Universidad de Valladolid

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Máster Universitario en Traducción en Entornos Digitales Multilingües

TRABAJO FIN DE MÁSTER

Análisis de la traducción automática inglés-español de las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol: el caso de *ball*, *shot* y *referee*

Presentado por D. Alfonso Peñaranda Latorre

Tutelado por el Dr. Jaime Sánchez Carnicer

Soria, 2023

A mis padres,

por introducirme al turismo de polideportivo;

a mis hermanos,

por vivir el deporte conmigo;

a Jaime,

por sufrir mi lentitud y guiarme en el camino.

RESUMEN

El lenguaje de especialidad del fútbol y la traducción automática constituyen dos objetos de estudio trascendentes en el ámbito de la investigación académica de los Estudios de Traducción y de la Lingüística Contrastiva, pero todavía no se han abordado de manera conjunta. Estos dos marcos de investigación convergen en el presente trabajo fin de máster, que persigue el objetivo de comprobar si la traducción automática del inglés al español de las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol es adecuada. Para lograr este objetivo, se emplea una metodología basada en corpus para analizar la traducción al español de unidades terminológicas en inglés extraídas de crónicas futbolísticas en directo.

PALABRAS CLAVE: lenguaje de especialidad del fútbol; traducción automática; unidad terminológica nominal univerbal; crónica futbolística en directo.

ABSTRACT

The language of football and machine translation are significant research topics within Translation Studies and Contrastive Linguistics. However, they have not yet been addressed jointly. These research frameworks converge in this Master's Degree Final Project, which is aimed at testing whether the machine translation from English into Spanish of football, single-word nominal terminological units is adequate. In order to meet this objective, a methodology based on corpora is used to analyse the translation of terminological units extracted from minute-by-minute football commentaries.

KEUWORDS: language of football; machine translation; single-word nominal terminological unit; minute-by-minute football commentary.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE GRÁFICAS

INTRODUCCIÓN

El deporte constituye uno de los pilares centrales del ocio y ocupa una posición de prestigio en el imaginario colectivo de las sociedades contemporáneas. No obstante, no cabe duda de que la actividad deportiva representa mucho más que un mero divertimento, pues desempeña un papel clave en la vida de las personas que lo practican: contribuye a mantener unos niveles altos de salud física y metal, constituye un medio de socialización y está relacionada con la realización personal a largo plazo.

Por supuesto, detrás de la elección de este tema de investigación subyace una motivación intrínseca y muy personal, ya que el deporte siempre ha formado parte de mi vida, al menos desde que tengo uso de razón. He practicado muchos deportes, entre los que destaco el bádminton, el voleibol y el baloncesto, y los seguiré practicando. De hecho, es muy probable que este trabajo fin de máster nunca hubiera visto la luz de no ser por las horas de felicidad sufrida y de emoción desenfrenada que he pasado practicando y viviendo el deporte.

Más allá de estas implicaciones personales, este trabajo fin de máster responde también a un fin de investigación académica en materia del lenguaje de especialidad del deporte (más concretamente del lenguaje de especialidad del fútbol) y de la traducción automática.

La prensa deportiva se postula como la herramienta perfecta para abordar el estudio del lenguaje de especialidad del deporte y, más concretamente, el del fútbol. Esta idea se respalda en los numerosos trabajos de investigación publicados en los últimos años (Casasús i Guri y Núñez Ladeveze, 1991; Martínez Albertos, 1998; Nomdedeu Rull, 2004; Rost, 2006; Bonvin Faura, 2007; Parrat, 2008; Bergh, 2011; Naranjo de Arcos, 2011; Lewandowski, 2012; Gárciga Rodríguez y Gómez Masjuán, 2013; Kovljanin, 2018b) que han empleado la prensa deportiva, sus géneros y sus características lingüísticas como elemento para estudiar distintos fenómenos lingüísticos, tanto en inglés como en español. Asimismo, también hay bastante literatura acerca del estudio del lenguaje del deporte y del fútbol, que se ha abordado a través de diferentes perspectivas, como la lingüística, la semántica, la lexicológica, la terminológica y la traductológica.

Por otro lado, en los últimos años, la revolución tecnológica, personificada sobre todo en las TIC, Internet y la aparición y el desarrollo de las inteligencias artificiales (IA), ha supuesto un cambio total en el paradigma epistemológico imperante. No obstante, la tecnológica avanza mucho más rápido que el conocimiento, de manera que son muchos los nuevos frentes de investigación que se están abriendo. En el ámbito de la traducción, el desarrollo exponencial de la traducción automática es uno de los frentes de investigación más representativos de esta catarsis tecnológica, hecho que se refleja en el gran volumen investigaciones al respecto publicadas en los últimos años (Berner, 2003; Bowker y Fisher, 2010; Quah, 2006; Díaz Prieto, 2012; Koponen, 2015; Forcada *et al*, 2016; Ortego Antón y Seghiri, 2019; O'Hagan, 2020; Sánchez Ramos y Rico Pérez, 2020).

No obstante, pese a que estos dos marcos de investigación se caracterizan por su trascendencia en el ámbito de la investigación académica, todavía no se han abordado de manera conjunta. En el presente trabajo fin de máster convergen estos dos objetos de investigación: el lenguaje de especialidad del fútbol, que abordamos desde un enfoque terminológico, y la traducción automática.

HIPÓTESIS Y OBJETIVOS

Adoptamos la siguiente hipótesis inicial como referencia para desarrollar nuestro trabajo de investigación: la traducción automática del inglés al español de las unidades terminológicas nominales univerbales pertenecientes al subcampo de especialidad del fútbol es inadecuada. Con base en esta hipótesis, el objetivo principal del presente trabajo de investigación es realizar un análisis de la traducción automática del inglés al español de las unidades terminológicas nominales univerbales pertenecientes al subcampo de especialidad del fútbol.

No obstante, este objetivo principal se cimenta sobre varios objetivos secundarios, que se exponen a continuación:

- aproximarse al lenguaje de especialidad del deporte y del fútbol;
- definir la crónica futbolística en directo y describir las características de este género textual;
- aproximarse a la traducción automática y a los principales sistemas de traducción automática;
- definir el concepto de corpus y estudiar los sistemas de clasificación de corpus y la metodología relativa al empleo de esta herramienta;
- compilar un corpus comparable bilingüe (EN-ES) equilibrado y representativo;
- analizar, gestionar y explotar dicho corpus de manera sistemática y adecuada mediante programas informáticos punteros, y
- comparar el tratamiento que se hace en lengua española de las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol en la crónica futbolística en directo y en los textos traducidos por medio de traducción automática.

En lo que se refiere a las competencias recogidas el Proyecto/Guía docente del Trabajo Fin de Máster, cumplimos con todas las competencias generales (G1, G2, G3, G4 y G5) y transversales (T1, T2, T3 y T4); asimismo, desarrollamos las competencias específicas relacionadas con los procesos de documentación, de adquisición de conocimientos específicos y de síntesis, correlación y reformulación de información en lengua española y en lengua inglesa (E2, E3 y E4), con la redacción de textos para fines específicos en español y en inglés (E6), con el empleo de herramientas informáticas para llevar a cabo un análisis interlingüístico e intralingüístico (E11, E12

y E13) y con el uso de sistemas de traducción automática (E14). El alumno ha adquirido y desarrollado estas competencias, generales, transversales y específicas, en las asignaturas Metodología de la Investigación Aplicada a la Traducción y Redacción Multilingüe, Últimos Avances Tecnológicos para la Traducción y la Redacción Multilingüe y Traducción y Redacción Multilingüe para los Sectores del Ocio y el Deporte EN-ES/ES-EN del Máster en Traducción en Entornos Digitales Multilingües.

METODOLOGÍA

En este capítulo, explicamos de manera resumida y razonada el proceso de delimitación del objeto de estudio y describimos la organización interna del presente trabajo de investigación.

Antes de delimitar el objeto de estudio, tuvimos que seleccionar un tema de investigación acorde a la línea temática de nuestro trabajo: la traducción en los sectores del ocio y del deporte. Así, nos decantamos por tratar el lenguaje futbolístico, que forma parte del lenguaje deportivo. Decidimos adoptar un enfoque terminológico para abordar el estudio de este lenguaje de especialidad y, posteriormente, determinamos que las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol constituirían nuestro objeto de estudio. En línea con esta decisión, concluimos que la crónica futbolística en directo era el género textual ideal para tratar nuestro objeto de estudio. Finalmente, incorporamos a la ecuación la traducción automática y establecimos el objetivo principal: «realizar un análisis de la traducción automática del inglés al español de las unidades terminológicas nominales univerbales pertenecientes al subcampo de especialidad del fútbol».

Hemos organizado nuestro trabajo de investigación de acuerdo con una macroestructura dividida en dos partes principales: el marco teórico y el marco práctico.

El marco teórico presenta una estructura interna conformada por tres bloques que sientan la base teórica sobre la que se cimenta el marco práctico. El primer bloque versa sobre la crónica futbolística en directo, el género textual que empleamos para abordar nuestro objeto de estudio; el segundo bloque trata sobre la traducción automática; por último, el tercer bloque comprende el concepto de corpus y la metodología relativa a la aplicación de esta herramienta de análisis lingüístico.

El marco práctico consta de un único apartado, «1.4. Metodología del marco práctico», en el que se exponen de manera pormenorizada los procesos de compilación y explotación del corpus comparable bilingüe FÚTBOL_MXM, que vertebra el análisis práctico, así como la metodología de dicho análisis.

A partir de la explotación del análisis práctico, que se refleja en los resultados, hemos extraído las conclusiones. Finalmente, hemos elaborado una lista de referencias bibliográficas con las fuentes consultadas a lo largo de la elaboración del trabajo.

MARCO TEÓRICO

1. La crónica futbolística en directo

En este apartado, abordamos de manera breve y sintética la crónica futbolística en directo[1] (CFED), un género periodístico de reciente aparición. Puesto que se trata de un género textual novedoso, apenas existen trabajos académicos dedicados específicamente a su estudio y descripción, así como al establecimiento de sus principales características formales, que no lingüísticas. Por tanto, nos basamos en el trabajo de Kovljanin (2018a) para definir este género.

1.1. Clasificación de la crónica futbolística en directo como género periodístico

La crónica futbolística en directo es, en realidad, una variante de la crónica deportiva en directo, que, a su vez, representa una modalidad de la crónica en directo. Se considera un género ciberperiodístico (Kovljanin, 2018a: 92) que se enmarca en el periodismo deportivo, de manera que no se publica en la prensa escrita, sino que su único medio de difusión es Internet. Atendiendo a la clasificación de los géneros periodísticos propuesta por Casasús i Guri y Núñez Ladeveze (1991) en géneros informativos, géneros interpretativos, géneros argumentativos y géneros instrumentales, la CFED se puede considerar un género informativo-interpretativo Kovljanin (2018b), puesto que al componente informativo se le suma las valoraciones subjetivas y personales con las que el cronista acompaña el relato.

1.2. Definición y características comunicativas principales

La crónica futbolística en directo consiste, a grandes rasgos, en la narración sincrónica y cronológica de un partido de fútbol en formato escrito. En consecuencia, la CFED es el resultado de la hibridación entre la crónica pospartido tradicional, las retransmisiones radiofónicas y televisivas en directo y la comunicación por medio de Internet (Kovljanin, 2018a: 107), por lo que reúne características de todos estos elementos:

- hereda de la crónica pospartido tradicional su carácter escrito (Bergh, 2011: 86) y la combinación de los componentes informativo, interpretativo y de opinión (Gárciga

[1] Nuestro objeto de estudio es la crónica futbolística en directo que se publica en la prensa española, por lo que el género equivalente en la prensa inglesa no comparte todas sus características textuales y convenciones.

Rodríguez y Gómez Masjuán, 2013: 924), aunque el componente informativo prevalece sobre los dos restantes (*ibid.*: 924; Kovljanin, 2018a: 106);

- comparte con las retransmisiones radiofónicas y televisivas la reproducción del lenguaje oral, que se pone de manifiesto en el uso de un registro informal (Bergh, 2011: 86), y la naturaleza síncrona y cronológica de la narración (narración en tiempo real) de los hechos (Rost, 2006: 159; Bergh, 2011: 86), de manera que se producen actualizaciones constantes de la información según el desarrollo de la acción y de los acontecimientos (Bonvin Faura, 2007: 179);
- toma de Internet su dimensión social, ya que el público puede interactuar con el cronista por medio de un foro (Lewandowski, 2012: 68).

Por ende, si abordamos la crónica futbolística en directo como un acto de comunicación en sí mismo, este género comprende dos funciones principales (Lewandowski, 2012): primero, informar de la acción que se está desarrollando, y después, entretener al público.

1.3. Características formales de la crónica futbolística en directo

La crónica futbolística en directo presenta una macroestructura interna que cuenta, además de con el titular, con tres partes bien diferenciadas (Gárciga Rodríguez y Gómez Masjuán, 2013; Kovljanin, 2018a): la previa, o entrada, el cuerpo y el cierre. En algunos casos, se prescinde de la previa o del cierre.

En la previa, que comprende el tiempo que precede al comienzo del partido, el cronista saluda al público y se publica información relativa al encuentro que se va a disputar (Kovljanin, 2018a, 112), como las alineaciones posibles o definitivas, el estado en el que llegan los dos equipos y los antecedentes y las posibles consecuencias del partido.

En el cuerpo, o narración en directo del partido[2], se narra la acción, de manera que la información más reciente se sitúa en la parte superior (Kovljanin, 2018a: 112). Cabe destacar que el cronista puede invertir el tiempo de descanso para aportar valoraciones personales o interactuar con el público (*ibid.*: 112).

En el cierre, el cronista lleva a cabo una breve recapitulación del partido (Kovljanin, 2018a: 112) y, en ocasiones, añade una conclusión de carácter subjetivo.

2. La traducción automática

El propósito del presente apartado es realizar una breve contextualización teórica acerca de la traducción automática (TA), una herramienta que, sin duda, ha revolucionado los Estudios de

[2] A lo largo de nuestro trabajo, nos referimos al cuerpo de la CFED con la denominación «narración en directo», puesto que consideramos que es más adecuada.

Traducción y la práctica profesional de la traducción en los últimos años y que muchos autores (Berner, 2003; Bowker y Fisher, 2010; Quah, 2006; Díaz Prieto, 2012; Koponen, 2015; Forcada *et al.*, 2016; Ortego Antón y Seghiri, 2019; O'Hagan, 2020; Sánchez Ramos y Rico Pérez, 2020) han abordado, abordan y abordarán desde una perspectiva investigadora, puesto que la investigación y el desarrollo de la TA están todavía en ciernes.

2.1. Definición y concepto de traducción automática

Desde su nacimiento en la primera mitad del siglo pasado hasta la actualidad, la traducción automática ha sido objeto de avances inmensos y de numerosos cambios de paradigma, una metamorfosis impulsada por el desarrollo tecnológico que, sin duda, resultaría inconcebible para los ingenieros y los lingüistas que idearon los primeros sistemas de TA. Sin embargo, estos cambios drásticos en la forma no han sido tales en el fondo, ya que el concepto de TA no ha variado demasiado con el tiempo.

La traducción automática se puede definir como «la aplicación de la tecnología informática a la traducción de textos de una lengua a otra sin intervención humana» (Sánchez Ramos y Rico Pérez, 2020). Consideramos que esta definición, aunque breve y sintética, es acertada, dado que a partir de ella se pueden inferir la función principal, o propósito final, de la TA y el medio para cumplir con dicha función. La TA persigue el objetivo de poder llevar a cabo el proceso de traducción entre lenguas por medio de un sistema informático, de manera que se prescinda de la traducción humana.

La TA presenta un diseño y un funcionamiento complejos, circunstancia que se refleja en su componente interdisciplinar y transversal, pues en ella confluyen disciplinas de naturaleza muy diversa (Quah, 2006: 57; Sánchez Ramos y Rico Pérez, 2020: 1-2), como la lexicografía, la lingüística, la lingüística computacional, la ingeniería informática, la inteligencia artificial y la estadística. La TA comparte esta característica esencial, su complejidad, con la traducción. Esto se debe, en gran parte, a que reducir el proceso traslativo a una especie de trasvase de palabras equivalentes entre lenguas constituye una simplificación y, en consecuencia, un error, puesto que requiere de la aplicación de un conocimiento lingüístico complejo e integral en relación con aspectos como la morfología, la sintaxis, la semántica y la capacidad de comprender y procesar conceptos abstractos como la ambigüedad (Berner, 2003: 6), así como de la consideración de otros factores relacionados con la pragmática y la situación comunicativa. Por tanto, la abstracción y el contexto adquieren una importancia considerable en el ámbito de la TA, dado que representan parte de las limitaciones actuales de los sistemas de traducción automática.

2.2. Nacimiento y evolución de la traducción automática

Pese a que todavía existe cierto debate en torno a cuándo se concibió la traducción automática como idea, sí que parece haberse alcanzado un consenso casi generalizado acerca de su

nacimiento en la práctica, que tuvo lugar a mediados del siglo XX, más concretamente a finales de la década de 1940, como una de las aplicaciones principales de los primeros ordenadores (Díaz Prieto, 2012: 141).

La historia de la traducción automática está llena de claroscuros y cuenta con periodos de optimismo seguidos por otros de contrariedad y desilusión; sin embargo, se puede resumir en los hitos presentados de aquí en adelante (Sánchez Ramos y Rico Pérez, 2020: 3-6).

En la década de 1950, las dos potencias mundiales del momento, Estados Unidos y la Unión de Repúblicas Socialistas Soviéticas (URSS), decidieron empezar a destinar fondos públicos a la investigación científica y el desarrollo de la TA, por lo que se inició un periodo de esplendor en lo que se refiere a la investigación y el desarrollo de esta tecnología y de optimismo acerca de las posibilidades que brindaría.

En 1966, se publicó el informe ALPAC (Automatic Language Peocessing Advisory Comittee), en el que se concluía que la TA no podía competir con la traducción humana debido, en gran parte, a los resultados imprecisos que ofrecía, por lo que EE. UU. retiró la financiación pública que destinaba al estudio y el desarrollo de la TA.

En la década de 1990, la irrupción de Internet revolucionó tanto el acceso a la información como la capacidad de procesamiento de los ordenadores, circunstancia que se trasladó al ámbito de la TA a través del proyecto Candie, impulsado por IBM en 1994. Se produjo un cambio de paradigma en el planteamiento conceptual del proceso de traducción de la TA promovido por la implementación de los métodos estadísticos.

A comienzos del siglo XXI, los sistemas de TA basados en estadística, encabezados por Google Translate (traductor en línea gratuito lanzado por Google en 2006) y Moses (proyecto de código abierto financiado por la Unión Europea), alcanzaron su máximo potencial y ratificaron el proceso de democratización y accesibilidad a los sistemas de TA a través de Internet que se había iniciado en la década de 1990. Además, en 2016 Google desarrolló y lanzó un sistema inédito de TA basado en redes neuronales. Con la irrupción de este y de otros sistemas de TA neuronal, como DeepL, la brecha de calidad entre la traducción automática y la traducción humana se ha estrechado considerablemente.

2.3. Sistemas de traducción automática

A lo largo de sus ocho décadas de existencia, los sistemas de TA han sido objeto de varios cambios de paradigma fomentados por el desarrollo técnico y tecnológico. Estos cambios han promovido la adopción de diferentes planteamientos conceptuales en relación con la arquitectura de estos sistemas.

Las clasificaciones de los sistemas de TA, así como sus denominaciones, son diferentes en función del autor[3]. No obstante, tomaremos como referencia la propuesta de Sánchez Ramos y Rico Pérez (2020), que distinguen cinco sistemas diferentes: sistemas de traducción automática basados en reglas, sistemas de traducción automática basados en estadística, sistemas de traducción automática basados en ejemplos, sistemas de traducción automática híbridos y sistemas de traducción automática basados en redes neuronales.

2.3.1. Sistemas de traducción automática basados en reglas

Los sistemas de traducción automática basados en reglas, o *rule-based machine translation*, representan un modelo conceptual de TA basado principalmente en la lingüística. Así, estos sistemas recurren a reglas gramaticales y a diccionarios seleccionados previamente por un ser humano para llevar a cabo el proceso de traducción, que concretan a través de la aplicación de técnicas de transferencia morfológica, sintáctica y semántica (Sánchez Ramos y Rico Pérez, 2020: 11-12).

El proceso de traducción ejecutado por los sistemas de TA basados en reglas consta de tres fases (Sánchez Ramos y Rico Pérez, 2020: 13; Maldonado González y Liébana González, 2021: 191-192; Forcada *et al.*, 2016: 174): análisis, transferencia y generación.

1. Análisis: se lleva a cabo un análisis morfológico, sintáctico y semántico del texto origen para convertirlo en una representación abstracta que contiene información monolingüe en la lengua origen.
2. Transferencia: se convierte esta representación abstracta en la lengua origen en otra equivalente en la lengua meta.
3. Generación: se origina el texto meta a partir de la representación abstracta en la lengua meta obtenida en la fase anterior.

Aunque los sistemas de TA basados en reglas pueden llegar a ofrecen buenos resultados en relación con la sintaxis cuando operan con pares de lenguas similares (Maldonado González y Liébana González, 2021: 192), su modelo conceptual basado en la lingüística, que data de la década de 1970, presenta problemas de base considerables: en primer lugar, la traducción que ofrecen tiende a ser demasiado literal (Maldonado González y Liébana González, 2021: 192); en segundo lugar, los errores de transferencia relacionados con las ambigüedades semánticas y contextuales constituyen un problema de base (Sánchez Ramos y Rico Pérez, 2020: 15); por

[3] Ortego Antón y Seghiri (2019), por ejemplo, describen cuatro tipos sistemas de TA: sistemas de TA basados en reglas, sistemas de TA basados en estadística, sistemas de TA basados en motores híbridos y sistemas de TA neuronal. En cambio, Sánchez Ramos y Rico Pérez (2020) y Metola Navaridas (2022) describen cinco clases de sistemas de TA: sistemas de traducción automática basados en reglas, sistemas de traducción automática basados en estadística, sistemas de traducción automática basados en ejemplos, sistemas de traducción automática híbridos y sistemas de traducción automática basados en redes neuronales.

último, la capacidad traslativa de estos sistemas depende de la calidad de los diccionarios y de las reglas que rigen su funcionamiento, por lo que requieren de una actualización humana constante, ardua y poco rentable desde el punto de vista económico (*ibid.*: 15). En consecuencia, los sistemas basados en reglas han quedado relegados a un segundo plano en el ámbito de la TA.

2.3.2. Sistemas de traducción automática basados en estadística

Los sistemas de traducción automática basados en estadística suponen el paso de un modelo conceptual basado en la lingüística a otro fundamentado en la estadística y la probabilidad. Estos sistemas se alimentan por medio de un corpus paralelo alineado de grandes dimensiones (millones de palabras) y emplean un proceso estadístico basado en algoritmos para llevar a cabo el proceso de traducción (Sánchez Ramos y Rico Pérez, 2020: 16).

Estos sistemas adoptan la oración, y no el texto, como unidad de traducción y operan en tres fases (Sánchez Ramos y Rico Pérez, 2020: 16; Forcada *et al*., 2016: 177):

1. primero, se segmenta el texto origen en unidades de traducción (en oraciones);
2. a continuación, se comparan estas unidades con el corpus paralelo alineado que alimenta el sistema, y
3. finalmente, se generan diferentes hipótesis de traducción y se aplica un modelo probabilístico para escoger la más adecuada.

Para poder llevar a cabo el proceso de traducción, los sistemas de TA basados en estadística cuentan con tres componentes principales (Maldonado González y Liébana González, 2021: 193; Sánchez Ramos y Rico Pérez, 2020: 17-18): el modelo de lengua, que calcula la probabilidad de que una unidad de traducción sea correcta y adecuada en la lengua meta mediante el entrenamiento con un corpus monolingüe en dicha lengua; el modelo de traducción, que establece la correspondencia entre el par de lenguas de trabajo mediante el entrenamiento con un corpus paralelo alineado; el decodificador, que analiza las hipótesis de traducción generadas y selecciona aquella con mayor probabilidad mediante la aplicación de un algoritmo.

En la primera década del siglo XXI, los sistemas de TA basados en estadística alcanzaron su máximo potencial y experimentaron un periodo de auge gracias a que brindaban resultados de calidad y disponían de un alto nivel de aceptación (Sánchez Ramos y Rico Pérez, 2020: 18). Sin embargo, el grado de calidad de las traducciones que ofrecen estos sistemas es demasiado dependiente del par de lenguas de trabajo y de la calidad de los corpus con los que se entrenen (*ibid.*: 18). Por consiguiente, si estos corpus no están bien alimentados, la hipótesis de traducción más probable puede no ser la más adecuada, dando lugar a errores de diversa naturaleza (Maldonado González y Liébana González, 2021: 194).

2.3.3. Sistemas de traducción automática basados en ejemplos

Al igual que los sistemas basados en estadística, los sistemas basados en ejemplos también presentan un concepto fundamentado en el entrenamiento con corpus paralelos alineados. Sin embargo, estos dos sistemas se diferencian en el procedimiento de conversión de la información del corpus en la traducción final. Así, los sistemas basados en estadística utilizan modelos estadísticos y probabilísticos, mientras que los sistemas basados en ejemplos llevan a cabo un proceso de traducción por analogías (Sánchez Ramos y Rico Pérez, 2020: 19), de manera que se sirven de segmentos (oraciones) traducidos previamente como ejemplo para realizar otras traducciones.

En general, los sistemas basados en ejemplos presentan los mismos inconvenientes que los sistemas basados en estadística, aunque su entrenamiento es menos costoso (Sánchez Ramos y Rico Pérez, 2020: 20).

2.3.4. Sistemas de traducción automática híbridos

Por sistemas de traducción automática híbridos entendemos «aquellos que combinan los sistemas de TA basados en reglas y en estadística» (Ortego Antón y Seghiri, 2019: 333). La idea que subyace tras estos sistemas es subsanar, en la medida de lo posible, los errores de las traducciones generadas por los sistemas basados en estadística mediante la incorporación de reglas y entradas de diccionarios (Sánchez Ramos y Rico Pérez, 2020: 20), como ocurre en los sistemas basados en reglas.

2.3.5. Sistemas de traducción automática basados en redes neuronales

Los sistemas de traducción automática basados en redes neuronales representan el concepto de traducción automática más novedoso. Como consecuencia de la aparición de estos sistemas, que ocupan a día de hoy una posición central en lo concerniente a la investigación en el ámbito de la traducción automática, los sistemas de TA basados en estadística han quedado relegados a un segundo plano.

Al contrario que los sistemas comentados anteriormente, los sistemas basados en redes neuronales introducen dos métodos relacionados con la inteligencia artificial inéditos hasta el momento (Sánchez Ramos y Rico Pérez, 2020: 21): las redes neuronales artificiales (Forcada *et al*., 2016: 152) y el aprendizaje profundo, o *deep learning*.

Las redes neuronales artificiales constituyen el núcleo de la arquitectura de estos sistemas. Estas redes emulan el funcionamiento del cerebro humano (Ortego Antón y Seghiri, 2019: 333), puesto que están compuestas por neuronas (o nodos) interconectadas entre sí con la capacidad de activarse o no al recibir un estímulo y de recibir un aprendizaje profundo mediante un entrenamiento guiado por un corpus paralelo alineado de gran tamaño (Sánchez Ramos y Rico Pérez, 2020: 21-23). Además, la información lingüística, sean palabras u oraciones, que transita

por estas redes se codifica y se representa de manera vectorial (*ibid.*: 21); esto es, mediante números.

La arquitectura de los sistemas basados en redes neuronales presenta un «modelo de percepción multicapa» que se fundamenta en la activación o no activación de las neuronas de las diferentes capas al recibir información por medio de un estímulo (Sánchez Ramos y Rico Pérez, 2020: 23). Este modelo está dividido en tres capas diferentes, todas ellas compuestas por neuronas (*ibid.*: 23): la capa de entrada, la capa oculta y la capa de salida.

- En la capa de entrada, las neuronas reciben la información y la transmiten a las neuronas de la capa oculta.
- En la capa oculta, se produce el entrenamiento de las neuronas que conforman la capa mediado por un corpus paralelo alineado de grandes dimensiones.
- En la capa de salida, las neuronas reciben la información proporcionada por las de la capa oculta y generan un resultado definitivo a fin de valorar si el entrenamiento es el adecuado.

Aunque los sistemas de traducción automática basados en redes neuronales están lejos de alcanzar la perfección, pues requieren de corpus paralelos de grandes dimensiones para su entrenamiento y solo pueden desarrollarse y operar si se dispone de un *hardware* muy potente (Sánchez Ramos y Rico Pérez, 2020: 25-26), estos sistemas brindan muchas ventajas frente a los sistemas basados en estadística (Lommel, 2017, en Sánchez Ramos y Rico Pérez, 2020: 26):

- son garantía de calidad en lo referente a la fluidez del texto meta que generan;
- requieren de un volumen de datos menor para proporcionar un resultado de la misma calidad, y
- se pueden extrapolar entre pares de lenguas sin tener que disponer de datos explícitos para alguna de ellas.

Google Translate[4] y DeepL[5] son dos motores de traducción automática en línea y de acceso gratuito que operan por medio de sistemas de traducción automática basados en redes neuronales.

2.4. Evaluación de la traducción automática

Con evaluación de la traducción automática nos referimos al análisis de la traducción proporcionada por un sistema de traducción automática (Sánchez Ramos y Rico Pérez, 2020: 23). En realidad, este concepto nace de la necesidad de desarrollar métodos objetivos para analizar y evaluar la calidad del resultado generado por los sistemas de TA atendiendo a un propósito concreto. Por consiguiente, el sistema de evaluación adoptado ha de presentar un diseño, un

[4] https://translate.google.es/?hl=es.
[5] https://www.deepl.com/es/translator.

enfoque analítico y unas características acordes al objetivo que se pretende conseguir a través de dicha evaluación.

Sánchez Ramos y Rico Pérez (2020: 31) distinguen dos modelos principales de evaluación de la TA: la evaluación manual y la evaluación automática.

2.4.1. Evaluación manual

La evaluación manual de la TA es aquella que llevan a cabo profesionales humanos y presenta un enfoque holístico y analítico (Sánchez Ramos y Rico Pérez, 2020: 34).

Por un lado, se trata un modelo de evaluación holístico porque aborda el texto meta de manera integral con base en dos parámetros (Sánchez Ramos y Rico Pérez, 2020: 34): la fidelidad (*adecuacy*), que hace referencia a la transferencia semántica, y la fluidez (*fluency*), en relación con la adecuación lingüística en la lengua meta, que se manifiesta, por ejemplo, en la corrección gramatical y ortográfica y en la conservación de la coherencia terminológica.

Por otro lado, se considera un modelo de evaluación analítico debido a que su aplicación práctica consiste en identificar los errores cometidos por el sistema de TA, clasificarlos según una plantilla preestablecida en la que se apliquen tanto una tipología de errores como una escala de gravedad para dichos errores y, finalmente, llevar a cabo una valoración numérica general del texto meta proporcionado por el sistema de TA (Sánchez Ramos y Rico Pérez, 2020: 34).

El modelo DQF (Dynamic Quality Framewok), propuesto por TAUS (Translation Automation User Society), es un ejemplo representativo de sistema de evaluación manual de la TA. La concepción de este modelo responde a la necesidad de establecer un medio que permita comparar los criterios de evaluación de la calidad de las traducciones adoptados por diferentes empresas (Sánchez Ramos y Rico Pérez, 2020: 40). Para cumplir con este objetivo, el modelo DQF cuenta con una tipología de errores dividida en siete categorías (precisión, fluidez, terminología, estilo, diseño, convenciones y veracidad) a las que se les debe asignar un valor numérico y con una escala de gravedad de los errores conformada por cinco niveles (grave, importante, poco importante, neutro y excelente) (*ibid.*: 40-41). Tras identificar los errores del texto meta generado por un sistema de TA, clasificarlos y atribuirles un valor numérico, se puede calcular la calidad ponderada de la traducción por medio de fórmulas matemáticas (*ibid.*: 41-43).

2.4.2. Evaluación automática

La evaluación automática de la TA es aquella que se realiza por medio de programas informáticos (Sánchez Ramos y Rico Pérez, 2020: 34).

Por lo general, el funcionamiento de la evaluación automática consiste en la comparación mediante métricas de las traducciones proporcionadas por el sistema de TA con una traducción de referencia realizada por un traductor humano (Sánchez Ramos y Rico Pérez, 2020: 43), de modo que la traducción más adecuada es la que más se acerca a la traducción de referencia.

La evaluación automática presenta ciertas ventajas frente a la evaluación manual (Sánchez Ramos y Rico Pérez, 2020: 43):

- brinda una mayor productividad y permite abaratar costes;
- no requiere de la intervención de traductores nativos, y
- presenta un enfoque más objetivo.

Sánchez Ramos y Rico Pérez (2020: 43) distinguen tres clases de métricas de evaluación automática: las métricas basadas en la precisión, como BLEU, las métricas basadas en la exhaustividad, como METEOR, y las métricas basadas en la distancia de Levenshtein, como TER.

2.5. Posedición

Aunque en los últimos años la traducción automática ha experimentado un salto cualitativo que ha cerrado de manera considerable la brecha de calidad que la privaba de competir frente a la traducción humana, los textos meta generados por los sistemas de traducción automática continúan estando lejos, al menos por norma general, de cumplir con los estándares de calidad del sector y de satisfacer las necesidades de los clientes sin precisar de una mediación humana. Esta tarea de revisión y corrección de los errores lingüísticos (ortográficos, gramaticales, semánticos, contextuales y estilísticos) cometidos por los sistemas de TA que desempeña un traductor humano se conoce como posedición.

La posedición se puede definir como «the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria» (O'Brien, 2011: 197). De esta definición se pueden resaltar varios elementos. En primer lugar, la posedición es una técnica que consiste en corregir la traducción generada por un sistema de TA llevando a cabo solo aquellas modificaciones que resulten esenciales (Forcada *et al.*, 2016: 107). En segundo lugar, ese proceso de corrección está a cargo de un traductor profesional. Por último, la posedición se lleva a cabo de acuerdo con un umbral de calidad preestablecido que se ajuste a la finalidad del texto meta y a las necesidades del cliente (Sánchez Gijón, 2016: 160; Sánchez Ramos y Rico Pérez, 2020: 76).

La decisión de optar por el doble proceso de traducción automática y posedición en remplazo de la traducción humana responde a la relación que se establece entre tres parámetros (Nunes Vieira, 2020: 320; Sánchez Ramos y Rico Pérez, 2020: 78): la capacidad de alcanzar la calidad esperada del texto meta, el aumento de la productividad y la reducción de los costes de traducción. De esta forma, poseditar el texto meta generado por un sistema de TA merece la pena siempre y

cuando sea garantía de ajustarse a los estándares de calidad esperados y brinde ventajas en relación con el tiempo y el capital invertidos con respecto a la traducción humana.

En cuanto a los tipos de posedición, Allen (2003) propone dos: la posedición completa (*full-post-editing*) y la posedición rápida (*light post-editing*). La elección del tipo de posedición adecuado está determinada por la calidad esperada del texto meta (Sánchez Gijón, 2016: 160) y por el esfuerzo de corrección humana requerido para alcanzar esta calidad en función del resultado ofrecido por el sistema de TA (Nunes Vieira, 2020: 325).

La posedición completa es la más adecuada cuando se pretende alcanzar un estándar de calidad similar al que ofrece la traducción humana (Sánchez Ramos y Rico Pérez, 2020: 78), de manera que implica una revisión y corrección integral del resultado generado por el sistema de TA. En cambio, la posedición rápida es idónea para que el receptor pueda extraer la idea general del texto meta y aproximarse a su contenido (*ibid.*: 78), por lo que priman la comprensibilidad y la transferencia semántica (Nunes Vieira, 2020: 325) sobre los aspectos formales, como la ortografía, la sintaxis, el estilo y el formato.

3. El corpus

El presente apartado va a estar dedicado al corpus, pues es una herramienta que desempeña un papel fundamental en el desarrollo de nuestra investigación.

Comenzaremos por definir y delimitar el concepto de corpus; para ello, comentaremos definiciones actuales y aportaremos nuestra propia definición en línea con el planteamiento inicial y el desarrollo de nuestro trabajo. Asimismo, trataremos algunos de los sistemas de clasificación de corpus más trascendentes de los últimos años. Por último, nos adentraremos en los protocolos de compilación de corpus y analizaremos los conceptos de equilibrio y representatividad, de gran relevancia a la hora de trabajar con corpus.

3.1. Definición de corpus

El corpus constituye una herramienta esencial en los campos de la traducción y del análisis contrastivo. Sin embargo, el concepto de corpus no es estático ni inmutable, ya que ha ido evolucionando, desarrollándose y adaptándose a lo largo del tiempo desde que se acuñó a finales del siglo XX. Este cambio ha estado motivado, principalmente, por el desarrollo exponencial de las tecnologías de la información y de la comunicación, que han ido modelando esta herramienta, su funcionamiento y sus aplicaciones.

Son muchos los autores (Francis, 1982; Sinclair, 1991 y 2005; Baker, 1995; Sánchez, 1995; McEnery y Wilson, 1996; Bowker y Pearson, 2002; McEnery *et al.*, 2005; Zanettin, 2012; Rojo, 2016; Ortego Antón, 2019; Sánchez Carnicer, 2022) que, desde la década de 1980, han

contribuido al desarrollo teórico de la noción de corpus mediante definiciones que respondían a actualizaciones conceptuales motivadas, sobre todo, por los avances en la informática, tales como los siguientes: el cambio generalizado en el soporte de los textos de papel a formato electrónico; la democratización del acceso a la información a través de Internet; la sistematización, sea mediante procesos automáticos o semiautomáticos, del análisis de textos (Baker, 1995: 225), así como de su compilación; la gestión instantánea de grandes cantidades de información (Corpas Pastor y Seghiri, 2009: 77); la aparición y el desarrollo de herramientas y aplicaciones que catalizan los diferentes procesos inherentes o complementarios a la metodología relativa a los corpus, como son la compilación, la gestión y la explotación de corpus.

No obstante, pese a toda la literatura existente acerca del concepto de corpus, consideramos conveniente restringir nuestro análisis a dos definiciones actuales y completas a fin de, por un lado, deducir las características principales de los corpus y, por otro lado, presentar nuestra propia definición de corpus, que tomaremos como referencia para llevar a cabo nuestra investigación.

Así, nos hemos decantado por las definiciones propuestas por Sinclair (2005) y Sánchez Carnicer (2021). Sinclair (2005: 16) define «corpus» como «*a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or a language variety as a source of data for linguistic research*». Por su parte, Sánchez Carnicer (2021: 73) hace la siguiente propuesta:

> [...] entendemos «corpus» como el conjunto de textos, independientemente de su formato y longitud, en formato electrónico en una o varias lenguas que se compilan con el objetivo de estudiar y analizar la lengua o parte de la lengua de la que son representativos.

Sendas definiciones comparten algunos elementos que constituyen, en definitiva, la mayor parte de las características principales de los corpus. En primer lugar, ambas propuestas coinciden en el establecimiento del texto (o del fragmento textual) como unidad básica; es decir, el corpus consta, por definición, de una compilación de textos o de fragmentos textuales. En segundo lugar, se alude al componente electrónico de los textos que conforman el corpus, pues están publicados en soporte electrónico, no en formato físico. En tercer lugar, se describe una característica que todos los textos que integran el corpus han de compartir: la representatividad, bien de la lengua, bien de una variedad o dominio de la lengua en concreto. Por último, en las dos propuestas se hace referencia a la finalidad o función última del corpus como herramienta de investigación lingüística[6].

[6] Pese a que los corpus se compilan con el objetivo de contribuir a la investigación lingüística, existen dos enfoques metodológicos contrapuestos en relación con el papel que desempeña el corpus en el seno de dicha investigación: *corpus-based approach* (Baker, 1993) y *corpus-driven approach* (Tognini-Bonelli, 2001).

Además de estos elementos comunes presentes en las dos definiciones, cada autor pone énfasis en una característica definitoria de los corpus. De esta forma, Sinclair (2005: 16) pone de manifiesto que la selección de los textos que conforman el corpus se realiza en base a criterios predeterminados, mientras que Sánchez Carnicer (2021: 73) precisa que un corpus puede estar compuesto por textos publicados en una, dos o más lenguas.

Ahora, procedemos a plantear nuestra definición de corpus, que toma como referencia toda la información expuesta anteriormente y la completa con el objetivo de construir una base teórica que se ajuste a nuestra investigación. Así, podemos definir «corpus» de la siguiente manera:

> herramienta de investigación lingüística compuesta por un conjunto de textos o fragmentos textuales en soporte electrónico en una o varias lenguas que se compilan atendiendo a criterios predeterminados con el propósito de llevar a cabo un estudio objetivo del uso de la lengua o de la variedad o dominio de la lengua de la que son representativos.

3.2. Clasificación de los corpus

De forma paralela a la evolución del concepto de corpus, muchos autores (Baker, 1995; Laviosa, 1997; Torruella y Llisterri, 1999; Corpas Pastor, 2001; Granger, 2003; Faya Ornia, 2015) han realizado sus propuestas taxonómicas en relación con esta herramienta en base, sobre todo, a sus características principales y a su aplicación práctica. Por norma general, estas clasificaciones comparten los mismos criterios de base empleados en su confección, como son la variedad de la lengua analizada, el grado de especialización de la lengua, el número de lenguas implicadas, los tipos de textos que conforman el corpus y la temporalidad; no obstante, cada autor adopta un enfoque distinto para abordar esta cuestión y aporta, por consiguiente, un planteamiento, una configuración y unos matices personales que definen su clasificación y la diferencian del resto. En este sentido, pese a que el catálogo taxonómico de los corpus es amplio e interesante, dedicaremos este apartado a estudiar las clasificaciones establecidas por Baker (1995), Corpas Pastor (2001) y Faya Ornia (2015).

3.2.1. Baker (1995)

Esta autora realiza dos aportaciones en relación con la clasificación de los corpus. Por un lado, propone seis criterios de clasificación de corpus y, por otro lado, establece tres tipos de corpus atendiendo a la clase de textos, originales o traducciones, que componen un corpus.

En lo que se refiere a estos seis criterios de clasificación de los corpus, estos responden a otros seis parámetros predeterminados, que son los siguientes (Baker, 1995):

1. el grado de especialización de la lengua (lengua general o lengua de especialidad);
2. la modalidad de la lengua (lengua escrita o lengua oral);

3. la temporalidad (corpus sincrónicos o corpus diacrónicos);
4. las fuentes y los géneros;
5. la variedad geolectal de la lengua, y
6. el número de lenguas implicadas.

Además de estos seis criterios, Baker (1995) plantea una primera división de los corpus según si los textos son originales o traducciones. Así, aplicada esta distinción, la autora distingue tres tipos de corpus: corpus paralelos (*parallel corpora*), corpus multilingües (*multilingual corpora*) y corpus comparables (*comparable corpora*).

Los corpus paralelos constan de textos originales en una lengua y de sus respectivas traducciones en otra lengua (Baker, 1995). Los corpus multilingües, por su parte, están compuestos por dos o más corpus monolingües en diferentes lenguas para cuya compilación se han aplicado criterios parecidos (*ibid.*). Finalmente, los corpus comparables reúnen características de los otros dos tipos, pues son «two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translation in that language from a given source language or languages» (*ibid.*).

3.2.2. Corpas Pastor (2001)

En línea con la estructuración y la sistematización mediante criterios preestablecidos del diseño de propuestas taxonómicas de corpus, Corpas Pastor (2001) construye su clasificación a partir de características compartidas por los textos que conforman el corpus. Esta autora filtra dichas características compartidas, como se muestra en la Tabla 1, mediante la aplicación de cinco criterios:

CRITERIOS		CLASIFICACIÓN
I.	Porcentaje y la distribución de los tipos de textos del corpus	corpus grande o extenso
		corpus equilibrado
		corpus piramidal
		corpus monitor
		corpus paralelo
		corpus comparable
II.	Especificidad de los textos del corpus	corpus general
		corpus especializado
		corpus genérico
		corpus canónico
		corpus periódico o cronológico
		corpus diacrónico
III.	Cantidad de texto recogida en cada uno de los textos del corpus	corpus textual
		corpus de referencia
		corpus léxico
		corpus no anotado

IV.	**Codificación y anotación de los corpus**	corpus anotado
V.	**Documentación que acompaña a los textos del corpus**	corpus documentado
		corpus no documentado

Tabla 1. Clasificación de los corpus según Corpas Pastor (2001).

El primer criterio (Corpas Pastor, 2001: 157-158) se basa en el porcentaje y la distribución de los tipos de textos contenidos en el corpus. Así, en el *corpus grande* priman la extensión y el número de palabras, parámetros que suelen ser muy elevados, sobre otras características como el equilibrio y la representatividad; el *corpus equilibrado*, por el contrario, se fundamenta en la distribución equitativa de las variedades de la lengua representadas; en el *corpus piramidal*, los textos se distribuyen por niveles; el *corpus monitor* posee un volumen textual constante y actualizado; el *corpus paralelo* consta de textos en una lengua origen y de sus traducciones en una lengua meta (o varias); por último, el *corpus comparable* incorpora textos originales en dos o más lenguas que se han compilado de acuerdo a criterios similares.

El segundo criterio (Corpas Pastor, 2001: 158) está determinado por distintos niveles de especificidad de los textos contenidos en el corpus. Atendiendo a este criterio, son seis los tipos de corpus resultantes. El *corpus general* y el *corpus especializado* mantienen una relación de contraposición, pues el primero es representativo de la lengua general, mientras que el segundo lo es de una lengua de especialidad concreta; el *corpus genérico*, como su nombre indica, es representativo de un género textual determinado; el *corpus canónico*, por su parte, consta exclusivamente de textos de un autor; finalmente, los dos últimos, el *corpus periódico o cronológico* y el *corpus diacrónico* presentan una especificidad de carácter temporal, de forma que el primero reúne textos producidos a lo largo de un periodo de tiempo definido y el segundo está compuesto por textos de periodos de tiempo sucesivos a fin de estudiar la evolución lingüística.

El tercer criterio (Corpas Pastor, 2001: 158-159) se fundamenta en la porción de texto contenida en el corpus de cada uno de los textos que lo conforman. De la aplicación de este criterio se originan tres nuevos tipos de corpus: el *corpus textual*, que incorpora textos completos, el *corpus de referencia*, que consta de fragmentos de textos, y el *corpus léxico*, que está compuesto por fragmentos de texto pequeños de una extensión similar.

El cuarto criterio (Corpas Pastor, 2001: 159) implica una distinción dual atendiendo al etiquetado (o anotación) de los textos entre el *corpus no anotado*, que consta de textos sin etiquetar, y el *corpus anotado*, que está formado por textos etiquetados siguiendo un método sistemático.

El quinto criterio (Corpas Pastor, 2001: 159) alude a la presencia o ausencia de documentación relativa a la procedencia de los textos del corpus. Así pues, en el *corpus documentado* se presenta

información sobre la procedencia de los textos, mientras que en el *corpus no documentado* no se manifiesta dicha información.

3.2.3. Faya Ornia (2015)

La propuesta taxonómica de Faya Ornia (2015) destaca por su carácter unificador e integrador, ya que aúna clasificaciones de otros autores bajo una única. Además, presenta una organización en forma de árbol que brinda la posibilidad de establecer relaciones entre todos los tipos de corpus resultantes, una característica novedosa y distintiva.

Faya Ornia (2015: 344) estructura su clasificación en dos fases: la primera hace referencia a los aspectos formales de los corpus y la segunda, a sus aspectos lingüísticos. Estas dos fases constan, a su vez, de varias divisiones ulteriores, que llamaremos subfases, definidas con base en criterios acotados, de forma que los diferentes tipos de corpus se originan a partir de la aplicación de estos criterios. No obstante, los tipos de corpus resultantes de las dos fases y de las diferentes subfases no son independientes, sino que mantienen entre sí una relación de jerarquía. En consecuencia, el grado de vinculación que comparten dos tipos de corpus está determinado por la posición relativa que ocupan en la ramificación.

En la siguiente tabla, se muestra la primera fase, que atañe a los aspectos formales, con sus diferentes subfases por orden de jerarquía:

N.º DE LA SUBFASE: CRITERIO DE LA SUBFASE	TIPOS DE CORPUS DE LA SUBFASE	DESCRIPCIÓN DE LOS TIPOS DE CORPUS DE LA SUBFASE
1.1: medio de difusión	corpus oral	compuesto solo por textos orales
	corpus escrito	compuesto solo por textos escritos
	corpus oral y escrito	compuesto por textos orales y escritos
1.2: grado de amplitud o restricción del análisis	corpus de referencia	representativo de la lengua general
	corpus de no referencia	no representativo de la lengua general
1.3: restricción de carácter temporal	corpus sincrónico	textos restringidos a nivel temporal
	corpus diacrónico	textos no restringidos a nivel temporal
1.4: posibilidad de adición y actualización de información	corpus monitorizado	posibilidad de añadir nuevos textos
	corpus no monitorizado	imposibilidad de añadir nuevos textos
1.5: extensión de los textos[7]	corpus de textos completos	compuesto solo por textos completos

[7] Subfase basada en la aplicación de un criterio de la clasificación de Laviosa (1997).

	corpus de muestras	compuesto solo por fragmentos de texto
	corpus mixto	compuesto por fragmentos de texto y por textos completos
1.6: publicación de los textos	corpus de textos publicados	compuesto solo por textos publicados
	corpus de textos no publicados	compuesto solo por textos no publicados
	corpus de textos publicados y no publicados	compuesto por textos publicados y no publicados
1.7: grado de especialización de los textos	corpus generalista	textos pertenecientes a la lengua general
	corpus especializado	textos pertenecientes a una lengua especializada

Tabla 2. Primera fase de la clasificación de los corpus según Faya Ornia (2015: 345-349): los aspectos formales.

Se ha de tener en cuenta que, como consecuencia del modelo de interdependencia característico de este sistema de clasificación, los tipos de corpus resultantes de cada subfase son, en cierta manera, dependientes de los de la subfase anterior. Esto se debe, en realidad, al carácter secuencial del orden de aplicación de los criterios de cada subfase, que da lugar a una característica definitoria de este sistema de clasificación: la verticalidad de su aplicación, que atañe tanto a la primera fase, en sentido 1.1-1.7, en la que se abordan los aspectos formales del corpus, como a la segunda, en sentido 2.1-2.4 (como se muestra más adelante), en relación con los aspectos lingüísticos del corpus. Por consiguiente, aunque la clasificación se divida en dos fases diferentes, la relación de verticalidad que mantienen las subfases de la primera se conserva y prosigue en las cuatro que conforman la segunda.

La segunda fase se enfoca en los aspectos lingüísticos del corpus y cuenta con cuatro subfases. Dado que esta fase presenta una disposición más asimétrica que la primera, consideramos que el sistema de representación más adecuado es el diagrama de árbol, el mismo que utiliza Faya Ornia (2015: 352).

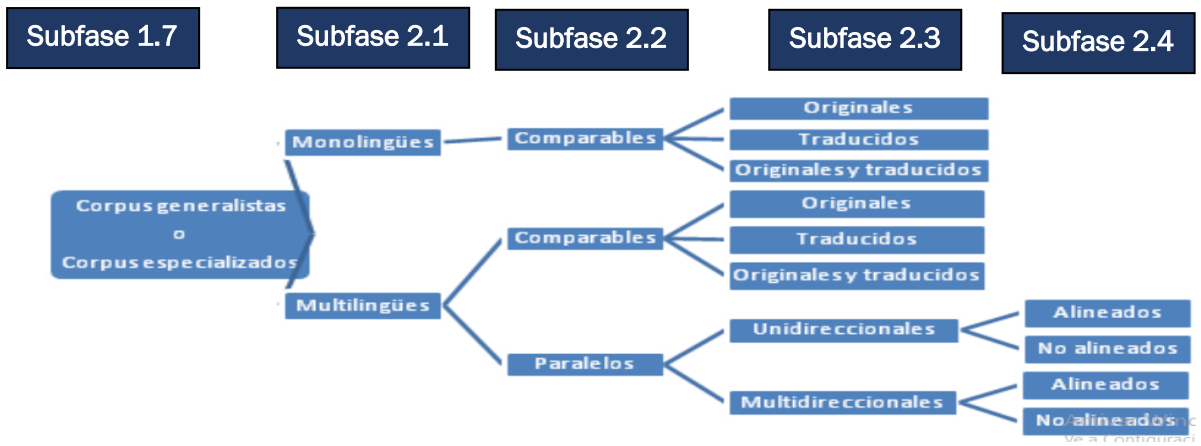


Ilustración 1. Segunda fase de la clasificación de los corpus según Faya Ornia (2015: 352): los aspectos lingüísticos.

3.3. Metodología de compilación de corpus

Con «metodología de compilación de corpus» nos referimos, en realidad, a los pasos que se han de seguir para la correcta compilación de un corpus. En este apartado, reflexionaremos sobre los elementos que ha de incluir una metodología de compilación de corpus, estudiaremos en detalle dos metodologías de referencia, la de Vargas Sierra (2005) y la de Seghiri (2011), y, por último, explicaremos en detalle la metodología que vamos a adoptar para la compilación del corpus que presentamos en este trabajo, al que hemos denominado FÚTBOL_MXM.

Antes de abordar la metodología de compilación de corpus, resulta necesario determinar los elementos que la constituyen. Así, bajo nuestro punto de vista, esta consta de dos elementos: el diseño de corpus y el protocolo de compilación de corpus. La decisión de incluir estos dos elementos, y no solo el segundo, se fundamenta en la idea de que ambos constituyen pasos previos a la explotación de corpus; es decir, estos procesos forman parte de la confección de corpus y sientan la base de su aplicación práctica real, su explotación.

Por norma general, los autores que han tratado esta materia desglosan esta metodología en fases secuenciales y relativamente estancas, probablemente con fines descriptivos y de comprensión. Sin embargo, la metodología de compilación de corpus también se puede considerar, hasta cierto punto, un proceso continuo, ya que las fronteras de las fases que conforman toda metodología, así como el orden canónico de su aplicación, pueden desvirtuarse y alterarse en función de las características y las necesidades específicas del proyecto de investigación lingüística al que sirven. En todo caso, en el presente trabajo nos decantamos por el primer planteamiento, pues responde a un enfoque más estructurado, sistemático y descriptivo.

En lo que respecta a propuestas metodológicas de compilación de corpus concretas, consideramos idóneas para el desarrollo de nuestra investigación las publicadas por Vargas Sierra (2005) y Seghiri (2011), que explicaremos en los siguientes apartados.

3.3.1. Vargas Sierra (2005)

La propuesta metodológica de esta autora se caracteriza por integrar tres elementos bien diferenciados a los que ya se ha hecho alusión anteriormente: el diseño de corpus, el protocolo de compilación de corpus y la explotación de corpus. Así, Vargas Sierra (2015: 547) incluye estos tres elementos en una metodología que se estructura en cuatro fases:

1. las especificaciones y diseño;
2. el soporte lógico y físico (*hardware* y *software*);
3. la adquisición textual, y
4. el procesamiento informático de los textos.

De acuerdo con la división de los elementos integradores de esta metodología mencionada anteriormente, la primera y la segunda fase se corresponderían con el diseño de corpus, la tercera conformaría el protocolo de compilación de corpus y, por último, la cuarta representaría la explotación de corpus. Puesto que la cuarta fase, el procesamiento informático de los textos, no forma parte de la metodología relativa a la compilación de corpus, sino a su explotación, nos centraremos en explicar de manera resumida las tres primeras.

3.3.1.1. Las especificaciones y diseño

Vargas Sierra (2005: 548-575) señala que en esta primera fase se tienen que abordar, principalmente, dos cuestiones. En primer lugar, el diseño del corpus debe estar, ante todo, al servicio del objetivo que se persigue con el proyecto de investigación, de forma que los componentes lingüístico (o textual) e informático del diseño deben adecuarse a dicho proyecto. En segundo lugar, en el diseño del corpus se debe tener en cuenta su definición, entendida como el establecimiento de sus características principales; para ello, resulta conveniente aplicar un sistema de clasificación de corpus adecuado y aclarar otros factores relevantes como el equilibrio y la representatividad.

3.3.1.2. El soporte lógico y físico (hardware y software)

Una vez completada la fase de diseño del corpus, hay que realizar una valoración de la tecnología requerida para su compilación, su procesamiento y su explotación (Vargas Sierra, 2005: 575-577). Esta tecnología puede comprender desde un ordenador personal donde compilar y almacenar el corpus y descargar las aplicaciones necesarias para su procesamiento y explotación (*software*) hasta los propios componentes internos del ordenador (*hardware* interno), como el procesador o la memoria RAM, y otros componentes externos, pero directamente relacionados con el mismo (*hardware* periférico), como un ratón, una impresora o un escáner.

3.3.1.3. La adquisición textual

Atendiendo a nuestra propuesta metodológica de compilación de corpus dividida en dos elementos, el diseño de corpus y el protocolo de compilación de corpus, esta fase se corresponde con el segundo elemento. La fase de adquisición textual propuesta por Vargas Sierra (2005: 577-638) comprende las siguientes tareas: la búsqueda documental, el establecimiento y la aplicación de los criterios de selección de los textos, que se dividen en criterios internos (como el tema y el estilo) y criterios externos (como el género textual, la autoría, la facticidad y la lengua), y, finalmente, el registro sistemático e informatizado de los textos compilados. Además, esta autora resalta el papel que desempeña Internet en el proceso de adquisición textual.

3.3.2. Seghiri (2011)

La metodología protocolizada de compilación de corpus planteada por Seghiri (2011) destaca, sobre todo, por su practicidad y por su reproducibilidad. Por un lado, se trata de una metodología sencilla y breve, pero, al mismo tiempo, completa y sistemática, todas ellas características que denotan que se diseñó y creó específicamente para su aplicación en la práctica. Por otro lado, esta metodología ha tenido muy buena acogida en el ámbito académico, puesto que otros autores (Ortego Antón, 2019; Sánchez Carnicer, 2022) la han adoptado para llevar a cabo sus propias investigaciones.

Esta autora estructura su propuesta metodológica de compilación de corpus, previo diseño, en cuatro fases (Seghiri, 2011: 17):

1) búsqueda y acceso a la información;
2) descarga de datos;
3) normalización, y
4) almacenamiento.

3.3.2.1. Búsqueda y acceso a la información

La primera fase consiste en buscar, identificar y acceder a la información que sea susceptible de formar parte del corpus a través de Internet. Seghiri (2011: 18) describe dos tipos de búsqueda principales: la búsqueda institucional, a través de las páginas web oficiales de instituciones, asociaciones u organizaciones nacionales e internacionales, y la búsqueda por palabras clave, que se fundamenta, sobre todo, en el uso de operadores *booleanos*, como los truncamientos y las comillas (“”), para concretar la búsqueda.

3.3.2.2. Descarga de datos

Por lo general, la segunda fase, dedicada a la descarga de datos, se lleva a cabo de forma manual, aunque también se puede realizar una descarga automática en lotes de la información mediante programas como GNU Wget o GetBot (Seghiri, 2011: 22). En determinadas ocasiones, los documentos descargados pueden estar codificados en formatos de archivo diferentes, circunstancia que se ha de corregir en la siguiente fase.

3.3.2.3. Normalización

Una vez completada la descarga de los documentos que conforman el corpus, resulta necesario homogeneizar los formatos de los archivos que los codifican bajo una única extensión (.txt, por ejemplo). En caso de que se encuentren trabas en el proceso de homogeneización de los archivos, se puede recurrir a programas de conversión de archivos (Seghiri, 2011: 22).

3.3.2.4. Almacenamiento

Al proceso de homogeneización de los documentos descargados, por lo menos en lo que respecta al formato de los archivos, le sigue su almacenamiento. Esta última fase consiste en adoptar un sistema lógico de identificación, denominación y relación para todos los documentos. Asimismo, Seghiri (2011: 22) propone crear una carpeta principal, denominada de tal manera que se explicite la temática del corpus, y varias subcarpetas donde guardar dichos documentos atendiendo a diferentes criterios, como son la lengua y el formato de los mismos. El propósito de esta fase es guardar y almacenar los documentos del corpus de manera organizada y sistemática, de forma que se consiga una identificación rápida de los documentos del corpus y se posibiliten ampliaciones futuras a otras lenguas o tipos textuales (Seghiri, 2011: 22).

3.4. Equilibrio y representatividad de un corpus

El equilibrio y la representatividad son dos características del corpus que atienden a los componentes cualitativo y cuantitativo de la información que contiene, por lo que constituyen dos parámetros clave para determinar si un corpus se adecúa al propósito y a las características de un proyecto de investigación.

McEnery y Hardie (2012: 8-9) llevan a cabo una aproximación al concepto de equilibrio al concluir que «proportions of data in our corpus reflect, in some way, the numbers of each type of interaction of interest that actually occur». No obstante, estas interacciones de interés varían en función del corpus.

En cuanto a la representatividad, Villayandre Llamazares (2008: 341) indica que «la selección de los textos, además de a unos criterios adecuados, debe responder a parámetros estadísticos que garanticen que los textos 'representan' la variedad de lengua objeto de estudio ('muestra representativa'».

MARCO PRÁCTICO

1. Metodología del marco práctico

En este apartado, abordaremos la metodología adoptada para llevar a cabo el análisis práctico, obtener los resultados y extraer las conclusiones. Esta metodología está dividida en cuatro partes que coinciden con los cuatro pilares principales sobre los que se cimienta la parte práctica del presente trabajo de investigación: la compilación del corpus FÚTBOL_MXM, el equilibrio y la representatividad del corpus FÚTBOL_MXM, la explotación del corpus FÚTBOL_MXM y la metodología del análisis práctico.

1.1. Compilación del corpus FÚTBOL_MXM

Como se manifiesta en el apartado «3.3. Metodología de compilación de corpus», creemos conveniente integrar en la metodología de compilación de nuestro corpus dos elementos: el diseño del corpus y el protocolo de compilación del corpus. Esta decisión se fundamenta en la idea de que ambos elementos constituyen la confección del corpus y preceden a su explotación.

En línea con este planteamiento, consideramos que la opción que más se adecúa a la compilación del corpus FÚTBOL_MXM y a los objetivos que se persiguen con el desarrollo de nuestro trabajo es confeccionar una metodología mixta, resultado de la combinación de las propuestas de Vargas Sierra (2005) y Seghiri (2011). Así, la metodología de compilación del corpus FÚTBOL_MXM consta de las cinco fases presentadas a continuación:

1. diseño del corpus FÚTBOL_MXM;
2. búsqueda y acceso a la información;
3. descarga de datos;
4. normalización, y
5. almacenamiento.

Por tanto, nuestra metodología aúna una primera fase, «diseño del corpus FÚTBOL_MXM», que es equivalente a la primera fase descrita por Vargas Sierra (2005), «las especificaciones y diseño», y las cuatro fases, «búsqueda y acceso a la información», «descarga de datos», «normalización» y «almacenamiento», que Seghiri (2011) atribuye al protocolo de compilación de corpus.

1.1.1. Diseño del corpus FÚTBOL_MXM

La compilación del corpus FÚTBOL_MXM responde a la naturaleza contrastiva del presente trabajo de investigación. Asimismo, como queda patente en el apartado «3.3. Metodología de compilación de corpus», la génesis y el diseño de este corpus sirven al objetivo principal del trabajo: «realizar un análisis de la traducción automática del inglés al español de las unidades terminológicas nominales univerbales pertenecientes al subcampo de especialidad del fútbol». Esta subordinación del diseño del corpus al objetivo principal se manifiesta, por tanto, en los componentes más relevantes de dicho diseño, las características principales y la clasificación del corpus, así como, a la postre, en los parámetros que conforman el análisis cualitativo y cuantitativo que sigue al proceso de compilación, como son la representatividad y el equilibrio del corpus.

1.1.1.1. Características principales

Las características del corpus FÚTBOL_MXM se reflejan en los criterios de compilación adoptados. Estas características se establecen específicamente para alcanzar los objetivos que se persiguen con nuestro trabajo.

En primer lugar, delimitamos el género textual, de manera que todos los textos del corpus pertenezcan al mismo; en este caso, a la versión en línea y en formato escrito de la crónica futbolística en directo.

En segundo lugar, fijamos una restricción de contenido en relación con los fragmentos de texto que conforman la crónica futbolística en directo. Así, en lugar de compilar los textos completos, incorporamos únicamente los fragmentos textuales que comprenden la narración en directo del partido, desde el minuto cero hasta el pitido final (minuto noventa más el tiempo añadido).

En tercer lugar, establecemos las lenguas de trabajo: el español de España y el inglés de Reino Unido. En cuarto lugar, nos aseguramos de que los textos sean originales en lo que a su lengua de publicación se refiere, por lo que quedan excluidos los textos traducidos. En línea con este criterio, consideramos oportuno establecer otra restricción que está relacionada con el género textual y con las lenguas de trabajo: las crónicas en español han de tratar partidos de LaLiga Santander, mientras que las crónicas en inglés deben tratar partidos de la Premier League. En quinto lugar, los textos tienen que extraerse de periódicos generalistas de tirada nacional.

Por último, consideramos conveniente establecer una restricción común, tanto en español como en inglés, en relación con la fecha de publicación de los textos. En este sentido, puesto que LaLiga Santander y la Premier League son ligas profesionales de fútbol cuyas temporadas discurren de forma paralela, la delimitación temporal de la publicación de los textos está determinada por el número de jornadas incluidas en la primera vuelta de la temporada 2022-2023 de cada liga. Por tanto, incluimos exclusivamente crónicas de los partidos disputados entre las jornadas 1 y 19 (ambas incluidas) de sendas ligas.

1.1.1.2. Clasificación

La aplicación de un sistema de clasificación de corpus adecuado constituye un medio para reunir, reflejar y concretar sus características definitorias de forma sistemática y organizada. En nuestro caso, vamos a adoptar, salvo contadas excepciones, la clasificación en dos fases propuesta por Faya Ornia (2015). La elección de esta clasificación obedece a varias razones: presenta una disposición en forma de árbol, por lo que todos los tipos de corpus se encuentran interrelacionados entre sí; se caracteriza por ser práctica, sistemática y completa, de manera que permite precisar de forma pormenorizada y ordenada las características del corpus; posee cierta flexibilidad, que se manifiesta en la posibilidad de añadir nuevas ramificaciones al diagrama de árbol de partida.

De acuerdo con esta taxonomía y en base a las características principales expuestas anteriormente, la definición del corpus FÚTBOL_MXM es la siguiente:

- escrito, puesto que está compuesto exclusivamente por textos escritos;
- de no referencia, al no ser representativo de la lengua general;
- sincrónico, ya que la compilación de textos está restringida a nivel temporal;
- monitorizado, dado que admite la adición de nuevos textos en caso de que fuera necesario;
- de muestras, pues está compuesto únicamente por fragmentos textuales y no por textos completos;
- de textos publicados, dado que los textos que lo conforman están publicados y se puede acceder a ellos de forma libre y gratuita;
- especializado, puesto que en los textos que lo conforman se hace uso de una lengua especializada;
- multilingüe, más concretamente bilingüe, debido a que está compuesto por textos en español y en inglés, y
- comparable, porque incluye textos originales en dos lenguas que se han compilado de acuerdo a criterios similares.

Una vez definido el corpus FÚTBOL_MXM, consideramos oportuno realizar una aclaración en relación con el concepto de corpus comparable, puesto que existe cierto debate acerca de su definición. Así, aunque los posicionamientos al respecto varían en función del autor, nos vamos a alinear con el enfoque de Corpas Pastor (2001), autora que sostiene que es un «corpus que, en relación a otro u otros corpus de lenguas distintas, incluyen tipos similares de textos originales» (Corpas Pastor, 2001: 158).

1.1.2. Búsqueda y acceso a la información

Tras completar el diseño del corpus, comenzamos con el desarrollo del protocolo de compilación a través de esta primera fase, que, en nuestro caso, está enfocada hacia la búsqueda de periódicos generalistas de tirada nacional en España y en Reino Unido que dispongan de crónicas futbolísticas en directo de libre acceso.

Finalmente, optamos por seleccionar *La Vanguardia*, en España, y *The Guardian*, en Reino Unido, ya que cumplen con los requisitos expuestos anteriormente. Sin embargo, estos dos periódicos difieren en el volumen de crónicas futbolísticas en directo publicadas entre las jornadas 1 y 19 de la temporada 2022-2023. *La Vanguardia* dispone de una base de datos amplia y completa donde se recogen las crónicas de todos los partidos disputados en la primera vuelta de LaLiga. Por el contrario, *The Guardian* brinda una cobertura informativa más reducida de los partidos de la Premier League por medio de este género textual, dado que solo se publican entre dos y seis crónicas en directo por jornada de liga. En vista de esta circunstancia, adoptamos el siguiente procedimiento para realizar el proceso selectivo:

- en el caso de *The Guardian*, seleccionamos todas las crónicas publicadas entre las jornadas 1 y 19 de la Premier League de la presente temporada 2022-2023, que suman un total de 80 crónicas;
- en lo que respecta a *La Vanguardia*, a fin de igualar las 80 crónicas en lengua inglesa, seleccionamos cinco crónicas para las jornadas 1, 3, 5 y 7 y cuatro crónicas para el resto de jornadas de la primera vuelta de LaLiga Santander de la presente temporada 2022-2023, de manera que siempre se incluyen las crónicas de los partidos del Real Madrid CF y del FC Barcelona, así como las primeras crónicas, por orden cronológico de publicación, de cada jornada.

1.1.3. Descarga de datos

Una vez realizada la selección de crónicas en ambas lenguas, identificamos los fragmentos de texto que nos interesan y los descargamos de forma manual. En realidad, esta descarga manual consiste en copiar el fragmento de texto correspondiente a la narración en directo del partido de cada crónica y pegarlo individualmente en un documento de texto sin formato (TXT). Así, generamos 80 archivos en lengua española y otros 80 en lengua inglesa, que, sumados, hacen un total de 160 TXT.

1.1.4. Normalización

Pese a que en este caso no es necesario recurrir a procesos o programas de conversión de archivos para unificarlos bajo una misma extensión, hay que asegurarse de que todos los TXT cuentan con la codificación interna UTF-8 para garantizar su compatibilidad con los programas empleados en la fase de explotación del corpus y de selección de la muestra de análisis y en el análisis práctico.

1.1.5. Almacenamiento

El almacenamiento del corpus consiste en la adopción de un sistema común de almacenamiento, codificación y denominación de los TXT que permita identificarlos y recuperar la información que contienen de forma rápida y sistemática. Esta fase se puede desglosar en cinco sencillos pasos.

Primero, creamos una carpeta principal con el mismo nombre que el corpus, FÚTBOL_MXM, que desempeña una función doble: por un lado, representa el eje central del almacenamiento del corpus; por otro lado, alberga todos los archivos que están directamente relacionados con el corpus.

A continuación, dentro de esta carpeta principal, generamos otras dos subcarpetas, ES y EN, que se corresponden con los subcorpus en español y en inglés, respectivamente, y cuyo propósito es almacenar los textos compilados.

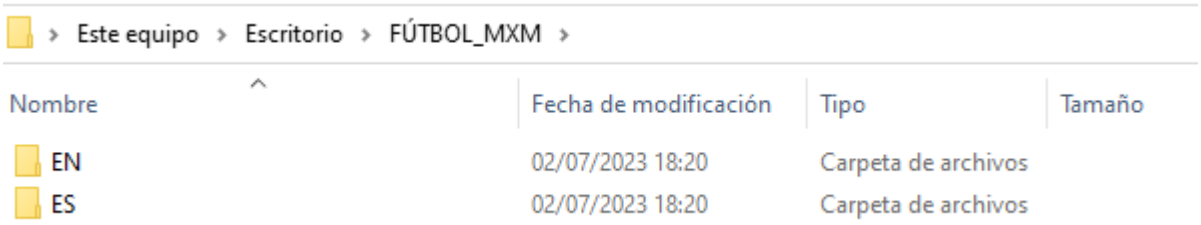


Ilustración 2. Carpeta principal del corpus FÚTBOL_MXM con los subcorpus EN y ES.

Después, diseñamos un sistema de codificación de los TXT. En esta ocasión, este sistema consta de un total de siete etiquetas que proporcionan diferentes tipos de información.

1. Nombre del corpus completo: FÚTBOL_MXM.
2. Lengua de publicación del fragmento de texto contenido en el TXT: ES para los fragmentos en español y EN para los fragmentos en inglés. Esta etiqueta desempeña un papel fundamental, puesto que indica la subcarpeta, EN o ES, en la que debe almacenarse el TXT en cuestión.
3. Periódico del que se ha extraído el fragmento de texto contenido en el TXT: LV, *La Vanguardia*, para los fragmentos en español y TG, *The Guardian*, para los fragmentos en inglés.
4. Liga a la que pertenece el partido que trata la crónica en directo de la que se ha extraído el fragmento de texto contenido en el TXT: LL, LaLiga Santander, para los fragmentos en español y PL, Premier League, para los fragmentos en inglés.
5. Temporada de liga y jornada a la que pertenece el partido que trata la crónica en directo de la que se ha extraído el fragmento de texto contenido en el TXT, en formato 22_23_n.º de jornada de liga; por ejemplo, *22_23_01*, en caso de un partido que se dispute en la primera jornada de la temporada 2022-2023.
6. Equipos que se enfrentan en el partido que trata la crónica en directo de la que se ha extraído el fragmento de texto contenido en el TXT, en formato equipo local_equipo visitante; por ejemplo, *BAR_RMA*, en caso de un partido disputado entre el FC Barcelona y el Real Madrid CF.

Una vez establecido el sistema de codificación, denominamos los 160 TXT adecuadamente. Así, por ejemplo, el archivo del subcorpus EN que contiene la narración en directo en inglés del partido de la jornada uno de la Premier League que enfrenta al Crystal Palace y al Arsenal se denomina FÚTBOL_MXM_EN_TG_PL_22_23_01_CPA_ARS, mientras que el archivo del subcorpus ES donde se almacena la narración en directo en español del clásico, el partido que disputaron el Real Madrid y el FC Barcelona en la jornada nueve de LaLiga, se denomina FÚTBOL_MXM_ES_LV_LL_22_23_09_RMA_BAR.

Por último, guardamos en la subcarpeta que corresponda, ES o EN, todos y cada uno de los 160 TXT tras cerciorarnos de que están debidamente denominados.

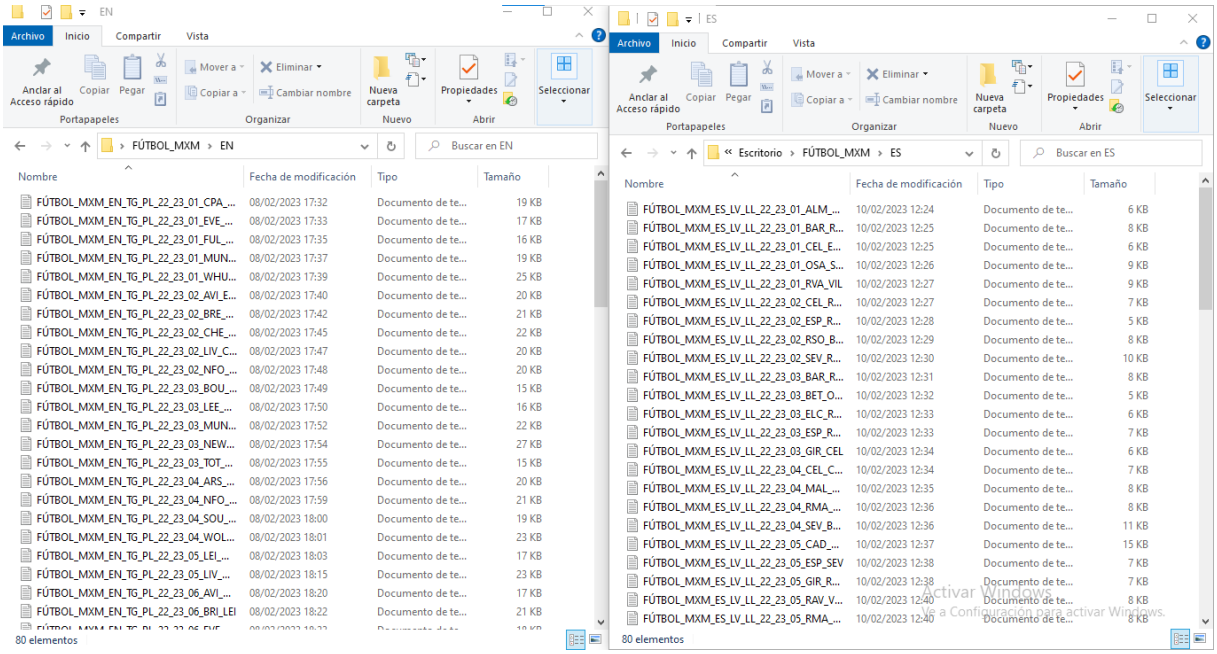


Ilustración 3. Carpetas de los subcorpus EN y ES.

1.2. Equilibrio y representatividad del corpus FÚTBOL_MXM

Tras completar el proceso de compilación, analizamos los parámetros de representatividad y equilibrio a fin de conocer si el corpus cumple con los estándares de calidad y se ajusta a los objetivos y a las necesidades específicas del trabajo al que sirve.

1.2.1. Equilibrio

Con el objetivo de determinar si nuestro corpus es equilibrado, estudiamos tres variables cuantitativas de los dos subcorpus que lo conforman: el número de textos, el número de palabras totales[8] y la media de palabras por texto. De este modo, extraemos de manera individual los datos relativos a estas variables de los dos subcorpus, EN y ES para, después, comparar los resultados de los subcorpus entre sí.

	EN	ES
N.º de textos	80	80
N.º de palabras totales	243 438	90 767
Media de palabras por texto	3042,98	1135,59

Tabla 3. Equilibrio del corpus FÚTBOL_MXM en base a variables cuantitativas.

[8] El valor de esta variable se obtiene a partir de los datos facilitados por Sketch Engine.

Como se puede apreciar en la Tabla 3, la brecha de volumen textual entre los dos subcorpus, EN y ES, es evidente, pues así se manifiesta en los resultados obtenidos en relación con el número de palabras totales y la media de palabras por texto. No obstante, este defecto se compensa con la variable relativa al número de textos, cuyo valor es igual para los dos subcorpus. Por tanto, se puede considerar que el corpus FÚTBOL_MXM es equilibrado.

1.2.2. Representatividad

Para verificar que nuestro corpus es representativo, empleamos el programa informático ReCor[9], ideado por Seghiri (2006). La función de ReCor se puede explicar de la siguiente manera:

> Con este método pretendemos plantear una solución eficaz para determinar, por primera vez, *a posteriori* el tamaño mínimo de un corpus o colección textual, independientemente de la lengua o tipo textual de dicha colección, estableciendo, por tanto, el umbral mínimo de representatividad a partir de un algoritmo (N-Cor) de análisis de la densidad léxica en función del aumento incremental del corpus (Corpas Pastor y Seghiri, 2007: 166).

Este programa permite conocer si un corpus compilado previamente es o no representativo mediante un análisis cuantitativo que aborda el número de documentos del corpus y el número de *tokens*[10], o casos, del corpus, dos variables directamente relacionadas con el tamaño del corpus. Puesto que el método de análisis que emplea ReCor es complejo, nos limitamos a exponer la información esencial que permite determinar si un corpus es representativo.

Tras introducir el corpus en el programa y una vez realizado el análisis, el programa genera dos gráficas (Seghiri, 2006: 389), el Estudio Gráfico A y el Estudio Gráfico B, de manera que en el eje de abscisas de la primera se representa el número de documentos del corpus y en el de la segunda, el número de *tokens*, las dos variables cuantitativas relacionadas con el tamaño del corpus que el programa utiliza para establecer el umbral de representatividad del corpus. En ambas gráficas, se representan dos funciones, una en rojo y otra en azul, que presentan un descenso exponencial: se considera que el punto aproximado del eje de abscisas en el que estas funciones se estabilizan constituye el umbral de representatividad aproximado para cada variable.

Tras esta breve contextualización teórica acerca del funcionamiento de ReCor, utilizamos este programa para determinar si el corpus Fútbol_MXM es representativo. Por consiguiente, llevamos a cabo el análisis de representatividad de los subcorpus ES y EN.

[9] https://www.lexytrad.es/es/recursos/recor-2/.
[10] Seghiri (2006: 376) indica que los *tokens* hacen referencia al número total de palabras

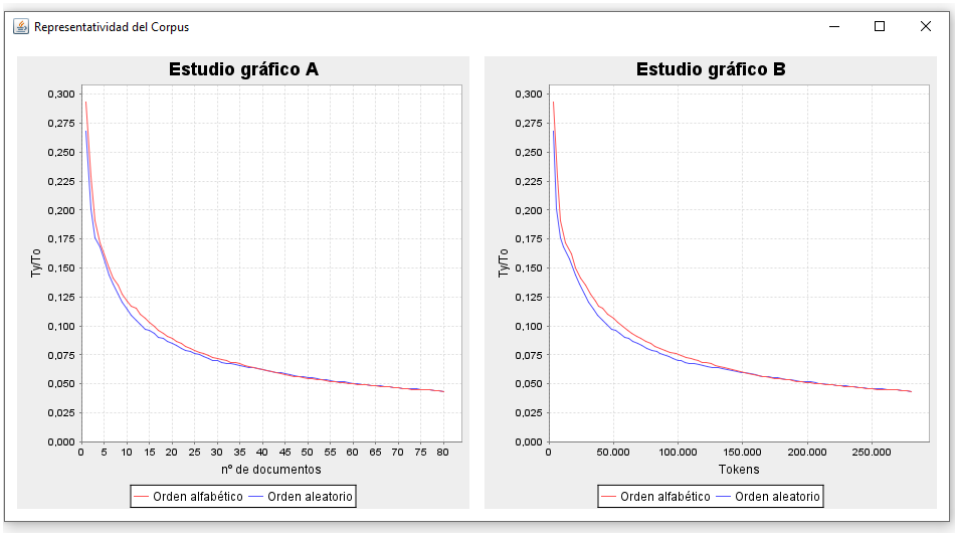


Ilustración 4. Umbrales de representatividad del subcorpus EN calculados por medio de ReCor.

De estas dos gráficas se pueden extraer los umbrales de representatividad relativos al número de documentos (Estudio gráfico A) y al número de *tokens* (Estudio gráfico B) para el subcorpus EN, así como determinar la representatividad de este subcorpus según estas dos variables.

	Umbral de representatividad calculado por ReCor	En el subcorpus EN
N.º de documentos	55	80
N.º de *tokens*	175 000	243 438

Tabla 4. Representatividad del subcorpus EN.

De acuerdo con estos resultados, el subcorpus EN es representativo atendiendo a estas dos variables cuantitativas. Continuamos ahora con el análisis de la representatividad del subcorpus ES.

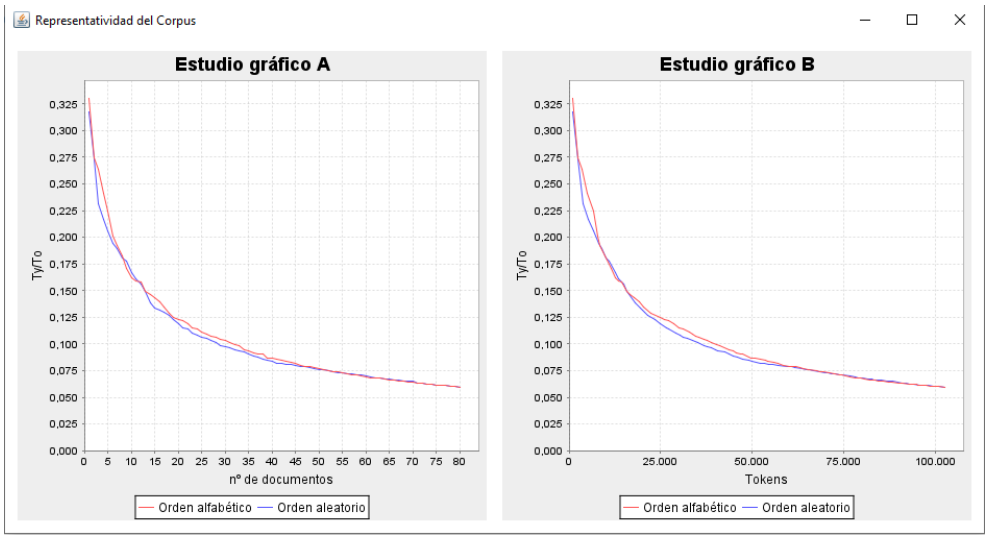


Ilustración 5. Umbrales de representatividad del subcorpus ES calculados por medio de ReCor.

De forma paralela al procedimiento seguido con el subcorpus EN, extraemos los umbrales de representatividad para el subcorpus ES mediante la interpretación de los resultados del análisis.

	Umbral de representatividad calculado por ReCor	En el subcorpus ES
N.º de documentos	55	80
N.º de *tokens*	60 000	90 767

Tabla 5. Representatividad del subcorpus ES.

Según los umbrales de representatividad establecidos por ReCor, que responden a estas dos variables cuantitativas, el subcorpus ES también es representativo. En consecuencia, se puede confirmar la representatividad del corpus FÚTBOL_MXM, puesto que los dos subcorpus que lo conforman son representativos.

1.3. Explotación del corpus FÚTBOL_MXM y selección de la muestra de análisis

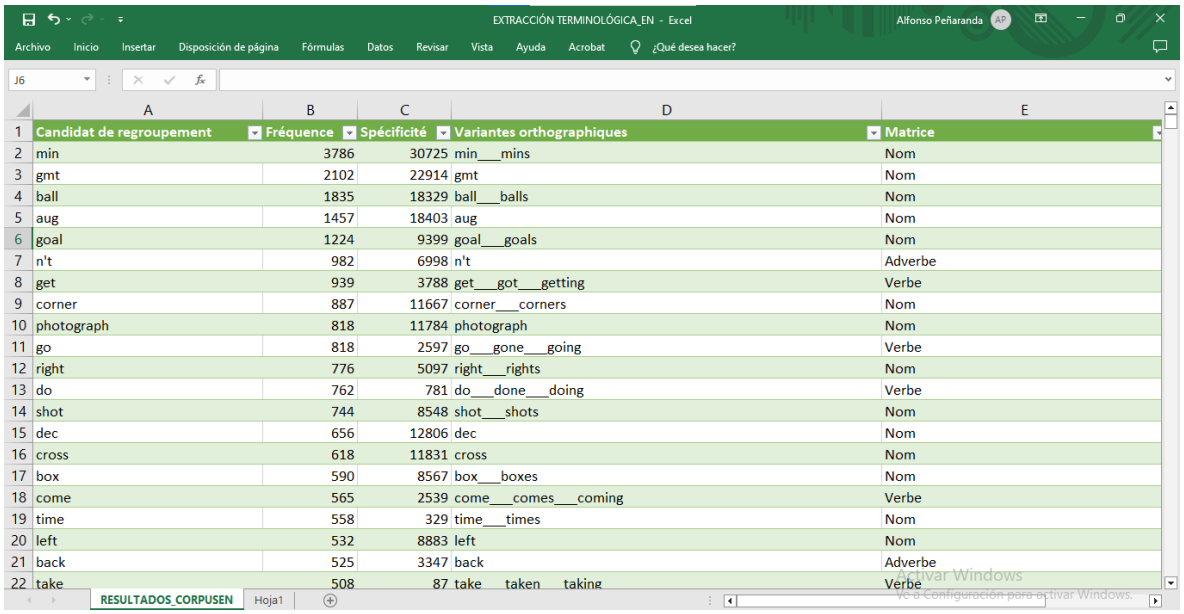
Tras definir el corpus FÚTBOL_MXM, explicar su proceso de compilación y corroborar su equilibrio y representatividad, nos enfocamos en su explotación.

Abordamos la explotación del corpus FÚTBOL_MXM desde una perspectiva terminológica. La terminología, como indica Cabré (1993, en Santamaría Pérez, 2015: 79), es «el campo de conocimiento que se encarga del estudio, la descripción y la recopilación de las unidades terminológicas (UT) utilizadas en las distintas áreas de especialidad».

La definición del concepto de unidad terminológica, o término, todavía es objeto de debate, pues varía en función del enfoque empleado y del autor. No obstante, tomamos como referencia el trabajo de Santamaría Pérez (2009), quien considera que las unidades terminológicas son unidades léxicas del lenguaje natural que adquieren un significado especializado al designar conceptos propios de un campo de especialidad concreto.

Partiendo de esta base teórica, nuestro objeto de estudio son las unidades terminológicas, o términos, del campo de especialidad del deporte, más concretamente del subcampo de especialidad del fútbol.

Primero, llevamos a cabo la extracción terminológica del corpus FÚTBOL_MXM. Para ello, nos servimos de TermoStat Web, «a software tool dedicated to term extraction. The system identifies not only complex terms, but also simple terms, which often tend to be ignored by automated systems» (Drouin, 2003: 99-100). Mediante esta herramienta semiautomática, extraemos de manera independiente las unidades terminológicas contenidas en los subcorpus EN y ES. De este modo, una vez completado el proceso de extracción terminológica correspondiente a cada subcorpus, el programa genera una lista con las unidades terminológicas extraídas. Descargamos esta lista en un archivo TXT que, a continuación, convertimos en un documento Excel.



	A	B	C	D	E
1	Candidat de regroupement	Fréquence	Spécificité	Variantes orthographiques	Matrice
2	min	3786	30725	min___mins	Nom
3	gmt	2102	22914	gmt	Nom
4	ball	1835	18329	ball___balls	Nom
5	aug	1457	18403	aug	Nom
6	goal	1224	9399	goal___goals	Nom
7	n't	982	6998	n't	Adverbe
8	get	939	3788	get___got___getting	Verbe
9	corner	887	11667	corner___corners	Nom
10	photograph	818	11784	photograph	Nom
11	go	818	2597	go___gone___going	Verbe
12	right	776	5097	right___rights	Nom
13	do	762	781	do___done___doing	Verbe
14	shot	744	8548	shot___shots	Nom
15	dec	656	12806	dec	Nom
16	cross	618	11831	cross	Nom
17	box	590	8567	box___boxes	Nom
18	come	565	2539	come___comes___coming	Verbe
19	time	558	329	time___times	Nom
20	left	532	8883	left	Nom
21	back	525	3347	back	Adverbe
22	take	508	87	take taken taking	Verbe

Ilustración 6. Extracción terminológica del subcorpus EN proporcionada por TermoStat Web en formato Excel.

Además de llevar a cabo el barrido terminológico, que se refleja en la columna A, *Candidat de regroupement*, TermoStat Web proporciona información adicional de gran utilidad sobre las unidades terminológicas identificadas:

- las ordena por frecuencia de aparición, como se aprecia en la columna B, *Fréquence*;
- les asigna una puntuación de especificidad, como se muestra en la columna C, *Spécificité*;
- incluye las posibles variantes de cada unidad terminológica, como se observa en la columna D, *Variantes ortographiques*, y
- recoge la categoría gramatical de cada unidad terminológica, como se refleja en la columna E, *Matrice*.

A continuación, en línea con el diseño y los objetivos de nuestra investigación, establecemos como objeto de estudio las unidades terminológicas nominales univerbales[11] del subcampo de especialidad del fútbol. Asimismo, los sustantivos seleccionados han de ser comunes, concretos y contables.

Desde el punto de vista semántico, esta delimitación del objeto de estudio permite analizar unidades terminológicas que designan conceptos denotativos y unívocos, de manera que se reduce considerablemente el riesgo de contaminación de los resultados por cuestiones de pragmática o de ambigüedad semántica.

Después, establecemos un volumen muestral de tres unidades terminológicas en lengua inglesa procedentes de la extracción terminológica del subcorpus EN. A fin de garantizar la

[11] Con unidades terminológicas nominales univerbales nos referimos a unidades terminológicas conformadas por una única palabra, que es un sustantivo.

representatividad cualitativa y cuantitativa de la muestra, las unidades terminológicas seleccionadas deben cumplir con los criterios de selección relativos a la delimitación del objeto de estudio, así como contar con un mínimo de 150 ocurrencias[12] en el subcorpus EN.

Por último, llevamos a cabo la selección manual de la muestra de análisis a partir de la extracción terminológica del subcorpus EN proporcionada por TermoStat Web. Al término de este proceso, la muestra de análisis está conformada por las tres unidades terminológicas nominales univerbales en lengua inglesa presentadas a continuación:

Unidad terminológica	N.º de ocurrencias en el subcorpus EN	Categoría gramatical
ball	1835	sustantivo
shot	744	sustantivo
referee	166	sustantivo

Tabla 6. Muestra de análisis seleccionada a partir de la extracción terminológica del subcorpus EN.

1.4. Metodología del análisis práctico

Una vez seleccionada la muestra de análisis, planteamos la metodología sobre la que se cimenta el análisis práctico para, después, poder presentar los resultados obtenidos de su aplicación.

En primer lugar, realizamos un segundo proceso de extracción terminológica del subcorpus ES por medio de la opción *Wordlist*[13] de Sketch Engine, una herramienta en línea que brinda la posibilidad de realizar muchos procesos relacionados con los corpus (compilación, gestión y explotación de corpus) y con la terminología (extracción y gestión terminológica). Este proceso adicional responde al objetivo de contrastar y completar la extracción terminológica proporcionada por TermoStat Web.

[12] En este caso, la frecuencia de aparición es equivalente al número de ocurrencias, puesto que ambos conceptos hacen referencia al número de veces que una unidad terminológica, incluidas sus variantes, está presente en el corpus.

[13] Dirigimos y agilizamos la extracción terminológica mediante la aplicación de un criterio de búsqueda adicional, el filtro *noun*, que permite mostrar solo aquellas unidades terminológicas cuyo núcleo sea un sustantivo.

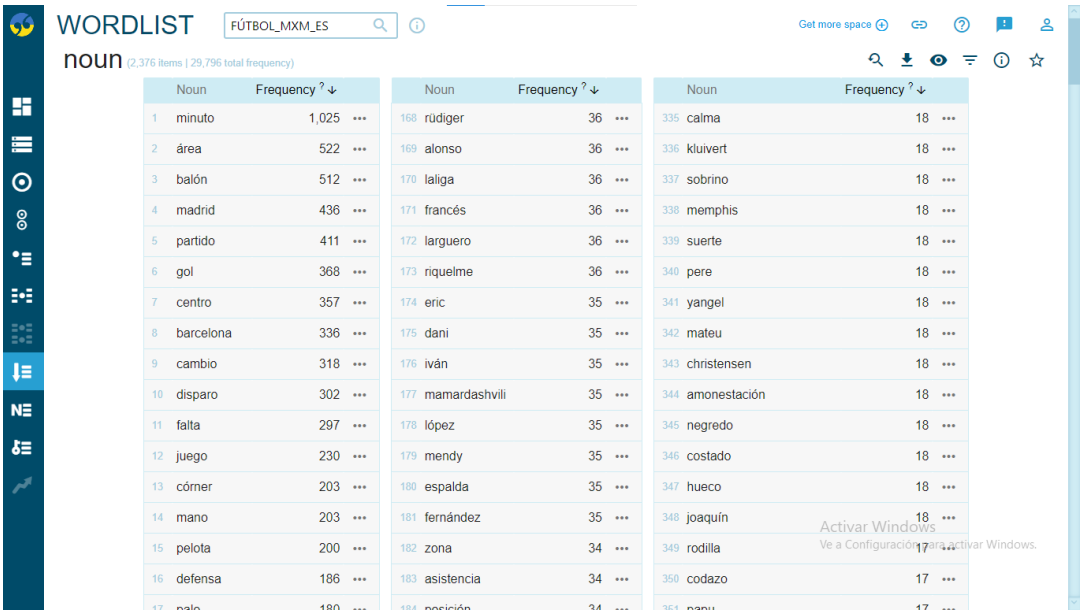


Ilustración 7. Extracción terminológica del subcorpus ES proporcionada por Sketch Engine.

En segundo lugar, a partir de la lista de unidades terminológicas facilitada por Sketch Engine, preseleccionamos de forma manual las unidades terminológicas en español susceptibles de ser equivalentes de traducción de las que conforman la muestra de análisis; de aquí en adelante, emplearemos la denominación «equivalentes de traducción potenciales» para referiremos a estas unidades terminológicas en español. Posteriormente, sometemos estos equivalentes a un proceso de filtrado, también manual, mediante la opción Concordance que brinda esta herramienta. Con este proceso de filtrado de los equivalentes de traducción potenciales preseleccionados perseguimos tres objetivos.

- Excluir de la selección definitiva aquellos equivalentes de traducción potenciales que dan lugar a ambigüedades semánticas de manera sistemática. Por ejemplo, las unidades terminológicas «envío» y «lanzamiento» pueden considerarse equivalentes de traducción potenciales de «*shot*» en determinados contextos; sin embargo, en muchas ocasiones resulta muy complicado discernir si estas unidades terminológicas se emplean como sinónimos de «tiro», de «centro» o de «pase».
- Excluir de la selección definitiva aquellas unidades terminológicas que podrían considerarse equivalentes de traducción potenciales de alguna de las unidades de la muestra de análisis porque designan el mismo concepto, pero, a su vez, presentan precisiones semánticas que las convierten en inadecuadas para formar parte de la selección definitiva. Por ejemplo, las unidades terminológicas «zapatazo», «misil», «latigazo» y «chutazo» pueden tratarse como equivalentes de traducción potenciales de «*shot*», mas hacen referencia a disparos con mucha potencia, no al concepto «disparo» en general.
- Descartar de la selección definitiva aquellas ocurrencias de los equivalentes de traducción potenciales preseleccionados que, por su naturaleza polisémica, designan un

concepto no especializado u otro concepto especializado diferente.

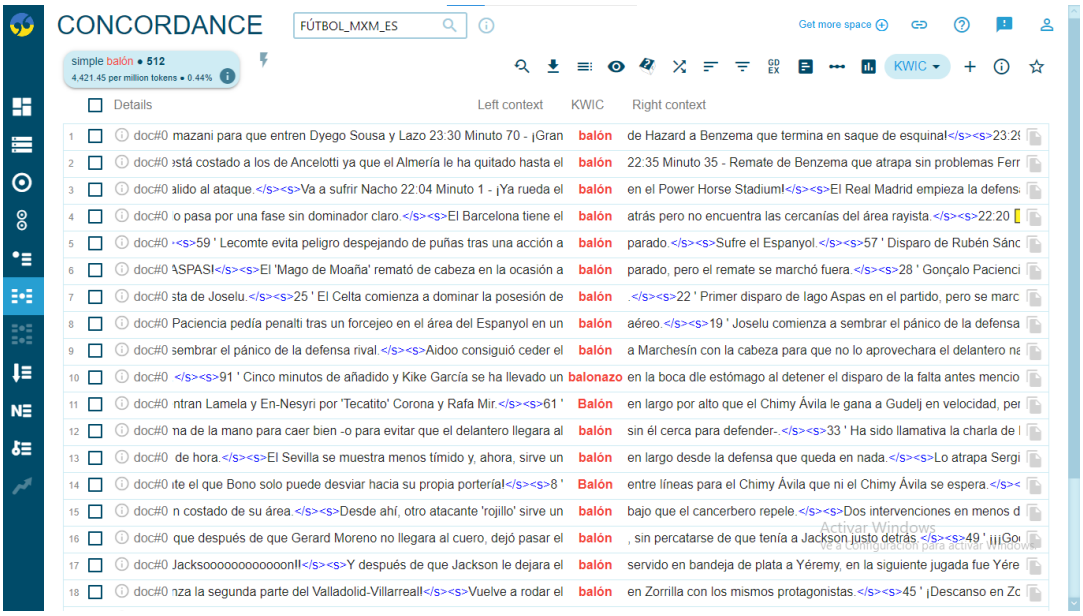


Ilustración 8. Proceso de filtrado de «balón», equivalente de traducción potencial en español de la unidad terminológica «ball».

Tras el proceso de filtrado, cerramos la selección definitiva de los equivalentes de traducción potenciales en español de las unidades terminológicas «*ball*», «*shot*» y «*referee*». Después, recogemos y anotamos el número de ocurrencias correspondientes a cada uno de estos equivalentes.

Unidad terminológica en inglés	Selección definitiva de equivalentes de traducción potenciales en español
ball	balón
	pelota
	bola
	cuero
	esférico
shot	disparo
	remate
	tiro
	chut
	golpeo
referee	colegiado
	árbitro

Tabla 7. Equivalentes de traducción potenciales en español de las unidades terminológicas en inglés que conforman la muestra de análisis.

En tercer lugar, realizamos la traducción automática al español de los 80 textos originales en inglés que conforman el subcorpus EN. Para ello, recurrimos a DeepL Pro, un motor de traducción automática que opera de acuerdo con un sistema basado en redes neuronales.

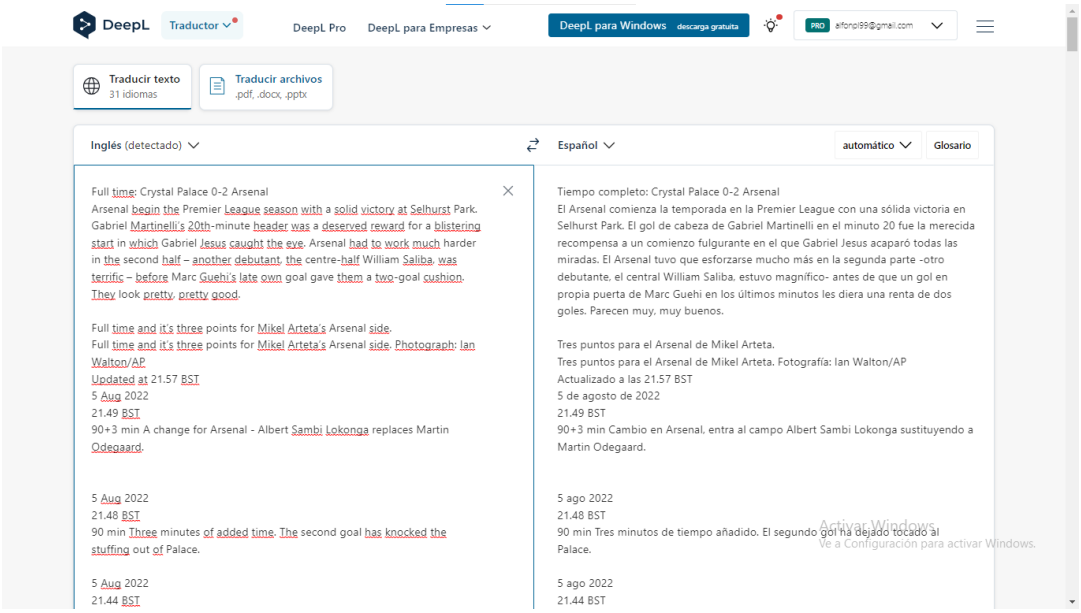


Ilustración 9. Traducción automática del inglés al español.

De forma paralela al proceso de traducción automática, creamos un corpus paralelo, FÚTBOL_MXM_PARALELO, que compilamos en torno a una carpeta principal que recibe el mismo nombre. Dentro de esta carpeta principal, generamos dos subcarpetas, EN_SOURCE y ES_TARGET, de modo que la subcarpeta EN_SOURCE se corresponde con el subcorpus en inglés, y la subcarpeta ES_TARGET constituye el subcorpus en español. Sin embargo, puesto que FÚTBOL_MXM_PARALELO es un corpus paralelo, los subcorpus que lo constituyen contienen textos de naturaleza opuesta en lo que se refiere a su originalidad:

- el subcorpus EN_SOURCE es una copia exacta del subcorpus EN, pues consta de los mismos 80 textos originales en inglés;
- el subcorpus ES_TARGET, en cambio, está formado por las traducciones al español proporcionadas por DeepL Pro, por lo que contiene 80 textos meta.

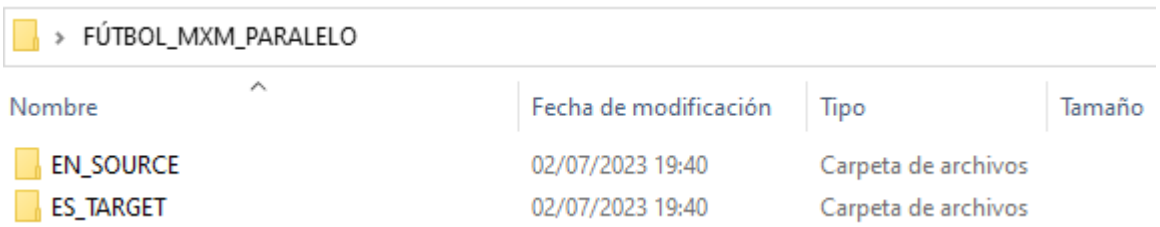


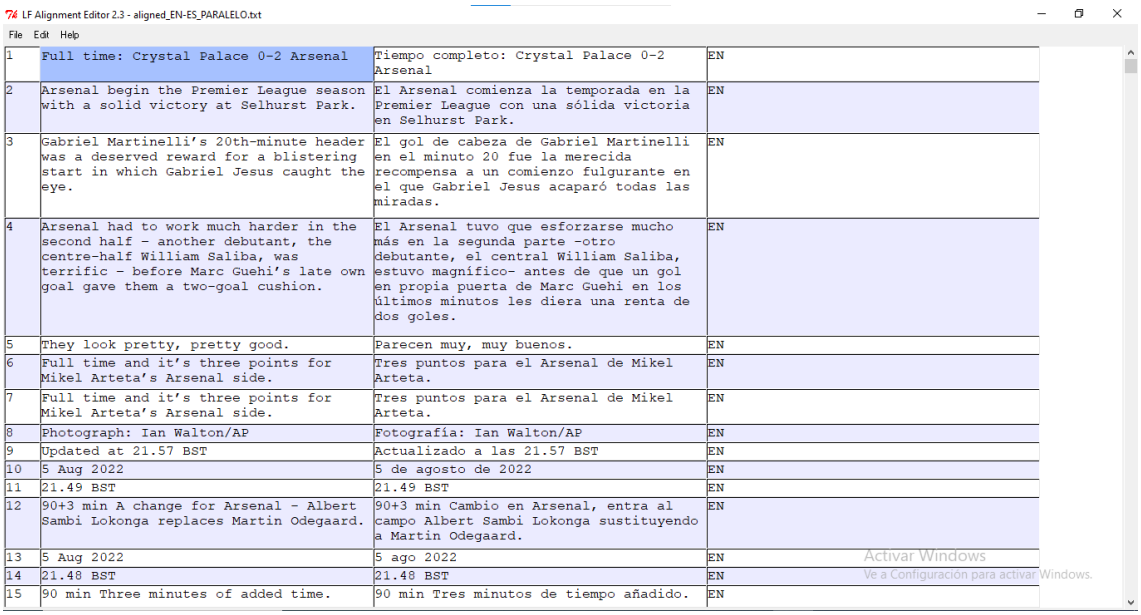
Ilustración 10. Carpeta principal del corpus FÚTBOL_MXM_PARALELO con los subcorpus EN_SOURCE y ES_TARGET.

En lo que se refiere al sistema de codificación de los TXT, aplicamos el mismo que con el corpus FÚTBOL_MXM, a excepción de dos etiquetas:

- la primera, que hace alusión al nombre completo del corpus, por lo que se reemplaza FÚTBOL_MXM por FÚTBOL_MXM_PARALELO;
- la segunda, que, en este caso, además de hacer referencia a la lengua de publicación del texto (EN para la lengua inglesa y ES para la lengua española), también especifica si el texto es original o una traducción. Por consiguiente, los 80 TXT del subcorpus EN_SOURCE presentan la etiqueta ENSL («*English, source language*») en su denominación, mientras que los 80 TXT que conforman el subcorpus ES_TARGET portan la etiqueta ESTL («español, *target language*»).

Por ejemplo, dado el texto origen en inglés FÚTBOL_MXM_PARALELO_ENSL_TG_PL_22_23_01_CPA_ARS, el texto meta en español se denomina FÚTBOL_MXM_PARALELO_ESTL_TG_PL_22_23_01_CPA_ARS.

En cuarto lugar, alineamos de manera automática el subcorpus EN_SOURCE con el subcorpus ES_TARGET por medio del programa LF Aligner, que nos proporciona un archivo TMX con la alineación bilingüe. Este procedimiento consiste en alinear uno a uno los segmentos en la lengua origen, el inglés, con los segmentos en la lengua meta, el español.


LF Alignment Editor 2.3 - aligned_EN-ES_PARALELO.txt

File Edit Help

1	Full time: Crystal Palace 0-2 Arsenal	Tiempo completo: Crystal Palace 0-2 Arsenal	EN
2	Arsenal begin the Premier League season with a solid victory at Selhurst Park.	El Arsenal comienza la temporada en la Premier League con una sólida victoria en Selhurst Park.	EN
3	Gabriel Martinelli's 20th-minute header was a deserved reward for a blistering start in which Gabriel Jesus caught the eye.	El gol de cabeza de Gabriel Martinelli en el minuto 20 fue la merecida recompensa a un comienzo fulgurante en el que Gabriel Jesus acaparó todas las miradas.	EN
4	Arsenal had to work much harder in the second half – another debutant, the centre-half William Saliba, was terrific – before Marc Guehi's late own goal gave them a two-goal cushion.	El Arsenal tuvo que esforzarse mucho más en la segunda parte –otro debutante, el central William Saliba, estuvo magnífico– antes de que un gol en propia puerta de Marc Guehi en los últimos minutos les diera una renta de dos goles.	EN
5	They look pretty, pretty good.	Parecen muy, muy buenos.	EN
6	Full time and it's three points for Mikel Arteta's Arsenal side.	Tres puntos para el Arsenal de Mikel Arteta.	EN
7	Full time and it's three points for Mikel Arteta's Arsenal side.	Tres puntos para el Arsenal de Mikel Arteta.	EN
8	Photograph: Ian Walton/AP	Fotografía: Ian Walton/AP	EN
9	Updated at 21.57 BST	Actualizado a las 21.57 BST	EN
10	5 Aug 2022	5 de agosto de 2022	EN
11	21.49 BST	21.49 BST	EN
12	90+3 min A change for Arsenal - Albert Sambi Lokonga replaces Martin Odegaard.	90+3 min Cambio en Arsenal, entra al campo Albert Sambi Lokonga sustituyendo a Martin Odegaard.	EN
13	5 Aug 2022	5 ago 2022	EN
14	21.48 BST	21.48 BST	EN
15	90 min Three minutes of added time.	90 min Tres minutos de tiempo añadido.	EN

Ilustración 11. Revisión de la alineación bilingüe (EN-ES) proporcionada por LF Aligner.

En este punto del análisis práctico, consideramos conveniente realizar una recapitulación de los corpus y los subcorpus que intervienen en el análisis práctico. Por un lado, distinguimos el subcorpus EN y el subcorpus ES, que conforman el corpus comparable bilingüe FÚTBOL_MXM, de manera que el subcorpus EN contiene 80 textos originales en inglés y el subcorpus ES está compuesto por otros 80 textos originales en español. Por otro lado, encontramos el subcorpus EN_SOURCE y el subcorpus ES_TARGET, que constituyen el corpus paralelo bilingüe FÚTBOL_MXM_PARALELO; en este caso, el subcorpus EN_SOURCE cuenta con los mismos 80

textos originales en inglés que el subcorpus EN, mientras que el subcorpus ES_TARGET consta de las traducciones al español realizadas por DeepL Pro de dichos textos, por lo que está compuesto por 80 textos meta en español.

En quinto lugar, analizamos el contenido del subcorpus ES_TARGET desde un punto de vista terminológico. Realizamos una búsqueda dirigida de los equivalentes de traducción potenciales que conforman la selección definitiva a través de la opción Concordance de Sketch Engine con el propósito de anotar el número de ocurrencias correspondiente a cada equivalente tras llevar a cabo un proceso de filtrado manual.

Ilustración 12. Anotación manual de las ocurrencias de «disparo», equivalente de traducción potencial en español de la unidad terminológica «shot»*, mediante la opción Concordance de Sketch Engine.*

Finalmente, revisamos el resultado de la traducción automática ofrecida por DeepL Pro. Para ello, recuperamos el archivo TMX con la alineación bilingüe de los subcorpus EN_SOURCE y ES_TARGET generado por el programa LF Aligner y lo importamos en Sketch Engine. A continuación, mediante la opción *Parallel Concordance*, analizamos de manera manual las traducciones que DeepL Pro propone para las cuatro unidades terminológicas en inglés que constituyen la muestra de análisis a fin de conocer cuáles de los equivalentes de traducción potenciales en español emplea este motor y determinar cuántas veces recurre a cada uno de ellos. Sin embargo, resulta necesario aclarar que esta última fase del análisis práctico no se corresponde con la fase anterior, dado que, si bien en ambos casos trabajamos con el subcorpus ES_TARGET, en esta ocasión abordamos su análisis desde su alineación bilingüe con el subcorpus EN_SOURCE, mientras que en la fase anterior lo estudiamos de manera individual e independiente, sin tener en cuenta dicha alineación.

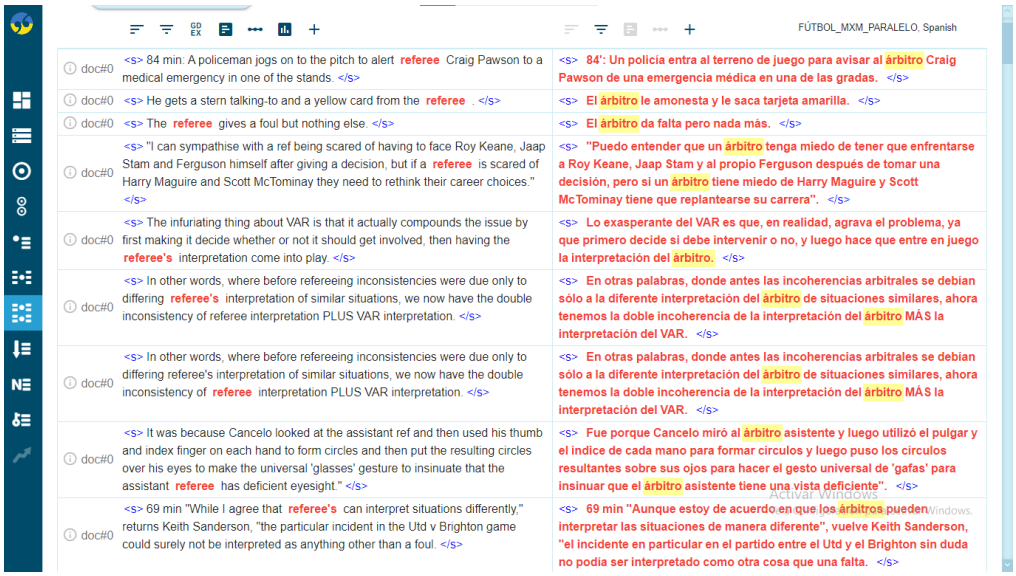


Ilustración 13. Análisis mediante la opción Parallel Concordance de Sketch Engine de los equivalentes de traducción en español proporcionados por DeepL Pro para la unidad terminológica «referee».

En línea con los objetivos de nuestro trabajo de investigación y con el par de lenguas seleccionado, inglés y español, realizamos un análisis contrastivo entre la muestra textual representativa del lenguaje de especialidad del fútbol, el subcorpus ES, y los textos meta generados por DeepL Pro, el subcorpus ES_TARGET. Este análisis se cimenta sobre tres parámetros relacionados con los equivalentes de traducción potenciales: la representación, la distribución y el grado de representación.

- Un equivalente alcanza la representación cuando cuenta con, al menos, una ocurrencia. Se trata de un parámetro mixto, ya que aúna rasgos cuantitativos y cualitativos.
- La distribución de un equivalente hace referencia a la relación porcentual de sus ocurrencias con respecto a las del resto. Este parámetro se calcula dividiendo las ocurrencias del equivalente en cuestión entre el sumatorio de las ocurrencias de todos los equivalentes, por lo que se trata de un parámetro cuantitativo.
- El grado de representación de un equivalente se obtiene mediante la aplicación de una escala que refleja su distribución desde un enfoque cualitativo.

Intervalo de distribución	Grado de representación del equivalente
0 %	NULO
(0 - 5 %]	MUY BAJO
(5 % - 35 %]	BAJO
(35 % - 65 %]	MEDIO
(65 % - 95 %]	ALTO
(95 % - 100 %)	MUY ALTO
100 %	MÁXIMO

Tabla 8. Escala cualitativa de grado de representación en función del valor de distribución.

Planificamos el análisis contrastivo de acuerdo con el desarrollo del análisis práctico y con los datos recabados en relación con estos tres parámetros. Así pues, estructuramos este análisis en tres secciones, una segmentación que responde al objetivo de estudiar las unidades terminológicas que conforman la muestra de análisis por separado y de manera independiente. Asimismo, cada una de estas secciones está dividida en cuatro partes dedicadas a la interpretación, mediante la aplicación de los parámetros de representación, distribución y grado de representación, del número de ocurrencias de los equivalentes de traducción potenciales extraídas en el análisis práctico.

- En la primera parte, se analiza la representación, la distribución y el grado de representación de los equivalentes de traducción potenciales en el subcorpus ES.
- En la segunda parte, se aborda la representación, la distribución y el grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET;
- En la tercera parte, también se trata la representación, la distribución y el grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET, pero tomando como referencia su alineación bilingüe con el subcorpus EN_SOURCE. Por tanto, de ahora en adelante nos referiremos a este objeto de estudio, el subcorpus ES_TARGET cuando se aborda desde su alineación con el subcorpus EN_SOURCE, como subcorpus ES_TARGET alineado.
- En la cuarta parte, se comparan los valores relativos a la representación, la distribución y el grado de representación de los equivalentes de traducción potenciales obtenidos en el subcorpus ES con los extraídos en el subcorpus ES_TARGET alineado.

RESULTADOS

En este capítulo, exponemos e interpretamos los resultados obtenidos a partir de la explotación del análisis práctico. Para la presentación de los resultados, adoptamos como modelo la estructura del análisis contrastivo, que consta de tres secciones, «*ball*», «*shot*» y «*referee*», una por cada unidad terminológica en inglés. Asimismo, estas secciones se estructuran en cuatro partes según el subcorpus que se analice: «Subcorpus ES», «Subcorpus ES_TARGET», «Subcorpus ES_TARGET alineado» y «Comparación entre el subcorpus ES y el subcorpus ES_TARGET alineado». Por último, incluimos, a modo de resumen, una recapitulación del análisis de cada subcorpus.

1. *Ball*

Según el barrido terminológico realizado por TermoStat Web, la unidad terminológica «*ball*» cuenta con 1835 ocurrencias en el subcorpus EN, por lo que encabeza la muestra de análisis en número total de ocurrencias.

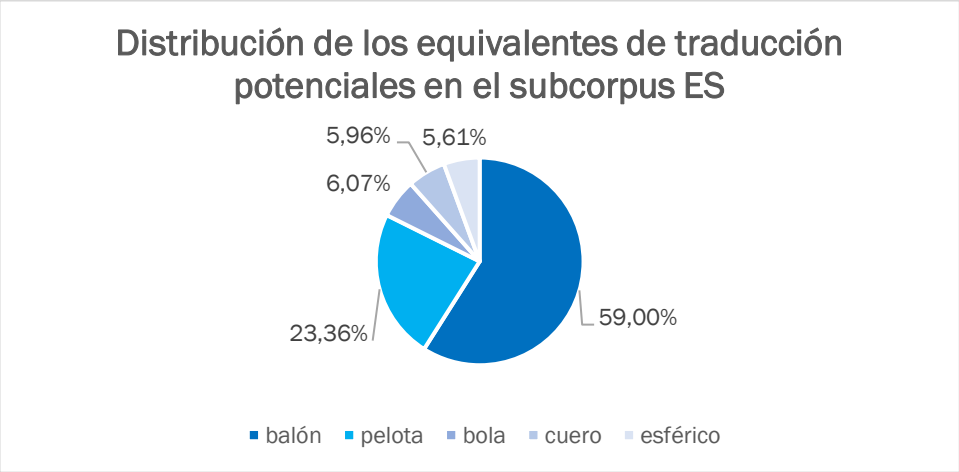
1.1. Subcorpus ES

A través de un análisis manual y pormenorizado de la lista de unidades terminológicas en español extraídas del subcorpus ES por medio de la función *Wordlist* de Sketch Engine, determinamos que esta unidad terminológica en inglés, «*ball*», cuenta con cinco equivalentes de traducción potenciales en español: «balón», «pelota», «bola», «cuero» y «esférico».

Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico	
N.º de ocurrencias	505	200	52	51	48	Σ = 856

Tabla 9. Equivalentes de traducción potenciales en español de la unidad terminológica «ball» *con sus ocurrencias en el subcorpus ES.*

Como se muestra en la Tabla 9, todos los equivalentes de traducción potenciales, a excepción de «balón» y de «pelota», cuentan con una distribución relativamente equilibrada en el subcorpus ES.



Gráfica 1. Distribución de los equivalentes de traducción potenciales en el subcorpus ES.

En la Gráfica 1, se pone de manifiesto que en la crónica futbolística en directo en lengua española coexisten varias opciones terminológicas para designar el concepto «*ball*», aunque resulta evidente que el equivalente predominante es «balón», ya que prácticamente dobla en ocurrencias al resto de equivalentes combinados.

Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico
Grado de representación del equivalente	medio	bajo	bajo	bajo	bajo

Tabla 10. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES.

La amplia brecha cuantitativa entre equivalentes contrasta, no obstante, con la reducida distancia de representación cualitativa, pues el salto máximo de representación entre equivalentes es de un grado según la escala cualitativa de grado de representación.

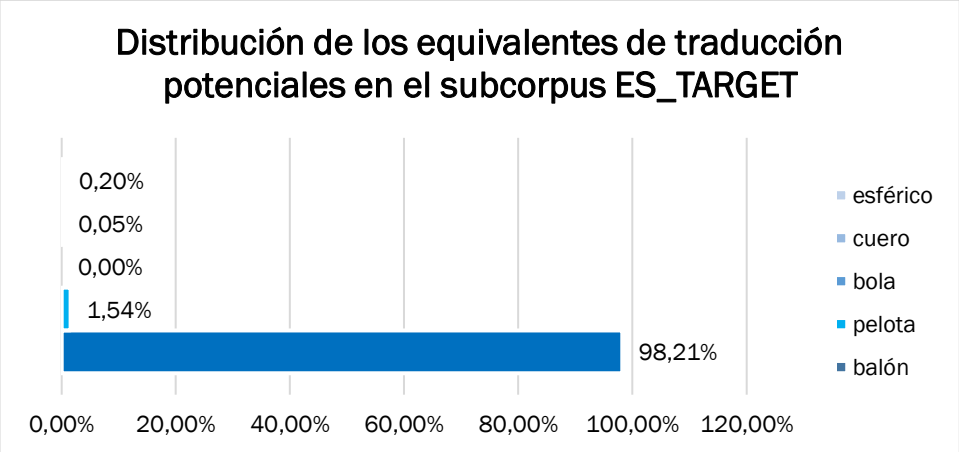
1.2. Subcorpus ES_TARGET

En el subcorpus ES_TARGET, conformado por textos meta traducidos por DeepL Pro, el panorama cambia por completo.

Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico	
N.º de ocurrencias	1975	31	0	1	4	Σ = 2011

Tabla 11. Equivalentes de traducción potenciales en español de la unidad terminológica «ball» *con sus ocurrencias en el subcorpus ES_TARGET.*

En primer lugar, uno de los equivalentes de traducción potenciales, «bola», que ocupa la tercera posición en número de ocurrencias en el subcorpus ES, no tiene representación.



Gráfica 2. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

En segundo lugar, la distribución de los cuatro equivalentes representados está totalmente descompensada: por un lado, con un valor de distribución del 98,21 %, el equivalente «balón» aúna casi todas las ocurrencias; por otro lado, los equivalentes «cuero» y «esférico» representan, aun si se combinan, un porcentaje muy inferior al 1 % de las ocurrencias.

Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico
Grado de representación del equivalente	muy alto	muy bajo	nulo	muy bajo	muy bajo

Tabla 12. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

Finalmente, la polarización observada en relación con la distribución de los equivalentes, que es el resultado de la aplicación de un parámetro cuantitativo, también se refleja al adoptar un enfoque cualitativo, como es el grado de representación. Así, como se demuestra en la Tabla 12, el equivalente «balón» presenta un grado de representación muy alto, que se corresponde con el segundo intervalo de distribución; en cambio, el resto de equivalentes representados, «pelota», «cuero» y «esférico» disponen de un grado de representación muy bajo, de manera que se sitúan en el penúltimo intervalo de distribución.

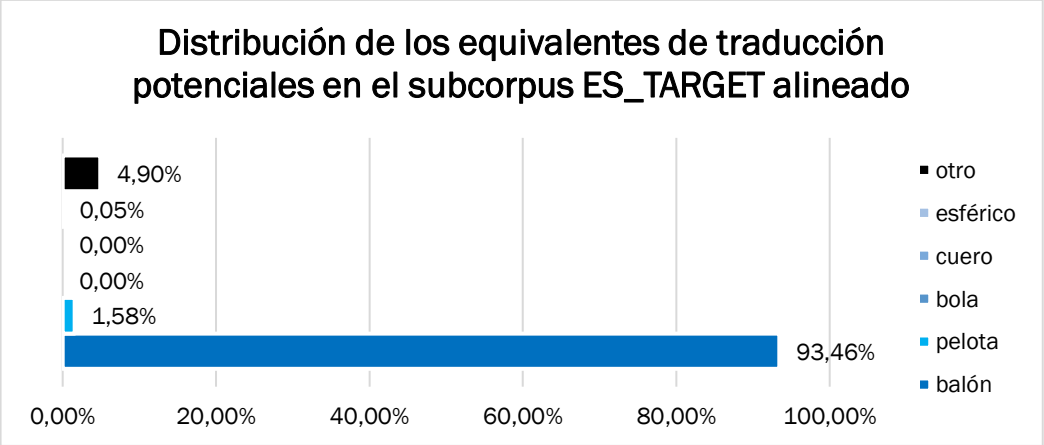
1.3. Subcorpus ES_TARGET alineado

Si tomamos como referencia la alineación bilingüe del subcorpus ES_TARGET con el subcorpus EN_SOURCE, se corrobora y se acentúa esta tendencia a la reducción de equivalentes representados y a la polarización de la distribución y del grado de representación de los equivalentes restantes.

Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico	otro	
N.º de ocurrencias	1715	29	0	0	1	90	Σ = 1835

Tabla 13. Equivalentes de traducción potenciales en español de la unidad terminológica «ball» con sus ocurrencias en el subcorpus ES_TARGET alineado.

En la Tabla 13, se muestra que de los cinco equivalentes de traducción potenciales iniciales solo están representados tres: «balón», «pelota» y «esférico». Esto significa que el sistema de traducción automática basado en redes neuronales con el que opera DeepL Pro ha establecido como equivalentes de traducción estas tres unidades terminológicas y ha descartado las dos restantes, «bola» y «cuero».



Gráfica 3. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

En lo que se refiere a la distribución, destacan, además de la preponderancia de «balón» con respecto a los otros dos equivalentes representados, la exigua presencia de la unidad

terminológica «esférico», que se manifiesta en su valor de distribución del 0,05 %, y la importancia cuantitativa que adquiere la categoría «otro».

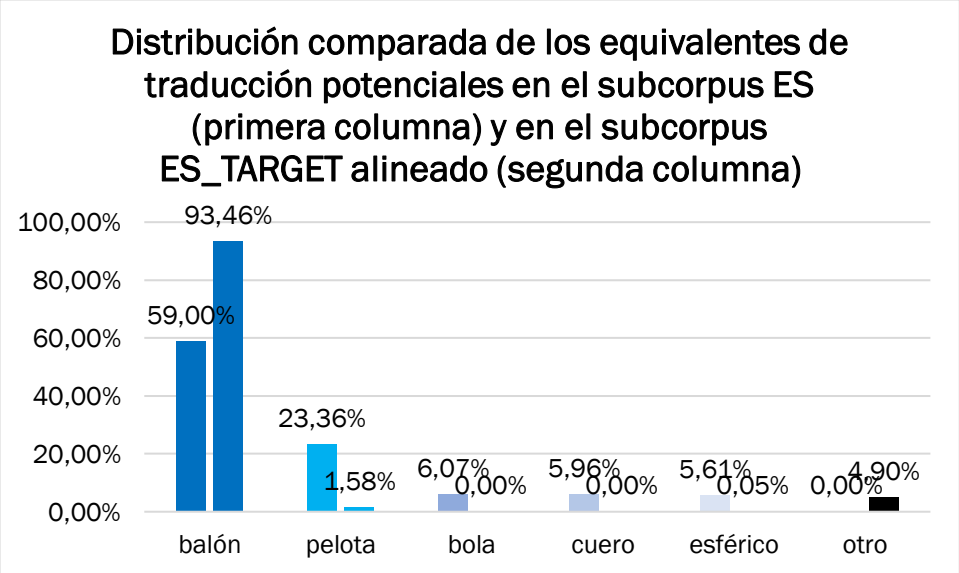
Equivalente de traducción potencial	balón	pelota	bola	cuero	esférico	otro
Grado de representación del equivalente	alto	muy bajo	nulo	nulo	muy bajo	muy bajo

Tabla 14. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

La interpretación cualitativa de los resultados respalda los valores relativos a la distribución, puesto que refleja la prevalencia del equivalente «balón» sobre los otros dos equivalentes, «pelota» y «esférico».

1.4. Comparación entre el subcorpus ES y el subcorpus ES_TARGET alineado

Al contrastar los resultados obtenidos de la aplicación de los parámetros de representación, distribución y grado de representación de los equivalentes en estos dos corpus, el primer elemento discordante es el número de equivalentes que cuentan con representación. Así pues, mientras que el subcorpus ES los cinco equivalentes de traducción potenciales están representados, en el subcorpus ES_TARGET alineado solo tres equivalentes, «balón», «pelota» y «esférico», cuentan con representación.



Gráfica 4. Distribución comparada de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

Si se comparan las distribuciones de los equivalentes, se puede apreciar que «balón» y «pelota» ocupan las posiciones de liderazgo en ambos corpus, mientras que «esférico» asciende de la quinta

posición que ocupa en el subcorpus ES a la tercera que ostenta en el subcorpus ES_TARGET debido a la ausencia de representación de los equivalentes «bola» y «cuero» en este corpus. Asimismo, «balón» es el único equivalente que cuenta con una relación porcentual de ocurrencias mayor en el subcorpus ES_TARGET alineado que en el subcorpus ES.

	Subcorpus ES	**Subcorpus ES_TARGET alineado**
balón	medio	alto
pelota	bajo	muy bajo
bola	bajo	nulo
cuero	bajo	nulo
esférico	bajo	muy bajo

Tabla 15. Grado de representación comparado de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

En la Tabla 15, donde se comparan los grados de representación de los equivalentes en los dos subcorpus, se pone de manifiesto que el sistema de traducción automática se decanta en español, sin duda, por la unidad terminológica «balón», que considera el equivalente de traducción más adecuado para la unidad terminológica «*ball*», en detrimento del resto de equivalentes.

2. *Shot*

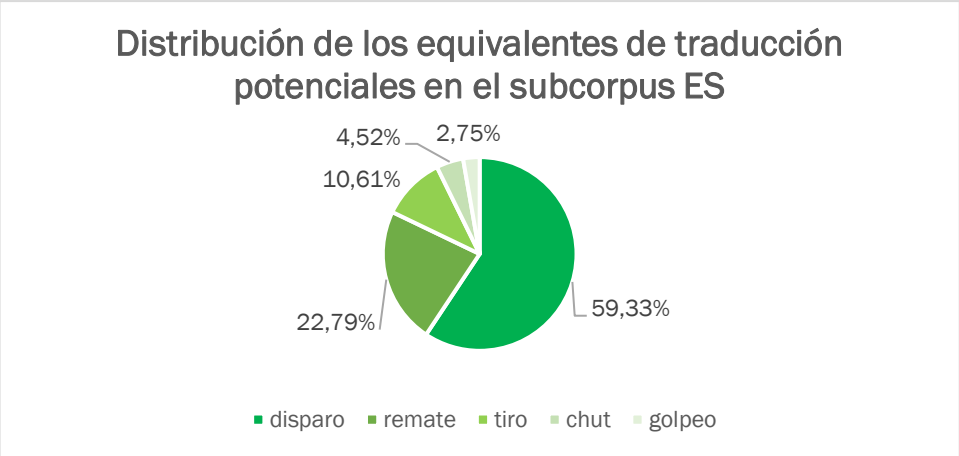
La unidad terminológica «*shot*» dispone de 744 ocurrencias en el subcorpus EN, de modo que ocupa el segundo lugar en la muestra de análisis según el número total de ocurrencias.

2.1. Subcorpus ES

Tras inspeccionar de forma manual la extracción terminológica del subcorpus ES mediante Sketch Engine, seleccionamos un total de cinco equivalentes de traducción potenciales en español: «disparo», «remate», «tiro», «chut» y «golpeo».

Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo	
N.º de ocurrencias	302	116	54	23	14	Σ = 509

Tabla 16. Equivalentes de traducción potenciales en español de la unidad terminológica «shot» *con sus ocurrencias en el subcorpus ES.*



Gráfica 5. Distribución de los equivalentes de traducción potenciales en el subcorpus ES.

De forma paralela a lo observado con la unidad terminológica «*ball*», «*shot*» presenta un abanico amplio de equivalentes de traducción potenciales en español, aunque el equivalente «disparo» posee un valor de distribución muy superior al resto.

Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo
Grado de representación del equivalente	medio	bajo	bajo	muy bajo	muy bajo

Tabla 17. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES.

La interpretación cualitativa de la distribución contrasta con los valores obtenidos para este parámetro cuantitativo, puesto que ninguno de los equivalentes de traducción potenciales presenta un grado de representación alto, muy alto o máximo.

2.2. Subcorpus ES_TARGET

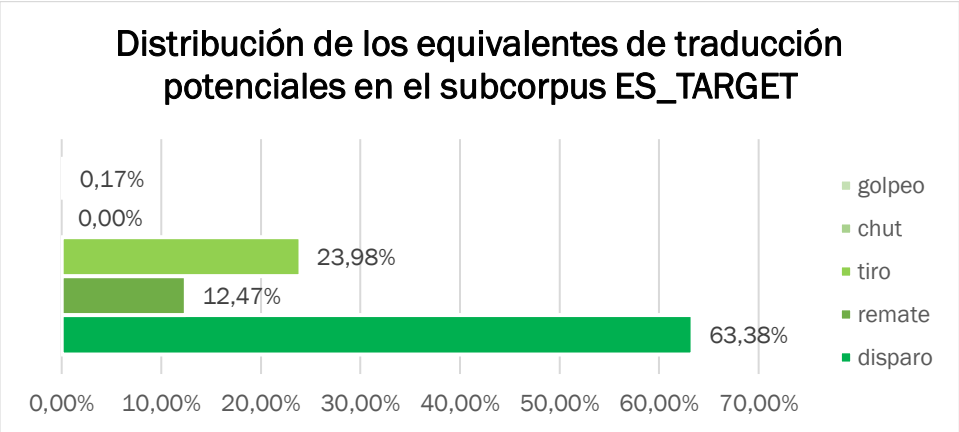
En el subcorpus ES_TARGET, solo obtienen representación cuatro de los cinco equivalentes de traducción potenciales seleccionados inicialmente, puesto que el término «chut» no aparece en ninguno de los 80 textos meta que lo conforman.

Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo	
N.º de ocurrencias	727	143	275	0	2	Σ = 1147

Tabla 18. Equivalentes de traducción potenciales en español de la unidad terminológica «shot» *con sus ocurrencias en el subcorpus ES_TARGET.*

De nuevo, el resultado generado por el sistema de traducción automática se caracteriza por conceder mucha importancia a una unidad terminológica en concreto, ya que «disparo» excede con creces en número de ocurrencias al resto de equivalentes combinados. Sin embargo, en contraste

con la tendencia observada con la unidad terminológica «*ball*», los equivalentes «remate» y «tiro» también logran un volumen de ocurrencias significativo.



Gráfica 6. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

La distribución de los equivalentes de traducción potenciales corrobora esta idea: cuando se le concede un valor porcentual al volumen relativo de ocurrencias de los equivalentes «remate» y «tiro», se pone de manifiesto que poseen una significación cuantitativa elevada.

Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo
Grado de representación del equivalente	medio	bajo	bajo	nulo	muy bajo

Tabla 19. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

De hecho, desde una perspectiva cualitativa, los grados de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET se asemejan a los obtenidos en el subcorpus ES, salvo para el equivalente «chut», que no está representado en el subcorpus ES_TARGET. Por consiguiente, mediante la interpretación cualitativa de los datos recabados en relación con el subcorpus ES_TARGET, podemos deducir que el sistema de traducción automática otorga un grado de representación a los equivalentes de traducción potenciales que se adecúa casi por completo a su uso real, definido en la muestra textual recogida en el subcorpus ES.

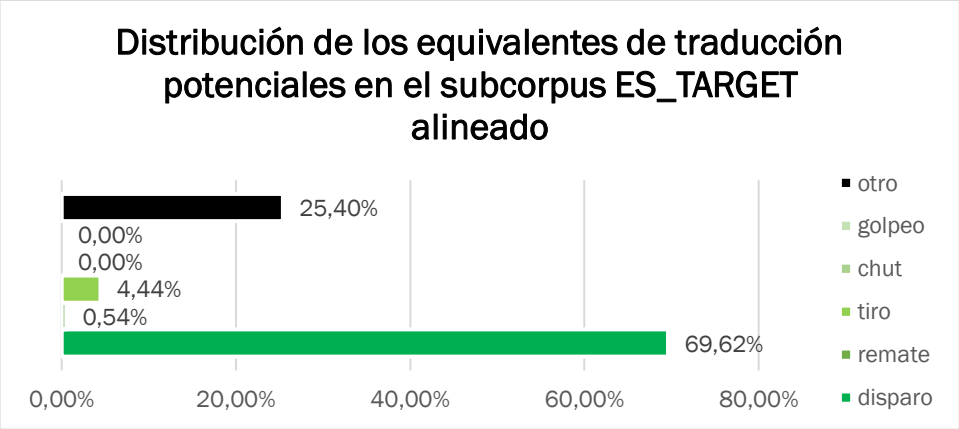
2.3. Subcorpus ES_TARGET alineado

En el subcorpus ES_TARGET alineado, hay dos equivalentes que no cuentan con representación: «chut» y «golpeo».

Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo	otro	
N.º de ocurrencias	518	4	33	0	0	189	Σ = 744

Tabla 20. Equivalentes de traducción potenciales en español de la unidad terminológica «shot» con sus ocurrencias en el subcorpus ES_TARGET alineado.

Atendiendo al número de ocurrencias de los equivalentes representados, se amplía de forma considerable el margen entre «disparo» y los otros dos equivalentes, «remate» y «tiro». Asimismo, «tiro» supera a «remate» en número de ocurrencias y «otro», que hace alusión a otras opciones de traducción de la unidad terminológica original en inglés (omisiones, por lo general) se coloca en segundo lugar.



Gráfica 7. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

En la Gráfica 7, que muestra la distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado, se confirma el cambio de tendencia de los resultados cuantitativos en comparación con los obtenidos para el subcorpus ES_TARGET que anticipábamos al interpretar los valores relativos al número de ocurrencias.

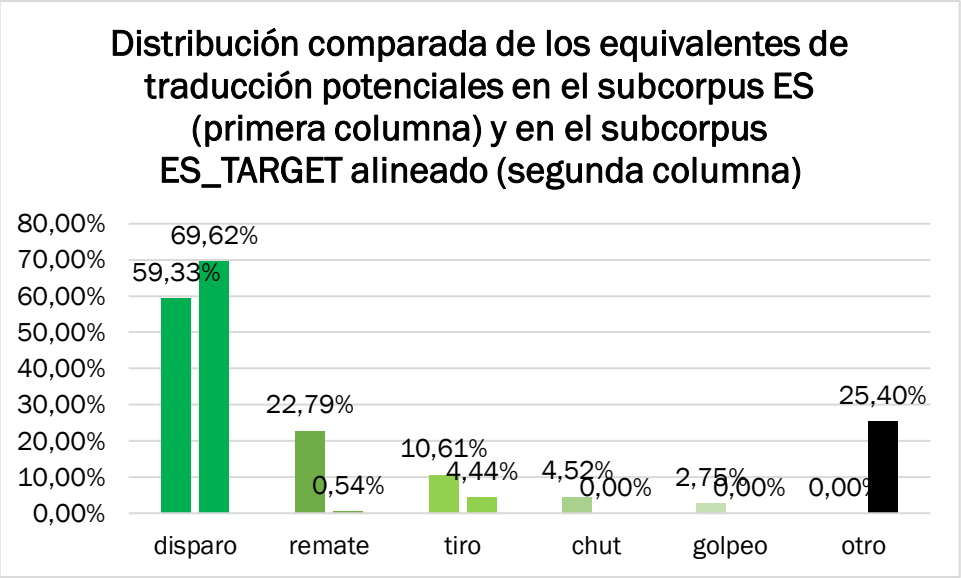
Equivalente de traducción potencial	disparo	remate	tiro	chut	golpeo	otro
Grado de representación del equivalente	alto	muy bajo	muy bajo	nulo	nulo	bajo

Tabla 21. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

Este cambio de tendencia reflejado en la distribución se reproduce también al adoptar un enfoque cualitativo. Los resultados relativos al grado de representación de los equivalentes de traducción potenciales de «*shot*» parecen alinearse en cierta manera con el patrón observado en el análisis de la unidad terminológica «*ball*», pues existe una brecha significativa en cuanto al grado de representación entre «disparo» y los otros dos equivalentes representados, «remate» y «tiro».

2.4. Comparación entre el subcorpus ES y el subcorpus ES_TARGET alineado

En el subcorpus ES_TARGET alineado, solo se conservan tres de los cinco equivalentes de traducción potenciales extraídos a través de la explotación del subcorpus ES, exactamente el mismo resultado que para la unidad terminológica «*ball*».



Gráfica 8. Distribución comparada de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

La distribución comparada muestra cómo se produce un fenómeno de redistribución del volumen de ocurrencias relativo que presentan los equivalentes «remate», «tiro», «chut» y «golpeo» hacia el equivalente «disparo» y la opción «otro» en el subcorpus ES_TARGET alineado; desde un punto de vista cuantitativo, esto se traduce en la ausencia de representación de «chut» y «golpeo», así como en una pérdida notable de significación cuantitativa de los equivalentes «remate» y «tiro», cuyos valores de distribución son objeto de una reducción del 22,25 % y del 6,17 %, respectivamente, con respecto a los valores obtenidos en el subcorpus ES.

	Subcorpus ES	Subcorpus ES_TARGET alineado
disparo	medio	alto
remate	bajo	muy bajo
tiro	bajo	muy bajo
chut	muy bajo	nulo
golpeo	muy bajo	nulo

Tabla 22. Grado de representación comparado de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

Mediante la comparación del grado de representación que presentan los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado, comprobamos que el sistema de traducción automática concede un nivel de preferencia elevado a la unidad terminológica «disparo» como equivalente de traducción de «*shot*», mientras que relega a los otros dos equivalentes representados, «remate» y «tiro», a un rol residual. Por ende, podemos concluir que DeepL Pro está cerca de establecer la unidad terminológica «disparo» como equivalente de traducción único en lengua española de la unidad terminológica «*shot*».

3. *Referee*

La unidad terminológica «*referee*» dispone de 166 ocurrencias, de modo que se posiciona en tercer lugar en la muestra de análisis atendiendo al número total de ocurrencias.

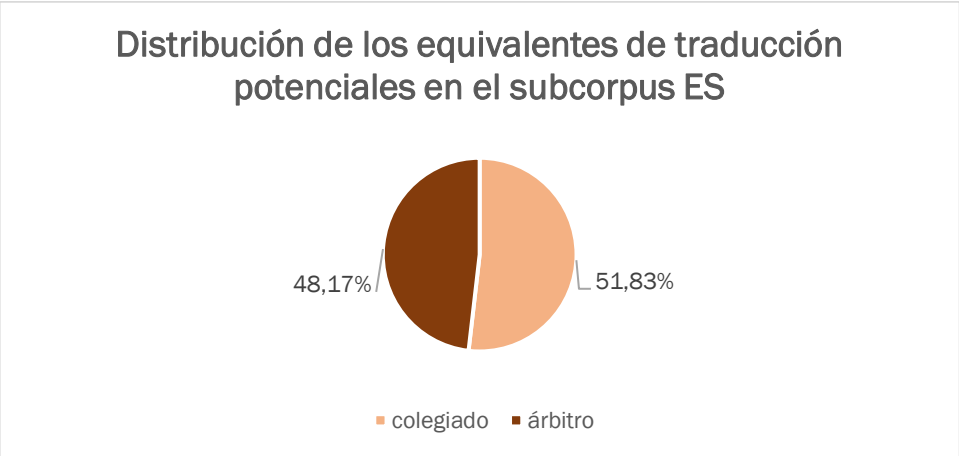
3.1. Subcorpus ES

Esta unidad terminológica tiene la peculiaridad de ser la única unidad analizada con menos de cinco equivalentes de traducción potenciales en español, puesto que solo se recogen dos, «colegiado» y «árbitro», en la selección definitiva resultante de la explotación del subcorpus ES.

Equivalente de traducción potencial	colegiado	árbitro	
N.º de ocurrencias	113	105	Σ = 218

Tabla 23. Equivalentes de traducción potenciales en español de la unidad terminológica «referee» con sus ocurrencias en el subcorpus ES.

Como se muestra en la Tabla 23, el número de ocurrencias de los dos equivalentes de traducción potenciales es similar, lo que constituye otro rasgo distintivo frente a los resultados obtenidos para los equivalentes de las unidades terminológicas estudiadas hasta el momento.



Gráfica 9. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

El equilibrio descrito en relación con el número de ocurrencias de cada equivalente se extrapola a su valor de distribución, dado que solo difieren en un 3,66 %, una distancia de distribución inédita hasta el momento.

Equivalente de traducción potencial	colegiado	árbitro
Grado de representación del equivalente	medio	medio

Tabla 24. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES.

Asimismo, la codominancia de ambos equivalentes, «colegiado» y «árbitro», se demuestra desde una perspectiva cualitativa, puesto que los dos presentan un grado de representación medio.

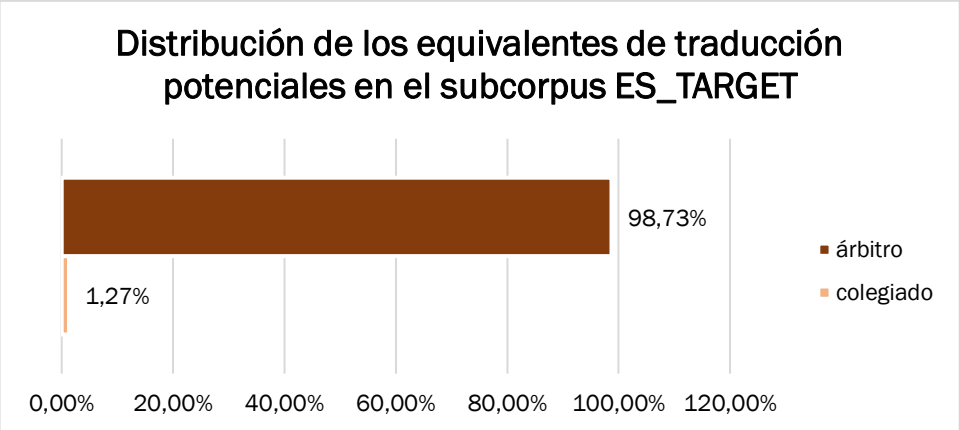
3.2. Subcorpus ES_TARGET

Tanto «colegiado» como «árbitro» obtienen representación en el subcorpus ES_TARGET.

Equivalente de traducción potencial	colegiado	árbitro	
N.º de ocurrencias	3	233	Σ = 236

*Tabla 25. Equivalentes de traducción potenciales en español de la unidad terminológica «*referee*» con sus ocurrencias en el subcorpus ES_TARGET.*

En contraste con el número de ocurrencias registrado para los dos equivalentes en el subcorpus ES, en el subcorpus ES_TARGET se abre una brecha cuantitativa significativa entre ellos. Además, los equivalentes intercambian posiciones, ya que «colegiado», que supera ligeramente a «árbitro» en número de ocurrencias en el subcorpus ES, se coloca en segunda posición en lo concerniente al número de ocurrencias, de forma que «árbitro» asciende a la primera posición.



Gráfica 10. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

La gran distancia cuantitativa entre los dos equivalentes se refleja a la perfección mediante sus valores de distribución: las ocurrencias de «colegiado» representan el 1,27 % del total de ocurrencias; en cambio, las ocurrencias de «árbitro» comprenden el 98,73 %. Así, la estrecha brecha de distribución del 3,66 % que separa los dos equivalentes en el subcorpus ES se contrapone con la amplia brecha del 97,46 % que los distancia en el subcorpus ES_TARGET.

Equivalente de traducción potencial	colegiado	árbitro
Grado de representación del equivalente	muy bajo	muy alto

Tabla 26. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET.

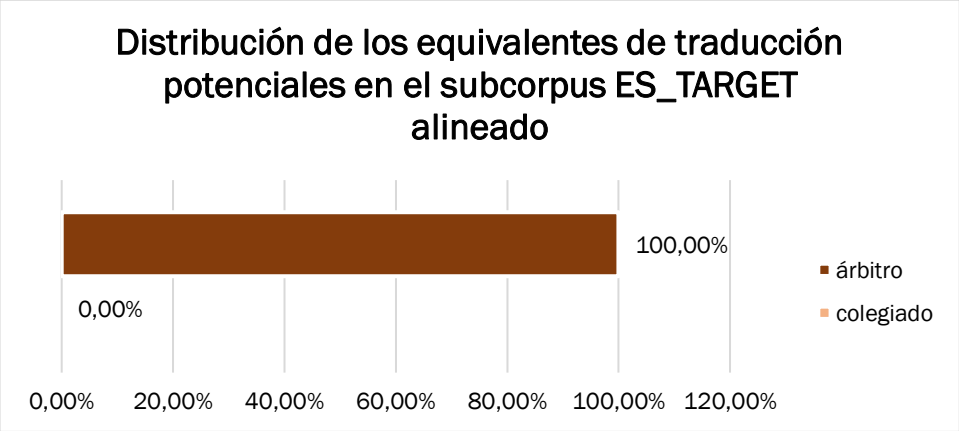
La interpretación cualitativa de los valores de distribución corrobora la amplia diferencia de representación que existe entre los dos equivalentes. En efecto, «colegiado» presenta un grado de representación muy bajo y «árbitro» dispone de un grado de representación muy alto, por lo que se ubican en extremos opuestos de la escala que mesura dicho parámetro cualitativo.

3.3. Subcorpus ES_TARGET alineado

En el subcorpus ES_TARGET alineado, solo obtiene representación el equivalente «árbitro»; además, tampoco se muestra la opción «otro».

Equivalente de traducción potencial	colegiado	árbitro	
N.º de ocurrencias	0	166	Σ = 166

*Tabla 27. Equivalentes de traducción potenciales en español de la unidad terminológica «*referee*» con sus ocurrencias en el subcorpus ES_TARGET alineado.*



Gráfica 11. Distribución de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

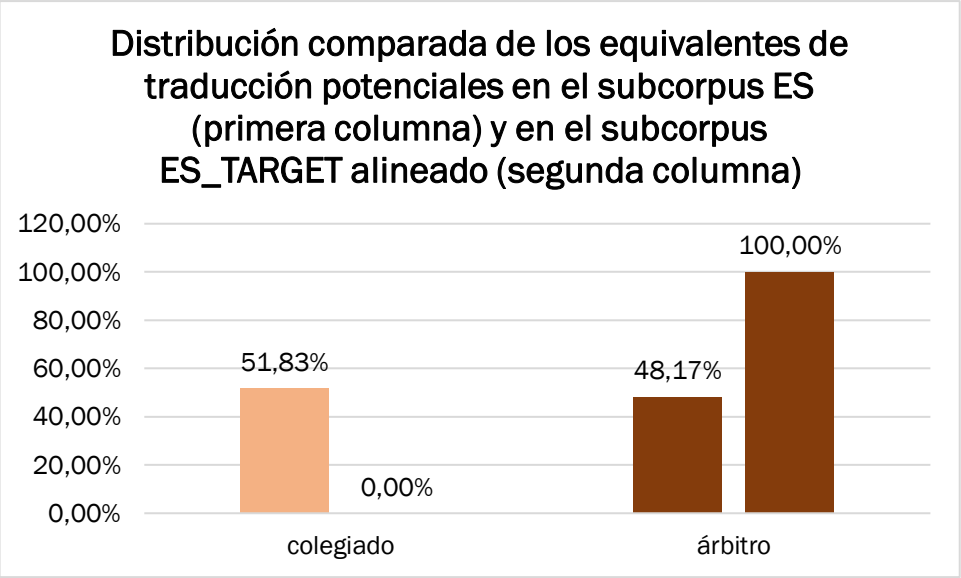
Equivalente de traducción potencial	colegiado	árbitro
Grado de representación del equivalente	nulo	máximo

Tabla 28. Grado de representación de los equivalentes de traducción potenciales en el subcorpus ES_TARGET alineado.

Por ende, las 166 ocurrencias de unidad terminológica «*referee*» en el subcorpus EN_SOURCE se traducen por «árbitro» en el subcorpus ES_TARGET. Se trata de un resultado trascendente, ya que implica que DeepL Pro establece la unidad terminológica «árbitro» como equivalente de traducción único en lengua española de la unidad terminológica «*referee*».

3.4. Comparación entre el subcorpus ES y el subcorpus ES_TARGET alineado

De los dos equivalentes, «colegiado» y «árbitro», que conforman la selección definitiva en el subcorpus ES solo se representa «árbitro» en el subcorpus ES_TARGET alineado.



Gráfica 12. Distribución comparada de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

	Subcorpus ES	Subcorpus ES_TARGET alineado
colegiado	medio	nulo
árbitro	medio	máximo

Tabla 29. Grado de representación comparado de los equivalentes de traducción potenciales en el subcorpus ES y en el subcorpus ES_TARGET alineado.

Los datos reflejados en la Gráfica 12 y en la Tabla 29 muestran cómo «colegiado», que es el equivalente de «*referee*» más representado en el subcorpus ES, ya que cuenta con un valor de distribución del 51,83 % y con un grado de representación medio, desaparece en el subcorpus ES_TARGET alineado; por el contrario, el equivalente «árbitro», pese a presentar un valor de distribución ligeramente inferior en el subcorpus ES, aúna todas las ocurrencias en el subcorpus ES_TARGET alineado. Por tanto, «árbitro» es el primero y el único de los equivalentes de traducción potenciales analizados en alcanzar el grado de representación máximo. Como mencionábamos anteriormente, esto quiere decir que DeepL Pro ha proporcionado una traducción unívoca al español, «árbitro», de la unidad terminológica «*referee*».

4. Recapitulación

Partimos de una muestra de análisis compuesta por tres unidades terminológicas nominales univerbales en inglés con más de 150 ocurrencias en el subcorpus EN: «ball» cuenta con 1835, «*shot*» dispone de 744 y «*referee*» presenta 166.

Este apartado se divide en tres subapartados, «Subcorpus ES», «Subcorpus ES_TARGET» y «Subcorpus ES_TARGET alineado», uno por cada corpus analizado en los «RESULTADOS».

4.1. Subcorpus ES

A partir de la explotación del subcorpus ES, establecemos la selección definitiva de los equivalentes de traducción potenciales en español de cada unidad terminológica en inglés: «balón», «pelota», «bola», «cuero» y «esférico» para «*ball*»; «disparo», «remate», «tiro», «chut» y «golpeo» para «*shot*», y «colegiado» y «árbitro» para «*referee*».

Cabe recordar que el subcorpus ES contiene una muestra textual extraída de crónicas futbolísticas en directo en lengua española redactadas por cronistas, por lo que este subcorpus constituye una muestra real y fiable de la lengua en uso. En este sentido, la selección de más de un equivalente de traducción potencial, cinco para «*ball*» y «*shot*» y dos para «*referee*», pone de manifiesto la diversidad y la riqueza de opciones terminológicas aptas en lengua española para designar cada uno de estos conceptos futbolísticos especializados. Esta abundancia de sinónimos intercambiables responde, con casi toda probabilidad, a las funciones apelativa y poética del lenguaje, ya que cubre la necesidad de adecuar el estilo a fin de focalizar la atención del lector en el texto.

En lo que respecta a la distribución y el grado de representación, los equivalentes de «*ball*» y de «*shot*» parecen regirse por el mismo patrón, ya que en ambos casos hay un equivalente, «balón» y «disparo», respectivamente, que obtiene una representación superior al resto, mientras que los demás presentan un valor de distribución y un grado de representación inferiores, pero relativamente equilibrados. En cambio, los equivalentes de traducción potenciales de «*referee*», «colegiado» y «árbitro», disponen de una representación equilibrada, tanto desde el punto de vista cualitativo como cuantitativo.

4.2. Subcorpus ES_TARGET

El subcorpus ES_TARGET está compuesto por las traducciones proporcionadas por DeepL Pro de los 80 textos originales en inglés contenidos en el subcorpus EN_SOURCE. Hay que tener en cuenta que los resultados correspondientes a este subcorpus son el resultado de su explotación directa y, en consecuencia, no dependen de su alineación bilingüe con el subcorpus EN_SOURCE. Así, el número total de ocurrencias que se le computan a cada equivalente no se obtiene mediante el análisis de su alineación con la unidad terminológica de partida en inglés, sino que representa el número total de veces que el sistema de traducción automática ha optado por dicho equivalente, sea cual sea la razón subyacente.

En relación con las unidades terminológicas «*ball*» y «*shot*», se aprecia una reducción del número de equivalentes representados, ya que no se registra ninguna ocurrencia de los equivalentes «*bola*» y «*chut*». En el caso de «*ball*», la prevalencia del equivalente «balón» sobre el resto se acentúa de manera muy significativa con respecto a los resultados correspondientes al subcorpus ES, lo que se manifiesta tanto en su valor de distribución, un parámetro cuantitativo, como en su grado de representación, un parámetro cualitativo. En cambio, en lo que se refiere a

la unidad terminológica «*shot*», su equivalente de traducción potencial «disparo» también experimenta un ligero cambio al alza en cuanto a su valor de distribución, que no se refleja cualitativamente; por el contrario, el resto de los equivalentes continúan contando con una representación equilibrada. Por último, en cuanto a la unidad terminológica «*referee*», aunque sus dos equivalentes están representados en el subcorpus ES_TARGET, se reproduce el patrón observado con «*ball*», puesto que el equivalente «árbitro» aúna casi todas las ocurrencias, mientras que «colegiado» queda relegado a un papel residual; estos cambios drásticos en la presencia de los dos equivalentes quedan patentes tanto a nivel cuantitativo como cualitativo.

4.3. Subcorpus ES_TARGET alineado

El subcorpus ES_TARGET alineado se origina a partir de la alineación bilingüe del subcorpus ES_TARGET, que consta de 80 textos meta en español traducidos por DeepL Pro, con el subcorpus EN_SOURCE, que contiene los 80 textos originales en inglés. El proceso de alineación uno a uno de los segmentos en la lengua origen con los segmentos correspondientes en la lengua meta persigue el objetivo de conocer de manera numérica y con precisión las traducciones que el sistema de traducción automática ofrece para las unidades terminológicas «*ball*», «*shot*» y «*referee*». De esta forma, podemos realizar un análisis dirigido de estas traducciones con base en los equivalentes de traducción potenciales en español de cada una de estas unidades.

Tras completar este análisis, deducimos que DeepL Pro sigue un patrón para abordar la transferencia al español de las tres unidades terminológicas.

- No emplea todos los equivalentes de traducción potenciales a su disposición para traducir las unidades terminológicas, puesto que en el subcorpus ES_TARGET alineado no se representan los equivalentes «bola», «cuero», «chut», «golpeo» ni «colegiado». Así, DeepL Pro recurre a tres de cinco equivalentes para traducir las unidades terminológicas «*ball*» y «*shot*» y a uno de dos equivalentes en el caso de la unidad terminológica «*referee*».
- Siempre prioriza un equivalente de traducción sobre los demás: «balón» cuenta con un valor de distribución del 93,46 % y con un grado de representación alto; «disparo» presenta un valor de distribución del 69,62 % y un grado de representación alto; «árbitro» dispone de un valor de distribución del 100 % y de un grado de representación máximo. En lo que respecta al resto de equivalentes de traducción representados, ninguno posee un grado de representación superior a muy bajo, por lo que adquieren una significación mínima y juegan un papel residual.

Por consiguiente, atendiendo a la interpretación de los resultados obtenidos, concluimos que el sistema de traducción con el que opera DeepL Pro muestra una tendencia relativa a la univocidad a la hora de traducir al español las unidades «*ball*», «*shot*» y «*referee*».

CONCLUSIONES

En el presente trabajo de investigación, hemos llevado a cabo un análisis de la traducción automática del inglés al español de las unidades terminológicas «*ball*», «*shot*» y «*referee*». Para ello, hemos empleado una metodología basada en corpus. De hecho, nos hemos servido de dos corpus, el corpus comparable bilingüe FÚTBOL_MXM (EN-ES) y el corpus paralelo bilingüe FÚTBOL_MXM_PARALELO (EN>ES), que hemos compilado, analizado y explotado para obtener los resultados presentados en el capítulo anterior (v. el capítulo «RESULTADOS») y, en consecuencia, las conclusiones extraídas de la interpretación de estos resultados que abordamos en el presente capítulo.

El objeto de nuestra investigación son las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol, una elección que responde a que estas unidades designan conceptos denotativos y unívocos, lo que facilita en gran medida su estudio al no tener que lidiar con dificultades relacionadas con la ambigüedad semántica y la pragmática.

Atendiendo a este objeto de estudio, hemos desarrollado nuestro trabajo de investigación en torno a la siguiente hipótesis: «la traducción automática del inglés al español de las unidades terminológicas nominales univerbales pertenecientes al subcampo de especialidad del fútbol es inadecuada». Con base en los resultados presentados en el capítulo anterior, corroboramos esta hipótesis. En esencia, consideramos inadecuada la traducción al español que el motor de traducción automática escogido para este trabajo, DeepL Pro, ofrece para las unidades terminológicas nominales univerbales estudiadas porque el uso que hace de los equivalentes de traducción potenciales no se corresponde con lo observado en la muestra textual del subcorpus ES, que constituye una muestra real y fiable de la lengua en uso. Esta inadecuación se justifica a través de los resultados correspondientes al subcorpus ES_TARGET alineado:

- el sistema de traducción automática no recurre en ningún caso a todos los equivalentes de traducción potenciales a su alcance para transferir al español las unidades terminológicas que conforman la muestra de análisis, puesto que utiliza tres de cinco equivalentes para traducir las unidades «*ball*» y «*shot*» y uno de dos en el caso de «*referee*»;
- el valor de distribución y el grado de representación de los equivalentes representados dista mucho de lo registrado en el subcorpus ES, de manera que todos los equivalentes representados son objeto de un proceso de polarización cuantitativa y cualitativa, bien sea hacia la sobrerrepresentación o hacia la infrarrepresentación.
- el sistema de traducción automática muestra una tendencia a la univocidad a la hora de traducir las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol, de manera que otorga mucha preferencia a un único equivalente de traducción potencial en detrimento del resto de equivalentes.

Asimismo, el análisis del contenido lingüístico del subcorpus ES nos ha permitido comprobar que, en español, el lenguaje de especialidad del fútbol se caracteriza por su diversidad terminológica, que se pone de manifiesto en el gran número de variantes aptas para designar un único concepto especializado. Esta riqueza terminológica, que no da lugar a ambigüedades semánticas, parece cumplir con una doble función apelativa y poética; es decir, el uso de opciones terminológicas sinónimas puede responder al objetivo de adecuar el estilo para llamar la atención del lector.

Más allá de las conclusiones derivadas de la interpretación de los resultados, hemos confeccionado un protocolo de compilación de corpus mixto a partir de la combinación de las propuestas de Vargas Sierra (2005) y Seghiri (2011). También hemos compilado dos corpus equilibrados y representativos, uno comparable y otro paralelo, que podemos emplear en futuras investigaciones.

A través de este trabajo, se abren varias líneas de investigación nuevas: continuar profundizando en el estudio de las unidades terminológicas nominales univerbales del subcampo de especialidad del fútbol a fin de corroborar o desmentir las conclusiones aquí presentadas y de subsanar posibles carencias y defectos relacionados con la obtención de los resultados; entrenar el sistema de traducción automática con el que opera DeepL Pro para adecuar el resultado que ofrece, o extrapolar esta investigación a otros tipos de unidades terminológicas.

BIBLIOGRAFÍA

Allen, J. H. (2003). Post-editing. En H. Somers (Ed.), *Computers and Translation: A translator's guide* (297-318). John Benjamins.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. En M. Baker, G. Francis y E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (233-250). John Benjamins.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7 (2), 223-243. https://benjamins.com/online/target/articles/target.7.2.03bak.

Bergh, G. (2011). Football is war: a case study of minute-by-minute football commentary. *Veredas – Revista de Estudos Linuísticos*, 15 (2), 83-93.

Berner, S. (10-12 de septiembre, 2003). *Lost in Translation: Cross-Lingual Communication, and Virtual Academic Communities* [Comunicación en congreso]. 5th Annual Conference on World Wide Web Applications, Durban. http://general.rau.ac.za/infosci/www2003/Papers/Berner,%20S%20Lost%20in%20Translation.pdf.

Bonvin Faura, M. A. (2007). *La prensa digital: lenguaje y características* [Tesis doctoral, Universidad de Granada]. DIGIBUG. https://digibug.ugr.es/handle/10481/1700.

Bowker, L., y Fisher, D. (2010). Computer-aided translation. En Y. Gambier y L. van Doorslaer (Eds.), *Handbook of Translation Studies. Volume 1* (60-65). John Benjamins.

Bowker, L. y Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Editorial Empúries.

Casasús i Guri, J. M. y Núñez Ladeveze, L. (1991). *Estilo y géneros periodísticos*. Ariel Comunicación.

Corpas Pastor, G. (2001). Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada. *TRANS : Revista de Traductología*, 5, 155-184. https://revistas.uma.es/index.php/trans/article/view/2916/2710.

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

Corpas Pastor, G. y Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish). En A. Beeby, P. Rodríguez Inés y P. Sánchez Gijón (Eds.), *Corpus Use and Translating* (75-107). John Benjamins.

Díaz Prieto, P. (2012). Luces y sombras en los 75 años de traducción de automática. En J. J. Lanero Fernández y J. L. Chamosa (Eds.), *Lengua, traducción, recepción: en honor de Julio César Santoyo. Vol. 2* (139-175). Servicio de Publicaciones de la Universidad de León.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9 (1), 99–115.

Faya Ornia, M. G. (2015). Propuesta de clasificación de corpus textuales. En M. T. Sánchez Nieto, S. Álvarez Álvarez, V. Arnáiz Uzquiza, M. T. Ortego Antón, L. Santamaría Ciordia y R. Fernández Muñiz (Eds.), *Metodologías y aplicaciones en la investigación en traducción e interpretación con corpus* (339-355). Universidad de Valladolid.

Francis, N. W. (1982). Problems of assembling and computerizing large corpora. En S. Johansson (Ed.), *Computer corpora in English language research* (7-24). Norwegian Computing Centre for the Humanities.

Forcada, M. L., Sánchez Martínez, F. y Pérez Ortiz, J. A. (2016). *Manual de informática y de tecnologías para la traducción*. Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante.

Gárciga Rodríguez, M. C. y Gómez Masjuán, M. E. (2013). Redefiniciones de la crónica en el mundo online. Estudio de la crónica hipermedia en directo en los cibermedios 20 minutos, El Comercio y RTVE. *Palabra Clave,* 16 (3), 913-943.

Granger, S. (2003). The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies? En S. Granger, J. Lerot y S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies* (17–30). Editions Rodopi B.V.

Koponen, M. (2015). How to teach machine translation post-editing? Experiences from a post-editing course. En S. O'Brien y M. Simard (Eds.), *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (2-15). Association for Machine Translation in the Americas.

Kovljanin, S. (2018a). La crónica futbolística en directo en el panorama de los géneros periodísticos. Komunikacija i Kultura Online, 9 (9), 92-118.

Kovljanin, S. (2018b). El lenguaje y el estilo de la crónica futbolística. *Beoiberística*, 2 (1), 73-85. https://beoiberistica.fil.bg.ac.rs/index.php/beoiberistica/article/view/35.

Laviosa, S. (1997). How Comparable Can 'Comparable Corpora' Be? *Target*, 9 (2), 287-317. https://benjamins.com/online/target/articles/target.9.2.05lav.

Lewandowski, M. (2012). The Language of Online Sports Commentary un a Comparative Perspective. *Lingua Posnaniensis*, LIV (1), 65-76.

Lommel, A. (2017). *Neural MT: Sorting Fact from Fiction*. Common Sense Advisory.

Maldonado González, M. C. y Liébana González, M. (2021). Los motores de traducción automática y su uso como herramienta lexicográfica en la traducción de unidades léxicas aisladas. *Círculo de Lingüística Aplicada a la Comunicación*, 88, 189-212. https://doi.org/10.5209/clac.77002.

Martínez Albertos, J. L. (1998). *Curso general de redacción periodística*. Edición revisada. Paraninfo.

Metola Navaridas, M. (2022). *La traducción automática neuronal inglés-español de las recetas de cocina: análisis de errores* [Trabajo fin de grado, Universidad de Valladolid]. UVaDOC. https://uvadoc.uva.es/handle/10324/54877.

McEnery, T., Xiao, R. y Tono, Y. (2005). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.

McEnery, T. y Hardie, A. (2012). *Corpus linguistics: Method, Theory and Practice*. John Benjamins.

McEnery, T. y Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press.

Naranjo de Arcos, A. (2011). Tratamiento de la información deportiva en la prensa: la crónica como género prevalente. El caso de los encuentros de fútbol entre Real Madrid y F. C. Barcelona [Tesis doctoral, Universidad de Málaga]. RIUMA. https://riuma.uma.es/xmlui/handle/10630/4848.

Nomdedeu Rull, A. (2004). *Terminología del fútbol y diccionarios: elaboración de un diccionario de especialidad para el gran público* [Tesis doctoral, Universitat Autònoma de Barcelona]. TDX. https://www.tdx.cat/handle/10803/4872#page=1.

Nunes Vieira, L. (2020). Post-editing of machine translation. En M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology* (319-335). Routledge.

O'Hagan, M. (2020). *The Routledge Handbook of Translation and Technology*. Routledge.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, 25 (3), 197-215.

Ortego Antón, M. T. (2019). *La terminología del sector agroalimentario (español-inglés) en los estudios contrastivos y de traducción especializada basados en corpus: los embutidos*. Peter Lang.

Ortego Antón, M. T. y Seghiri, M. (2019). La traducción automática de locuciones nominales del español al inglés: a pain in the neck? En C. Carrasco, M. Cantarero Muñoz y C. Díez Carbajo (Eds.), *Traducción y sostenibilidad cultural: sustrato, fundamentos y aplicaciones* (331-342). Ediciones Universidad de Salamanca.

Parrat, S. (2008). *Géneros periodísticos en prensa*. Quipus, CIESPAL.

Quah, C. K. (2006). *Translation and Technology*. Palgrave Macmillan.

Rojo, G. (2016). Los corpus textuales del español. En J. Gutiérrez-Rexach (Ed.), *Enciclopedia de lingüística hispánica* (285-296). Routledge.

Rost, A. (2006). *La interactividad en el periódico digital* [Tesis doctoral, Universitat Autònoma de Barcelona]. TDX. https://www.tdx.cat/handle/10803/4189#page=1.

Sánchez, A. (1995). Definición e historia de los corpus. En A. Sánchez, R. Sarmiento, P. Cantos, y J. Simón (Eds.), *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis* (7-25). SGEL.

Sánchez Ramos, M. M. y Rico Pérez, C. (2020). *Traducción automática. Conceptos clave, procesos de evaluación y técnicas de posedición*. Comares.

Sánchez Carnicer, J. (2021). *Las connotaciones presentes en la estructuración de la terminología de la discapacidad. Un estudio comparado inglés-español del uso que de ella realizan los medios de comunicación escritos de referencia* [Tesis doctoral, Universidad de Valladolid]. UVaDOC. https://uvadoc.uva.es/handle/10324/59069.

Sánchez Carnicer, J. (2022). *Traducción y discapacidad: un estudio comparado de la terminología inglés-español en la prensa escrita*. Peter Lang.

Sánchez Gijón, P. (2016). La posedición: hacia una definición competencial del perfil y una descripción multidimensional del fenómeno. *Sendebar*, 27, 151-162. https://revistaseug.ugr.es/index.php/sendebar/article/view/4016/5057.

Santamaría Pérez, M. I. (2009). *La terminología: definición, funciones y aplicaciones*. Repositorio Institucional de la Universidad de Alicante. http://rua.ua.es/dspace/handle/10045/12770.

Santamaría Pérez, M. I. (2015). Diseño, implementación y elaboración de una terminología multilingüe del ámbito del turrón, mazapanes y otros dulces. *Cuadernos AISPI*, 6, 75-94. http://rua.ua.es/dspace/handle/10045/53080.

Seghiri, M. (2006). Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad [Tesis doctoral, Universidad de Málaga]. RIUMA. https://riuma.uma.es/xmlui/handle/10630/2715.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *RLA. Revista de Lingüística Teórica y Aplicada*, 49 (2), 13-30.

Seghiri, M. (2017). Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. *Babel*, 63 (1), 43–64.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. (2005). Corpus and Text - Basic Principles. En M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (1-16). Oxbow Books.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins.

Torruella, J. y Llisterri, J. (1999). Diseño de corpus textuales y orales. En J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos* (45-77). Editorial Milenio.

Vargas Sierra, C. (2005). *Aproximación terminográfica al lenguaje de la piedra natural: propuesta de sistematización para la elaboración de un diccionario traductológico* [Tesis Doctoral, Universidad de Alicante]. Repositorio Institucional de la Universidad de Alicante. http://rua.ua.es/dspace/handle/10045/13272.

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios Humanísticos. Filología*, 30, 329–349. https://revpubli.unileon.es/ojs/index.php/EEHHFilologia/article/view/2847/2024.

Zanettin, F. (2012). *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Routledge.