Artículo recibido: 13/02/2021
Artículo aceptado: 26/04/2021
DOI: 10.17533/udea.mut.v14n2a11

MUTATIS MUTANDIS
ISSN 2011-799X

Extracción de contextos definitorios de tecnologías biomédicas en corpus especializado francés[1]

Manuel Cristóbal Rodríguez Martínez
manuelcristobalrm@gmail.com
https://orcid.org/0000-0001-5644-7483
Universidad Europea del Atlántico, España

Resumen

Ante nuevos campos de conocimiento, los traductores se enfrentan a lagunas contextuales y terminológicas que los diccionarios especializados no cubren. Gracias al potencial que demuestran las herramientas de análisis y gestión de corpus, estas se perfilan como indispensables a la hora de estudiar las relaciones gramaticales entre palabras concretas y sus vínculos semánticos, y extraer información definitoria que supla las carencias de los diccionarios en el proceso de traducción. Por ello, esta investigación extrae contextos definitorios en el ámbito de las tecnologías biomédicas, ámbito de conocimiento en continua expansión y con desarrollo constante de técnicas, instrumentos, metodologías y productos, mediante lenguaje de interrogación de corpus (Corpus Query Language, CQL) y expresiones regulares. En otras palabras, se pretende establecer unas pautas para la creación de búsquedas que combinen CQL y expresiones regulares, para localizar información que defina, reformule o matice terminología presente en este ámbito. Para ello, se compila un corpus de 100 artículos en francés de investigación sobre ingeniería genética y biotecnología en la herramienta en línea de gestión de corpus Sketch Engine. Las ecuaciones de búsqueda y los ejemplos muestran la utilidad de esta estrategia para localizar contextos ricos en conocimiento que podrían resultar de interés no solo en el ámbito investigador, sino también en el ámbito profesional de la traducción especializada del francés al español, que actualmente cuenta con pocos recursos terminológicos.
Palabras clave: contextos definitorios; lingüística de corpus; tecnologías biomédicas; traducción especializada.

1 Este artículo se desprende de la investigación independiente, realizada en 2021, titulada "Extracción de contextos definitorios de ingeniería genética mediante corpus comparable especializado francés-español", y la institución que lo avala es la Universidad Europea del Atlántico.

Extracting Bioengineering Definitional Contexts in a Specialized French Corpus

Abstract
Before new developments of certain fields of knowledge, translators face context and terminology gaps that specialized dictionaries are not able to cover. Thanks to the potential shown by corpus management and text analysis software, corpora are emerging as essential tools to study not only the grammatical relations between specific words but also their semantic relations and extract defining contexts which could compensate the shortcomings of dictionaries during the translation process. Therefore, this study extracts definitional contexts in biomedical technology, continuously expanding areas with ongoing techniques, tools, methodologies, and products using CQL (Corpus Query Language) and RegEx. This study seeks hence to establish patterns in order to build searches combining CQL and RegEx to retrieve defining, rewording, and/or refining the terminology of these fields. Hence, a corpus made of 100 research articles, written in French, about bioengineering and biotechnology is compiled using the corpus management online tool Sketch Engine. The suggested searches and the results confirm the usefulness of this strategy to retrieve knowledge-rich contexts. These contexts are interesting for researchers and specialized translation professionals from French to Spanish, language combination with scarce terminological resources.
Keywords: corpus linguistics, knowledge-rich contexts, definitional contexts, biomedical technologies, specialized translation.

Extraction de contextes définitoires de génie génétique dans un corpus spécialisé français

Résumé
Face à des nouveaux domaines, les traducteurs se trouvent confrontés avec des lacunes contextuelles et terminologiques, lesquelles ne peuvent pas être couvertes par les dictionnaires spécialisés. Grâce au potentiel des outils de gestion et d'analyse de corpus, ils s'imposent comme des outils essentiels pour analyser les relations grammaticales entre des mots particuliers, leurs relations sémantiques et pour extraire des informations définitoires afin de combler les lacunes des dictionnaires au cours du procès de traduction. Pour cette raison, cette étude réalise une extraction de contextes définitoires dans le domaine des technologies biomédicales, domaine de connaissance toujours croissante et en développement continu de techniques, instruments, méthodologies et produits avec la combinaison de CQL (Corpus Query Language) et expressions régulières (RegEx). Ainsi, cette étude vise à établir certaines stratégies pour la création de recherches basée sur CQL et expressions régulières afin de récupérer des informations qui définissent, reformulent ou nuancent la terminologie présente dans ce domaine. Pour ce faire, un corpus a été compilé avec 100 articles de recherche écrits en français (génie génétique et biotechnologie) avec l'outil en ligne de gestion de corpus Sketch Engine. La syntaxe de recherche et les exemples montrent l'utilité de cette stratégie pour la localisation de contextes riches en connaissances qui pourraient être d'intérêt non seulement dans le domaine de la recherche scientifique mais aussi dans le cadre professionnel de la traduction spécialisée du français vers l'espagnol, actuellement avec des ressources terminologiques limitées.
Mots-clés : linguistique de corpus ; contextes riches en connaissance ; contextes définitoires ; technologies biomédicales ; traduction spécialisée.

1. Introducción

A pesar del avance en la lingüística de corpus, la localización de contextos definitorios (CD) y contextos de uso aún no se ha implementado de manera automática en diccionarios y otros recursos terminológicos empleados por profesionales de la traducción o de los ámbitos de conocimiento objeto de traducción (Corpas y Roldán, 2014). La práctica profesional, en particular la traducción del ámbito biosanitario (Corpas y Roldán, 2014), se ve dificultada, a su vez, por la ausencia de contextos de los términos en los diccionarios especializados (Barceló y Delgado, 2015).

Gracias al desarrollo e implementación de los lenguajes de interrogación de corpus (Corpus Query Language, CQL) y expresiones regulares[2] en corpus, la investigación en lingüística de corpus se ha incrementado considerablemente (Corpas y Seghiri, 2007; Nazar y Galdames, 2020), lo cual podría ayudar en la búsqueda de los CD y los contextos de uso en textos de especialidad. Sin embargo, estas investigaciones suelen centrarse en el inglés como lengua de trabajo, debido a la hegemonía de aquel como *lingua franca* de la comunicación científica (Bordons, 2004; Haße *et al.*, 2011; Navarro, 2001, 2006; Ruiz Rosendo, 2007), por lo que los patrones pragmáticos, léxicos y gramaticales que conforman las ecuaciones de búsqueda se han aplicado poco en el análisis de corpus franceses (Boré y Malrieu, 2017; Cartier y Viaux, 2018); no obstante, en español hay diversos estudios que aplican estos patrones para recuperar información definitoria (Sierra, 2009; Sierra *et al.*, 2010).

Para cubrir la carencia de estudios que contemplen el francés a la hora de emplear la metodología de corpus en la literatura científica, en este trabajo se propone localizar los CD y los contextos de uso en un corpus francés de terminología del ámbito de las tecnologías biomédicas, para comprobar la utilidad de los patrones léxicos y sintácticos recurrentes, relativos a los contextos ricos en conocimiento en francés, que se incorporarán a una serie de ecuaciones de búsqueda que contemplen esta variación mediante CQL y expresiones regulares soportadas por la herramienta en línea de gestión de corpus Sketch Engine (Kunilovskaya y Koviazina, 2017). Para ello, se compila el corpus especializado comparable de 100 artículos de investigación sobre ingeniería genética y biotecnología en francés, para identificar si las temáticas localizadas en el corpus proporcionan información definitoria, y verificar, así, que la metodología de corpus seguida (restricciones diasistemáticas, bases de datos empleadas, etc.) ha sido la adecuada.

2. Contextos definitorios

La *lingüística de corpus* es una disciplina que ha crecido de manera exponencial en los últimos años en términos de investigación, herramientas disponibles y aplicaciones tanto teóricas como prácticas (Clark *et al.*, 2010; McEnery *et al.,* 2006; Sierra, 2017). Esto se debe a queuna de las tecnologías más utilizadas en el estudio del lenguaje desde hace décadas es el corpus (Egbert *et al.,* 2020). En este sentido, la utilización de herramientas de gestión y análisis de corpus ha ayudado a que la investigación en lingüística y traducción aumente el volumen de textos que se pueden analizar y extienda el alcance de sus resultados, con una metodología replicable en numerosos idiomas.

En la actualidad, estas herramientas de gestión y análisis de corpus son muy variadas y soportan diversos lenguajes de interrogación —como

2 Las expresiones regulares no son un lenguaje exclusivo de las herramientas de corpus, sino un tipo de lenguaje formal que se emplea en múltiples contextos para procesar el lenguaje natural. Este lenguaje formal se puede emplear, junto con CQL, en funciones de búsqueda para ampliar el alcance de las búsquedas.

es el caso del CQL o las expresiones regulares— que, combinados en lo que se conoce como una ecuación de búsqueda, pueden localizar no solo terminología, sino también estructuras sintácticas con variantes léxicas, para extraer información semántica y estructural (León-Araúz *et al.*, 2016). Así, la metodología de corpus se ve complementada por estas ecuaciones de búsqueda avanzadas para localizar una mayor información, por medio delas cuales se identifican relaciones gramaticales complejas.

Gracias a estas relaciones gramaticales, se puede automatizar la búsqueda y la recuperación de información semántica, por ejemplo, debido a los *contextos ricos en conocimiento (Knowledge-Rich Context*, KRC), muy útiles para la labor terminográfica y para la traducción especializada por igual (Condamines y Rebeyrolle, 1997; Meyer, 2001; Picton *et al.*, 2017). Para Meyer,

> *Por contexto rico en conocimiento*, designamos un contexto que indica al menos un elemento de conocimiento de área que podría ser útil para el análisis conceptual. En otras palabras, el contexto debe indicar al menos una característica conceptual, ya sea un atributo o una relación (2001, p. 281).

Como se indica en esta propuesta definitoria, estos contextos ricos en conocimiento son de gran utilidad para extraer información contextual (colocaciones, régimen preposicional, registro, etc.) y proporcionar información sobre relaciones gramaticales que servirían para ampliar el conocimiento sobre un ámbito específico o, incluso, como base para formular definiciones (Valero y Alcina, 2010).

No obstante, al utilizar contextos, debido a la utilidad de sus relaciones conceptuales, suelen entenderse como CD. Estos se definen, para este estudio, como un fragmento de texto de un documento especializado, en el que aparece, mediante una serie de patrones lingüísticos, un término y su definición, u otros elementos semánticos contextualizados que ayuden a la conceptualización del término (Alarcón *et al.,* 2007; Sierra, 2009).

La investigación acerca de los CD suele estar vinculada a la terminografía, debido a la necesidad de localizar definiciones para acompañar a los términos, la cual se ha visto potenciada por avances tecnológicos que permiten la extracción automática de terminología (Cabré *et al.*, 2001; Sierra, 2011). Sin embargo, la extracción automática de definiciones presenta mayor complejidad, en cuanto a las características que muestran diferentes tipologías textuales a nivel sintáctico y pragmático, que resultan en una gran variedad de relaciones que se dan entre el término y la definición, o los rasgos que se pueden extraer de dicho término en el texto (Sierra, 2009).

La dificultad que se identifica de los CD, como se enuncia en el párrafo anterior, radica en la manera en la que se relacionan los diferentes elementos que los conforman. Estos elementos que configuran el CD son el término (T), la definición (D) y el patrón definitorio (PD), con variaciones potenciales del PD, que podrían incluir una predicación verbal definitoria (PVD), marcadores reformulativos definitorios (MRD), marcadores tipográficos definitorios (MTD) o patrones pragmáticos (PP) (Alarcón *et al.,* 2007; Sierra, 2009).

El T se entiende como un signo lingüístico que hace referencia a un concepto especializado (Cabré, 2001), que puede estar presente o no mediante relación anafórica con otro elemento del texto o, como es habitual, por medio de una estructura nominal, con un sustantivo como núcleo (Alarcón y Sierra, 2006). Aunque no se puede desdeñar la idea de que los términos se concreten como unidades de significación especializada no léxicas (Estopà, 2001), en función del campo de conocimiento que se analice (Alarcón *et al.*, 2007).

La D, por otro lado, se entiende como la descripción del concepto que dicho término representa (Cabré, 2001), que podría darse mediante una definición aristotélica o por medio de la identificación de rasgos propios del término (Alarcón *et al.*, 2007).

La relación que se mantiene entre T y D suele facilitar la interpretación de ambos elementos (Alarcón y Sierra, 2006), mediante el patrón definitorio (PD) y sus variaciones potenciales ya enunciadas, lo cual constituirían una serie de patrones sintácticos recurrentes que facilitan la recuperación de esta información lingüística (Sierra, 2009), y que se explicitan a continuación.

Las PVD son ciertos patrones verbales recurrentes, generalmente verbos metalingüísticos, como *definir,* u otros verbos que suelen introducir definiciones, como *entender* o *ser* (Alarcón *et al.,* 2007), que pueden aparecer acompañados de otros elementos lingüísticos según el régimen preposicional de dichos verbos.

Por otro lado, los MRD son aquellas estructuras sintácticas que proporcionan un indicador al lector del texto de una reformulación o ejemplificación de términos o definiciones anteriores, como en las expresiones *como es el caso, es decir,* etc. (Sierra *et al.,* 2003).

Estas indicaciones en el texto pueden ser de carácter lingüístico (MRD) o de carácter ortotipográfico, lo que se consideraría un MTD, como serían los paréntesis, los dos puntos o recursos estilísticos de énfasis como la cursiva, empleada como resalte y con función metalingüística.

Finalmente, cabe mencionar los PP, los cuales sirven para contextualizar diferentes aspectos de los términos y las definiciones, con alusiones a autoría, temporalidad, contextos de uso, frecuencia de uso, etc. (Alarcón *et al.*, 2007; Sierra *et al.*, 2008), lo que ayudaría a extraer información analítica sobre el término.

Así, estas estructuras son variadas y utilizan sintagmas diferentes. Además, las propuestas de búsqueda para recuperar gran cantidad de información han de ser abiertas, con el fin de localizar estructuras sintácticas recurrentes, en lugar de palabras específicas y sus respectivas colocaciones (López, 2020). De lo contrario, la información recuperada estaría más centrada en la terminología empleada que en patrones definitorios.

3. Método

Para llevar a cabo esta investigación, se ha compilado un corpus de 100 artículos de investigación para comprobar la extracción de información definitoria en estos géneros textuales. A continuación se muestran las características principales del corpus empleado, así como la herramienta usada.

3.1. Corpus

Para esta investigación, se ha decidido compilar un corpus monolingüe francés de textos especializados. Este corpus se ha creado con el fin de ampliarlo en un futuro a un corpus comparable bilingüe francés-español, el cual se caracteriza, principalmente, porque los textos que lo componen en ambas combinaciones lingüísticas son textos originales, no traducciones, lo que garantiza que se evitarán las potenciales pérdidas de sentido y estructuras sintácticas idiomáticas que se originan en el proceso de traducción, y permite una extracción de resultados más segura (Condamines *et al.,* 2013; Corpas, 2001; Laursen y Arinas-Pellón, 2012). Además, de esta manera se puede realizar una investigación tanto monolingüe como contrastiva sobre los diferentes patrones definitorios que se emplean en ambos idiomas en el ámbito de la ingeniería genética y la biotecnología. No obstante, por lo pronto, la presente investigación se centra únicamente en la lengua francesa.

Los textos que componen el corpus se han seleccionado tras proponer las siguientes restricciones diasistemáticas y pragmáticas: los textos han de pertenecer al género artículo de investigación y han de tratar sobre ingeniería genética y biotecnología; en este sentido, se han incluido textos sobre ingeniería genética y biotecnología a nivel general, nanotecnología, neuroprótesis, biónica, transgénesis, edición genética, entre otros temas, aplicados a humanos, animales y vegetales. Esta pluralidad temática se debe a las limitaciones impuestas por la combinación lingüística (francés), que en el ámbito de la investigación suponen una reducción considerable de la investigación disponible sobre estos ámbitos científicos (Haße *et al*., 2011; Martínez, 2010; Rodríguez, 2017).

El idioma en el que habrían de estar redactados los textos para su publicación es francés, en su variante europea, para evitar así la variación terminológica y garantizar una mayor homogeneidad al localizar expresiones mediante palabras clave específicas. Además, los textos deben estar publicados con posterioridad a 2010, con lo que se busca eliminar la variación diacrónica que puede tener un ámbito como la biotecnología y la ingeniería genética, así como evitar la inclusión de artículos de carácter ético, muy recurrentes en años anteriores a 2010, debido a los tímidos avances y con tantas implicaciones ideológicas, que podrían alterar los resultados (transhumanismo, ética de la ingeniería genética, etc.).

Para la localización de estos textos se ha recurrido a buscadores especializados como Google Académico, PubMed y Medicina en Español (MEDES). Estos buscadores han permitido utilizar los diferentes filtros y la lógica booleana para aplicar las restricciones diasistemáticas presentadas anteriormente (año de publicación, lugar de publicación, idioma, género, temática, etc.).

Como se mencionaba anteriormente, una de las mayores limitaciones que se ha tenido durante la compilación del corpus es la selección textual de artículos en francés, lengua que a pesar de tener una presencia relevante en internet (Pimienta y Prado, 2016), se suele utilizar con menos frecuencia en la divulgación científica (Navarro, 2001). Esto se puede ver en los resultados recuperados en la base de datos PubMed, una base de datos específica de literatura científica biosanitaria con más de 30 000 000 millones de publicaciones, que solo recupera 759 831 resultados en el caso del francés al buscar únicamente por el idioma y sin aplicar restricciones diasistemáticas.

Una vez que se han respetado los anteriores criterios de inclusión textual, el corpus compilado posee las características que se enuncian en la Tabla 1.

Tabla 1. Características del corpus objeto de estudio

Textos compilados	100
Género textual	Artículo de investigación
Temática	Biotecnología
Idioma	Francés
Restricciones diatópicas	Francés europeo
Originalidad del texto	Textos originales
Etiquetado	Gramatical
Palabras (*tokens*)	632 388
Formas (*types*)	466 847
Ratio *type* / *token* por idioma	73,82 %

Como se puede apreciar, los textos tienen una longitud media de 6323 palabras en francés, y la ratio *type* / *token* indica que no hay mucha variación léxica en el corpus compilado (Capsada y Torruella, 2017). Esta ratio también se puede considerar elevada, debido al número de palabras (*tokens*) del corpus, que apenas supera el medio millón de palabras, así como de la condición especializada del corpus objeto de estudio, que

implicaría temas recurrentes en los diferentes textos y similitud en metodologías, instrumentos y otros elementos que incluyen estos artículos, con una consecuente repetición léxica.

El formateo de los textos se ha reducido a la conversión de formatos editables (ODT, DOC/X) y formatos electrónicos (HTML) a un formato no editable (PDF) que permita un mejor procesamiento, manejo y gestión. La elección del formato PDF se debe a la herramienta seleccionada para la compilación y explotación del corpus.

3.2. Herramienta empleada

Las herramientas de gestión y análisis de corpus actuales presentan similitudes a nivel funcional (Rodríguez y Ortega, 2020). Sin embargo, la herramienta seleccionada para realizar esta investigación es Sketch Engine (s. f.). Esta se ha seleccionado por su versatilidad, la variedad de formatos textuales que soporta (PDF entre ellos), los análisis que pueden llevarse a cabo gracias a su lenguaje de interrogación (CQL) y a las expresiones regulares que admite (Kilgarriff *et al.,* 2014), como también por incluir la función de etiquetado gramatical automático con una elevada calidad (Fromm *et al., 2*020). Este etiquetado, además, varía en función de cada idioma, por lo que las búsquedas mediante la utilización de CQL han de ser específicas de cada idioma, debido a las estructuras oracionales propias de cada lengua, así como al comportamiento textual y macroestructural que tienen diversos géneros textuales.

Esta variación provoca que las estrategias de búsqueda en corpus cambien según el idioma objeto de estudio. En el caso del francés, se hace uso de las etiquetas gramaticales francesas que proporciona FreeLing, basado en las premisas del grupo EAGLES (1996). Estas etiquetas gramaticales se utilizan integradas en ecuaciones de búsqueda con CQL, así como con expresiones regulares, para verificar que su uso devuelve contextos definitorios de ayuda para la traducción especializada, teniendo en consideración las estructuras particulares de los diferentes patrones que presentan estos contextos.

Estos lenguajes formales, tanto CQL como las expresiones regulares, se emplean para la interrogación avanzada de corpus, con el fin de extraer información sintáctica (Jakubíček *et al.,* 2010; Ryšavá *et al.*, 2015).

4. Resultados

Para mejorar la localización de los resultados, se establece una clasificación de las ecuaciones de búsqueda propuestas para la extracción de contextos definitorios en el corpus francés objeto de estudio, basada en los elementos de los CD. Así pues, las ecuaciones propuestas se estructuran en función de si los elementos principales de las ecuaciones de búsqueda son PVD, MRD, MTD o PP.

4.1. Predicaciones verbales definitorias

Como se mencionaba anteriormente, las PVD son patrones verbales recurrentes que introducen definiciones (Alarcón *et al., 2007).* Los verbos que se suelen emplear a la hora de proponer la definición de un término suelen ser *être, s'entendre, désigner, dénommer, etc.* Estos verbos pueden aparecer en forma de participio pasado o en una construcción pasiva en francés (*est entendu, dénommé,* etc.), en contraposición al español, que suele redactarse en forma reflexiva con la partícula *se* (*se define, se denomina, se entiende)* (Alarcón y Sierra, 2017). Teniendo esto en cuenta, a continuación se presenta una ecuación de búsqueda basada en CQL.

```
[tag="D.*"]?[tag="N.*"][tag="N.*|A.*"]{0,2}
    [lemma="être"][tag="D.*"]?[tag="N.*"]
```

Esta ecuación de búsqueda es básica, en el sentido en que utiliza el verbo *être* únicamente como núcleo de la PVD. Sin embargo, ya arroja 645 resultados en el corpus interrogado.

La ecuación de búsqueda está estructurada para localizar un determinante opcional, seguido de un sustantivo como elemento obligatorio, que se considera el núcleo del elemento que se pretende definir. Al sustantivo pueden seguirle una concatenación de entre 0 y 2 sustantivos o adjetivos (o una alternancia de ellos), con lo que se pretenden recuperar términos poliléxicos.

Tras esta estructura, se introduce el verbo *être* lematizado, seguido de un determinante opcional y un sustantivo que, como en la parte del término, se considera también núcleo de la definición. Así, esta ecuación arroja resultados como los siguientes:

> Par définition, *les marqueurs tumoraux sont les indicateurs* biologiques de ces changements moléculaires survenant lors du processus tumoral
> *Le mélanome cutané est une tumeur* maligne du système pigmentaire
> Les altérations génétiques de ces *éléments sont une cause* majeure de pathologies humaines

Como se puede observar en estos resultados, con esta simple ecuación de búsqueda se localizan numerosos contextos definitorios vinculados al verbo *être*.

Una de las limitaciones que se entrevé en la ecuación de búsqueda es que no se logró recuperar la totalidad del término o del evento que se pretende definir (*altérations génetiques de ces élements*). Una manera de solventar estas limitaciones es realizar una ampliación de lo que se entiende por "término" o "evento definible". Así, se propone la siguiente ecuación de búsqueda, que contempla esta limitación:

[tag="N.*"]{0,2}[tag="A.*"]{0,1}[lemma="de|comme|par"]?[tag="D.*"]?[tag="N.*"]
[tag="N.*|A.*"]{0,3}[lemma="être"]
[tag="D.*"]?[tag="N.*"]

Con esta propuesta, se incluyen términos poliléxicos que pueden estar constituidos por varios sustantivos concatenados *(Candida maltosa*), seguidos de adjetivos (*thrombopénie alloimmune*) y vehiculados por las palabras *de, comme* o *par,* con el siguiente sustantivo (*système de traitement biologique*). Esta búsqueda arroja 907 resultados en el corpus, lo que supone un incremento de resultados considerable respecto a la anterior ecuación de búsqueda, al ampliar únicamente el término en la estructura del CD.

Por otro lado, al incluir otros verbos empleados comúnmente en las PVD y mencionados anteriormente, se observa un incremento anecdótico en los resultados (924), con la siguiente ecuación de búsqueda:

[tag="N.*"]{0,2}[tag="A.*|V.*"]{0,1}
[lemma="de|comme|par"]?[tag="D.*"]?[-
tag="N.*"][tag="N.*|A.*"]{0,3}
[word="est"]?[lemma="être|définir|dénom-
mer|entendre"][tag="D.*"]?

Esta ecuación arroja nuevos resultados, con presencia esta vez de estructuras introductorias como *dénommé* o *définit.*

> *un coronavirus dénommé SARSCoV-1* a été reconnu responsable d'un SRAS ayant causé…
> *La norme EN149 définit trois* classes d'efficacité de filtration pour ces masques, à savoir FFP1, FFP2 et FFP3

Con estos resultados se identifican ciertos patrones verbales recurrentes en el proceso de definición en el contexto académico y científico en francés, que ayudarían a los investigadores a localizar tanto terminología como relaciones gramaticales y semánticas con las cuales ampliar su conocimiento al respecto.

4.2. Marcadores reformulativos definitorios

Los MRD incluían elementos que introducían al lector reformulaciones o ejemplificaciones de términos, con ayuda generalmente de locuciones, entendidas como unidades fraseológicas con fijación interna, unidad de significado y fijación externa pasemática, que suelen funcionar como elementos oracionales (Corpas, 1996), de uso común en estos contextos como *c'est-à-dire, par exemple, autrement dit,* etc.

Algunas propuestas de ecuación de búsqueda que incluyan estos elementos son las siguientes: en primer lugar, se propone la ecuación de búsqueda que localice la locución *c'est-à-dire:*

```
[word="c"][][word="est"][word="à"]
[word="dire"]
```

En esta ecuación se pueden intuir algunas de las limitaciones que presenta el francés respecto al inglés en cuanto al uso de CQL, un lenguaje de interrogación que está basado en la sintaxis inglesa, a pesar de tener un etiquetado gramatical específico para la localización de información en francés. En el ejemplo se ve que la contracción de *ce* y *est* en *c'est* no se localiza como una búsqueda lematizada de *ce* ([lemma="ce"]), sino que se ha de buscar con el atributo [word] y la letra *c* para localizar esta contracción. Esta búsqueda nos permite localizar reformulaciones como las siguientes:

> Dans le cas de protéines inconnues, *c'est à dire* non répertoriées dans les banques de données protéiques

En esta ecuación de búsqueda se realiza únicamente la búsqueda de la locución que, en estos contextos, introduce estas reformulaciones o acotaciones semánticas, como la que se muestra en el ejemplo anterior. En el corpus objeto de estudio arroja un resultado, por lo que no es necesario acotar más; sin embargo, en un corpus con mayor volumen de palabras y, por consiguiente, de contextos definitorios, sería necesario acotar esta búsqueda con una propuesta como la siguiente:

```
[tag="N.*|A.*"]{0,3}[word="\,"]?[word="c"]
[][word="est"][word="à"][word="dire"]
[word="\,"]?
```

Así, la estructura localizaría una serie de sustantivos o adjetivos (o una alternancia de ellos), seguida de una coma opcional que precede ortotipográficamente a la locución *c'est-à-dire* en francés, que se cierra con otra coma opcional.

Las convenciones ortotipográficas se han introducido como optativas, debido al posible desconocimiento de las normas de redacción de los investigadores; de esta manera, al hacerlas opcionales, se recuperarían aquellos casos con errores ortotipográficos.

De igual modo, y con otro MRD similar, se encuentra la expresión *à savoir*, que también puede ir precedida por signos ortotipográficos como comas y dos puntos. Para recuperar esta expresión dentro de un CD, se propone la siguiente ecuación de búsqueda:

```
[tag="N.*|A.*"]{1,3}
[lemma="de"]?[tag="A.*|V.*|N.*"]{0,2}
[word="\,|\:"]?[word="à"][word="savoir"]
[word="\,|\:"]?
```

Esta búsqueda, lanzada en el corpus objeto de estudio, devuelve 21 resultados. En esta ocasión, la expresión *à savoir* ([word="à"][word="savoir"]) se enmarca entre dos elementos opcionales, una coma o dos puntos, que vendría tras la localización de una serie de elementos que podrían conformar un término monoléxico, poliléxico o uno que incluya el lema de en su estructura:

> 1*'analyse protéomique du sérum, à savoir* la dynamique 726

> la *méthode bactériologique classique, à savoir* le test de double synergie
> l'absence d'un traitement efficace peut causer *différentes complications: à savoir* la rétinopathie; la néphropathie,

Otra estructura similar considerada un MRD sería *autrement dit*. Para recuperar esta estructura, se propone la siguiente ecuación de búsqueda:

```
[tag="N.*|A.*"]{1,3}[lemma="de"]?[tag="A.*|V.*|N.*"]{0,2}
[word="\,|\:"]?[word="autrement"]
[word="dit"][word="\,|\:"]?
```

De igual modo que con la búsqueda que localiza la estructura *à savoir*, se enmarca la expresión *autrement dit* entre signos ortotipográficos opcionales, para poder recuperar los potenciales contextos con usos de puntuación no normativos. La búsqueda recupera 4 resultados, ambos con esta variación ortotipográfica, como se ve en los siguientes ejemplos:

> Il en résulte une entrave à la *progression des polymérases, autrement dit* une inhibition de réplication ou de transcription
> Dans la population caucasienne, les deux variants les plus fréquents sont CYP2C9*2 (Arg144Cys) et CYP2C9*3 (Ile359Leu), présents chez, respectivement, 8 à 19 % et 6 à 10 % *des sujets : autrement dit*, près d'un tiers de la population générale possède au moins un allèle muté.

Un caso similar se presenta con la construcción *par exemple,* que se podría localizar con la siguiente ecuación de búsqueda:

```
[tag="N.*|A.*"]{0,3}[lemma="de"]?[tag="A.*|V.*|N.*"]{0,2}
[word="\,|\:"]?[word="par"][word="exemple"]
[word="\,|\:"]?
```

Los resultados que arroja esta búsqueda son de 479, entre los que se localiza, como en los casos previos, una variación a nivel ortotipográfico, que se recupera gracias a la inclusión de la coma y los dos puntos opcionales en la ecuación de búsqueda.

> pouvant ensuite être étudiés en détail par *des techniques immunocytochimiques, par exemple* pour une délimitation précise de la tumeur au sein du tissu
> L'étape de fixation (1) du virus sur le récepteur cellulaire (*CD4/CCR5 par exemple*) peut être étudiée
> permettant d'interrompre la fonction ou de détruire les C*AR-T cells, par exemple* un gène « suicide » activable par un médicament ou un de ses métabolites

Tras analizar estos ejemplos anteriores, se observa que los MRD forman parte de los CD, pero que la relación que mantienen con el término no es tanto de definición, sino de matización o de ampliación de la información propuesta en el texto. Estos elementos, si bien no proporcionan una definición aristotélica, añaden matices informativos que ayudarían al receptor de estos textos a interpretar correctamente el contenido por medio de ejemplificaciones o reformulaciones. Además, se puede ver un mayor uso de la expresión *par exemple e*n detrimento de *c'est-à-dire* o *à savoir*, que podrían considerarse poco adecuadas a nivel de registro en el ámbito científico.

4.3. Marcadores tipográficos definitorios

Los MTD, al igual que los MRD, introducen en el texto pautas para identificar el comienzo de una reformulación o definición, esta vez con elementos de carácter ortotipográfico, como pueden ser los paréntesis, los dos puntos o recursos como la cursiva. A continuación se muestran las propuestas de ecuación de búsqueda con estos elementos.

El primer MTD que se plantea es el uso de los dos puntos. Para ello, se proporciona la siguiente ecuación de búsqueda:

[tag="N.*|A.*]{0,3}[lemma="de|en|comme"]?[tag="A.*|V.*|N.*"]{0,2}[word="\:"] [tag="D.*"] [tag="N.*|A.*"]{1,3}

Esta simple ecuación busca eliminar ruido informativo mediante la localización de un sintagma nominal precedido del signo ortotipográfico de dos puntos y un determinante. Esta búsqueda arroja 1062 resultados, de los cuales la mayoría introducen potenciales definiciones o reformulaciones, como se observa en los siguientes ejemplos:

> relié à un spectromètre de masse (MS) e*n tandem: la première colonne* est un échangeur fort de cations (CE), la deuxième une phase inverse (PI)
> *pharmacogénétique : le lien* entre gènes et réponse aux médicaments
> deux aspects clés de la *cytocinèse : les microtubules* du fuseau mitotique dictent la position du sillon de clivage; et l'invagination de la membrane est associée à la contraction d'un anneau cortical interne d'actine/myosine II

El uso de este signo ortotipográfico está muy extendido a nivel de redacción en el ámbito académico. Un ejemplo de ello se encuentra en las interferencias que no devuelven los resultados, ya que se localizan títulos de publicaciones científicas presentes en las referencias bibliográficas (elemento indispensable del género textual objeto de estudio), como los siguientes ejemplos:

> La thérapie génique dans la *maladie de Parkinson : un traitement* d 'avenir ?
> *LE MATÉRIEL : LA PLANTE TRANSGÉNIQUE* ET SES PRODUITS DÉRIVÉS

No obstante, y sabiendo cómo funciona la creación de títulos de publicaciones científicas y las recomendaciones para ello, se construye la siguiente ecuación de búsqueda basada en la anterior:

[tag="N.*|A.*"]{0,3}[lemma="de|en|comme|par"]?[tag="N.*|A.*|V.*"]{0,3} [lemma="de|en|comme|par"]?[tag="A.*|V.*|N.*"]{0,2}[word="\:"][tag="D.*"]{1,2} [tag="N.*|A.*"][word!="\.|\?|\:"]{6}

De esta manera, se establece como principio creador de títulos en publicaciones científicas la recomendación de la American Psychology Association (APA), una referencia en muchos ámbitos científicos actuales, que recomienda no incluir títulos de más de 12 palabras (APA, 2020). Así, teniendo como referencia que la posición de los dos puntos sería intermedia en el título, añadimos una expresión que amplía la búsqueda con 8 palabras más a partir de los dos puntos, de las cuales han de ser 1 o 2 determinantes, seguidos de 1 sustantivo o 1 adjetivo, y 6 palabras que no incluyan un signo ortotipográfico que sea un punto, un cierre de interrogación o, de nuevo, dos puntos. Con la inclusión de esta última parte se garantiza que la expresión no localiza varias frases, ya que se limita a devolver la frase que incluye los dos puntos. Esta ecuación de búsqueda devuelve 693 resultados, entre los cuales ya no se detectan títulos y solo hay frases presentes en el texto del artículo.

Otro MTD recurrente en el ámbito académico es el uso de paréntesis. La localización de estos se puede realizar con la siguiente ecuación de búsqueda (7083 resultados):

[word="\(*"][word!="\.|\("]{1,9} [word="\)*"]

Para cribar los resultados que pueden resultar de la información en el género textual (como, por ejemplo, la identificación de un ISSN digital o impreso), se añade la etiqueta de sustantivo antes de la apertura de paréntesis:

[tag="N.*"][word="\(*"][word!="\.|\("]{1,9}[word="\)*"]

No obstante, y a pesar de que se reducen considerablemente los resultados (4012), siguen

apareciendo resultados que no son CD, sino elementos como citas y referencias bibliográficas. Para filtrar también estos resultados en una única ecuación de búsqueda, se propone la siguiente:

[tag="N.*|A.*" & word!="(A-Z)"] [tag="N.*|A.*" & word!="\."] [word="\(*"] [word!="\.|\("] {1,9} [word="\)*"]

Con esta sintaxis de búsqueda, los resultados se reducen a más de un tercio desde la primera ecuación a un total de 2256.

En la anterior ecuación de búsqueda, los resultados devuelven una estructura compuesta por dos sustantivos, adjetivos o una combinación de ambos, excluyendo la estructura que precede el paréntesis bibliográfico: una letra inicial (una letra mayúscula seguida de un punto). En esta ocasión, los ejemplos de paréntesis son aquellos que se pretenden localizar en esta investigación, es decir, aquellos que introducen una definición, reformulación o matización de un término que precede el paréntesis, como se ve en los siguientes ejemplos:

> Dans la *technologie* SELDI *(surface enhanced laser desorption-ionization)*, ce qui est retenu sur la surface chromatograpique est analysé en spectrométrie de masse
> L'existence de *métaboliseurs ultrarapides (MUR)* (activité augmentée) ou *intermédiaires (MI)* (activité réduite)
> Les médecins généralistes calculaient le SAM *score (Self Assessement Melanoma Risk Score)* et repéraient ainsi les patients à risque
> la production d'enzymes ou d'autres *protéines recombinantes (insuline, interférons, hormone de croissance)*, mais seule Pichia pastoris est actuellement mise à contribution pour la production d'AcM

Estos ejemplos ilustran la función de los paréntesis reformulativos en este género textual, con un léxico especializado que, en numerosas ocasiones, utiliza siglas que se desarrollan usando estos recursos ortotipográficos, así como ejemplificaciones para detallar el contenido expuesto.

4.4. Patrones pragmáticos

Los PP sirven para contextualizar algunos aspectos que afectan al término o a la definición. Para ello, se propone una serie de ejemplos que podrían recuperar información contextual. La primera propuesta de PP es la estructura *par définition:*

[tag="N.*"]?[word="\,"]?[lc="par"][word="définition"][word="\,"]?[tag="D.*"]?[tag="N.*|A.*|V.*"]{1,3}

Con esta ecuación de búsqueda, se localizan unos 13 casos, que reafirman las sospechas iniciales a la hora de construirla.

La estructura *par définition* puede ser un elemento de inicio de frase, por lo que una mala elección de atributo podría reducir los resultados recuperados; para evitar esto, se ha introducido la primera palabra de la estructura (*par*) con el atributo [lc] (*lowercase*), que convierte la búsqueda en indiferente a la capitalización y localizaría esta forma en mayúscula (comienzo de frase) y minúscula.

Para ubicar también esta estructura, se hace opcional la presencia de un sustantivo tras la estructura *par définition*, así como la inserción de la coma para recuperar, como se mencionaba anteriormente, estructuras con incorrecciones ortotipográficas.

Tras la estructura *par définition* se recupera un determinante opcional, que puede preceder un sintagma nominal, seguido de uno o varios (hasta 3) sustantivos, adjetivos, verbos o una combinación de ellos.

> *Par définition, les marqueurs tumoraux sont* les indicateurs biologiques de ces changements

moléculaires survenant lors du processus tumoral
Le *diabète par définition peut être regroupé* en deux types majeurs : le diabète de type 1 appelé aussi diabète insulino-dépendant qui est causé par la destruction des cellules bêta du pancréas,

Así, se evita la pérdida de resultados, al mantener una búsqueda no marcada por la capitalización de sus elementos; si el atributo [lc] se cambiase por el atributo [word], los resultados se reducirían a 3, lo que implica una ineficaz recuperación de la información.

Otra estructura considerada un PP es la partícula *comme*, que introduciría matices o definiciones por parte de los autores. Para realizar la propuesta de búsqueda se ha de tener en cuenta que *comme* es una partícula muy empleada en lenguaje académico y, también, en géneros textuales menos especializados, por lo que su utilización está muy extendida y puede devolver demasiados resultados. Se propone así la siguiente ecuación de búsqueda:

```
[tag="N.*|A.*"]{0,3}[tag="V.*"]?[tag="R.*"]?[tag="V.*"]{1,3}[word="comme"]
[tag="D.*"]?[tag="N.*|A.*|V.*"]
```

Esta ecuación de búsqueda plantea una estructura en la que la partícula *comme* está precedida de uno o varios verbos; de esta manera, se contempla el uso de tiempos verbales compuestos muy recurrentes en francés. Además, se alterna el uso de verbos con adverbios, recurso estilístico igualmente muy empleado en la lengua que se analiza en la presente investigación.

Como se ha mencionado en estrategias de búsqueda anteriores, la limitación de los resultados se basa en la creación de un patrón de término que incluya un sintagma nominal compuesto por uno o varios sustantivos, adjetivos o la combinación de ellos, para identificar así terminología poliléxica o estructuras complejas que pretenden ser definidas. Esta ecuación de búsqueda devuelve 338 resultados, con unas estructuras compuestas por los verbos *considérer* (74), *utiliser* (33), *reconnaître* (27), *définir* (16), *apparaître* (15), *présenter* (13), *décrire* (11) y otros muchos verbos con una menor frecuencia de uso.

l'établissement d'un *profil protéique spécifique peut être considéré comme fiable* et reproductible si le protocole utilisé pour le recueil des échantillons, leur analyse et l'interprétation des résultats
Les *principaux facteurs reconnus comme cancérogènes* sont l'amiante, la poussière de bois, les rayonnements ionisants, le radon, la silice, les métaux lourds
Un site précis du génome de la *cellule préalablement caractérisé comme étant* transcriptionnellement actif, peut être également ciblé

De estos ejemplos se desprende la necesidad de considerar estos PP como parte de los CD, que añaden información adicional a los términos que se emplean en los textos especializados, como condición, causa, enumeración, etc. Los PP se toman como elementos periféricos de los CD; sin embargo, estas estructuras ayudan a reformular y matizar conceptos en ámbitos especializados, un recurso muy necesario al intentar extraer la mayor cantidad de información posible de un término.

5. Conclusiones

Tras realizar las pertinentes ecuaciones de búsqueda en el corpus francés, se dejan entrever las limitaciones específicas que, respecto a la lengua inglesa, tiene la lengua francesa en cuanto a procesamiento de la información y la recuperación mediante un corpus etiquetado, dado que la primera es la que toman como base la mayoría de programas y expresiones regulares.

Sin embargo, estas mismas expresiones regulares son las que ayudan en la recuperación de la

información cuando se dan este tipo de problemas, como es el caso de la contracción de *ce* seguida del verbo *être* en tercera persona del singular (*c'est*), que se ha de recuperar como dos palabras (*c* + *est*), al no contemplar la letra *c* como una variante morfológica del demostrativo *ce*. Este error es frecuente en el etiquetado gramatical francés, el cual, según la literatura científica, suele confundir contracciones de artículos y demostrativos, así como alternar categorías gramaticales entre, por ejemplo, participios y adjetivos (Denis y Sagot, 2012).

No obstante, estas ecuaciones de búsqueda permiten deducir que la lengua francesa tiene el mismo potencial que la lengua inglesa, de acuerdo con los numerosos estudios que emplean el inglés como lengua de trabajo (Cabezas-García y Faber, 2018; Evert *et al.*, 2020; Ježek y Melloni, 2011; León-Araúz y Reimerink, 2019, 2020), al ser procesada formalmente por medio de CQL y expresiones regulares para extraer contextos definitorios mediante patrones recurrentes de PVD, MRD, PP e incluso MTD, con un procesamiento adecuado de la sintaxis y la morfología francesa. De hecho, los estudios que toman como base la lengua francesa son muy necesarios aún en la actualidad, para potenciar la investigación científica sobre esta combinación lingüística e incentivar así una utilización de estrategias que se puedan seguir en la traducción profesional, con el fin de extraer información contextualizada y definitoria de textos especializados que pueda ayudar en el trasvase lingüístico.

Además, uno de los principales problemas que se da en la traducción especializada del francés al español es precisamente la interferencia lingüística que se produce entre lenguas romances (Cagnolati, 2015), que se ve potenciada en la traducción de esta combinación lingüística en el ámbito biosanitario, debido a la presencia de falsos amigos o calcos léxicos y sintácticos, entre muchos otros problemas de traducción (Ortega, 2003). Esta estrategia de búsqueda ha demostrado ser de utilidad en la extracción de información para el proceso de traducción, si se aplica al ámbito profesional, y una gran fuente de información contextual y definitoria para generar más recursos terminológicos especializados en la combinación lingüística francés-español, si se aplica al ámbito investigativo. En cualquier caso, se muestra como una línea de trabajo con gran potencial aplicada a diferentes áreas de especialización temática.

En vista de los resultados, se plantea que si bien la cantidad de palabras puede ser reducida debido a que no es un corpus de gran tamaño, este corpus se ha configurado, como se ha mencionado anteriormente, como un corpus abierto, por lo que admite la incorporación de nuevos textos, siempre y cuando respeten las restricciones diasistemáticas planteadas en la metodología. Así, el corpus aumentaría en palabras (*tokens)* y tipos *(types),* con lo que se presentarían, de este modo, unos resultados con mayores frecuencias.

Por otro lado, se ha detectado cierto potencial para ampliar esta línea de investigación en futuras publicaciones. Una de estas líneas está vinculada a Sketch Engine, la herramienta empleada para comprobar la utilidad de las ecuaciones de búsqueda basadas en CQL y en expresiones regulares. Sketch Engine permite generar patrones gramaticales que se pueden incorporar en la plataforma y localizar colocaciones gramaticales y léxicas con la función *Word sketch*. Estas ecuaciones de búsqueda podrían aumentarse y catalogarse según las relaciones gramaticales y semánticas que se establecen entre los elementos de la búsqueda (causa-efecto, hiperonimia, etc.) (León-Araúz *et al.*, 2016). Así, la propia herramienta podría establecer estas relaciones automáticamente, al introducir el término que se pretende analizar en un entorno más intuitivo y con una mayor cantidad de datos cuantitativos.

Además, se han identificado ciertas limitaciones en cuanto al tamaño del corpus objeto de estudio, ya que las frecuencias son, en ocasiones, demasiado bajas para poder generalizar resultados. Sin embargo, con un corpus que contenga un mayor número de palabras *(tokens)* y siguiendo la metodología llevada a cabo, se podría analizar si la ausencia de estas estructuras se debe al género textual (artículo de investigación) y registro (académico) que se localizan en los textos o, por el contrario, y al contrastarlos con esta investigación, se debe a la cantidad de palabras con las que se inicia este estudio.

No obstante, y a pesar de las frecuencias de uso expuestas, los ejemplos del estudio demuestran el potencial de estas búsquedas no solo para la investigación en traductología y terminografía, sino también para la práctica profesional de la traducción especializada del francés al español, que con tan pocos recursos sigue contando en la actualidad.

Referencias

Alarcón, R., Bach, C. y Sierra, G. (2007). Extracción de contextos definitorios en corpus especializados: hacia la elaboración de una herramienta de ayuda terminográfica. *Revista Española de Lingüística (RSEL), 37*(1), 247-277. https://doi.org/10.31810/RSEL.37.1

Alarcón. R. y Sierra G. (2006). Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios. En A. Hernández y J. L. Zechineli (Eds.), *Avances en la ciencia de la computación, VII Encuentro Internacional de Computación ENC'06 18 al 22 de septiembre 2006, San Luis Potosí, México* (pp. 242-247). MSCC.

Alarcón, R. y Sierra, G. (2017). El rol de las predicaciones verbales en la extracción automática de conceptos. *Estudios de Lingüística Aplicada*, (38), 129-144. https://ela.enallt.unam.mx/index.php/ela/article/view/753

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7.ª ed.). APA.

Barceló Martínez, T. y Delgado Pugés, I. (2015). La traducción de la preposición *sur* en el lenguaje jurídico francés: un estudio de caso. *Çédille*, (11), 51-67.

Boré, C. y Malrieu, D. (2017). Approche textométrique des séquences de discours direct dans un corpus de contes. A textometric approach of direct speech sequences in a tales corpus. *Corpus*, (17), 1-16. https://doi.org/10.4000/corpus.2856

Bordons, M. (2004). Hacia el reconocimiento internacional de las publicaciones científicas españolas. *Revista Española de Cardiología, 57*(9), 799-802. https://www.revespcardiol.org/es-hacia-el-reconocimiento-internacional-publicaciones-articulo-13065646

Cabezas-García, M. y Faber, P. (2018). Phraseology in specialized resources: An approach to complex nominals. *Lexicography, 5*, 55-83. https://doi.org/10.1007/s40607-018-0046-x

Cabré, M. T. (2001). *La terminología. Representación y comunicación*. IULA.

Cabré, M. T., Estopà, R. y Vivaldi, J. (2001). Automatic term detection. A review of current systems. En D. Bourigault, C. Jaquemin y M. C. l'Homme (Eds.), *Recent advances in computational terminology* (pp. 53-87). John Benjamins.

Cagnolati, B. E. (2015). Interferencia en la traducción francés-español de textos de ciencias sociales. *Hikma, 14*, 55-74. https://doi.org/10.21071/hikma.v14i.5200

Capsada Blanch, R. y Torruella Casañas, J. (2017). Métodos para medir la riqueza léxica de los textos. Revisión y propuesta. *Verba, 44*, 347-408. https://doi.org/10.15304/verba.44.3155

Cartier, E. y Viaux, J. (2018). Étude de la pénétration des anglicismes de type N ou ADJ(-)Ving à partir d'un corpus contemporain journalistique : les exemples de *bashing* et *shaming* en français contemporain. En C. Jacquet-Pfau, A. Napieralski y J.-F. Sablayrolles (Eds.), *Emprunts néologiques et équivalents autochtones : études interlangues* (pp. 11-34). Presses Universitaires de Łódź.

Clark, A., Fox, C. y Lappin, S. (2010). *The handbook of computational linguistics and natural language processing*. Wiley-Blackwell.

Condamines, A., Josselin-Leray, A., Fabre, C., Lefeuvre, L., Picton, A. y Rebeyrolle, J. (2013). Using comparable corpora to characterize knowledge-rich contexts for various kinds of users: Preliminary steps. *Procedia – Social and Behavioral Sciences, 95*, 581-586. https://doi.org/10.1016/j.sbspro.2013.10.685

Condamines, A. y Rebeyrolle, J. (1997). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. En *Actes Journées Ingénierie des Connaissances et Apprentissage Automatique, JICAA'97* (pp. 191-206). Institut National de Recherche en Informatique et en Automatique.

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

Corpas Pastor, G. (2001). Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada. *Trans*, 5, 155-184. https://doi.org/10.24310/TRANS.2001.v0i5.2916

Corpas Pastor, G. y Roldán Juárez, M. (2014). Análisis de necesidades documentales y terminológicas de médicos y traductores médicos como base para el diseño de un diccionario multilingüe de nueva generación. *MonTI,* 6, 167-202. http://dx.doi.org/10.6035/MonTI.2014.6.6

Corpas Pastor, G. y Seghiri Domínguez, M. (2007). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del Lenguaje Natural*, (39), 165-172. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2670

Denis, P. y Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation, 46*, 721-736. https://doi.org/10.1007/s10579-012-9193-0

EAGLES (1996). *Preliminary recommendations on corpus typology*. EAGLES DOCUMENT, EAG-TCWG-FR-2. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Egbert, J., Larsson, T. y Biber, D. (2020). *Doing linguistics with a corpus*. Cambridge.

Estopà, R. (2001). Elementos lingüísticos de las unidades terminológicas para su extracción automática. En M. T. Cabré y J. Feliu (Eds.), *La terminología científico-técnica. Reconocimiento, análisis y extracción de información formal y semántica* (pp. 67-80). Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Evert, S., Harlamov, O., Heinrich, P. y Banski, P. (2020). Corpus Query Lingua Franca part II: Ontology. En N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk y S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference, May 2020, Marseille* (pp. 3346-3352). European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.410/

Fromm, G., Grama, D. F., Bilke, N. S. y Guarato Santos, C. (2020). Wordsmith Tools and Sketch Engine: An analytical comparative study for scientific research with corpora manipulation. *Revista de estudos de linguagem, 28*(3), 1101-1248. http://dx.doi.org/10.17851/2237-2083.28.3.1191-1248

Haße, W., Peters, S. y Fey, K. H. (2011). ¿*Lingua franca* impuesta o lenguas europeas de la ciencia en medicina? La opción del multilingüismo. *Panace@, 12*(34), 267-372. https://www.tremedica.org/wp-content/uploads/n34-tribuna-habepetersfey_ESP.pdf

Jakubíček, M., Kilgariff, A., McCarthy, D. y Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. En R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto y Y. Harada (Eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (pp. 741-747). Waseda University, Institute of Digital Enhancement of Cognitive Processing. https://www.aclweb.org/anthology/Y10-1086/

Ježek, E. y Melloni, C. (2011). Nominals, polysemy and co-predication. *Journal of Cognitive Science, 12*(1), 1-31. https://doi.org/10.17791/jcs.2011.12.1.1

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. y Suchomel, V. (2014). The Sketch Engine: ten

years on. *Lexicography, 1*, 7-36. https://doi.org/10.1007/s40607-014-0009-9

Kunilovskaya, M. y Koviazina, M. (2017). Sketch engine: A toolbox for linguistic discovery. *Journal of Linguistics, 68*(3), 503-507. https://doi.org/10.2478/jazcas-2018-0006

Laursen, A. y Arinas-Pellón, I. (2012). Text corpora in translator training. A case of study of use of comparable corpora in classroom teaching. *The Interpreter and Translator Trainer, 6*(1), 45-70. https://doi.org/10.1080/13556509.2012.10798829

León-Araúz, P., San Martín, A. y Faber, P. (2016). Pattern-based word sketches for the extraction of semantic relations. En P. Drouin, N. Grabar, T. Hamon, K. Kageura y K. Takeuchi (Eds.), *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)* (pp. 73-82). ACL. https://www.aclweb.org/anthology/W16-4709/

León-Araúz, P. y Reimerink, A. (2019). High-density knowledge rich contexts. *Argentinian Journal of Applied Linguistics, 7*(1), 109-130. https://redib.org/Record/oai_articulo2061846-high-density-knowledge-rich-contexts

León-Araúz, P. y Reimerink, A. (2020). Improved knowledge rich context extraction for terminography. En T. Read, S. Montaner y B. Sedano (Eds.), *Technological innovation for specialized linguistic domains. Languages for digital lives and cultures. Proceedings of TISLID'18* (pp. 69-84). Éditions universitaires européennes.

López Rodríguez, C. I. (2020). Marcos predicativos asociados al concepto signo y síntoma en textos sobre medicina en español. *Revista Signos, 53*(103), 392-418. https://doi.org/10.4067/S0718-09342020000200392

Martínez López, A. B. (2010). La terminología médica en francés, inglés y español: problemas que se derivan de la presencia del inglés como *lingua franca* de la comunicación científica a escala internacional. *Anales de Filología Francesa*, (18), 393-404. https://revistas.um.es/anales-ff/article/view/117041

McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-based language studies. An advanced resource book.* Routledge.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework. En D. Bourigault, C. Jacquemin y M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 279-302). John Benjamins.

Navarro, F. A. (2001). El inglés, el idioma internacional de la medicina. *Panace@, 3*(2), 35-51. https://www.tremedica.org/wp-content/uploads/n3_Panacea3_Marzo2001.pdf

Navarro, F. A. (2006). La anglización del español: mucho más allá de *bypass, piercing, test, airbag, container* y *spa*. En L. González y P. Hernúñez (Coords.), *Traducción: contacto y contagio*. Actas del III Congreso Internacional "El Español, lengua de traducción". Puebla (México) (pp. 231-232). ESLEtRA. https://cvc.cervantes.es/lengua/esletra/pdf/03/017_navarro.pdf

Nazar, R. y Galdames, A. (2020). Formalización de reglas para la detección del plural en castellano en el caso de unidades no diccionarizadas. *Linguamática, 11*(2), 17-32. https://doi.org/10.21814/lm.11.2.285

Ortega Arjonilla, E. (2003). Aspectos metodológicos de la traducción científica y técnica. Aplicaciones al ámbito francés-español. En M. Á. García Peinado y E. Ortega Arjonilla (Coords.), *Panorama actual de la investigación en traducción e interpretación* (pp. 199-234). Atrio.

Picton, A., Josselin-Leray, A. y Planas, E. (2017). Defining knowledge-rich contexts for specialized translation: Uses and limitations of a mixed-method approach. En I. Simonnæs, Ø. Andersen y K. Schubert (Eds.), *21st Conference on Language for Specific Purposes, Bergen (Norway), June 28-30* (pp. 289-307). Frank & Timme.

Pimienta, D. y Prado, D. (2016). Medición de la presencia de la lengua española en la internet: métodos y resultados. *Revista Española de Documentación Científica, 39*(3), 1-14. https://doi.org/10.3989/redc.2016.3.1328

Rodríguez Martínez, M. C. (2017). Particularidades de la traducción en el ámbito biosanitario del francés al español. En E. Ortega, A. B. Martínez López y F. García Luque (Eds.), *Cartografía de la traducción, la interpretación y las industrias de la lengua. Mundo profesional y formación académica: interrogantes y desafíos* (pp. 151-166). Comares.

Rodríguez Martínez, M. C. y Ortega Arjonilla, E. (2020). Particularidades morfológicas en traducción biosanitaria del francés al español: el prospecto de medicamentos para uso humano. *Trans*, (24), 401-418. https://doi.org/10.24310/TRANS.2020.v0i24.6440

Ruiz Rosendo, L. (2007). El predominio del inglés en el lenguaje científico: características del lenguaje médico español en la actualidad. *Polissema*, (7), 85-113. https://core.ac.uk/reader/47137820

Ryšavá, D., Volková, N. y Rambousek, A. (2015). Converting the corpus query language to the natural language. En A. Hokák, P. Rychlý y A. Rambousek (Eds.), *Proceedings of recent advances in Slavonic natural language processing, RASLAN 2015* (pp. 43-48). Tribun EU. https://nlp.fi.muni.cz/raslan/2015/paper11-Rysava_Volkova_Rambousek.pdf

Sierra, G. (2009). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática, 1*(2), 13-37. https://www.linguamatica.com/index.php/linguamatica/article/view/38

Sierra, G. (2011). Lexicografía computacional en las búsquedas onomasiológicas con lenguaje natural. En M. E. Vázquez Laslop, K. Zimmerman y F. Segovia (Eds.), *De la lengua por sólo la extrañeza: estudios de lexicología, norma lingüística, historia y literatura en homenaje a Luis Fernando Lara* (pp. 445-464). El Colegio de México.

Sierra, G. (2017). *Introducción a los corpus lingüísticos*. UNAM, Instituto de Ingeniería.

Sierra, G., Alarcón, R., Aguilar, C., Barrón, A., Benítez, B. y Baca, I. (2008). Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología. En L. Fabbri (Ed.), *Terminología, conocimientos, sociedad y poder: X Simposio Iberoamericano de Terminología (RITERM): Montevideo, 7 a 19 de noviembre de 2006*. *riterm*; Intendencia Municipal de Montevideo.

Sierra, G., Medina, A., Alarcón, R. y Aguilar, C. A. (2003). Towards the extraction of conceptual information from corpora. En D. Archer, P. Rayson, A. Wilson y T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 691-697). UCREL.

Sierra, G., Torres-Moreno, J.-M. y Molina, A. (2010). Regroupement sémantique de définitions en espagnol. *ArXiv, abs/1501.04920*.

Sketch Engine (s. f). *Learn how language works*. https://www.sketchengine.eu/

Valero Doménech, E. y Alcina Caudet, A. (2010). Exploración de características conceptuales en contextos ricos en conocimiento mediante un programa de análisis cualitativo. *Revista de Lingüística y Lenguas Aplicadas*, *5*, 241-254. https://doi.org/10.4995/rlyla.2010.772

Cómo citar este artículo: Rodríguez Martínez, M. C. (2021). Extracción de contextos definitorios de tecnologías biomédicas en corpus especializado francés. *Mutatis Mutandis, Revista Latinoamericana de Traducción, 14*(2), 509-526. https://doi.org/10.17533/udea.mut.v14n2a11