ÉTUDES ROMANES DE BRNO 42/2021/1
HTTPS://DOI.ORG/10.5817/ERB2021-1-19

La variación fraseológica: análisis del rendimiento de los corpus monolingües como recursos de traducción

Idiom Variation: Analyzing the Performance of Monolingual Corpora as Translation Resources

Carlos Manuel Hidalgo-Ternero [cmhidalgo@uma.es]
Universidad de Málaga, España

Gloria Corpas Pastor [gcorpas@uma.es]
Universidad de Málaga, España

Resumen
Las múltiples manifestaciones con las que se pueden presentar las unidades fraseológicas en el discurso (variación, flexión gramatical, discontinuidad…) hacen especialmente compleja la creación de patrones de búsqueda apropiados que permitan recuperarlas en todo su esplendor discursivo sin que ello implique un excesivo ruido documental. En este contexto, a lo largo del presente estudio se analiza el rendimiento de diferentes sistemas de gestión de corpus disponibles para el español en la consulta de las variantes fraseológicas *tener entre manos*, *traer entre manos* y *llevar entre manos,* e *ir al pelo* y *venir al pelo*. De forma concreta, se someterán a examen dos corpus creados por la RAE (el CREA, en sus versiones tradicional y anotada, y el CORPES XXI), el Corpus del Español de Mark Davies (BYU) y Sketch Engine. Los resultados arrojados por este análisis permitirán vislumbrar qué sistema de gestión de corpus ofrece un mejor rendimiento para los traductores ante el desafío de la variación fraseológica.

Palabras clave
Variación fraseológica; discontinuidad; patrones de búsqueda; sistemas de gestión de corpus

Abstract
Idioms tend to vary significantly in discourse (variation, grammatical inflection, discontinuity…). This makes it especially difficult to create appropriate query patterns that obtain these units in all shapes and forms while avoiding excessive noise. In this context, this paper analyses the performance of different corpus management systems available for Spanish when searching phraseological variants such as *tener entre manos*, *traer entre manos* and *llevar entre manos,* as well as *ir al pelo* and *venir al pelo*. More specifically, we will examine two corpora created by the Real Academia Española (CREA, in its original and annotated version, and CORPES XXI), the Corpus del Español by Mark Davies (BYU), and Sketch Engine. The results of our study will shed some light on which corpus management system can offer a better performance for translators under the challenge of idiom variation.

Keywords
Idiom variation; discontinuity; query patterns; corpus management systems

RECIBIDO 2021–01–28; **ACEPTADO** 2021–04–22

Agradecimientos: La presente investigación se ha llevado a cabo en el marco de distintos proyectos de investigación en tecnologías de la lengua aplicadas a la traducción e interpretación (ref. FFI2016–75831-P, UMA18-FEDERJA-067, CEI-RIS3 y EUIN2017–87746). Asimismo, ha sido subvencionada por el Ministerio de Ciencia, Innovación y Universidades (FPU16/02032).

1. Introducción

Los corpus monolingües ofrecen un amplio abanico de posibilidades documentales en la consulta de unidades fraseológicas (UF) durante el proceso traslaticio, ya sea a fin de examinar el uso real de estas unidades en contexto, su rango colocacional, su paradigma flexivo, sus posibles variantes formales, así como la variedad diasistemática a la que pertenecen, entre muchas otras opciones. A este respecto, los corpus monolingües presentan una amplia ventaja comparativa con respecto a los paralelos, no solo porque ofrecen mayor cantidad de texto en lengua no traducida, sino también, y especialmente en lo que respecta a la lengua de llegada, porque esta tiende a mostrar rasgos de simplificación y normalización en lo que concierne a la fraseología (cfr. Leppihalme 2000, Lawick 2007, Marco Borillo 2009 y 2011, Corpas Pastor 2015 y 2018, Valencia Giraldo y Corpas Pastor 2019), por lo que resulta complicado detectar, en los textos meta, UF en todo su esplendor variacional.

A pesar de este potencial documental de los corpus monolingües, los sistemas de gestión de corpus no siempre cuentan con una interfaz de búsqueda que permita una consulta refinada y exhaustiva de las unidades fraseológicas, especialmente teniendo en cuenta el distinto grado de fijación de sus elementos integrantes y las múltiples manifestaciones con las que pueden presentarse en el discurso: variación, flexión gramatical, discontinuidad… (cfr. Philip 2008). Al mismo tiempo, la constante presión temporal a la que está sometida la labor traductora hace imprescindible la selección de herramientas documentales que con el menor esfuerzo heurístico[1] posible obtengan los mejores resultados en términos cuantitativos y cualitativos, a fin de evitar así la necesidad de realizar múltiples búsquedas para cada una de las posibles manifestaciones formales de una UF en el discurso.

En este contexto, el presente artículo tiene por objetivo evaluar y comparar el rendimiento de distintos sistemas de gestión de corpus disponibles para el español ante el desafío de la variación fraseológica. Con este fin, el trabajo queda estructurado de la siguiente manera. En el apartado 2, se ofrece una breve introducción al concepto de *variante fraseológica* a fin de presentar y describir

1 Con *esfuerzo heurístico*, hacemos referencia al volumen de trabajo que una determinada consulta le requiere al traductor. Entre los distintos parámetros para medir este esfuerzo se encuentran el número de búsquedas que son precisas para obtener una UF en todas sus formas (variación, discontinuidad…), la necesidad o no de consultar previamente la flexión de alguno de los constituyentes de la UF para redactar el patrón, etc.

las UF objeto de estudio. A continuación, en el apartado 3 se presentan las limitaciones que ofrecen los recursos lexicográficos en el tratamiento de estas variantes. Por su parte, en el apartado 4, se desgranan las posibilidades que ofrecen distintos sistemas de corpus disponibles para el español en la consulta de estas variantes, para posteriormente comparar el rendimiento de estos sistemas en el apartado 5. Finalmente, en el apartado 6 se exponen las conclusiones del presente estudio.

2. Sobre el concepto de variante fraseológica

Variación y fijación son dos caras de la misma moneda fraseológica, dado que toda variación requiere en esencia que exista de esa forma previa estable, reconocible y bien delimitada que conforman las UF:

> Si en un principio las UFs *[sic]* se definieron como unidades estables y fijas que mostraban rechazo a cualquier alteración léxica, semántica y morfosintáctica, en vista de la existencia de cambios reales y potenciales, no cabe hablar de la fijación como una propiedad absoluta (Hausermann, 1977; Burger, 1998) sino como una cualidad relativa (Fleischer, 1982; Glaser, 1986; Wotjak, 1992; Corpas Pastor, 1996; Burger, 1998) y, nunca mejor dicho, variable. (Corpas Pastor y Mena Martínez 2003: 183)

De esta manera, la variación es un rasgo inherente de las UF no "a pesar de" sino "gracias a" su carácter estable: "Die Polilexicalität ist ein Appell an die Fragmentierung, die Fixierheit an die Variabilität, die Figuriertheit an die Litteralisierung"[2] (Gréciano 1987: 196). A este respecto, Mellado Blanco (2004) señala que no solo es que las UF puedan sufrir variación, sino que estas, por su propia naturaleza, son más proclives a la variación que las unidades léxicas simples:

> Los FR [fraseologismos] están por naturaleza más expuestos a cambios evolutivos porque son doblemente asimétricos: sintagmáticamente – por su carácter plurilexemático – y paradigmáticamente – por su significado no composicional. La estructura plurimembre fomenta la variación formal, […] mientras la asimetría paradigmática está relacionada con la metaforización y con el cambio semántico […]. Si tenemos en cuenta que cada palabra está, en principio, expuesta a cualquier tipo de variación, con mayor probabilidad aparecerá esta en cadenas de palabras. (Mellado Blanco 2004: 155)

Esta propensión hacia la variación no quiere decir que las variantes de una UF sean ilimitadas e impredecibles, sino todo lo contrario, estas están prefijadas en el sistema, es decir, incluso en la variación hay fijación. En este contexto, con el concepto de *variante fraseológica* haremos referencia a aquellas alteraciones formales que presenta una UF obtenidas de cambios estables e institucionalizados fruto de las opciones que ofrece el sistema lingüístico (y, concretamente, el fraseológico), como han venido afirmando autores como Zuluaga (1980), Corpas Pastor (1996), Montoro del Arco (2005) o García-Page (2008), entre otros.

2 "La polilexicalidad es una llamada a la fragmentación; la fijación, a la variabilidad; la figuratividad, a la literalidad" (la traducción es nuestra)

Según García-Page (2008), estos cambios institucionalizados pueden ser, principalmente, de naturaleza fónica (*a volapié*/*a vuelapié*), gráfica (*a regañadientes*/*a regaña dientes*), morfológica (*meter la nariz*/*meter las narices*), gramatical (*sentar cabeza*/*sentar la cabeza*), sintáctica (*hacer oídos sordos*/*hacer oídos de mercader*) o léxica (*meter la pata*/*meter la gamba*). En nuestro estudio, abordaremos de forma concreta las variantes léxicas (denominadas asimismo como *variantes por relación paradigmática* [Koike 2007]), en UF como *tener entre manos [algo]*, *traer(se) entre manos [algo]* y *llevar(se) entre manos [algo]*, con el significado de 'estar ocupándose [en ello]' (Seco y otros 2017: 492), e *ir al pelo* y *venir al pelo*, con el significado de "[ir] a la medida de la necesidad o el deseo" (Seco y otros 2017: 633). Cada una de estas variantes consta, a su vez, de una parte que admite flexión (el verbo) y otra fija (los bigramas *entre manos* y *al pie*, respectivamente). También conviene señalar que dichas UF admiten la inclusión de elementos que pueden dividir la secuencia, es decir, pueden presentarse en los textos tanto en sus formas continuas como discontinuas (Anastasiou 2010). Así, el principal objetivo del presente estudio será analizar las posibilidades que ofrece la interfaz de búsqueda de distintos sistemas de gestión de corpus para la detección de todas estas posibles alteraciones formales que pueden sufrir las UF (variación léxica, flexión gramatical, discontinuidad…).

3. Variación fraseológica y limitaciones de los recursos lexicográficos

Los recursos lexicográficos constituyen una herramienta fundamental en la labor traductora tanto para la consulta de las acepciones de una UF como de posibles correspondencias primarias en el plano léxico en otras lenguas. No obstante, aún queda un largo camino por recorrer en el tratamiento lexicográfico de las UF dado que la mayoría de los recursos disponibles en línea suelen recoger exclusivamente aquellas con mayor frecuencia de uso y, dentro de ellas, principalmente solo sus formas canónicas, sin reflejar el amplio espectro de posibilidades que presenta la variación fraseológica (cfr. Corpas Pastor 2015 y 2018, Dal Maso 2020).

A modo de ejemplo, analizaremos en este apartado distintos diccionarios en línea tanto bilingües español-inglés como *Wordreference* (Kellog), *Collins Dictionary* (HarperCollins Publishers), *Cambridge Dictionary* (Cambridge University Press) y *Oxford Spanish Dictionary* (Oxford University Press) como monolingües en español (el *Diccionario de la Lengua Española* [DLE]), a fin de determinar en qué medida pueden ser recursos documentales útiles en la consulta de las siguientes variantes de UF, que aparecen en fragmentos extraídos de diversas publicaciones digitales:

1. Con la mirada de un arquitecto acostumbrado a observar detenidamente todos los elementos, Moneo comprende que la situación actual del país es "muy difícil", pero espera que se resuelva, a una "escala supranacional", ya que el mundo ha pasado por "muchas convulsiones", de las que finalmente está saliendo. Entre los proyectos que lleva actualmente entre manos se encuentra la remodelación urbana de un hotel en Málaga y un laboratorio para la Universidad de Princeton (New Jersey) en Estados Unidos, país en el que ha desarrollado numerosos trabajos.[3]

3 Este fragmento, extraído del periódico Diario de Navarra (02/03/2013) se encuentra disponible en la siguiente dirección URL: https://www.diariodenavarra.es/noticias/mas_actualidad/sociedad/2013/03/02/rafael_moneo_confia_repunte_arquitectura_109409_1035.html

2. Una de esas bodegas que puedes recomendar a todo el mundo y nunca defraudan, y encima ¡que sean de Madrid! Bueno, de Madrid y de Ávila, que para este cronista (nacido en Ávila y habitante de Madrid) va ni que al pelo…[4]

En el ejemplo 1 es posible detectar la UF *llevar entre manos* en su forma discontinua. Al consultar esta UF en los diccionarios mencionados encontramos resultados dispares. En *Wordreference* y *Oxford Spanish Dictionary* solo se recogen las UF *traer(se) entre manos* y *tener entre manos,* ambas con correspondencia en inglés (en el caso de *Wordreference* se ofrece además definición y 2 ejemplos de uso, y, en *Oxford Spanish Dictionary*, 1 ejemplo de uso y sin definición). En *Collins Dictionary* y *Cambridge Dictionary* exclusivamente se recoge la variante *traerse entre manos* con una información lexicográfica desigual: mientras que en *Collins Dictionary* únicamente se ofrece ejemplos de uso y traducción al inglés, en *Cambridge Dictionary* solo se muestra la definición de la UF. Finalmente, en el DLE solo se recogen las variantes *traer entre manos* y *tener entre manos*, con definición pero sin ejemplo de uso. A modo de resumen, en Tabla 1 puede comprobarse el tratamiento de las variantes *traer entre manos*, *tener entre manos* y *llevar entre manos* en los diccionarios objeto de consulta.

Diccionarios	Variantes	Definición	Ejemplos	Equivalentes
Wordreference	*traer entre manos, tener entre manos*	Sí	2	Sí
Oxford Spanish Dictionary	*traerse entre manos, tener entre manos*	No	1	Sí
Collins Dictionary	*traerse entre manos*	No	2	Sí
Cambridge Dictionary	*traerse entre manos*	Sí	0	No
Diccionario de la Lengua Española	*traer(se) entre manos, tener entre manos*	Sí	0	No

Tabla 1. Tratamiento lexicográfico de las UF *traer/tener/llevar entre manos* en los diccionarios objeto de estudio

Por su parte, en el ejemplo 2 es posible detectar la UF *ir al pelo*, también en su forma discontinua. Al consultar esta variante en los distintos diccionarios nos encontramos nuevamente con una situación similar. En *Wordreference* y *Cambridge Dictionary* no aparece ni *ir al pelo* ni *venir al pelo*. En el caso de *Collins Dictionary*, solo está registrada la UF *venir al pelo*, sin definición pero con 3 ejemplos de uso y traducción al inglés de estos ejemplos. En *Oxford Spanish Dictionary* se ofrece la UF *venir al pelo*, con definición, un ejemplo de uso y equivalente en inglés. Finalmente, en el DLE solo se recoge el bigrama *al pelo* (sin verbo), con definición pero sin ejemplo de uso. En Tabla 2 pueden observarse estos resultados esquematizados.

4 Este fragmento, extraído del periódico El Mundo (05/11/2012), se encuentra disponible en la siguiente dirección URL: http://elmundovino.elmundo.es/elmundovino/noticia.html?vi_seccion=25&vs_fecha=201211&vs_noticia=1352072804

Diccionarios	Variantes	Definición	Ejemplos	Equivalentes
Wordreference	—	—	—	—
Oxford Spanish Dictionary	*venir al pelo*	Sí	1	Sí
Collins Dictionary	*venir al pelo*	No	3	Sí
Cambridge Dictionary	—	—	—	—
Diccionario de la Lengua Española	*al pelo*	Sí	0	No

Tabla 2. Tratamiento lexicográfico de las UF *ir/venir al pelo* en los diccionarios objeto de estudio

Así, podemos comprobar como los diccionarios ofrecen un tratamiento muy dispar para las variantes objeto de estudio en lo que se refiere tanto a la UF registrada como a la inclusión de información lexicográfica (definición, ejemplo de uso, equivalente de traducción al inglés…). El único parámetro común a todos ellos es que en ninguno de estos diccionarios ha sido posible recuperar, de forma concreta, las variantes *llevar entre manos* ni *ir al pelo*. Por todo ello, en el siguiente apartado analizaremos en qué medida el traductor puede salvar este obstáculo y obtener información útil sobre estas UF y sus variantes en otra herramienta esencial en la labor documental: los corpus lingüísticos.

4. Posibilidades y limitaciones de los sistemas de gestión de corpus

En esta sección analizamos la detección automatizada de las variantes fraseológicas objeto de estudio. Con este fin, hemos seleccionado cuatro sistemas de gestión de corpus: dos de ellos (CREA y CORPES XXI) constituyen sistemas integrados específicos (*in-built*) para los corpus del español compilados por la RAE; el tercero es también de tipo integrado, específico para la consulta de diversos corpus monolingües del español, el inglés y el portugués, alojados en la plataforma de gestión de corpus desarrollada por Mark Davies en la Brigham Young University (BYU), Estados Unidos; y el cuarto es Sketch Engine (Kilgarrif y otros 2003), un sistema de gestión de corpus monolingües y multilingües (paralelos), que se puede utilizar como los anteriores (es decir, integrado para la consulta específica de los corpus alojados por defecto en dicho sistema) o bien de forma autónoma (*stand-alone*) para consultar cualquier corpus subido por el usuario.

En lo que concierne al tamaño, el Corpus de Referencia del Español Actual (CREA) está disponible tanto en su versión tradicional (con más de más de 160 millones de formas ortográficas en la última actualización en junio de 2008) como en la anotada (publicada en noviembre de 2015, con 126 millones de formas), mientras que el Corpus del Español del Siglo XXI (CORPES XXI) cuenta con más de 312 millones de formas. Respecto a los corpus que tiene alojados el Corpus del Español de Mark Davies (BYU), se examinó de forma concreta el NOW (News on the Web) (Davies 2018), que incluye más de 7 mil millones de palabras. Finalmente, en Sketch Engine, se analizó el rendimiento de uno de los corpus de español que aloja, concretamente el esTenTen, perteneciente a la familia de corpus TenTen, que cuenta con más de 17 500 millones de palabras.

Además del tamaño, también es diferente la naturaleza de estos corpus (cfr. Corpas Pastor, 2015 y 2017). Los corpus de referencia del español creados por la RAE están compuestos en su

mayoría por libros (49 % en el caso del CREA y 40 % en el CORPES XXI) y publicaciones periódicas (nuevamente 49 % en CREA y 40 % en CORPES XXI); en menor medida encontramos material diverso extraído de internet (1 % en CREA y 7,5 % en CORPES XXI). Cada uno de los textos incluidos se encuentra perfectamente documentado (con información sobre país, año, zona geográfica, tema, tipo de texto, etc.) y la variedad de textos es equilibrada. El Corpus NOW, por su parte, está compuesto por periódicos y revistas electrónicos durante el periodo comprendido entre 2012 y 2019. Para cada uno de los textos se ofrece diferente información tal como fecha de publicación y país, a fin de poder recuperar y contrastar las búsquedas atendiendo a estos parámetros. Por último, esTenTen está compilado de forma automática de internet, según el dominio de nivel superior (.es, .ar, .cu, etc.), y no está equilibrado ni tampoco documentado. Únicamente se ha eliminado el código reutilizable (como por ejemplo las marcas de html) y se han suprimido documentos o párrafos repetidos. Asimismo, los textos pueden contener erratas, dado que buena parte de ellos está conformada por contenido generado por el usuario. Sin embargo, como podremos comprobar a lo largo de la publicación, es precisamente la inclusión de esta tipología textual, unida a la gran cantidad de textos que ofrece este corpus, lo que facilita la aparición de la variación fraseológica en todo su potencial. Finalmente, en lo que respecta a las principales diferencias en la interfaz de búsqueda de los distintos sistemas de gestión de corpus, estas serán analizadas en los próximos subapartados.

4.1. CREA (versión tradicional)

La versión tradicional del CREA no permite la búsqueda por lemas, formas, ni categorías gramaticales. A fin de obtener el verbo en toda su flexión verbal, será necesario el uso de los únicos comodines que permite el sistema de concordancias: "?", para obtener la aparición simple de un carácter en la posición que se inserta, o "*", para recuperar cualquier número de caracteres (incluso ninguno) desde la posición en la que se incluye. No obstante, esta versión tradicional no permite el uso simultáneo de varios comodines, por lo que no es posible introducir el patrón "t*n* la cabeza" a fin de obtener concordancias con el verbo *tener* en toda su flexión gramatical (*tengo, tienes, tendrá*…). El sistema sí permite el uso de los operadores booleanos (Y, O, NO), por lo que podremos hacer una búsqueda con las principales raíces que presenta la flexión verbal de *tener*, es decir, podemos emplear el patrón "(ten*) O (tien*) O (tuv*) entre manos". Estos operadores booleanos también nos permitirán la obtención simultánea de concordancias con las restantes variantes, lo que ampliaría el patrón a "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*) entre manos". Finalmente, emplearemos el operador "dist/" a fin de también obtener concordancias de la UF en su forma discontinua y, de esta manera, aumentar la exhaustividad del sistema de concordancias. Por ejemplo, para obtener resultados con la secuencia "tener/traer/llevar [3 *tokens*] entre manos" podemos emplear el patrón "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*) dist/3 entre manos".

Como podemos observar en la Figura 1, con este patrón obtenemos 280 concordancias, 276 de las cuales coinciden con las UF objeto de búsqueda (por tanto, alcanza una precisión del 98,6 %). Los restantes resultados se corresponden principalmente con ejemplos del bigrama *entre manos* antepuesto a la secuencia "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*)" sin guardar relación con esta. Esto se debe a que "dist/3" no permite especificar si la distancia es hacia la izquierda o hacia

la derecha, lo que ha provocado que algunas concordancias se hayan obtenido por duplicado si alguno de los elementos posteriores incluía la secuencia "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*)", como podemos observar en los siguientes fragmentos (en negrita la UF concreta y en subrayado el secuencia que no guarda relación con la UF):

3. Ni uno ni otro son hombres del equipo económico. Ambos tienen su respaldo en la Jefatura de Gabinete. Pero los temas que **tienen entre manos** son tan transcendentes que desataron una guerra santa en el Gobierno, y una lluvia de munición pesada contra los funcionarios desde la actividad privada.
4. Hace 24 horas comenzaba a darme cuenta que el trabajo que **tenía entre manos** me iba a llevar terminarlo algo más de lo que había previsto.

Figura 1. Consulta de las UF *tener entre manos*, *traer entre manos* y *llevar entre manos* en la versión tradicional del CREA

Comprobaremos ahora la precisión de este sistema de gestión de corpus en la consulta de UF cuyos elementos integrantes presenten gran irregularidad en su flexión gramatical, como es el caso del verbo *ir* en la UF *ir al pelo* y su variante *venir al pelo*. Así, a fin de obtener toda la flexión verbal tanto de *ir* (*voy*, *ibais*¸ *fuimos, yendo*…) como de *venir* (*vienen*, *venías*, *viniendo*...) es necesario emplear el siguiente patrón "(v*) O (i*) O (f*) O (y*) dist/3 al pelo". Con él, obtenemos 77 casos, solo 32 de los cuales se corresponden con la UF *ir/venir al pelo*, es decir, la precisión de la herramienta ha descendido a un 41,6 %. Dado que, fruto de la gran irregularidad en el paradigma flexivo del verbo *ir*, el patrón de búsqueda se ha tenido que reducir hasta la primera letra (*v**, *i**, *f**, *y**), ha aumentado exponencialmente el ruido documental, como podemos observar en los siguientes ejemplos:

5. En cuanto a la imagen, lo que realmente da morbo a la mayoría del público, aún no se tienen noticias. Aunque muchos apuestan por una vuelta al pelo rizado, castaño y los crucifijos.

6. Muchas plantas utilizan el viento para dispersar sus semillas. En otros casos, los frutos o las semillas se fijan al pelo de los animales para ser transportados a otros lugares.

Así, el 58,4 % de resultados restante se ha correspondido con una miscelánea de concordancias que no guardaban relación alguna con las UF objeto de búsqueda, dado que el único criterio de filtrado era la aparición de *al pelo* antepuesto o pospuesto a una distancia de hasta 3 *tokens* de cualquier palabra que comenzase por las letras *v*, *i*, *f* o *y*, de ahí la reducida precisión de la versión tradicional del CREA ante UF en las que alguno de sus elementos integrantes presente una elevada irregularidad flexiva.

En resumen, comprobamos que el CREA, en su versión tradicional, no permite determinar qué elementos constituyentes de la UF pueden sufrir flexión (y, por tanto, deben ser recuperados como *lemas*), o la categoría gramatical de los mismos, tampoco permite establecer un número mínimo de *tokens* que divida la secuencia (solo se puede fijar el máximo), ni especificar el orden concreto del patrón, a fin de evitar que, por ejemplo, el bigrama "entre manos" o "al pelo" aparezca antepuesto en secuencias que no guardan relación con la UF. Estas limitaciones hicieron necesaria, en primer lugar, la laboriosa tarea de consultar previamente toda la flexión gramatical de los verbos implicados (*tener, traer* y *llevar* en el caso de *tener/traer/llevar entre manos* e *ir/venir* en el caso de *ir/venir al pelo*), a fin de poder introducir manualmente cada una de las raíces irregulares: "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*) dist/3 entre manos" y "(v*) O (i*) O (f*) O (y*) dist/3 al pelo", respectivamente. Aun así, esta redacción manual del patrón no evitó un elevado ruido documental en el segundo caso (el de las UF *ir/venir al pelo*). Por ello, la precisión de su mecanismo de búsqueda dependerá enormemente del nivel de irregularidad del paradigma flexivo de los verbos que conformen las UF objeto de búsqueda.

4.2. CREA (versión anotada) y CORPES XXI

Analizamos conjuntamente la versión anotada del CREA y del CORPES XXI dado que presentan la misma interfaz. En lo que respecta a esta búsqueda de UF, en el manual de consulta en línea de ambos sistemas se menciona la opción de introducir *Texto libre* en la casilla *Forma*, donde se permite buscar secuencias de hasta cinco palabras. No obstante, este tipo de búsqueda solo obtiene un alto nivel de exhaustividad con UF fijas, como las locuciones ejemplificadas en el manual (*de tal palo tal astilla*, *de tomo y lomo*, etc.). Para la consulta de UF cuyos elementos integrantes no presenten una gran irregularidad en su flexión (como el caso de *venir al pelo*), en estas versiones sí es posible el uso de varios comodines de forma simultánea, por lo que se podría emplear la secuencia "v*n*" para obtener el verbo *venir* en toda su flexión gramatical.

Aun así, la búsqueda en esta casilla *Forma* no sería eficiente en aquellos casos en los que alguno de los elementos integrantes de la UF presente gran irregularidad flexiva (como en el caso de la otra variante, *ir al pelo*) ni cuando queramos asimismo obtener la forma discontinua de la UF. Tampoco sería posible la búsqueda simultánea de todas las variantes de la UF en la misma casilla de *Forma*. Finalmente, dado que, a diferencia de la tradicional, estas versiones anotadas no cuentan con el operador "dist/", en estos casos de discontinuidad sería necesario realizar tantas

búsquedas como número de *tokens* pueda dividir la secuencia: "v*n* al pelo", en caso de 0 *tokens* divisores; "v*n* * al pelo", en caso de 1; "v*n* * * al pelo", en caso de 2, y así sucesivamente.

Por ello, para la búsqueda de una UF que pueda presentar discontinuidad, resulta más conveniente emplear la opción de *Proximidad*, con la que es posible recuperar concordancias con hasta cuatro lemas o formas divididas por un máximo de diez palabras. Así, a diferencia de los casos anteriores, el manual recomienda el empleo de la opción de *Proximidad* para UF como *dar cuartel* ('dar tregua o descanso', Seco y otros 2017: 225), en la que verbo principal *dar* presenta una flexión irregular (*doy, di, des, daban*, etc.). Ello, unido al hecho de que la parte inalterada en la flexión de este verbo se reduce a una sola letra ("d"), provocaría un elevado ruido documental si se emplease la casilla de *Forma* con el verbo truncado ("d* cuartel"): no solo se obtendrían concordancias con *dar cuartel,* sino también con secuencias que no guardan relación con dicha UF, como *de cuartel*, *del cuartel*, *dicho cuartel*, etc. Asimismo, la búsqueda en *Proximidad* permite la obtención de concordancias con UF discontinuas, ya que con ella es posible establecer tanto la distancia exacta entre el primer y el último elemento constituyente de la UF como el intervalo de *tokens* que puede dividir la secuencia. Así, para encontrar concordancias con la forma continua y discontinua de la UF *tener entre manos*, es necesario introducir la primera parte de la secuencia (*tener*) en la casilla *Lema*, a fin de obtenerla en toda su flexión gramatical, y, en la opción *Clase de palabra*, seleccionar *verbo*. A continuación, tras pulsar en *Proximidad,* comprobamos que no es posible buscar el bigrama *entre manos* en una misma casilla, por lo que será necesario incluirlo en distintas. Así, a fin de obtener concordancias con la UF *tener entre manos* dividida por hasta 3 *tokens*, buscaremos *entre* en la casilla *Forma* como preposición en un intervalo de 4 palabras hacia la derecha y *manos* también en la casilla *Forma* (dado que no admite variación en esta UF) como sustantivo en un intervalo de 5 palabras hacia la derecha. A fin de obtener asimismo concordancias con las variantes *traer entre manos* y *llevar entre manos*, comprobamos que el sistema solo permite la búsqueda simultánea de distintos lemas opcionales en casillas distintas, empleando el operador booleano "O". Por ello, será necesario repetir el mismo patrón de búsqueda en *Proximidad* tras el lema *traer* y tras *llevar*. En la Figura 2 es posible observar el patrón de búsqueda completo en la versión anotada del CREA.

Así, al introducir este patrón en CREA, obtenemos 375 concordancias, de las cuales 231 se corresponden exactamente con las UF objeto de búsqueda (*tener/traer/llevar entre manos*), es decir, este sistema ha alcanzado una precisión del 61,6 %. En CORPES XXI, obtenemos 1161 resultados (total de secuencias obtenidas con los parámetros de búsqueda), de los cuales 591 son n-gramas relevantes (es decir, UF correctamente identificadas y extraídas por el sistema de gestión de corpus); y 570 son n-gramas irrelevantes (esto es, secuencias extraídas con los parámetros de búsqueda que no constituyen UF), lo cual supone una precisión del 50,9 %. Así, podemos observar cómo una parte importante de los resultados (el 38,4 % en CREA y el 49,1 % en CORPES XXI) se corresponden con n-gramas irrelevantes (véanse los ejemplos 7 y 8):

7. Necesitaba llevar algo entre las manos, un objeto capaz de hacerme compañía, de sostenerme y de animarme. Sentía que no podía volver con las manos vacías.
8. Bienvenido Arcéiz —de profesión barbero— se había colocado delante y le interrumpía el camino. Tenía las manos entre las piernas —con las palmas vueltas— a manera de cuenco, reía, dijo, ¡aquí hay mucho borde!, y avanzó más hacia José Gracia colocando entonces una de sus manos en el aire con el índice extendido.

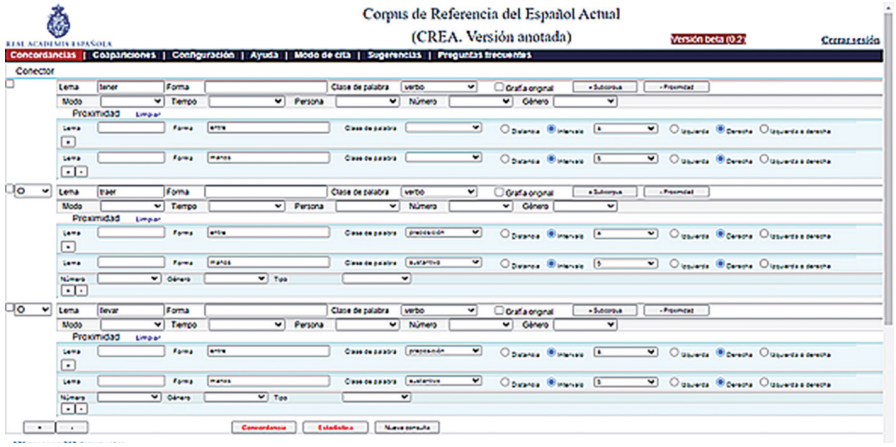

Figura 2. Consulta de las UF *tener entre manos*, *traer entre manos* y *llevar entre manos* en la versión anotada del CREA

Estas concordancias son una clara muestra de una de las principales limitaciones de las versiones anotadas del CREA y del CORPES XXI: dado que ninguna de ellas permite la búsqueda del bigrama invariable *entre manos* en una misma casilla de consulta en la opción de *Proximidad*, gran parte de los resultados irrelevantes se corresponden con concordancias que bien incluyen algún tipo de determinante o adjetivo modificando a *manos* (lo que conlleva una interpretación literal de la secuencia, como en *llevar algo entre las manos* o *tiene entre sus manos bañadas*) o bien contienen a *entre* y *manos* en sintagmas distintos, o incluso en un orden inverso (*tenía las manos entre las piernas*). Estos resultados muestran un considerable descenso en la precisión de las versiones anotadas del CREA y del CORPES XXI con respecto a la versión tradicional del CREA para las UF *tener/traer/llevar entre manos*: 61,6 % y 50,9 % frente a 98,6 %, respectivamente.

Pasamos ahora a analizar el rendimiento de ambas versiones anotadas para las UF *ir al pelo* y *venir al pelo*. En CREA, obtenemos 30 resultados, de los cuales 27 se corresponden con n-gramas relevantes, es decir, este sistema alcanza un 90 % de precisión. En CORPES XXI, se recuperan 56 concordancias, de las cuales 52 son n-gramas relevantes (92,9 % de precisión). Entre los n-gramas irrelevantes nuevamente encontramos casos con los elementos del bigrama (en este caso, *al pelo*) en sintagmas diferentes, e incluso en orden inverso:

9. ¡Te dejan ir al trabajo con el pelo largo! ¡Wow! ¡Aparentemente! Todavía no me han dicho lo contrario.
10. Durante la nueva fase experimentó alguna horripilación, pero de la misma manera que los mutilados sienten a veces dolor en el pie amputado, él sentía el cosquilleo a ocho o diez centímetros de su cuero cabelludo, en el extremo de unos largos pelos inexistentes, sin que sus amigos lo advirtieran. En esta ocasión Manena Abad se mostró menos entusiasmada: "No te va el pelo al rape; te hace cara de bilorro".

Así, con las UF *ir/venir al pelo* observamos una considerable mejora en las versiones anotadas del CREA (con una precisión del 90 %) y del CORPES XXI (92,9 %) en contraste con la versión tradicional del CREA (41,6 %). Esta notable diferencia se debe principalmente al hecho de que las versiones anotadas permiten la búsqueda por lema de los verbos irregulares *ir* y *venir*¸ lo que evita el elevado ruido documental de la versión tradicional del CREA al tener que emplear el patrón "(v*) O (i*) O (f*) O (y*) dist/3 al pelo".

4.3. Corpus de Mark Davies (BYU)

En este apartado analizamos el rendimiento del sistema de gestión de corpus creado por Mark Davies en la BYU. Para ello, vamos a realizar diversas búsquedas para las UF seleccionadas en el Corpus del Español, uno de los corpus que se alojan en dicha plataforma de gestión y consulta de corpus. Mediante la opción *Lista* se permite la búsqueda por lemas (poniendo la palabra en mayúsculas) y categorías gramaticales específicas. Asimismo, admite el uso de comodines y de operadores como "|" a fin de obtener de forma simultánea las distintas variantes de una UF. Esta opción no permite sin embargo especificar el número de *tokens* que puede dividir la UF, por lo que sería necesario emplear el operador "*" por cada uno de esos elementos divisores y realizar tantas búsquedas como *tokens* puedan aparecer en esa secuencia. Así, el patrón para la forma continua de *tener/traer/llevar entre manos* debería ser "TENER|TRAER|LLEVAR entre manos". No obstante, al probar este patrón (y sus variantes "TENER|LLEVAR|TRAER entre manos" y "TRAER|TENER|LLEVAR entre manos"), comprobamos que solo se recupera el elemento intermedio entre las alternativas (así en el patrón "TENER|TRAER|LLEVAR entre manos" solo se recuperan concordancias con "traer entre manos") y que no se recupera en toda la flexión gramatical (sino solo en infinitivo). Por ello, sería necesario realizar 3 búsquedas por separado con los patrones "TENER_v entre manos", "TRAER_v entre manos" y "LLEVAR_v entre manos". Mayores problemas presentaría aún en caso de querer encontrar, en esta opción de *Lista*, las UF divididas por hasta 3 *tokens*, dado que el número de búsquedas se cuadruplicaría.

Por ello, a fin de obtener la forma discontinua de estas secuencias, deberemos recurrir a la opción *Colocados*, la cual permite establecer el número de elementos que pueden dividir la UF. Así, deberemos introducir "entre manos" en la casilla *Palabra/frase* y "TENER_V" en la casilla *Colocados* a una distancia hacia la izquierda de hasta 4 *tokens*. Como podemos observar en la Figura 3, una de las ventajas que presenta este sistema de gestión de corpus es que permite visualizar de forma estadística cuáles son las formas verbales con mayor frecuencia de uso (*tiene*, *tienen* y *tenemos*, en el caso de *tener entre manos*), lo que podrá ser de gran utilidad especialmente cuando se desee comprobar si en una determinada UF el verbo presenta una flexión defectiva, o bien se persiga conseguir una traducción más idiomática y natural, que tenga en cuenta las preferencias de uso en la lengua meta.

Así, con esta búsqueda para *tener entre manos*, obtenemos 4267 resultados, de los cuales 4105 se corresponden con n-gramas relevantes (96,2 % de precisión). Entre los n-gramas irrelevantes encontramos casos con una interpretación literal de la secuencia:

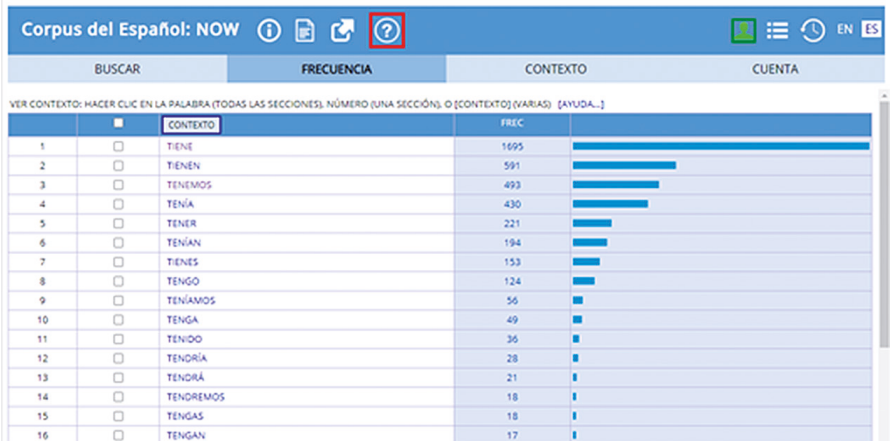

Figura 3. Consulta de la UF *tener entre manos* en el Corpus del Español

11. El globo de oro que tenía la Virgen entre manos se desvaneció y sus brazos se extendieron abiertos, mientras los rayos de luz continuaban cayendo sobre el globo blanco de los pies.
12. A una semana para que la nueva entrega de la popular franquicia llegue a las tiendas, se han marcado un espectacular vídeo de presentación que a buen seguro provocará que aumenten vuestras ganas de tener lo ya entre manos.

Tras esto, repetiremos el mismo proceso por separado para las restantes variantes. Con *traer entre manos*, obtenemos 1184 resultados, de los cuales solo 6 son n-gramas irrelevantes (por tanto, el sistema ha alcanzado una precisión del 99,5 %), que nuevamente se corresponden con una lectura literal de la secuencia, como las siguientes concordancias:

13. Recuerdo que cuando coincidíamos en una cafetería de la Calle de La Paz por el rumbo de San Ángel, siempre lo veía estudiando. Platicábamos brevemente. Y de inmediato regresaba a el texto que traía entre manos. Una ocasión vi que el ejemplar que cargaba estaba escrito en francés.
14. Jorge Vega, más conocido como Veguita, quien era considerado como el " gran bibliotecario de la prensa nacional ", falleció hoy domingo a los 77 años de edad, tras instruir a varias generaciones de periodistas ofreciendo los libros que siempre traía entre manos. # "

Finalmente, para *llevar entre manos,* se recuperan 172 concordancias, con solo dos n-gramas irrelevantes (98,8 % de precisión):

15. La pequeña está desilusionada, esperaba otra cosa. Frente a el agua, con la cabeza agachada, pregunta a su madre por los peces. "¿Dónde están?" Sólo conseguirá verlos tras desmenuzar parte de la barra de pan que lleva entre manos y arrojar los pequeños trozos a el agua.
16. Si la mitad no estuviera embargada, a día de hoy Jaume Matas podría conseguir hasta 1.23 millones de euros, dado que el precio cuadrado en Palma se paga a 1.876 euros. # Aunque no

> sería exactamente ese el dinero que se llevaría entre manos, ya que en su interior él y su mujer escondían una gran cantidad de posesiones a coste de oro.

Así, el cómputo global de resultados obtenidos para las variantes *tener/llevar/traer entre manos* es de 5623, de los cuales 5453 son n-gramas relevantes, por lo que la precisión media final para estas tres variantes asciende a un 97 %.

A continuación, analizaremos el rendimiento de este sistema de gestión de corpus con las variantes *ir/venir al pelo*, para las cuales será nuevamente necesario realizar dos búsquedas independientes. Lo primero que percibimos es que, al buscar ambas variantes, solo obtenemos resultados para *venir al pelo*, en concreto cinco n-gramas relevantes, por lo que el sistema ha ofrecido una precisión del 100 % para las escasas concordancias recuperadas. Esto se debe a que el resto de concordancias para ambas UF están compiladas sin el artículo contracto (por tanto, "a el pelo"), una construcción agramatical. Así, al buscar la secuencia "ir [3 *tokens*] a el pelo", ya sí obtenemos concordancias con esta UF, en concreto 46, de las cuales 40 son n-gramas relevantes (87 % de precisión). Para la secuencia "venir [3 *tokens*] a el pelo" se recuperan 590 concordancias, todas ellas n-gramas relevantes (100 % de precisión). Así, con estas 3 búsquedas ("ir a el pelo", "venir a el pelo" y "venir al pelo") se obtiene un total de 641 concordancias, de las cuales 635 son n-gramas relevantes, por lo que la precisión global asciende a 99,1 % para estas variantes.

4.4. Sketch Engine

El sistema de gestión de corpus Sketch Engine permite seis tipos principales de búsqueda, que presentan ventajas distintas según el grado de fijación de la UF:

a) *Simple*: trata cada uno de los elementos de la UF objeto de consulta como lema, por lo que los obtiene en toda su flexión gramatical. Es, por tanto, de especial utilidad para aquellas unidades en las que todos los elementos integrantes pueden sufrir flexión, como por ejemplo la UF *meter la nariz* (o *las narices*) *[en algo]*, con el significado de 'entrometerse [en ello]' (Seco y otros 2017: 551).
b) *Lemma:* permite obtener un determinado componente de la UF en toda su flexión gramatical, así como especificar su categoría gramatical, por lo que resulta de especial relevancia para refinar la búsqueda de UF como *cara conocida* ('Persona conocida', Seco y otros 2017: 135): a fin de evitar concordancias con el homófono *cara* como adjetivo, es posible buscar *cara* como sustantivo en la casilla *lema*, y en la casilla *context* buscar *conocida* también como *lema* un *token* a la izquierda de *cara*.
c) *Phrase*: recupera la expresión tal cual está introducida, de ahí que esta será la opción de búsqueda para obtener concordancias de UF con completa fijación, como por ejemplo *de cabo a rabo* ('del principio al fin, o totalmente', Seco y otros 2017: 112).
d) *Word*: obtiene concordancias con una palabra en la forma exacta en la que aparece introducida, así como también permite especificar la categoría gramatical de la misma. Este tipo de búsqueda podrá sernos de utilidad en UF bimembres como *plantar cara*, en la que uno de los componentes es invariable (*cara*) y, por tanto, podrá introducirse en la casilla de *Word* como

sustantivo y otro admite flexión gramatical (*plantar*) por lo que podrá buscarse en la casilla *context* a una distancia de hasta 4 *tokens* a la izquierda de *cara* (a fin de obtener también la forma discontinua de esta UF dividida por hasta 3 *tokens*).

e) *Character*: recupera *tokens* con un carácter o serie de caracteres específicos, lo que podrá ser de utilidad si se desea obtener UF con una secuencia fónica concreta y así crear un determinado efecto fonoestilístico en el texto meta (rima, paronomasia, aliteración…).

f) *CQL (Corpus Query Language):* es un lenguaje de búsqueda en corpus para la detección de patrones lexicogramaticales específicos con un nivel de precisión y exhaustividad mucho mayor que las opciones de búsqueda anteriormente descritas. Así, en Tabla 3 ofrecemos los esquemas CQL empleados para obtener la forma continua y discontinua de todas las variantes de UF analizadas en el presente artículo.

Secuencia	Esquema CQL
"tener/traer/llevar [0–3 tokens] entre manos"	[lemma="tener|traer|llevar" & tag = "V.*"][]{0,3}[word="entre" & tag = "S.*"][word="manos" & tag = "N.*"]
"ir/venir [0–3 tokens] al pelo"	[lemma="ir|venir" & tag = "V.*"][]{0,3}[word="al" & tag = "S.*"] [word="pelo" & tag = "N.*"]

Tabla 3. Esquemas CQL para las UF objeto de estudio

De esta manera, en los patrones CQL emplearemos el código *lemma* para aquellos elementos integrantes de la UF que puedan sufrir flexión gramatical y *word* para aquellos invariables (cfr. Hidalgo-Ternero 2020). Como indicábamos al comienzo de esta sección, hemos analizado el corpus esTenTen alojado en Sketch Engine. Para recuperar todas las posibles variantes, utilizaremos el operador "|" entre los elementos alternantes (por ejemplo, "tener|traer|llevar"). Asimismo, a fin de obtener concordancias con la UF en su forma continua y discontinua (dividida por hasta 3 *tokens*), se ha empleado el código CQL "[]{0,3}". Finalmente, también es posible especificar la categoría gramatical de cada uno de los componentes de la UF empleando el código *tag* seguido de la etiqueta correspondiente ("N.*" para sustantivo, "V.*" para verbo, "S.*" para preposición, etc.). A este respecto, conviene señalar que Sketch Engine ofrece la opción *CQL builder* para asistir al usuario en la creación del patrón según los parámetros que desee consultar (*tipo de token, distancia entre tokens, categoría gramatical*, etc.), lo que facilita enormemente el proceso de búsqueda dado que no es necesario saber de memoria ni introducir manualmente los distintos símbolos y códigos CQL.

Con el patrón "[lemma="tener|traer|llevar" & tag = "V.*"][]{0,3}[word="entre" & tag = "S.*"][word="manos" & tag = "N.*"]", obtenemos 25550 concordancias, de las cuales solo 173 son n-gramas irrelevantes (por tanto, Sketch Engine alcanza un 99,3 % de precisión). Entre las concordancias que no guardan relación con las UF, encontramos nuevamente casos con secuencias literales:

17. Cuando nos sentamos alrededor de una mesa larga de su departamento en París, Toni Negri, 84 años, tiene entre manos abundantes apuntes, mirada tensa, actitud exigente. </s><s> El resfrío que lo fastidia desde que regresó de un viaje a Brasil donde presentó Assembly, recién

publicado en inglés por Oxford University Press (cuarta parte de la investigación común escrita con el filósofo norteamericano Michael Hardt después de Imperio, Multitud y Común) lo tiene impaciente: "No logro trabajar como quisiera", dice.

18. Evidentemente cada uno tenemos nuestros gustos y preferencias, pero sí os animaría a, al menos, tener la edición entre manos. También os deseo que la disfrutéis, sea cual sea vuestra opción ;-). Buena compra. Aunque siendo sincero este steelbook en comparación con aquella preciosidad que salio para la primera pelí se queda muy pequeñito.

Con el patrón "[lemma="ir|venir" & tag = "V.*"][]{0,3}[word="al" & tag = "S.*"][word="pelo" & tag = "N.*"]" se recuperan 4358 resultados, de los cuales 56 son n-gramas irrelevantes, lo que supone una precisión del sistema del 98,7 %. Entre estos n-gramas irrelevantes, siguen predominando los casos con significado literal, especialmente en las concordancias con la secuencia en su forma discontinua:

19. Las pestañas que más trabajo y que más me demandan son las individuales. Me parecen la mejor opción porque se adaptan mejor a todos los tipos de ojos, no notas que las llevas puestas y además al ir pegadas al pelo su duración es mayor porque no toca la zona húmeda del ojo. ¿Dónde las compro? Desde hace años en la web americana Madame Madeline.
20. Se ha demostrado que la depilación láser va debilitando al pelo, de aquí que con el tiempo el mismo se caiga y no vuelva a crecer. Dentro de la depilación láser nos podemos encontrar diferentes tipos, los cuales pasamos a conocer a continuación.

5. Resultados y discusión

Una vez examinado el rendimiento de los distintos sistemas de gestión de corpus por separado, pasamos ahora a comparar sus resultados, que ofrecemos en la Tabla 4.

La frecuencia de aparición de las UF en los corpus se ofrece en términos absolutos (número total de n-gramas relevantes en el corpus) y normalizada (número total de n-gramas por millón de tokens en el corpus). En lo que respecta al número de n-gramas relevantes obtenidos para las UF objeto de consulta, se observa un claro contraste entre los corpus estudiados, fruto asimismo del número total de tokens que cada uno de ellos contiene. Por ello, se ofrece asimismo la frecuencia normalizada de estos corpus, la cual constituye una información esencial cuando se pretende contrastar corpus de distintos tamaños (Corpas Pastor 2018). Así, en lo que respecta a la forma continua y discontinua de las UF *tener entre manos*, *traer entre manos* y *llevar entre manos*, aquellos corpus que más n-gramas relevantes arrojaron fueron CORPES XXI (con 591 resultados), el Corpus del Español (con 5453) y Sketch Engine (con 25377). Por frecuencia normalizada, aquellos corpus que mejor rendimiento obtuvieron nuevamente fueron CORPES XXI (0,0019), el Corpus del Español (0,78) y Sketch Engine (1,45). Para las UF *ir al pelo* y *venir al pelo*, los corpus que más n-gramas relevantes recuperaron fueron CORPES XXI (con 52 resultados), el Corpus del Español (con 635) y Sketch Engine (con 4302). En cuanto a la frecuencia normalizada, ya sí observamos aquí una clasificación ligeramente distinta, con la versión anotada del CREA (0,00021), el Corpus del Español (0,091) y Sketch Engine (0,25).

Unidades fraseológicas	Sistema de gestión de corpus	N-gramas totales	N-gramas relevantes	Frecuencia normalizada	Precisión
llevar entre manos *tener entre manos* *traer entre manos*	CREA (tradicional)	280	276	0,0017	98,6 %
	CREA (anotado)	375	231	0,0018	61,6 %
	CORPES XXI	1161	591	0,0019	50,9 %
	Corpus del Español (NOW)	5623	5453*	0,78	97 %
	Sketch Engine (esTenTen)	25550	25377	1,45	99,3 %
ir al pelo *venir al pelo*	CREA (tradicional)	77	32	0,00020	41,6 %
	CREA (anotado)	30	27	0,00021	90 %
	CORPES XXI	56	52	0,00017	92,9 %
	Corpus del Español (NOW)	641	635**	0,091	99,1 %
	Sketch Engine (esTenTen)	4358	4302	0,25	98,7 %

Tabla 4. Rendimiento de los distintos sistemas de corpus objeto de estudio

* Dado que el Corpus del Español no permitía la búsqueda simultánea de las variantes *tener entre manos*, *llevar entre manos* y *traer entre manos*, los resultados que aquí se desglosan son el agregado de las 3 búsquedas realizadas por separado.

** Para estas variantes, con el Corpus del Español, nuevamente ha sido necesario realizar 3 búsquedas por separado ("venir al pelo", "venir a el pelo", "ir a el pelo"), de ahí que este resultado sea la suma global de esos 3 resultados.

No obstante, dada la premura que implica la labor traductora, a la hora de escoger un corpus u otro como herramienta documental, el número de resultados obtenidos (más allá de algunas decenas o cientos) no resulta un criterio tan determinante como el de la precisión, que pasamos a analizar a continuación. A este respecto, comprobamos que en la mayoría de los casos se ha obtenido un porcentaje igual o superior al 90 % a lo largo de los corpus, con solo tres excepciones. Para las UF *tener/traer/llevar entre manos*, la versión anotada del CREA y el CORPES XXI han alcanzado un 61,6 % y un 50,9 % de precisión, respectivamente. Este descenso se debe a que, como hemos observado anteriormente, la interfaz de ambos sistemas no permite la búsqueda del bigrama invariable *entre manos* en una misma casilla de consulta en la opción de *Proximidad*, por lo que se ha obtenido un elevado número de n-gramas irrelevantes con concordancias que bien incluyen algún tipo de determinante o adjetivo modificando a *manos* (lo que implica una interpretación literal de la secuencia) o bien contienen a *entre* y *manos* en sintagmas distintos, o incluso en un orden inverso. En cuanto a las UF *ir al pelo* y *venir al pelo*, la elevada irregularidad en el paradigma flexivo de *ir* unida a la ausencia de casilla *lema* en la versión tradicional del CREA hicieron necesario crear un patrón de búsqueda del verbo reducido hasta la primera letra (*v**, *i**, *f**, *y**), lo que elevó considerablemente el ruido documental y, por tanto, conllevó un descenso importante en la precisión del sistema hasta el 41,6 %.

En lo referente a los sistemas de gestión de corpus que han ofrecido un mejor rendimiento, en el caso de *tener/traer/llevar entre manos*, destacamos el Corpus del Español (con un 97 %), la versión tradicional del CREA (con un 98,6 %) y Sketch Engine (con un 99,3 %). A este respecto, conviene añadir que tanto el Corpus del Español como la versión tradicional del CREA han requerido de un mayor esfuerzo heurístico a fin de obtener esa elevada precisión: mientras que en el Corpus

del Español ha sido necesario realizar 3 búsquedas por separado para cada una de esas variantes, en la versión tradicional del CREA se ha precisado analizar todo el paradigma flexivo de cada uno de los verbos a fin de poder establecer el patrón de búsqueda: "(ten*) O (tien*) O (tuv*) O (tra*) O (llev*) dist/3 entre manos". En Sketch Engine, por su parte, solo fue necesario saber qué elementos de la UF permitían flexión y cuáles no, así como la categoría gramatical de los mismos.

Para las UF *ir/venir al pelo*, los sistemas de gestión de corpus que mejor rendimiento proporcionaron han sido la versión anotada del CREA (con un 90 %) y la del CORPES XXI (92,9 %), Sketch Engine (98,7 %) y el Corpus del Español (99,1 %). Aquí conviene también señalar que Sketch Engine ha sido el sistema que, si bien ha quedado segundo en precisión, nuevamente ha ofrecido un menor esfuerzo heurístico, dado que ha requerido un único patrón de búsqueda. En este sentido, si bien solo existían aquí dos variantes (*ir al pelo* y *venir al pelo*), el Corpus del Español ha requerido realizar 4 búsquedas, en este caso dado que, además de "ir al pelo" y "venir al pelo", fue necesario introducir las secuencias agramaticales (y, por tanto, no intuitivas) "ir a el pelo" y "venir a el pelo" a fin de aumentar considerablemente la exhaustividad. En lo concerniente a las versiones anotadas del CREA y del CORPES XXI, fue necesario introducir 2 patrones distintos ("ir [intervalo de 3 *tokens*] al pelo" y "venir [intervalo de 3 *tokens*] al pelo") a fin de alcanzar esa elevada precisión. Por todo ello, comprobamos que Sketch Engine es nuevamente el que mejor rendimiento ofrece tanto en términos cuantitativos como cualitativos (en relación precisión/esfuerzo heurístico) ante el desafío de la variación fraseológica.

6. Conclusiones

Los resultados obtenidos en el presente estudio han permitido vislumbrar cómo Sketch Engine es el sistema que ha ofrecido un mejor rendimiento en esta ardua labor de recuperar toda la información necesaria sobre las variantes de una UF con el menor número de búsquedas posible y evitando, al mismo tiempo, un elevado ruido documental. Estos prometedores resultados invitan a continuar analizando el rendimiento de estos sistemas ante el desafío que puede presentar, por ejemplo, la consulta de variantes diatópicas, como *agarrar/coger con las manos en la masa*; diastráticas, como *tocar las narices/los cojones*; diafásicas, como *sentidísima condolencia* o *sentido pésame*, o incluso diacrónicas, como *poner cual digan dueñas* y *poner verde*.

Asimismo, Sketch Engine es el sistema de gestión de corpus que ofrece un mayor número de opciones de búsqueda, desde aquellas que pueden resultar de mayor facilidad de empleo para el usuario no especialista (*Simple, Lemma, Phrase, Word, Character*) hasta aquellas que pueden requerir unos conocimientos más avanzados en el uso de corpus (*Corpus Query Language*, *CQL*). A fin de salvar este obstáculo, como hemos indicado anteriormente, Sketch Engine ofrece la opción *CQL builder* para asistir al usuario en la creación del patrón según los parámetros que desee consultar (*tipo de token, distancia entre tokens, categoría gramatical*, etc.), lo que facilita enormemente el proceso de búsqueda dado que no es necesario conocer de memoria ni introducir manualmente los distintos símbolos y códigos CQL, de ahí que también puede ser empleado por el traductor novel.

Por otro lado, el presente estudio pone asimismo de manifiesto el largo camino que aún queda por recorrer para un tratamiento lexicográfico exhaustivo de la fraseología en todas sus manifes-

taciones (variación, discontinuidad…). Así, en los diccionarios analizados, solo se han detectado instancias de las variantes *tener entre manos* (con una frecuencia normalizada de 0,92 por millón y con un porcentaje de concordancias discontinuas del 20,8 % en el esTenTen), *traer entre manos* (0,27 por millón, 14,6 % de discontinuidad) y *venir al pelo* (0,18 por millón, 25,7 % de discontinuidad), no quedando, por tanto, registradas ni las variantes *llevar entre manos* (0,07 por millón, 8,8 % de discontinuidad) ni *ir al pelo* (0,03 por millón, 11,7 % de discontinuidad). En lo que respecta a esa discontinuidad, se ha podido comprobar como solo se ofrece un ejemplo de estas variantes en sus formas discontinuas en los diccionarios *Wordreference* (*Por su mirada, puedo ver que Tomás trae un plan malvado entre manos*) y Collins (*Un tema que viene muy al pelo en esta discusión*). De esta manera, debido a la exclusión, en el proceso de repertorización lexicográfica, de las variantes con menor frecuencia de uso, así como a la escasez de una mayor ejemplificación de las UF en sus formas discontinuas, la elección del traductor se ve enormemente limitada y condicionada al consultar los diccionarios; lo que, por ende, contribuye a perpetuar los rasgos de simplificación y normalización en lo que respecta a la fraseología de la lengua meta detectados en anteriores estudios.

En este contexto, consideramos que una posible solución a estas limitaciones detectadas en el tratamiento lexicográfico de la variación fraseológica sería el diseño de diccionarios híbridos de última generación que permitan a los traductores el acceso directo a corpus de calidad a fin de poder analizar en contextos de uso real las posibles variantes de una UF. Resulta, por tanto, de vital importancia promover, en la práctica lexicográfica, el desarrollo de enfoques eclécticos que permitan combinar lo mejor de ambos mundos: la concisión y claridad expositiva de los diccionarios junto con la exhaustividad y riqueza textual de los corpus.

Referencias bibliográficas

Anastasiou, D. (2011). *Idiom treatment experiments in machine translation*. Newcastle upon Tyne: Cambridge Scholars.

Cambridge University Press (n. d.). *Cambridge Dictionary.* <https://dictionary.cambridge.org/>. Fecha de última actualización: 2021.

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

————. (2015). Translating English Verbal Collocations into Spanish: on Distribution and other Relevant Differences related to Diatopic Variation. *Lingvisticae Investigationes Special Issue 'Spanish Phraseology. Varieties and variations'*, 38, 2, 229–262. <doi: https://doi.org/10.1075/li.38.2.03cor>.

————. (2017). Collocational Constructions in Translated Spanish: What Corpora Reveal. In R. Mitkov (Ed.), *Computational and Corpus-Based Phraseology*. *Europhras 2017* (pp. 29–40). Berlin: Springer. doi: <https://doi.org/10.1007/978–3-319–69805–2>.

————. (2018). Laughing one's head off in Spanish subtitles: a corpus-based study on diatopic variation and its consequences for translation. En P. Mogorrón Huerta, & A. Albadalejo-Martínez (Eds.), *Fraseología, Diatopía y Traducción, Colección "IVITRA Research in Linguistics and Literature*, (pp. 54–106). Amsterdam: John Benjamins.

Corpas Pastor, G.; & Mena Martínez, F. (2003). Aproximación a la variabilidad fraseológica de las lenguas alemana, inglesa y española. *ELUA*, 17, 181–201.

[CORPES XXI] Real Academia Española. *Corpus del Español del Siglo XXI (CORPES XXI)*. <https://www.rae.es/banco-de-datos/corpes-xxi/>.

[CREA] Real Academia Española. *Corpus de Referencia del Español Actual (CREA)*. <https://www.rae.es/banco-de-datos/crea>.

Dal Maso, E. (2020). Sinonimia y variación léxica en la fraseología española e italiana: propuesta para un diccionario bilingüe bidireccional en línea. In M. De Beni (Ed.), *Representación de la fraseología en herramientas digitales: problemas, avances, propuestas*, *Círculo de Lingüística Aplicada a la Comunicación* 82 (pp. 27–40). <doi: https://dx.doi.org/10.5209/clac.689561>.

Davies, M. (2018). *Corpus del Español: NOW*. <http://www.corpusdelespanol.org/>.

[DLE] Real Academia Española (n. d.). *Diccionario de la Lengua Española.* <http://www.rae.es>.

Durán Muñoz, I.; & Corpas Pastor, G. (2020). Corpus-based multilingual lexicographic resources for translators: an overview. In M. J. Domínguez Vázquez, M. Mirazo Balsa, & C. Válcarcel Riveiro (Eds.), *Studies on Multilingual Lexicography, (Lexicographica, Series Maior)* (pp. 159–178). Berlin: De Gruyter. <doi: https://doi.org/10.1515/9783110607659–009>.

García-Page, M. (2008). *Introducción a la fraseología española*. Barcelona: Anthropos.

Gréciano, G. (1987). Idiom und Text. *Deutsche Sprache*, 15, 193–208.

HarperCollins Publishers (n. d.). *Collins Dictionary.* <https://www.collinsdictionary.com/>.

Hidalgo-Ternero, C. M. (2020). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. In P. Mogorrón Huerta (Ed.), *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue* 6, 154–177

Kellogg, M. (n. d.) *Wordreference*. <http://www.Wordreference.com/>.

Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P.; & Suchomel, V. (2003). *The Sketch Engine*. <https://www.sketchengine.eu>.

Koike, K. (2007). Relaciones paradigmáticas y sintagmáticas de las locuciones verbales en español. In J. Cuartero Otal, & M. Emsel (Eds.). *Vernetzungen Bedeutung in Wort, Satz und Text. Festschrift für Gerd Wotjak zum 65. Geburtstag* (pp. 263–275). Frankfurt am Main: Peter Lang.

Lawick, H. van. (2007). Phraseologie und Übersetzung unter Anwendung von Parallelkorpora. In M. Schlesinger, & R. Stolze (Eds.), *Translation Studies. Doubts and Directions*. (pp. 281–296). Amsterdam y Philadelphia: John Benjamins.

Leppihalme, R. (2000). Päätalo Idioms and Catchphrases in Translation. In P. Jauhola, O. Järvi, & D. Wilske (Eds.), *Erikoiskielet ja käännösteoria/LSP and Theory of Translation. VAKKI-symposium XX. Vaasa: University of Vaasa 11, 27*, 224–234.

Marco Borillo, J. (2009). Normalisation and the translation of phraseology in the COVALT corpus, *Meta*, 54, 4, 842–856.

Marco Borillo, J. (2011) Some Insights into the Factors Underlying the Translation of Phraseology in the COVALT corpus. In P. Kujamäki, L. Kolehmainen, E. Penttilä, & H. Kemppanen (Eds.), *Beyond Borders – Translations Moving Languages, Literatures and Cultures*. (pp. 197–214). Berlin: Frank & Timme.

Mellado Blanco, C. (2004). *Fraseologismos somáticos del alemán*. Frankfurt am Main: Peter Lang.

Montoro del Arco, E. T. (2005). Hacia una sistematización de la variabilidad fraseológica. In M. A. Pastor Millán (coord.), *Estudios lingüísticos en recuerdo del profesor Juan Martínez Marín*. (pp. 125–152). Granada: Universidad de Granada.

Oxford University Press (n. d.) *Oxford Spanish Dictionary.* <https://www.lexico.com/es>.

Philip, G. (2008). Reassessing the Canon: 'Fixed' Phrases in General Reference Corpora. In S. Granger, & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective*. (pp. 95–108). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Seco, M.; Ramos, G.; & Andrés, O. (2017). *Diccionario fraseológico documentado del español actual, locuciones y modismos españoles (2ª edición)*. Madrid: Aguilar.

Valencia Giraldo, M. V.; & Corpas Pastor, G. (2019). The Portrait of Dorian Gray: A Corpus-Based Analysis of Translated Verb + Noun (Object) Collocations in Peninsular and Colombian Spanish. In G. Corpas Pastor, & R. Mitkov (Eds.): *Computational and Corpus-Based Phraseology. Europhras 2019* (pp. 417–430). Cham: Springer Nature Switzerland AG. <doi: https://doi.org/10.1007/978–3-030–30135–4_13>.

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