiológica*, serie “El mínimo paremiológico”, no. 1, https://cvc.cervantes.es/lengua/biblioteca_fraseologica/ml_zurdo/el_minimo_paremiologico.pdf, última consulta el 07/11/2025.

STUDY ON THE FREQUENCY OF USE OF PROVERBS: REFLECTIONS FROM CORPUS LINGUISTICS

Abstract: This study examines the applicability of corpora as a computational tool for evaluating the frequency of use of proverbs. The concept of frequency of use, seemingly straightforward, is a complex variable to systematize due to difficulties in its cataloging process. To address this, in this work traditional methodologies are combined with corpus-oriented approaches. Both approaches present heterogeneous methodological aspects; however, despite their different natures, the results show a singular convergence in the frequency patterns. Aside from a few irregularities, data triangulation reveals a proportional consistency in the frequency of use of paremias across both screening methods. The corpus-linguistic approach, thus, emerges as a complementary method for extracting information that is not only quantitative but also qualitative and contextual, providing a richer and unconventional understanding of the proverbial phenomenon.

Keywords: digital paremiography, paremiographic survey, frequency of use, corpus linguistics, Italian proverbs

Vito Antonio Miccoli
Facultad de Filología
Universidad Complutense de Madrid
Ciudad Universitaria, C/Profesor Aranguren s/n, 28040 Madrid, España
Correo electrónico: vmiccoli@ucm.es

Copyright (c) 2026 Author(s)
This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License.

