CÓMO DOMINAR LA FRASEOLOGÍA Y AUTOMATIZAR EL PROCESO DE DOCUMENTACIÓN: UNA SOLUCIÓN TECNOLÓGICA PARA LA FORMACIÓN DE INTÉRPRETES EN LA COMBINACIÓN ESPAÑOL<>ÁRABE

Mahmoud Gaber

Universidad de Málaga, Departamento de Traducción e Interpretación,
Avda. Cervantes, 2. 29071 Málaga, España

Universidad de Málaga, IUITLM, C/ Severo Ochoa, 4,
Parque Tecnológico de Andalucía, 29071 Málaga, España

mahmoudgaber@uma.es

How to master phraseology and automate the documentation process: a technological solution for interpreter training in the Spanish<>Arabic combination

Abstract: Interpreting technologies are becoming more and more widespread thanks to the advances in Artificial Intelligence, the changes in the work environment caused by the COVID-19 pandemic, commercial interests, etc. However, there is a certain scarcity, in terms of quantity and quality, of technologies available to professional interpreters. This scarcity is even more acute in the field of interpreting training, especially in under-resourced languages, as in the case of Arabic. This study aims to: i) fill the existing gaps in interpreting training by introducing a methodology based on technological solutions for the documentation and mastery of phraseology from comparable corpora (spoken and written text), and ii) introduce a contrastive analysis of phraseology between Arabic and Spanish. In order to achieve our objectives, the design protocol and compilation process of the comparable corpora shall be described in detail. The corpus exploitation method used to carry out the documentation and phraseology extraction previous to the interpretation task is also illustrated. The results of the study reveal several advantages of our methodology for the trainee interpreter during documentation and preparation on a specialised topic (cybersecurity). Moreover, the study illustrates the most frequent collocations in both Arabic and Spanish subcorpora for the topic we have selected.

Keywords: semi-automatic corpus compilation; phraseology extraction: interpreting technologies; automatic speech recognition; conference interpreting

Resumen: Las tecnologías de la interpretación se están extendiendo cada vez más gracias a los avances en la inteligencia artificial, los cambios en los entornos de trabajo ocasionados

por la pandemia del COVID-19, los intereses comerciales, etc. No obstante, se nota cierta escasez, en términos de cantidad y de calidad, en las tecnologías disponibles para el uso de los intérpretes profesionales. Dicha escasez se acentúa aún más en el ámbito de la formación de intérpretes, sobre todo en lenguas de pocos recursos como el caso del árabe. Con ánimo de cubrir las lagunas existentes en la formación de intérpretes, este estudio tiene el doble objetivo de: i) establecer una metodología basada en soluciones tecnológicas para la documentación y el dominio de la fraseología a partir de corpus comparables (oral y textual), y ii) presentar un análisis contrastivo de patrones colocacionales entre el árabe y el español. Para lograr nuestros objetivos, se establecerá el protocolo de diseño y compilación del corpus comparable que nos servirá para la documentación y la extracción de las unidades fraseológicas y terminológicas y la adquisición del conocimiento temático. El método de explotación del corpus para la documentación y la extracción de la fraseología de cara a una tarea de interpretación se ilustra detalladamente en el presente estudio. Los resultados del estudio revelan varias ventajas que aporta nuestra metodología para el interprete en formación durante la fase de documentación y preparación sobre un ámbito especializado (ciberseguridad). Además, el estudio ilustra el porcentaje de frecuencia de colocaciones en el subcorpus de árabe y de español para el ámbito que hemos seleccionado.

Palabras clave: compilación semi-automática de corpus; extracción de fraseología; tecnologías de la interpretación; reconocimiento automático del habla; interpretación de conferencias

1. Introducción

Hoy día, es difícil imaginar un avance en el ámbito que sea sin contar con las Tecnologías de la Información y la Comunicación (TIC). En este contexto, la interpretación no es una excepción, ya que, según Sandrelli (2015), dichas tecnologías han tenido un impacto considerable tanto en la práctica profesional, como en la formación. De hecho, las tecnologías de la interpretación se están extendiendo cada vez más gracias a los avances de la inteligencia artificial, los cambios en los entornos de trabajo ocasionados por la pandemia del COVID-19, los intereses comerciales, entre otros factores (Gaber y Corpas Pastor 2022; en prensa). No obstante, Corpas Pastor (2022) apunta cierta escasez, en términos de cantidad y de calidad, de tecnologías disponibles para los intérpretes profesionales. Además, el estado del arte revela una situación nada halagüeña en lo que a las herramientas para la formación de intérpretes se refiere (*cf*. Gaber y Corpas Pastor 2022; en prensa). A modo de ejemplo, siguen existiendo carencias de corpus orales, herramientas de evaluación e interacción con el alumnado, soportes tecnológicos para la enseñanza de la interpretación de forma remota, entre otros. Además, la escasez de recursos tecnológicos es aún más acuciante en el caso de las lenguas de pocos recursos para la formación de intérpretes, como el árabe. A pesar de ser la lengua oficial de comunicación de 23 países y una de las seis lenguas oficiales de la Organización de las Naciones Unidas (ONU), la necesidad de desarrollar soluciones tecnológicas para la formación de intérpretes de la lengua árabe es más que notoria.

Con ánimo de cubrir las lagunas existentes en las herramientas tecnológicas de formación de intérpretes, sobre todo, en la combinación español<>árabe, este estudio tiene el doble objetivo de: i) establecer una metodología de documentación para

el intérprete en formación, y ii) presentar un análisis contrastivo de patrones colocacionales entre el árabe y el español. Dicha metodología consiste en compilar corpus comparables de forma semi-automática mediante transcripciones de discursos orales y de documentos escritos (compilación automática) sobre un mismo tema, facilitando así la documentación y la extracción de unidades fraseológicas al intérprete en formación, y, al mismo tiempo, realizar un análisis colocacional, aprovechando los corpus compilados.

Para lograr nuestro objetivo, la Sección 2 del presente estudio revisa la importancia de la documentación con corpus y el dominio de la fraseología para la preparación y la formación en interpretación. A continuación, se presentan los pasos de la metodología (Sección 3) que nuestro trabajo pretende ofrecer al intérprete en información. Para ello, se procede a describir los corpus comparables que hemos creado de forma semi-automática mediante RAH y el sistema de gestión de corpus Sketch Engine, así como los criterios de diseño y procedimiento de compilación. En la Sección 4, se presenta un análisis de patrones colocacionales para los subcorpus comparables árabe y español. Por último, la Sección 5 refleja las conclusiones del presente estudio.

2. El corpus como fuente de documentación: dominio de la fraseología y adquisición del conocimiento temático

La interpretación de por sí es una actividad muy compleja, y el uso de la fraseología y de la terminología especializada aumenta sobremanera la carga cognitiva del intérprete durante el proceso de la interpretación, sobre todo cuando estamos hablando de dos lenguas distantes en términos morfológicos, sintácticos y de origen como el español y el árabe. De hecho, los oradores a menudo emplean en sus discursos, sea cual sea el ámbito, unidades fraseológicas: locuciones, fórmulas, refranes, citas de obras literarias, expresiones rutinarias y ocasionales, etc. (Markič 2012). El conocimiento lingüístico que todo intérprete debe adquirir abarca, principalmente, la terminología especializada, así como la fraseología específica de un campo, que suelen emplear los expertos para intercambiar información (Fantinouli 2017). Cattaneo (2004) señala que los intérpretes gestionan mejor las expresiones idiomáticas cuando estas gozan de cierta «transparencia» o cuando los propios intérpretes están familiarizados con su uso. Sin embargo, cuando dichas expresiones son más «opacas», la reproducción de las expresiones idiomáticas viene asociada con mayor porcentaje de errores, omisiones e, incluso, problemas en la reproducción de los fragmentos del discurso que anteceden y preceden la expresión idiomática en cuestión *(ibid)*.

Pero la preparación no se limita únicamente al dominio de la fraseología y la terminología, sino también a la adquisición del conocimiento temático. El estado del arte considera la preparación una de las fases más importantes de cara a un encargo de interpretación, sobre todo cuando el tema a interpretar es muy especializado (Fantinouli 2017). De hecho, la preparación previa a la interpretación viene siendo una de las técnicas a la que recurren los intérpretes para reducir la carga cognitiva. Además, varios son los estudios que apuntan que el proceso de preparación y documentación influye, positivamente, en la calidad de la

interpretación (Gile 1995/2009; Fantinouli 2006, 2017; Seghiri 2017; Arce Romeral y Seghiri 2018; Pérez-Pérez 2018; Xu 2018, entre otros). Lamentablemente, dicha preparación, según Xu (2018), suele ocurrir bajo una considerable presión de tiempo. Adicionalmente, los intérpretes tienen que lidiar con una gran variedad de dominios en los que hace falta siempre una preparación avanzada. Sin embargo, los intérpretes no tienen este proceso de preparación sistematizado (Fantinouli 2017) y, por consiguiente, suelen recurrir a métodos, más o menos, manuales y tradicionales.

Así pues, el conocimiento temático, el dominio de la fraseología y la terminología especializada son clave en este proceso de preparación. Esto ayudará al intérprete a la recuperación inmediata de la información necesaria en el momento de la interpretación, y, por consiguiente, a la reducción de la carga cognitiva que, a su vez, le permite liberar espacio de tiempo a otros aspectos que requieren su atención (Nolan 2005; Tolosa-Igualada y Mezcua 2010; Markič 2012; Crezee y Grant 2013). Por ello, los estudios sobre la preparación basada en corpus (en inglés, *Corpus Driven Interpreters Preparation*) revelan una mejora en el rendimiento de los intérpretes, sobre todo, en dominios de conocimiento especializados (Fantinouli 2006; Bale 2013; Gallego Hernández y Tolosa 2012; Sánchez Ramos 2017; Pérez-Pérez 2018; Xu 2018). Por ello, la preparación terminológica basada en corpus textuales ha ido ganando terreno entre los intérpretes tanto profesionales como en formación (Pérez Pérez 2018; Xu 2018; Fantinuoli y Prandi 2018; Arce Romeral y Seghiri 2018, etc.). Por otro lado, varios autores han desarrollado métodos para llevar a cabo la preparación fraseológica basada en corpus (en este caso, orales) a los efectos de mejorar la competencia fraseológica de los intérpretes (Aston 2015; Corpas Pastor y Gaber 2021).

Dicho esto, se constata que los estudios de la traducción basados en corpus gozan de un mayor grado de desarrollo que los de interpretación (Shlesinger 1998; Bendazzoli y Sandrelli 2009; Bale 2013; Sánchez Ramos 2017). Además, se ha observado que los estudios de interpretación basados en corpus han sido canalizados, por lo general, en dos líneas:[1] a) el análisis del proceso de interpretación (estrategias, técnicas, carga cognitiva, comportamiento e interacción del intérprete, etc.) y la evaluación del producto final, comparando el discurso original con el discurso interpretado; y b) el uso del corpus como recurso de documentación para los intérpretes, basándose, según Corpas Pastor (2018), en la compilación de documentos escritos (es decir, el lenguaje escrito). Curiosamente, teniendo en cuenta las diferencias entre el lenguaje hablado y el escrito,[2] los intérpretes, según Corpas Pastor y Gaber (2021), procuran contar con material audiovisual en la fase de documentación para familiarizarse, además de otros aspectos, con el acento, la voz del orador que habrán de interpretar, así como las unidades fraseológicas. Todos estos factores demuestran la importancia de contar con corpus orales para la fase de documentación por parte de los intérpretes. No obstante, y a pesar de los beneficios que contienen los corpus orales, estos requieren un proceso clave para poder analizar su información

[1] Para información más exhaustiva sobre las orientaciones y metodologías de los estudios de interpretación basados en corpus, véase: Bendazzoli y Sandrelli (2009) y Setton (2011).

[2] No nos concierne en este estudio analizar las diferencias entre el lenguaje hablado y escrito, para información más detallada, véase: Akinnaso (1982) y Halliday (1989).

y extraer la terminología en cuestión. Estamos hablando precisamente del proceso de la transcripción, que, según Bendazzoli y Sandrelli (2009), es uno de los desafíos que ralentizan el avance en los estudios de interpretación basados en corpus orales.

Llegados a este punto, conviene aportar una solución tecnológica capaz de garantizar una documentación equilibrada y holística, contando con corpus no solo textuales sino también orales. De esta forma, el intérprete tendrá garantías de dominar la fraseología escrita y hablada, ya que ambos corpus aportan patrones diferentes (*cf.* Corpas Pastor y Gaber 2021). Así pues, el presente estudio aprovecha –aparte del corpus textual– el corpus oral basado en la transcripción automática del habla para romper la barrera que dificulta la explotación de los corpus orales.

3. Metodología

Teniendo en cuenta el doble objetivo del estudio, la metodología protocolizada que se persigue beneficia, sobre todo, al interprete en formación, ya que se trata de un encargo simulado de interpretación. En líneas generales, los docentes solemos encargar a los alumnos que se preparen para el ejercicio a interpretar, facilitando solamente el título del tema objeto de interpretación y algunos términos claves. En realidad, esta práctica ayuda a los futuros intérpretes a familiarizarse con los posibles encargos, conocidos como «interpretación a ciegas». Es decir, «una situación en que el organizador/cliente informa al intérprete solamente del tema de la conferencia sin facilitar más información sobre los oradores ni el tipo del discurso, etc.» (Corpas Pastor y Gaber 2021).

En este caso, la simulación sería una interpretación simultánea o consecutiva en una conferencia sobre «ciberseguridad» entre el español y el árabe como lenguas del evento (direccionalidad hacia el árabe). Mediante la metodología propuesta, se persigue dotar al intérprete en formación de dos corpus comparables que le servirán de fuente de documentación para preparar el encargo correspondiente. Al mismo tiempo, la metodología propuesta fomentará la adquisición de buenas prácticas para su futuro profesional. Al introducir aspectos de Lingüística de corpus en nuestro enfoque tecno-pedagógico, estaremos facilitando la adquisición de competencias tecnológicas y digitales en una combinación lingüística donde uno de los idiomas (el árabe) cuenta con escaso desarrollo computacional. Se da la circunstancia, además, de que apenas hay estudios de fraseología contrastiva en este par de lenguas, y mucho menos basados en corpus, por lo que nuestro trabajo constituye una de las primeras contribuciones para el avance de esta disciplina.

3.1. Protocolo y procedimiento de compilación de los corpus

El primer subcorpus (CIBERCOR_ORAL) es un corpus *ad hoc,* compilado con base en transcripciones realizadas de forma automática a partir de discursos orales en español. El segundo es un subcorpus textual (CIBERCOR_ES) que ha sido compilado de forma automática a través del sistema Sketch Engine y que integra textos escritos y las transcripciones del subcorpus oral, anteriormente compilado. El tercero

(CIBERCOR_AR) será un subcorpus en lengua árabe compilado con base en una selección de términos clave que nos proporcionó el CIBERCOR_ES. A continuación, describimos detalladamente el procedimiento de compilación de los corpus.

3.1.1. CIBERCOR_ORAL: criterios de diseño

CIBERCOR_ORAL es un subcorpus *ad hoc* obtenido mediante la transcripción automática de discursos orales que sirve de referencia para la documentación de cara a un ejercicio/encargo de interpretación. Además, es una fuente de información altamente relevante para la preparación previa al encargo (real o simulado), que permite al intérprete en formación extraer terminología y fraseología y adquirir el conocimiento experto de forma semiautomática. Para la compilación del CIBERCOR_ORAL se ha realizado un proceso de selección del material audiovisual que integra dicho subcorpus de acuerdo con los siguientes criterios:

- *Temática:* se ha optado por la «ciberseguridad» teniendo en cuenta su presencia en la agenda de la Comisión Europea como una de las áreas estratégicas (Consejo Europeo 2022). Este factor hace que dicha temática sea el enfoque de debate en casi 1500 congresos internacionales a lo largo del 2022[3].
- *Material*: discursos orales, considerando que este tipo de material es lo que más interesa a los intérpretes en el proceso de la preparación.
- *Lengua*: el español peninsular es la lengua utilizada en los discursos orales que conforman este corpus.
- *Grado de especialización*: varía entre bajo, medio y alto, con el objetivo de abarcar diferentes niveles de información. Esto permite al intérprete en formación ir adquiriendo el conocimiento temático de forma gradual para que pueda sobrellevar y procesar la información tanto general como experta que tocará interpretar.
- *Contexto*: se han incluido seminarios, ponencias, entrevistas y discursos unidireccionales.
- *Tipo*: se trata de un corpus *ad hoc* monolingüe.
- *Autoría*: siendo material audiovisual, YouTube nos ha servido como puerta de acceso a gran cantidad de discursos de los cuales hemos realizado una selección minuciosa. Al final, se han seleccionado canales oficiales de instituciones públicas: Centro Criptológico Nacional, Aprendemos Juntos, UNIR, URJC, etc. (*cf.* Gaber y Copas Pastor 2022; en prensa)

3.1.2. CIBERCOR_ORAL: protocolo de compilación

Acto seguido, se ha procedido al protocolo de compilación, que se compone de las siguientes fases:

i. Búsqueda del material: la fase de búsqueda y selección nos ha permitido recopilar un total de siete vídeos que cumplen con los criterios establecidos. La extensión de los vídeos varía entre 01:01:05 para el vídeo más largo y 00:04:48 para el más corto. Como resultado, la duración total del material audiovisual que conforma el CIBERCOR_ORAL es de 191 minutos y 17 segundos.

[3] Véase <https://infosec-conferences.com/>.

ii. *Descarga:* se ha realizado la descarga de los vídeos que no tenían habilitada por defecto la opción de transcripción automática en su propio canal de YouTube. Posteriormente, se han subido dichos vídeos a nuestro canal de YouTube para proceder a su transcripción de forma automática.

iii. *Transcripción automática:* es la fase más importante para la compilación del corpus oral. Así, con base en esta tarea de transcripción automática se obtienen los textos que integran el CIBERCOR_ORAL. Por ello, habrá que contar con un sistema robusto en términos de precisión y de recuperación informacional. Teniendo este factor en cuenta, se ha realizado un estudio previo comparativo (*cf.* Gaber y Copas Pastor 2022; en prensa) para evaluar diferentes sistemas de transcripción automática. El resultado de dicho estudio indica que YouTube es el sistema que mejor rendimiento ha tenido entre otros siete sistemas que han sido examinados.

iv. *Almacenamiento y gestión de corpus:* una vez realizada la transcripción automática mediante YouTube, se han guardado las transcripciones de los discursos orales en formato .txt. Dicho formato nos facilita la incorporación de los textos en cualquier aplicación de gestión de corpus.

3.1.3. CIBERCOR_ES: procedimiento de compilación

En cuanto al subcorpus CIBERCOR_ES, se ha aprovechado Sketch Engine para compilarlo de forma automática, recogiendo textos, previa revisión y filtro, desde páginas web. Para esta tarea, el CIBERCOR_ORAL nos ha servido de referencia para identificar los términos simples y multipalabra más frecuentes para utilizarlos como secuencias de búsqueda *(seedwords)* para compilar el CIBERCOR_ES (*cf.* Fig. 1): *ciberseguridad, cibercrimen, ciberdelincuencia, seguridad informática, aplicación maliciosa, factor de autentificación, inteligencia artificial y ciberdelincuencia organizada*.

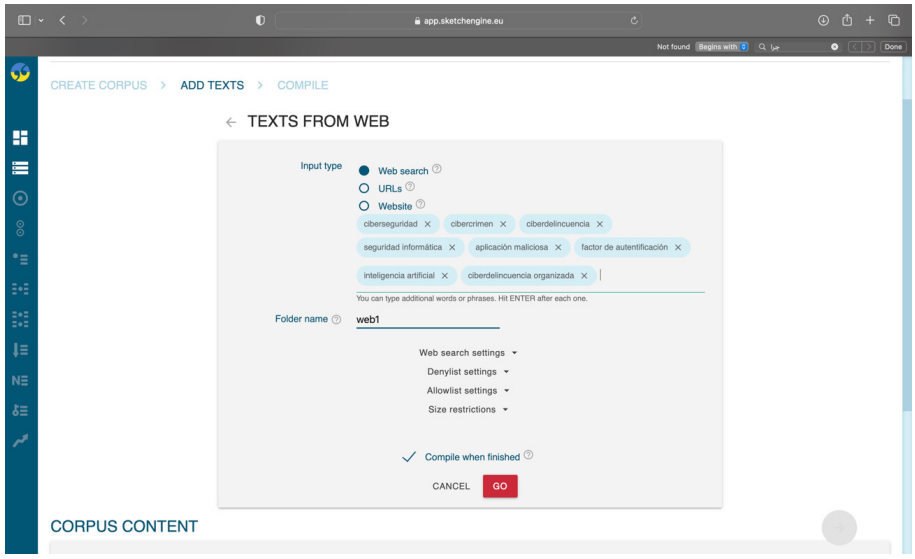

Fig. 1. Compilación del CIBERCOR_ES a través del sistema Sketch Engine

Ahora bien, somos conscientes de la importancia de tanto el corpus textual como el oral para una buena documentación previa a una interpretación por: i) los patrones diferentes que aporta cada uno (*cf*. Corpas Pastor y Gaber 2021); ii) el interés que tienen los intérpretes a recurrir al material audiovisual durante la fase de preparación *(ibid)* y iii) la diferencia entre el lenguaje escrito y hablado. Teniendo todos estos factores en cuenta, pensamos que sería muy útil fusionar tanto el CIBERCOR_ORAL como el CIBERCOR_ES en un único corpus. De esta manera, el intérprete tendrá la garantía de realizar una documentación equilibrada, holística y avanzada, contando con el corpus tanto textual como oral.

3.1.4 CIBERCOR_AR: procedimiento de compilación

Teniendo en cuenta que la metodología del presente estudio se enmarca en un encargo simulado de interpretación en la combinación español<>árabe, habrá que documentarse también con la ayuda de un corpus *ad hoc* en lengua de árabe. Por ello, se ha procedido a determinar la terminología clave que contiene el CIBERCOR_ES, que nos servirá como secuencias de búsqueda para crear el corpus árabe (CIBERCOR_AR). La terminología clave seleccionada ha sido traducida al árabe con el objetivo de introducirla de nuevo en el sistema Sketch Engine para compilar el CIBERCOR_AR. La Tabla 1 contiene la terminología simple y multipalabra en español y su correspondiente equivalente en árabe.

	Español	**Árabe**
1	ciberseguridad	الأمن السيبراني
2	ciberdelincuencia	جرائم الفضاء الإلكتروني
3	cibercrimen	جريمة سيبرانية
5	*phishing*	التصيد الاحتيالي
6	ciberdelincuentes	مجرمو الفضاء الإلكتروني
7	ciberataque	هجوم سيبراني
8	*malware*	البرامج الضارة
13	factor de autentificación	عامل المصادقة
15	aplicación maliciosa	برمجيات خبيثة

Tabla 1. Términos clave en español y su correspondiente equivalente en árabe

La Tabla 2 indica el tamaño de los dos corpus, según el número de ocurrencias *(tokens)*, así como el número de documentos que integran cada uno.

Corpus	**Documentos**	***Tokens***
CIBERCOR_ORAL	7	35,094
CIBERCOR_ES	192	514,385
CIBERCOR_AR	87	180,933

Tabla 2. *Tamaño de los corpus comparables*

4. Análisis colocacional

Una vez compilados los dos corpus comparables (CIBERCOR_ES y CIBERCOR_AR), el siguiente paso consiste en analizar las unidades fraseológicas que contiene

cada corpus. Si bien no se trata de un análisis meramente comparativo, teniendo en cuenta las diferencias entre ambos idiomas (árabe y español) y el tamaño de cada corpus, los datos proporcionan información ilustrativa y relevante sobre la frecuencia de patrones colocacionales en cada uno.

Para proceder con dicho análisis, nos valemos de los seis tipos de colocaciones que distingue Corpas Pastor (1996):

1. Sustantivo + Adjetivo
2. Adjetivo + Adverbio
3. Sustantivo + Preposición + Sustantivo
4. Verbo + Adverbio
5. Verbo + Sustantivo
6. Verbo + Preposición + Sustantivo

Acto seguido, realizaremos la búsqueda de los patrones colocacionales en ambos subcorpus a través de Sketch Engine. Cabe indicar que las etiquetas de búsqueda de patrones son distintas para cada lengua. La Tabla 3 nos indica las etiquetas que habrá que utilizar para el corpus de español.

N.º	Tipo de patrón	Etiqueta
1.	Sust + Adj	[tag=»N.*»] [tag=»A.*»]
2.	Adj + Adv	[tag=»A.*»][tag=»R.*»]
3.	Sust + Prep + Sust	[tag=»N.*»] [tag=»S.*»][tag=»N.*»]
4.	Verb + Adv	[tag=»V.*»][tag=»A.*»]
5.	Verb + Sust	[tag=»V.*»][tag=»N.*»]
6.	Verb + Prep + Sust	[tag=»V.*»][tag=»S.*»][tag=»N.*»]

Tabla 3. Etiquetas de búsqueda de colocaciones en español

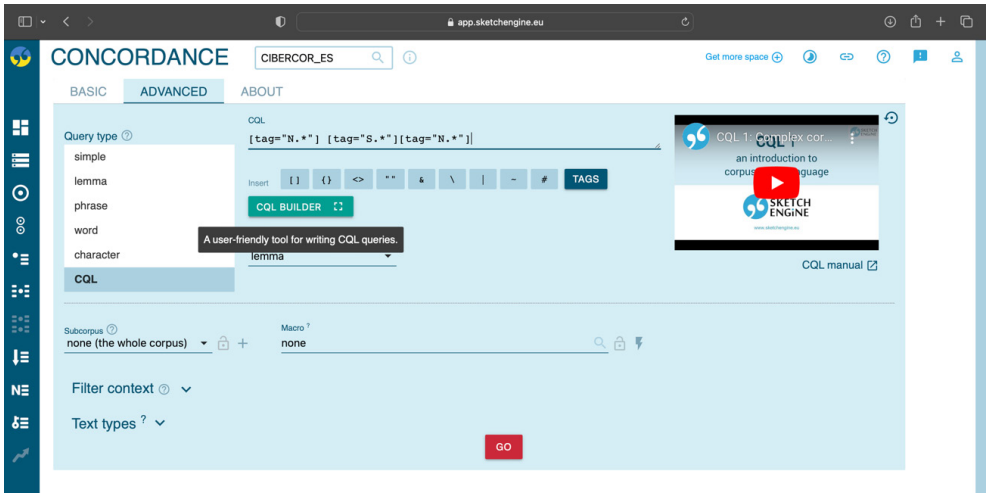

Fig. 2. Búsqueda de patrones en el Sketch Engine

De esta manera, hemos procedido a extraer los patrones en el subcorpus de español (CIBERCOR_ES). Este paso se ha llevado a cabo gracias a la funcionalidad de CQL (*Corpus Query Language*) que tiene Sketch Engine para la búsqueda de estructuras complejas (*cf.* Fig. 2).

A los efectos de extraer las apariciones de cada combinación, hemos tenido que repetir la búsqueda por cada patrón, asignando las etiquetas correspondientes según el caso. A la vista de los resultados, se destaca la combinación n.º 3 (Sust. + Prep. + Sust.) frente al resto de las combinaciones con un porcentaje del 38 % de apariciones, siendo la más utilizada en la temática de ciberseguridad que contiene el corpus *ah hoc* de lengua española que hemos compilado. A continuación, la Figura 3 resume los porcentajes de cada combinación en el CIBERCOR_ES.

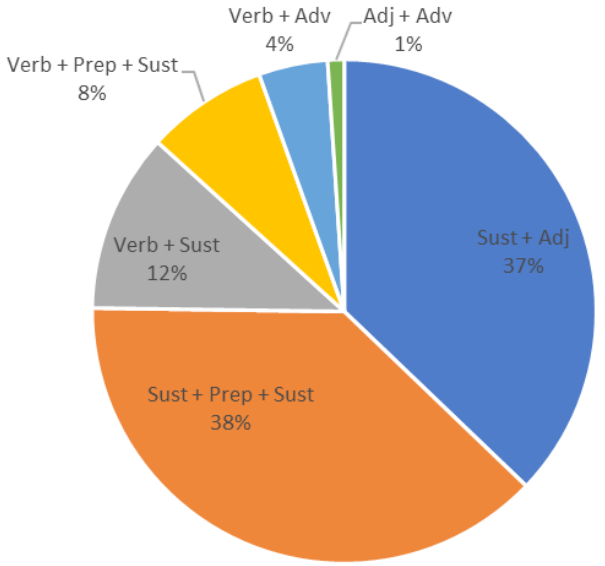

Fig. 3. Frecuencia de uso de colocaciones en el CIBERCOR_ES

N.º	Tipo de patrón	Etiqueta
1.	Sust + Adj	[tag=»(DT)?NN.*»] [tag=»(DT)?JJ.*»]
2.	Adj + Adv	[tag=»(DT)?JJ.*»] [tag=»W?RB»]
3.	Sust + Prep + Sust	[tag=»(DT)?NN.*»] [tag=»IN»] [tag=»(DT)?NN.*»]
4.	Verb + Adv	[tag=»VB.*»] [tag=»W?RB»]
5.	Verb + Sust	[tag=»VB.*»] [tag=»(DT)?NN.*»]
6.	Verb + Prep + Sust	[tag=»VB.*»] [tag=»IN»] [tag=»(DT)?NN.*»]

Tabla 4. Etiquetas de búsqueda de colocaciones en árabe

Este procedimiento, que determina las etiquetas de búsqueda de colocaciones, se ha repetido con el subcorpus de árabe (CIBERCOR_AR) para extraer los patrones más frecuentes en dicho subcorpus. Así pues, la Tabla 4 ilustra las etiquetas que hemos utilizado para la búsqueda en árabe.

En el caso del CIBERCOR_AR, se destaca la combinación n.º 1 (Sust. + Adj.) frente al resto de las combinaciones con un porcentaje del 43 % de apariciones, siendo la más utilizada en la temática de ciberseguridad que contiene el corpus *ah hoc* de lengua árabe que hemos compilado. La Figura 4 resume los porcentajes de cada combinación en el CIBERCOR_AR.

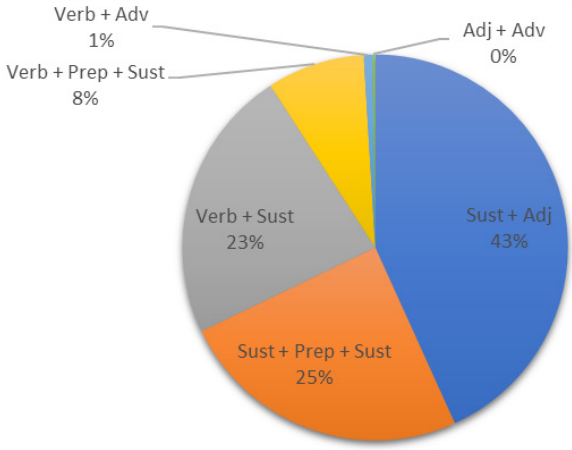

Fig. 4. Frecuencia de uso de colocaciones en el CIBERCOR_AR

A la luz de los datos que nos proporcionan las Figuras 3 y 4, se observa que existe una diferencia considerable en el porcentaje de apariciones en la combinación n.º 5 (Verb. + Sust.) en ambos subcorpus: 12 % para el subcorpus de árabe y 23 % para el subcorpus de español. No obstante, dicha diferencia sigue existiendo, pero en menor medida, en las otras combinaciones, a excepción de la combinación n.º 6 (Verb. + Prep. + Sust.) donde el porcentaje es igualado en ambos corpus: 8 %. Cabe señalar que las apariciones de la combinación n.º 2 (adj. + adv.) es casi nula en la lengua árabe, a rasgos generales, y en el corpus en cuestión en particular.

La Tabla 5 muestra una lista de equivalentes, partiendo del español hacia el árabe, de los patrones colocacionales más frecuentes localizados en los corpus analizados: Sust. + Prep. + Sust., Sust. + Adj. y Verb. + Sust.

Por otra parte, no solo hemos utilizado el CIBERCOR_ES para seleccionar la terminología clave para crear el CIBERCOR_AR, compilar el subcorpus comparable de árabe y analizar la frecuencia de uso de los patrones colocacionales, sino también para obtener una lista de términos simples y multipalabra en ambos subcorpus que gozan de alto grado de equivalencia. Esto, a su vez, nos facilita el proceso de búsqueda de equivalentes entre ambos idiomas para la lista de términos que se podrá generar automáticamente a los efectos de preparar y memorizar la terminología

antes del encargo de interpretación. La Tabla 6 ilustra una muestra de los primeros 25 términos simples y multipalabra del CIBERCOR_AR y los posibles equivalentes en el CIBERCOR_ES generados automáticamente mediante la funcionalidad de extracción de terminología en Sketch Engine.

Sust. + Prep. + Sust.			
Colocación (ES)	**Equivalente (AR)**	**Colocación (ES)**	**Equivalente (AR)**
vía de ataque	طريقة الهجوم	medida de protección	وسيلة الحماية
base de datos	قاعدة البيانات	fase de prevención	مرحلة الوقاية
alerta de seguridad	تنبيه أمني	robo de *tokens*	سرقة الرموز
certificado de validación	شهادة التحقق	manipulación de equipos	التلاعب بالأجهزة
factor de autentificación	عامل المصادقة	procesamiento de datos	معالجة البيانات
ataque a empresas	هجوم على الشركات	lenguaje de programación	لغة برمجة
fraude de pago	الاحتيال في الدفع	sistemas de navegación	أنظمة التصفح
sistema de delegación	نظام التفويض	robo de identidad	سرقة الهوية
clonación de tarjetas	استنساخ البطاقات	instalación de *malware*	تثبيت البرامج الضارة
vulnerabilidad de seguridad	ثغرة أمنية	investigación de ciberdelincuencia	التحقيق في الجرائم الالكترونية
intercepción de comunicaciones	اعتراض المراسلات	sabotaje de información	التخريب المعلوماتي
Sust. + Adj.			
seguridad nacional	الأمن القومي	delito informático	جريمة سيبرانية
contraseña insegura	كلمة مرور غير آمنة	huella digital	بصمة
sistema operativo	نظام التشغيل	ataque informático	هجوم سيبراني
aplicación maliciosa	تطبيق ضار	transformación digital	التحول الرقمي
inteligencia artificial	الذكاء الاصطناعي	analista forense	محلل الطب الشرعي
sociedad digital	المجتمع الرقمي	organización delictiva	منظمة إجرامية
seguridad informática	الأمن السيبراني	propiedad intelectual	الملكية الفكرية
Verb. + Sust.			
tomar medidas	اتخاذ التدابير	compartir contenido	مشاركة المحتوى
hacer *phishing*	القيام بالتصيد	mandar mensajes	إرسال رسائل
tener acceso	الولوج إلى	guardar información	حفظ المعلومات
robar datos	سرقة البيانات	diseñar tecnologías	تصميم تقنيات
distribuir *malware*	نشر البرامج الضارة	realizar ataques	القيام بهجمات

Tabla 5. Lista de ejemplos de los patrones colocacionales más frecuentes en español y sus equivalentes en árabe

	Términos simples (AR)	Términos simples (ES)	Términos multipalabra (AR)	Términos multipalabra (ES)
4				
5	السيبراني	ciberdelincuencia	الأمن السيبراني	seguridad informática
6	المخترق	ciberseguridad	أنظمة التشغيل الأمنة	doble factor
7	الفيروسات	cibercrimen	الذكاء الاصطناعي	delito informático
8	طروادة	autenticación	التوعية بالأمن السيبراني	factor de autenticación
9	فيروسات	*phishing*	المعلومات و الشبكات	aplicación maliciosa
10	الحاسوب	ciberdelincuentes	الحصول على شهادة	sistema informático
11	الاختراق	intrusión	مجال أمن المعلومات	inteligencia artificial
12	التشفير	acesso	أمن المعلومات	aprendizaje automático
13	انتحال	ciberdelito	المصرح به	pornografía infantil
14	الشبكات	ciberespacio	التطفل على أجهزة	delito cibernético
15	الدودة	autentificación	تخصص الأمن السيبراني	seguridad cibernética
16	المخترقين	*ransomware*	غير المصرح به	material pornográfico
17	السحابية	cibernético	أمن تكنولوجيا المعلومات	autenticación multifactor
18	الهكر	malicioso	سرقة بيانات	seguridad de la información
19	التصيد	*malware*	مجرمو الفضاء الإلكتروني	ataque cibernético
20	الترميز	*token*	جمعية المقرصنين	ciberdelincuencia organizada
21	الحوسبة	cibercriminales	سلامة و تكامل المعلومات	materia de ciberseguridad
22	الثغرات	interceptación	الدخول إلى جهاز	suplantación de identidad
23	الهجمات	ciberamenazas	الهجمات الالكترونية	autenticación de doble factor
24	العبث	informático	المعلومات و البيانات	software malicioso
25	الضارة	biométrico	عامل المصادقة	doble factor de autenticación
26	البرمجية	autenticador	المخترق المتطفل	robo de identidad
27	المقرصنين	ciberataque	انتحال شخصية	ataque informático
28	التطفل	multifactor	التحكم في الجهاز	pena privativa de libertad
29	المتطفلين	verificador	حماية الانظمة	programa malicioso
30	البيانات	suplantación	البنية التحتية	proceso de autenticación

Tabla. 6. Lista de posibles equivalentes entre el CIBERCOR_AR y el CIBERCOR_ES

Como se puede observar, todos los términos marcados en verde, que representan un número considerable, han encontrado su equivalente en el corpus comparable. A pesar de que los dos subcorpus son comparables y no son traducciones uno del otro, el simple hecho de contar con los términos clave como palabras de búsqueda para compilar el subcorpus de árabe nos ha facilitado un gran número de equivalencias. En este proceso de localización de equivalentes, se ha observado que, dado que muchos términos en español son compuestos, algunos ejemplos de los términos simples en el CIBERCOR_ES han encontrado su equivalente en términos multipalabra en el CIBERCOR_AR: ciberseguridad, cibercrimen, ciberdelincuente, ciberespacio / الأمن السيبراني، جريمة سيبرانية، مجرم الفضاء الالكتروني ، الفضاء الالكتروني, respectivamente.

Cabe subrayar que la extracción terminológica se utiliza para identificar una lista de términos especializados y unidades fraseológicas a partir de un corpus compilado para que así el intérprete pueda crear su glosario y empezar el proceso de aprendizaje (Fantinuoli 2017). La automatización de este proceso supone un ahorro de tiempo y de esfuerzo importantes para el intérprete durante el proceso de preparación y documentación.

5. Conclusiones

En el presente trabajo hemos ilustrado una metodología basada en la compilación de corpus comparables que facilitan al intérprete en formación el proceso de documentación y el dominio de la fraseología. La metodología ha hecho uso de un corpus oral

y otro textual para garantizar al intérprete una documentación avanzada y holística, familiarizándole con la terminología y fraseología usadas en la lengua hablada y escrita. Nos hemos valido del corpus en español para compilar un corpus en árabe que será de mucha utilidad para la preparación y la formación del intérprete. Además, teniendo interiorizada dicha metodología por parte del intérprete, se podrá agilizar el proceso de preparación, ahorrando tiempo y esfuerzo.

La metodología permite realizar un análisis colocacional que consideramos de mucha relevancia, ya que no hay casi ningún estudio de fraseología contrastiva entre el árabe y el español, y tampoco se cuenta con trabajos de fraseología contrastiva basados en corpus para esta combinación lingüística.

Este estudio contribuye, además, a promover el uso de los corpus orales, transcritos a través de la tecnología del reconocimiento automático del habla, evitando así las dificultades que implica la transcripción manual y el desafío de la extracción informacional. La metodología supone una mejora en la calidad de la interpretación, ya que, según Xu (2018), se observa mayor precisión terminológica por parte de los intérpretes cuando realizan la preparación contando con corpus frente a la preparación tradicional.

Además, se han establecido criterios de diseño y protocolo de compilación de un corpus *ad hoc* oral que podría ser de utilidad para futuros estudios.

Los pasos que hemos seguido para ilustrar el método de explotación del corpus comparable ayudarán al intérprete a sistematizar el proceso de documentación y extracción fraseológica, así como a desarrollar buenas prácticas para sus futuros encargos profesionales.

Por último, cabe subrayar que la metodología aquí presentada puede ser aplicada para la creación de otros corpus comparables (orales y textuales) que aborden otras temáticas.

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 (PID2020-112818GB-I00, PDC2021-121220-I00, UMA18-FEDERJA-067, D5-2021_03 y PIE22-135). Asimismo, ha sido subvencionada por el ministerio de ciencia, innovación y universidades (PIF Ref. BES-2017-082791, 2018). También, quisiera agradecer a la Dra. Corpas Pastor por sus valioso consejos, sugerencias e ideas y por las correcciones que ha realizado en este artículo. Igualmente, quisiera dar las gracias a los evaluadores anónimos por su gran esfuerzo y recomendaciones acertadas.

Referencias bibliográficas

Akinnaso, F. Niyi (1982), «On the differences between spoken and written language», *Language and speech* 25/2, 97-125.

Arce Romeral, Lorena – Seghiri, Míriam (2018), «Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters», *Current Trends in Translation Teaching and Learning E* 5, 1-46 [disponible en <http://www.cttl.org/uploads/5/2/4/3/5243866/cttl_e_2018_1.pdf>, 05/09/2022].

Aston, Guy (2015), «Learning phraseology from speech corpora», *Multiple affordances of language corpora for data-driven learning* 69, 65-84.

Bale, Richard (2013), «Undergraduate Consecutive Interpreting and Lexical Knowledge», *The Interpreter and Translator Trainer* 7/1, 27-50.

Bendazzoli, Claudio – Sandrelli, Annalisa (2009), «Corpus-based Interpreting Studies: Early work and future prospects», *Tradumàtica* 7, 1-9 [disponible en <https://raco.cat/index.php/Tradumatica/article/view/154835>, 18/09/2022].

Cattaneo, Elena (2004), *Idiomatic expressions in conference interpreting*, Graduation thesis, SSLMIT, Università degli Studi di Bologna, Sede di Forlì.

Consejo Europeo. *Europa digital* <https://www.consilium.europa.eu/es/topics/digital-europe/> [25/10/2022].

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

Corpas Pastor, Gloria (2018), «Tools for Interpreters: the Challenges that Lie Ahead», *Current Trends in Translation Teaching and Learning E* 5, 157-182 [disponible en <http://www.cttl.org/uploads/5/2/4/3/5243866/cttl_e_2018_5.pdf>, 20/10/2022].

Corpas Pastor, Gloria. (2021), «Technology Solutions for Interpreters: The VIP System», *Hermēneus. Revista de Traducción e Interpretación* 23, 91-123.

Corpas Pastor, Gloria (2022), «Interpreting Tomorrow? How to Build a Computer-Assisted Glossary of Phraseological Units in (Almost) No Time», *International Conference on Computational and Corpus-Based Phraseology*, Londres: Springer, 62-77.

Corpas Pastor, Gloria – Gaber, Mahmoud (2021), «Extracción de fraseología para intérpretes a partir de corpus comparables compilados mediante reconocimiento automático del habla», en Corpas Pastor, G – Bautista Zambrana, M. R. – Hidalgo Ternero, C. M. (eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*, Granada: Comares, 271-291.

Crezee, Ineke – Grant, Lynn (2013), «Missing the plot? Idiomatic language in interpreter education», *International Journal of Interpreter Education* 5/1, 17-33 [disponible en <https://cit-asl.org/missing-plot-vol5-1/>, 13/09/2022].

Fantinuoli, Claudio (2006), «Specialized Corpora from the Web for Simultaneous Interpreters», *Wacky! Working papers on the Web as Corpus*, Bolonia: GEDIT, 173-190.

Fantinuoli, Claudio. (2017), «Computer-assisted preparation in conference interpreting», *Translation and Interpreting* 9/2, 24-37 [disponible en <http://www.trans-int.org/index.php/transint/article/view/565>, 02/10/2022].

Fantinuoli, Claudio – Prandi, Bianca (2018), «Teaching information and communication technologies: a proposal for the interpreting classroom», *Transkom* 11/2, 162-182.

Gaber, Mahmoud – Corpas Pastor, Gloria (2022; en prensa), «La tecnología del reconocimiento automático del habla: recurso de documentación y apoyo tecnológico para la docencia en interpretación», Granada: Comares.

Gallego-Hernández, Daniel – Tolosa-Igualada, Miguel (2012), «Terminología bilingüe y documentación ad hoc para intérpretes de conferencias: Una aproximación metodológica basada en corpus», *Estudios de Traducción* 2, 33-46 [disponible en <https://revistas.ucm.es/index.php/ESTR/article/view/38976>, 25/09/2022].

Gile, Daniel (1995/2009), *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam, Philadelphia: John Benjamins.

Halliday, Michael Alexander Kirkwood (1989), *Spoken and written language*, Oxford: Oxford University Press.

Markič, Jasmina (2012), «Acerca de la (in)traducibilidad de las unidades fraseológicas en la interpretación de conferencias», en Jesenšek V. (ed.), *Phraseologie im Wörterbuch und Korpus = Phraseology in Dictionaries and Corpora*, Maribor: Univerza v Mariboru, 193-203.

Nolan, James (2005), *Interpretation: Techniques and exercises,* Clevedon, Reino Unido: Multilingual Matters.

Pérez-Pérez, Pablo (2018), «The Use of a Corpus Management Tool for the Preparation of Interpreting Assignments: A Case Study», *The International Journal for Translation and Interpreting Research* 10(1), 137-151 [disponible en <http://www.trans-int.org/index.php/transint/article/view/563>, 05/10/2022].

Sandrelli, Annalisa. (2015), «Becoming an interpreter: the role of computer technology», *MonTI* Special Issue 2, 111-138.

Sánchez Ramos, María del Mar (2017), «Interpretación sanitaria y herramientas informáticas de traducción: los sistemas de gestión de corpus», *Panace@* 18/46, 133-141.

Seghiri, Míriam (2017), «Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete (Corpora and medical interpreting: terminology extraction based on bitexts for the interpreter's documentation process)», *Panace@* 18/46, 123-132.

Setton, Robin (2011), «Corpus-based interpreting studies (CIS): Overview and prospects», en Kruger, A– Wallmach, K. – Munday, J. (eds.), *Corpus-based translation studies research and applications*, Londres: Continuum International, 33-75.

Shlesinger, Miriam (1998), «Corpus-based interpreting studies as an Offshoot of Corpus-based Translation Studies», *Meta* 43/4, 486-493.

Tolosa-Igualada, Miguel – Mezcua, Aurora Ruiz (2010), «La opacidad de las unidades fraseológicas y su tratamiento por intérpretes en formación», en *Opacidad, idiomaticidad, traducción*, Alicante: Universidad de Alicante; París: Université Paris 13; La Manouba: Université de la Manouba, 365-383.

Xu, Ran (2018), «Corpus-based terminological preparation for simultaneous interpreting», *Interpreting* 20/1, 33-62.