DOI 10.14672/2.2023.2312

OVIDIA MARTÍNEZ SÁNCHEZ Y MARÍA ISABEL SANTAMARÍA PÉREZ

DISEÑO Y COMPILACIÓN DE UN CORPUS EN EL ÁMBITO MÉDICO DE LA REPRODUCCIÓN ASISTIDA PARA EL PROYECTO *NEOTERMED*

Universidad de Alicante

Resumen

Este estudio es una presentación sobre la metodología empleada para el diseño y construcción de un corpus representativo de la Reproducción Asistida para el proyecto de investigación *NEOTERMED*. Se trata de un corpus digital monolingüe compuesto por dos subcorpus de textos con diferente nivel de especialización y con distintas funciones comunicativas según el destinatario. De este corpus textual dependerán en su mayor parte los resultados del Proyecto de Investigación *NEOTERMED*, un conjunto de infografías terminológicas para pacientes y una aplicación terminológica para estudiantes de Biomedicina.

palabras clave: Reproducción Asistida, corpus digital, compilación de corpus, herramientas terminológicas

Abstract

Design and compilation of a corpus in the medical field of Assisted Reproduction for the NEOTERMED project

This study is a presentation of the design and construction of a representative corpus of Assisted Reproduction for the NEOTERMED *research project. It is a monolingual digital corpus composed of two sub-corpora of texts with different specialisation levels and communicative functions depending on the target audience. The results of* NEOTERMED*, a set of terminology infographics for patients and a terminology application for students of Biomedicine, will largely depend on this textual corpus.*

keywords: Assisted Reproduction, digital corpus, corpus compilation, terminology tools

Recibido: 30 de junio de 2023
Aceptación definitiva: 26 de octubre de 2023

1. Introducción

La terminología de una nueva especialidad participa de los términos de la ciencia de la que emerge y, necesariamente, comprende nuevas unidades léxicas o neologismos que conforman con otras una red propia de la especialidad. El conocimiento médico, como es sabido, está en permanente evolución, y genera necesidades expresivas propias de su materia; esto es, a medida que avanza la medicina, la terminología también se actualiza para designar nuevas enfermedades, tratamientos o tecnologías, que carecían de denominación por su no existencia o por falta de conocimiento. Como señalan diversos autores (Cremades 2003; Barona 2004; Mayor 2008; Guardiola i Baños 2011; Estopà 2012, 2019; Estopà, Lorente 2022; Santamaría 2023; Domènech, Santamaría 2023), el lenguaje y la medicina son dos disciplinas estrechamente relacionadas: en el campo de la medicina, cada cambio o avance supone un cambio terminológico, de modo que, o bien se incorporan nuevos términos (neónimos), o bien desaparecen otros. Este avance científico en el contexto médico está fuertemente ligado a las necesidades sociales, ya que la población convive y se preocupa por la salud, quiere ser partícipe de los avances médicos y, en consecuencia, debe comprender el léxico especializado o la terminología que le permitirá acceder al conocimiento (Cremades 2003: 129). Como explican Estopà y Lorente (2019: 11) "el análisis del léxico en los textos médicos, y en concreto de la variación terminológica detectada a lo largo del tiempo, atendiendo a la presencia o la ausencia de unidades, a las modificaciones que se producen entre paradigmas o a los cambios que se detectan en cuanto al contenido, es determinante para representar la evolución (oculta) del conocimiento".

Asumimos que la neología y la terminología son disciplinas lingüísticas, que permiten, a través del estudio de las unidades léxicas tanto generales como especializadas, analizar la evolución cultural, científica, económica, política o ideológica que una sociedad ha experimentado a lo largo de su historia. Subscribimos la propuesta de Guerrero Ramos (2016) de que no es lo mismo abordar el estudio de los neologismos desde la perspectiva del hablante que desde la perspectiva del oyente, especialmente cuando se trata de neologismos terminológicos por las implicaciones pragmático-discursivas que conlleva. Distinguimos, así, entre neologismos de emisor y receptor, por una parte, y neologismos de receptor, por otra (Guerrero Ramos 2017: 1399). Pensemos, por ejemplo, en la pandemia del coronavirus a raíz de la cual hubo una proliferación de creaciones léxicas, neológicas tanto para el emisor como para el receptor, pues había una nueva realidad que

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

nombrar (*confinamiento, postcovid, desescalada, distancia social*, etc.). Pero otros términos como *coronavirus, PCR, incidencia acumulada* o *inmunidad de grupo*, ya existían y se utilizaban en determinados ámbitos temáticos, pero su difusión y divulgación por los medios de comunicación y redes sociales hizo que traspasaran la barrera de lo especializado y se convirtieran en neologismos para los receptores, neófitos en estos campos o materias. Pensamos que estos criterios pueden ayudar a fijar el grado de neologicidad de una palabra y, sobre todo a distinguir cuándo estamos realmente ante un neologismo propiamente dicho y cuándo ante un uso neológico únicamente en el orden de la recepción, lo que, por lo que respecta a la unidad terminológica, está en relación con su trasvase a la lengua general o a otro ámbito de especialidad, por un lado, y, por otro, con el grado de especialización del texto en el que aparece.

Teniendo en cuenta esta diferenciación, el objetivo principal del proyecto *NEOTERMED. Neología y terminología en ciencias de la salud: variación y análisis multidimensional del discurso biomédico. Aplicación al ámbito de la Reproducción asistida en la Comunidad Valenciana para la alfabetización en salud y la igualdad de género*[1] es indagar en el proceso cognitivo de la comprensión de los términos, al tiempo que se analiza la relación entre las representaciones conceptuales y el grado de neologicidad de una terminología a la que los usuarios y usuarias acceden por necesidad y que supone el acceso a un conocimiento especializado. Se trata de un estudio semántico-pragmático de los términos, en este caso aplicado a una especialidad de enorme interés social, con gran repercusión económica y en continua evolución, como es el caso de la Reproducción Asistida (RA), teniendo en cuenta el destinatario al que se dirige y su difusión en textos de distinto nivel de especialización. Así, para el estudio se han seleccionado dos grupos de receptores:

- Por un lado, los y las estudiantes universitarios del campo de la Biomedicina que acceden a esta terminología por primera vez. Para estos, la terminología desempeña una función cognitiva y discursiva importante para la transmisión del contenido especializado y la construcción de un texto cohesionado. Tras varios estudios anteriores (Santamaría 2021, 2022; Domènech, Estopà, Santamaría 2022; Domènech, Santamaría 2023) se percibe la necesidad de investigar sobre recursos terminológicos destinados a estudiantes, que los asistan en el desarrollo de habilidades cognitivas y comunicativas propias de su disciplina como, por ejemplo, la comprensión

1 Proyecto financiado por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Comunidad Valenciana en la Convocatoria AICO 2021, dirigido por M. Isabel Santamaría Pérez y Carmen Marimón Llorca [REF: CIAICO/2021/074]. Para más información, se puede visitar su página web: <https://www.neotermed.org/>

y producción de textos especializados.

- Por otro, los y las pacientes que inician un proceso largo de tratamientos y técnicas denominadas con unidades léxicas a las que acceden también por vez primera. Queremos medir el grado de legibilidad de la información a la que acceden estos destinatarios a través de páginas web de asociaciones y clínicas con el fin de educar y alfabetizar en salud para lograr el empoderamiento, la inclusión y la igualdad de este grupo, que viven una situación de estrés emocional, personal, y muchas veces acompañada de presión social y un importante esfuerzo económico.

Teniendo en cuenta estos dos grupos de destinatarios hemos elaborado dos cuestionarios, uno para cada uno de los receptores con el fin de conocer cómo perciben y acceden al conocimiento especializado a través de la terminología, lo que nos permitirá desarrollar pautas y desplegar mejores prácticas de comunicación basadas en evidencias, e implementar intervenciones para mitigar impactos e impulsar el bienestar de cada grupo. Por un lado, se ha elaborado un cuestionario para los y las estudiantes del campo de la Biomedicina con el fin de conocer cómo acceden al conocimiento de su especialidad a través de la terminología y si son capaces de percibir la variación terminológica, lo que les ayudará a entender mejor los textos de su ámbito, pero también a producir textos más adecuados, coherentes y cohesionados. Por otro lado, nos interesa el grado de cognición de pacientes que acceden a las técnicas de RA por primera vez y su nivel de comprensión de la terminología presente en los textos. Los resultados obtenidos servirán como base para el desarrollo y la implementación de dos aplicaciones prácticas:

- Una aplicación terminográfica en línea y de acceso libre sobre los términos de la Reproducción Asistida, adecuada al nivel cognitivo de los estudiantes universitarios en español-inglés. Esta aplicación incluirá la variación terminológica, relaciones conceptuales entre los términos, además de la información básica como la categoría gramatical, las definiciones, ejemplo de uso y la traducción al inglés.
- Un conjunto de infografías con los términos más frecuentes en español, inglés, francés, italiano y valenciano para facilitar el acceso al conocimiento de la materia con herramientas de consulta avaladas científicamente para el grupo de pacientes.

Para poder realizar ambas aplicaciones es necesario conocer la terminología *in vivo*, es decir, en su hábitat natural en una comunicación especializada, a través

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

del análisis de textos especializados, tal y como la usan los especialistas en la materia. Partimos de los principios de la Teoría Comunicativa de la Terminología (Cabré 1999) en la que una unidad léxica no es término *per se*, sino que adquiere ese valor especializado según su uso en un contexto comunicativo y según la finalidad comunicativa que se persiga, lo que es muy relevante para "etiquetar un neologismo como general o especializado" (Estopà 2022: 151). Lo que para un especialista no es nuevo, porque son términos conocidos desde hace tiempo, para alguien neófito en la materia sí lo sería. Un especialista en RA conoce términos como *fecundación in vitro, gestación subrogada* o *seminograma*, frente a un público general que no los conoce o no con la misma profundidad semántico-cognitiva. Desde esta concepción consideramos relevante partir de textos diferentes para establecer el vaciado terminológico, pues no precisan la misma terminología los aprendices de una especialidad que las y los pacientes que se acercan a ella por necesidad. Así, para la consecución de nuestro objetivo constituiremos un corpus textual conformado por dos subcorpus de textos con diferente nivel de especialización y con distintas funciones comunicativas, en los cuales nos centraremos en este trabajo para hablar de su diseño, compilación y características de cada uno.

2. El contexto: la Reproducción Asistida

La elección del campo biomédico de la RA no es aleatoria. Estamos ante un campo médico bastante reciente[2], en constante y vertiginosa evolución −las técnicas de Reproducción Asistida dan comienzo en los años setenta del siglo XX−, y genera nuevos términos constantemente. Actualmente, la infertilidad no se concibe como una enfermedad, sino como la "incapacidad para conseguir un recién nacido" (Sociedad Española de Infertilidad-SEF 2023). Como factores que causan infertilidad, la SEF señala aquellos que limitan de forma absoluta la capacidad reproductiva, como la ausencia de espermatozoides, junto a otros que disminuyen la probabilidad de embarazo espontáneo, como la endometriosis, los ovarios políquísticos, entre muchos otros (SEF 2011: 21). El éxito de las técnicas de Reproducción Asistida (TRA), entendidas como "conjunto de técnicas médicas

2 Este campo es bastante nuevo dentro del ámbito médico si tenemos en cuenta que el nacimiento del primer bebé probeta tiene lugar en 1978 en el Reino Unido. En España, el primer nacimiento por fecundación *in vitro* fue en 1984. Desde la década de los noventa las investigaciones sobre técnicas de reproducción asistida para solucionar problemas de fertilidad han ido en constante crecimiento, especialmente a partir de la Ley 14/2006, de 26 de mayo, sobre técnicas de reproducción humana asistida.

que favorecen la fecundación en caso de impedimentos fisiológicos del varón o de la mujer" (DLE) tiene que ver tanto con la solución de estas alteraciones como con factores sociales y culturales, por la modificación del concepto tradicional de familia y de patrones patriarcales de género. Esto es, el desarrollo de las TRA está íntimamente relacionado con los cambios en la sociedad, sobre todo por la incorporación de la mujer al ámbito laboral, las uniones tardías de las parejas, la competitividad laboral, la aparición de familias monoparenterales, parejas homosexuales, segundos matrimonios, etc. Asimismo, la sensibilización que se hace a la población de los éxitos en las TRA hace que las parejas retrasen la búsqueda de embarazo pensando que se van a quedar embarazadas fácilmente, y las conciben como una solución que supera todos los límites para la procreación (López, Moreno 2015: 244). Todas estas circunstancias actuales explican que las TRA hayan aumentado entre el 5 y el 10% en los países desarrollados.

Se trata, además, de un ámbito de interés científico y social y con gran impacto económico. De acuerdo con los datos obtenidos (https://www.reproduccionasistida.org/ y SEF) España es uno de los principales receptores de lo que se conoce como *turismo reproductivo* debido a una avanzada legislación y la mejora de la calidad de los tratamientos junto con la gran cantidad de profesionales formados y los años de experiencia. Muchas de las parejas que se someten a técnicas de reproducción asistida fuera de su país escogen España como destino: Francia es el primer país de procedencia con un 37,5% de pacientes, seguido de Italia con un 22% y Reino Unido, con un 8,3% (SEF 2021). De ahí que sea necesario incorporar estas lenguas (francés, italiano e inglés) en los productos de transferencia que se elaborarán como resultado aplicado del estudio.

Con respecto a la situación en la Comunidad Valenciana, si nos fijamos en las clínicas de fertilidad existentes actualmente en España (más de 280 repartidas principalmente en Cataluña, Madrid, Andalucía y Comunidad Valenciana), un importante número de ellas se encuentran en nuestra Comunidad: en la provincia de Alicante, 26 y en Valencia, 23. Disponemos de un elevado número de clínicas de fertilidad, con lo que el impacto económico y social para nuestra Comunidad es evidente y se hace necesario fomentar y divulgar este aspecto de la salud que tiene claras repercusiones en el desarrollo de nuestra región. De ahí que será importante también la recopilación, registro, difusión e implantación de estas unidades terminológicas en el marco general de la política de normalización del valenciano y su codificación.

Partiendo de este contexto en el que se conectan aspectos sociales, económicos, científicos y lingüísticos, el proyecto *NEOTERMED* se plantea como ob-

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

jetivo general el análisis multidimensional de la estructura conceptual y la red terminológica del discurso biomédico teniendo en cuenta quién es el destinatario. Nos interesa conocer cómo a través de la lengua se transmite la información y se realizará un análisis de este tipo de discurso desde el punto de vista comunicativo y lingüístico (pragmático-semántico, léxico, sintáctico, ortotipográfico).

En cuanto a la terminología de la RA, se han seguido tres procesos de estandarización internacional. El primer glosario fue resultado de un congreso celebrado en 2002 a iniciativa del Committee for Monitoring Assisted Reproductive Technology (ICMART) y se publicó con el título *The ICMART Glossary on ART Terminology* en 2006; posteriormente, en 2008, la OMS promueve una consulta terminológica en colaboración con el ICMART con el objetivo de mejorar y ampliar las definiciones de manera que se pudieran armonizar los datos internacionales, publicándose una primera revisión del *Glossary* en 2009, que se aumentó hasta 87 términos y que se tradujo a varias lenguas, entre otras el español (Zegers-Hochschild *et al*. 2017: 1787-88). El *Glossary* fue publicado simultáneamente, tanto en 2006 como en 2009, en las revistas *Human Reproduction* and *Fertility and Sterility* por Zegers-Hochschild. La traducción al español y el portugués fue realizada por la Red Latinoamericana de Reproducción y se presentó en 2010 en la revista *Jornal Brasileiro de Reproduçao Assistida*. Una segunda revisión, titulada *The International Glossary on Infertility and Fertility Care*, publicada en 2017 (283 términos), muestra la evolución conceptual y el progreso que ha experimentado este ámbito médico, tanto con relación a las tecnologías en sí mismas, como a los aspectos biológicos, médicos, sociológicos, jurídicos. Así, por ejemplo, en esta segunda y última revisión del glosario se han eliminado términos como *concepción* y sus derivados al no poder ser descritos biológicamente en el proceso de reproducción. (Zegers-Hochschild *et al.* 2017: 1790). Finalmente, la normalización acometida concierne a nombres, lo que, como ha indicado Cabré (2003: 41-42), se debe a la función específica de las unidades terminológicas de categoría nominal: denominar conceptos de especialidad. Para nuestro estudio se tendrán en cuenta diccionarios generales y especializados que recopilen esta terminología de la RA, además del *Glosario de Fertilidad Humana*[3], trabajo colaborativo entre

3 Este glosario es el primer resultado de una Red de Innovación Educativa entre estudiantes del Máster de Inglés/Español para Fines Específicos y el Máster de Fertilidad Humana de la Universidad de Alicante. Para su elaboración se ha contado con una ayuda del Vicerrectorado de Calidad e Innovación Educativa para el desarrollo de acciones de innovación educativa. Se trata de un glosario abierto que se irá implementando con más entradas por el alumnado de ambos másteres en nuevas ediciones de los mismos y que puede consultarse en el siguiente enlace: <https://terms.iulma.ua.es/es/glosario-de-fertilidad-humana>

estudiantes del Máster de Inglés/Español para Fines Específicos y del Máster de Fertilidad Humana de la Universidad de Alicante, que constituye el punto de partida de nuestro recurso terminológico.

El interés y la novedad de nuestro proyecto reside en la no existencia de ningún estudio de carácter multidimensional (dimensión lingüístico-terminológica, sociolingüística y pragmático-discursiva) aplicado a un campo de especialidad como la RA ni a los grupos de destinatarios elegidos y de los que nos ocuparemos. Tampoco encontramos estudios sobre cómo el sector de la población, no experta, comprende y asimila los términos de este campo haciendo hincapié en la perspectiva cognitiva y cómo se refleja esta en la información multimodal para el paciente. Otro de los aspectos novedosos del proyecto es la incorporación de otras lenguas además del español y del inglés como son el italiano, el francés y el valenciano, debido a factores socio-económicos de peso en este campo de la RA dentro de la Comunidad Valenciana.

3. Lingüística de corpus

El campo de la lingüística de corpus, que puede ser caracterizada como "rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus, esto es, muestras reales de uso de la lengua" (Martín Peris 2008: 335), ha transformado la comprensión y el análisis del lenguaje en la actualidad. Esto ha favorecido un incremento significativo de investigaciones tanto cualitativas como cuantitativas que abordan una amplia gama de aspectos del lenguaje (Goźdź-Roszkowski 2021). Esta transformación está relacionada con la rápida evolución de la comunicación web o *Web communication* (Collins 2019) y la constante evolución de los géneros digitales (Combe 2022), basados en el uso de internet como fuente de datos para la construcción de corpus lingüísticos. La intersección de esos campos es altamente productiva, caracterizada por diversas hipótesis, objetivos y enfoques. Este planteamiento resalta la estrecha relación entre la lingüística de corpus y la abundancia de información disponible en internet, así como la facilidad para acceder a ella.

Como es sabido, las características de cada corpus están vinculadas a los objetivos de su construcción o diseño, lo que tiene como resultado una variedad de tipos de corpus. En resumen, podemos identificar corpus generales o dialectales; sincrónicos o diacrónicos; orales o escritos; generales o especializados; monolingües o multilingües, y corpus codificados y anotados, por mencionar algunos

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

ejemplos. Estos conjuntos de textos sirven como fundamento de la lingüística de corpus. Además, una vez un corpus está terminado, se convierte en una herramienta esencial que puede ser explorada desde nuevas perspectivas, como el Procesamiento de Lenguaje Natural (PLN) y las nuevas tecnologías aplicadas a la lingüística. Estos corpus nos permiten examinar cómo se estructura el lenguaje en contextos reales, poner a prueba suposiciones lingüísticas, comprender los diversos significados que las unidades léxicas, ya sean especializadas o generales, adquieren en su contexto, e identificar las combinaciones terminológicas más representativas asociadas a un término específico.

En este contexto, abordamos el diseño y construcción del corpus representativo de RA desde un enfoque metodológico basado en la lingüística de corpus. Para lograrlo, tomamos decisiones previas para establecer las especificaciones, así como los criterios de diseño, los cuales se detallarán a continuación.

4. Consideraciones metodológicas

Tal y como hemos indicado, el objetivo del proyecto *NEOTERMED* es el análisis multidimensional de la estructura conceptual y la red terminológica del discurso biomédico de la Reproducción Asistida, teniendo en cuenta quién es el destinatario. Atendiendo a ese objetivo general, se han seleccionado los dos grupos de receptores, ya delimitados, y a partir de sus necesidades y teniendo en cuenta el supuesto de adecuación desde la aproximación teórica que asumimos, ha sido necesaria la constitución de un corpus textual compuesto por dos subcorpus: uno, de textos divulgativos y otro de textos especializados.

Consideramos que este corpus *ad hoc*, materia prima del proyecto, del que dependen en gran medida los resultados de los recursos terminológicos que vamos a elaborar, debe ser configurado detenidamente, y en ello nos detendremos a continuación.

4.1. *Las especificaciones y los criterios de diseño*

En esta sección se expondrán las especificaciones, así como los criterios de diseño empleados para crear nuestro corpus digital de Reproducción Asistida; por ende, el diseño de un corpus puede concebirse desde distintos puntos de vista. Sin embargo, nuestra aproximación a la construcción del corpus nace de su funcio-

nalidad (Acebes de la Arada 2018). De aquí se deriva la importancia que debe atribuirse a parámetros comunicativos a la hora de construir un corpus con un fin específico. Además, contamos con una sólida propuesta de investigación para su constitución, tal y como señala Reppen (2022: 13): "having a clearly articulated research question is an essential first step in corpus construction, since this will guide the design of the corpus".

En la búsqueda y selección de los textos que formarán parte del corpus, el cual se analizará lingüísticamente y explotará *a posteriori,* es necesario considerar dos tipos de criterios lingüísticos: los internos y los externos (Vargas-Sierra 2005). Por un lado, los criterios internos se refieren a factores puramente lingüísticos del texto y, por otro lado, los externos tienen que ver con cuestiones extralingüísticas como el tema o la autoría. A partir de estas premisas nos propusimos tener en cuenta los siguientes aspectos antes de delimitar las propiedades de nuestros textos (Atkins *et al.* 1992; Pearson 1998; Bowker, Pearson 2002; Sinclair 2004): relevancia (textos directamente relacionados con la Reproducción Asistida); diversidad (tipología textual con contenido especializado); representatividad (textos que representen un amplio rango de conocimiento dentro de la Reproducción Asistida) y calidad (textos escritos por expertos y especialistas que presentan precisión, claridad y fiabilidad).

Se han completado estos aspectos con los de Bowker y Pearson (2002: 54), basados en el tamaño, el medio, la temática, la tipología textual, la autoría, la fecha de publicación y las lenguas. Para algunos de los criterios, encontramos información idéntica tanto en el subcorpus divulgativo como en el subcorpus especializado. Dichos criterios son: *el medio*, todos los textos deben estar escritos y en formato digital[4]; la *temática* principal, que es la reproducción asistida, si bien contamos con descriptores temáticos (Vargas-Sierra 2008) o subtemas que fueron creados a partir de un árbol de campo para cada subcorpus, los cuales se detallan en la Tabla 1 y la Tabla 2; la *autoría*, todos los textos deben estar escritos por expertos y especialistas de la salud; la *fecha de publicación,* delimitamos los años de producción de los textos desde 2013 hasta junio de 2023 (hablamos de corpus sincrónico), aunque algunos de los textos del subcorpus divulgativo no tienen fecha de producción; y la *lengua,* todos los textos aparecen en una sola lengua, el español peninsular. De este modo, estamos ante un corpus monolingüe.

4 Todos los textos se encuentran en formato electrónico ya que encontramos todos los textos en internet. Sin embargo, uno de los manuales que componen el corpus especializado, el *Manual de Intervención Psicológica en Reproducción Asistida,* fue digitalizado con OCR por la Unidad de Digitalización de la Biblioteca Virtual Miguel de Cervantes en la Universidad de Alicante.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

En cuanto a las diferencias de selección de criterios entre un subcorpus y otro tenemos, en primer lugar, el *tamaño.* Para ambos subcorpus, fijamos inicialmente 1.000.000 de palabras. Actualmente contamos con 1.500.000 de palabras o 1.676.206 tokens[5] agrupados en 1370 muestras[6], para el subcorpus divulgativo, y 2.000.000 de palabras o 2.047.892 tokens concentrados en 175 muestras, para el subcorpus especializado. Para la *tipología textual*, los textos son expositivos, argumentativos e informativos en el subcorpus divulgativo y expositivos, argumentativos, informativos y didáctico-instructivos en el subcorpus especializado. Unido a esto, veremos las Figuras 2 y 4 con los diferentes géneros textuales para cada subcorpus.

A continuación, se presenta la Tabla 1, la cual muestra la plantilla que contiene los subtemas o descriptores temáticos para el subcorpus divulgativo. La tabla consta de tres columnas: la primera con los descriptores temáticos, los cuales suman un total de 10; la columna central con la palabra principal de cada descriptor, mientras que la tercera columna enumera los conceptos que abarca cada descriptor. Estos descriptores temáticos ayudan a trazar límites claros entre los contenidos que abarcan los textos, aunque en muchos casos se superpongan. Por ejemplo, si un texto aborda la donación de ovocitos como técnica de reproducción asistida, se incluirá en el descriptor *Técnicas/Tratamientos de Reproducción Asistida.* Si el texto también comenta los pasos a seguir para una persona donante de ovocitos, se incluirá en el descriptor *Donación de gametos o embriones*. El solapamiento es totalmente normal porque en los textos pueden aparecer diferentes subtemas. El subtema o descriptor temático con mayor número de muestras textuales es Técnicas/Tratamientos de Reproducción Asistida, con un total de 796 muestras textuales, mientras que el subtema con menos muestras textuales es Parto o posparto, con 6 muestras textuales.

Descriptor Temático/Subtema	Palabra del descriptor/subtema	Conceptos que abarca
Regulación jurídica española sobre la Reproducción Asistida	Regulación	Leyes sobre técnicas de RA en España, leyes sobre la donación de gametos o embriones en España: Ley 14/2006, de 26 de mayo, sobre técnicas de reproducción humana asistida.

Técnicas/Tratamientos de Reproducción Asistida	Técnicas/Tratamientos	Técnicas, tratamientos, pruebas, test, o métodos relacionados con la reproducción asistida, fecundación artificial o que indirectamente ayudan a que ocurra un embarazo.
Sexualidad con función reproductiva	Sexualidad	Sexualidad con carácter reproductivo, salud sexual y reproductiva, relaciones sexuales para concebir, tener sexo para quedar embarazada.
Donación de gametos o embriones	Donación	Donantes de óvulos, donantes de espermatozoides, donantes de gametos, donantes de embriones, proceso para donar gametos, receptores de gametos, pareja receptora de gametos.
Fertilidad humana	Fertilidad	Fertilidad femenina, fertilidad masculina, ovulación, días fértiles, regla, reproducción y fertilidad.
Enfermedades, trastornos o problemas relacionados con la fertilidad humana	Problemas	Infertilidad, esterilidad, enfermedades que causan infertilidad, problemas que causan infertilidad, trastornos que causan infertilidad, tratamientos para la infertilidad, consejos o hábitos para combatir la infertilidad.
Parto o posparto	Parto	Parto, lactancia, cambios físicos después del parto, fertilidad después del parto, emociones tras el parto, bebé, parto de riesgo.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

Aborto, aborto espontáneo o aborto recurrente	Aborto	Aborto inducido, aborto de repetición, aborto tras IA o FIV, gestación no evolutiva, fallo de implantación, tratamientos y abortos, aborto tardío, problemas en el embarazo, emociones y aborto.
Embarazo	Embarazo	Embarazo natural, problemas en el embarazo, problemas para conseguir el embarazo, estudio del feto o embrión, estudio del ADN, betaespera, consejos para el embarazo, estudio genético.
Duelo o emociones ante la infertilidad o la no llegada del embarazo	Procesos emocionales	Estado de ánimo e infertilidad, duelo en la betaespera, apoyo psicológico y problemas de fertilidad, culpabilidad e infertilidad, estrés e infertilidad.

Tabla 1: Plantilla descriptores temáticos subcorpus divulgativo

Relacionados con la tipología textual, encontramos los géneros textuales dentro del subcorpus divulgativo. Estos géneros textuales han sido extraídos de diversas fuentes en línea, como páginas web de clínicas de reproducción asistida, páginas web de asociaciones de reproducción asistida, blogs de reproducción asistida, etc. Hasta llegar a la cifra de 89 portales diferentes. Esta información se presenta en la Figura 1. En este conjunto de géneros textuales heterogéneo, destaca el artículo blog, con un total de 652 muestras textuales, y el artículo web, con 502 muestras. Este hecho parece evidente ya que nuestro punto de partida era contar con estos tipos de textos por ser los más consultados por los y las pacientes (Santamaría 2023). Asimismo, en este tipo de portales encontramos la comunicación asimétrica, experto-no experto (Faya 2016).

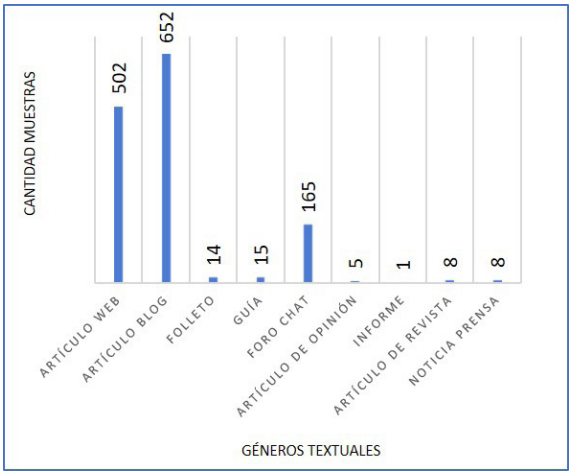

Figura 1: Gráfico de barras con la cantidad de muestras por géneros

Ahora pasamos a ilustrar los datos relacionados con el subcorpus especializado en cuanto a subtemas y géneros textuales. Primeramente, se detallan los datos sobre los 6 descriptores temáticos en la Tabla 2. Por ejemplo, uno de los descriptores temáticos para el subcorpus especializado es Técnicas/Tratamientos de Reproducción Asistida y procesos asociados, donde las palabras principales *son Técnicas/ Tratamientos* y abarca conceptos como *técnica TUNEL*, *fecundación in vitro*, *donación de gametos*, entre otros. En este caso, también puede haber superposición entre los subtemas que abarcan los textos. El subtema que más muestras textuales tiene es el de *Técnicas/ Tratamientos y procesos asociados*, con 100 muestras y el que menos es el de *Duelo o emociones (psicología) ante la infertilidad o la no llegada del embarazo* con 13 muestras.

Descriptor Temático/Subtema	**Palabra del descriptor/subtema**	**Conceptos que abarca**
Términos o procesos relacionados con la genética.	Genética	Genética, mutación, genes, genotipo, fenotipo, epigenética, ADN, cromosoma, diagnóstico genético, etc.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

Técnicas/Tratamientos de Reproducción Asistida y procesos asociados	Técnicas/Tratamientos	Técnicas, tratamientos, servicios, pruebas, test, o métodos relacionados con la reproducción asistida, fecundación artificial o que indirectamente ayudan a que ocurra un embarazo. Fecundación *in vitro*, donación de gametos/embriones, inseminación, swim up, técnica TUNEL…
Términos o procesos relacionados con la anatomía, fisiología o embriología	Anatomía, fisiología y embriología	Ovulación, ovarios, saco gestacional, vagina, útero, vesícula seminal, trompas de Falopio, menopausia, menstruación, aborto, cigoto, cuello uterino, hormonas, etc.
Enfermedades, trastornos o problemas relacionados con la fertilidad humana	Problemas	Infertilidad, esterilidad, enfermedades que causan infertilidad, problemas que causan infertilidad, trastornos que causan infertilidad, tratamientos para la infertilidad, consejos o hábitos para combatir la infertilidad. Cuidados de enfermería ante la infertilidad.
Duelo o emociones (psicología) ante la infertilidad o la no llegada del embarazo	Procesos emocionales	Estado de ánimo e infertilidad, apoyo psicológico y problemas de fertilidad, culpabilidad e infertilidad, estrés e infertilidad, intervención psicológica.
Términos o procesos relacionados con la biología celular (células humanas)	Biología celular	Óvulo, ovocito, gametos, flagelo espermático, espermatogénesis, citoesqueleto, blastocisto, acrosoma, etc.

Tabla 2: Plantilla subtemas corpus especializado

Del mismo modo que recogemos los datos sobre géneros textuales en un gráfico de barras para el subcorpus divulgativo, lo hacemos para el especializado en la Figura 2. Como era de esperar, no contamos con los mismos géneros textuales. En este caso los géneros textuales son: artículo científico, revista científica, cuaderno, guía, trabajo final de grado, trabajo final de máster, tesis doctoral, manual, artículo de revisión, carta al editor y caso clínico. Según se observa en la Figura 2, los géneros textuales prototípicos son el artículo científico con 61 muestras textuales, el TFG con 34 muestras textuales y la tesis doctoral con 29 muestras textuales. Estas muestras textuales se obtienen de fuentes como repositorios institucionales de Universidades españolas, revistas científicas o manuales, que en total suman 29 fuentes.

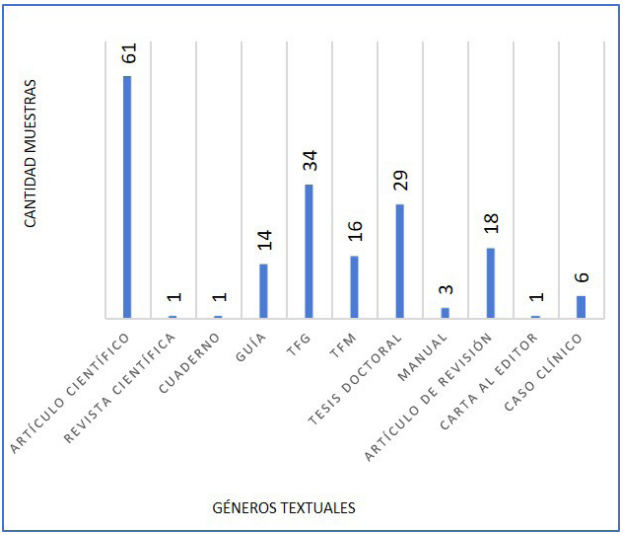

Figura 2: Gráfico de barras con la cantidad de muestras por géneros

Una vez descritos los parámetros que delimitan la preparación de nuestro corpus, veremos en el siguiente apartado las diferentes fases para construir cada subcorpus de reproducción asistida.

4.2. *Fase de compilación del corpus*

Para poder constituir el corpus, hay que tener en cuenta las diferentes etapas por las que pasan los textos, desde su descarga hasta su compilación final. En esta tarea seguimos la designación que propone Sánchez Ramos (2017) como *fase de compilación*.

Consideramos la fase de compilación común para cada subcorpus, aunque hay diferencias en algunas etapas. Plasmamos la información en la Tabla 3, para facilitar la comparación entre los subcorpus y visualizar las diferencias. Entre las diferencias destacan la *descarga de los documentos,* la etiqueta de *codificación* y la *limpieza 2ª parte*.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

Fases Subcorpus Divulgativo	Fases Subcorpus Especializado
1º Criterios y búsqueda y descarga de los documentos.	**1º** Criterios de búsqueda y descarga de los documentos. Formato original: PDF.
2º Almacenamiento (Navarro 2015): formato simple (.txt).	**2º** Almacenamiento (Navarro, 2015). Conversión a texto plano (.txt) .
3º Codificación bajo la misma etiqueta**:** *RAD00001, RAD00002, RAD00003…*	**3º** Codificación bajo la misma etiqueta: *RAE00001, RAE00002, RAE00003…*
4º Limpieza 1ª parte automática: eliminación de caracteres extraños, saltos de línea, HTML, etc. con *Tiny Tools by Jim Collective*.	**4º** Limpieza 1ª parte automática: eliminación de caracteres extraños, saltos de línea, HTML, etc. con *Tiny Tools by Jim Collective.*
5º Limpieza 2ª parte manual: eliminación de índices, bibliografía, símbolos, tablas con tasas de éxito, contacto.	**5º** Limpieza 2ª parte manual: eliminación de índices, bibliografía, símbolos, anexos, agradecimientos, tablas y figuras con cifras.
6º Creación del corpus y compilación en *Sketch Engine.*	**6º** Creación del corpus y compilación en *Sketch Engine.*
7º Base de Metadatos: información sobre el texto en hoja de cálculo. *Nombre codificado, tema, subtema, fuente, número de palabras, nivel de especialización, destinatarios, géneros, recursos audiovisuales, palabras clave y URL.*	**7º** Base de Metadatos: información sobre el texto en hoja de cálculo. *Nombre codificado, tema, subtema, fuente, número de palabras, nivel de especialización, destinatarios, géneros, información gráfica, palabras clave y URL.*

Tabla 3: Fase de compilación del corpus de *NEOTERMED*

El primer paso es la recuperación de la información, entendida como el "conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado". Buscamos la información directamente en internet y aplicamos los criterios para seleccionar contenidos web propuestos por Martínez (2016). Como buscador generalista de los textos divulgativos usamos *Google,* y para los textos especializados, usamos buscadores como *Google y Google Scholar* y bases de datos documentales, tales como *Dialnet*, *ScienceDirect* y repositorios institucionales de Universidades españolas, como *RUA* (Repositorio Institucional de la Universidad de Alicante).

Tras validar la información y seleccionar los documentos o textos que queremos, procedemos a realizar la descarga, dentro de la primera etapa de la compilación. Para el subcorpus divulgativo, descargamos los textos en texto plano en su mayoría, aunque algunos documentos están en formato HTML y PDF, frente al

subcorpus especializado, donde todos los archivos se descargan en formato original, es decir, como documento PDF.

En la segunda fase, almacenamos los textos en formato simple. Esto implica que debemos convertirlos a formato simple, al ser este un formato que podría considerarse estándar para utilizar en programas y herramientas de gestión de corpus. Por este motivo, para el subcorpus divulgativo sólo convertimos algunos documentos mientras que para el subcorpus especializado convertimos todos los documentos.

Posteriormente, guardamos los documentos bajo la misma codificación. Para el subcorpus divulgativo, la etiqueta es *RAD00000*, donde *RA* designa reproducción asistida, *D* divulgativo y los 00000, la posición que van teniendo conforme se van renombrando. Para el subcorpus especializado, cambiamos la *D* por la *E*, que se refiere a especializado, por ejemplo, *RAE00010.*

Una vez codificados los textos, nuestra siguiente tarea fue limpiar los textos automáticamente para reducir el ruido con la herramienta en línea *Tiny Tools by Jim Collection*, donde se eliminan caracteres extraños, saltos de línea, URL dentro del texto. Esta fase es común para ambos subcorpus. Sin embargo, en la siguiente limpieza, a la que denominamos *limpieza 2ª parte,* limpiamos las muestras textuales de forma manual y, en eso, sí hay diferencias. El propósito de realizar esta segunda limpieza es eliminar todas aquellas partes que acompañan al texto pero que no forman parte del contenido textual y, por tanto, tendríamos muestras textuales contaminadas. Los apartados comunes que se eliminan en ambos subcorpus son: índices, bibliografía y símbolos. Para el subcorpus divulgativo también limpiamos tablas con tasas de éxito y la información de contacto y para el subcorpus especializado suprimimos anexos, agradecimientos y tablas con cifras y figuras. Es muy importante que esta segunda limpieza se realice de manera minuciosa para que el texto quede lo más limpio posible en cuanto a contenido textual. En la Figura 3, podemos observar un fragmento de la muestra RAD00014 una vez se ha depurado:

```
El Endometrio: ¿Influye en mi fertilidad?
El endometrio es el tejido, o capa mucosa, que recubre la parte interior del útero.
El endometrio irá cambiando de grosor a lo largo del ciclo menstrual preparándose para la implantación del embrión y dando lugar al inicio de la gestación.
Esas variaciones del endometrio se deben a la acción de las hormonas sexuales: estrógenos y progesterona.
Si no se produce la fecundación del óvulo o la implantación del embrión, el endometrio se desprende y se inicia lo que conocemos como regla o menstruación.
En ese momento empieza un nuevo ciclo menstrual.
¿Qué es la Receptividad Endometrial del Endometrio?
La Receptividad Endometrial es el período de tiempo óptimo que tiene el endometrio para acoger al embrión.
Una vez que se produce la implantación y el embrión anida en la pared del útero, esa unión da lugar al desarrollo de la placenta, a la creación del saco gestacional y, finalmente, al desarrollo de un embarazo viable.
Para lograr el éxito de un tratamiento de fertilidad necesitamos seleccionar los embriones con mayor potencial de implantación, aquellos que son cromosómicamente normales y morfológicamente de buena calidad.
```

Figura 3: Ejemplo de fragmento de la muestra textual RAD00014

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

Cuando las muestras textuales ya están depuradas, pasamos a crear cada subcorpus, en primer lugar, el divulgativo, donde se crea y se compila en el gestor de corpus *Sketch Engine* y, en segundo lugar, el especializado, donde se sigue el mismo proceso.

Por último, creamos una base con los metadatos de cada texto para poder agrupar los datos más relevantes de los subcorpus. Esta base se crea en una hoja de cálculo, en nuestro caso, en EXCEL, y en total tenemos dos bases de datos, una para cada subcorpus. Los campos de las bases son comunes para cada subcorpus a pesar de que un campo se renombra. Este es el caso del campo *recursos audiovisuales*, para el subcorpus divulgativo, que pasa a ser *información gráfica*, en el subcorpus especializado. Para el resto de campos, introducimos la información en base a la siguiente explicación:

- *Origen:* se componen por cada una las fuentes de origen de cada uno de los textos, siendo el nombre de la clínica, el nombre de la institución, etc.
- *Número de palabras:* número total de palabras una vez el texto está limpio.
- *Nivel de especialización:* representación del corpus en cuanto a la naturaleza de los textos ya que estos son el reflejo de una situación comunicativa concreta, es decir, del nivel de especialización. Aun así, los textos, dentro de su etiqueta divulgativa o especializada, pueden tener diferentes grados de especialización. Es por este motivo que hemos creado una escala 1, 2 y 3 donde etiquetamos cada uno de los textos con un nivel de *divulgación* o *especialización* donde 1 corresponde a totalmente divulgativo (foros, blogs, folletos, artículos de opinión, etc.) o totalmente especializado (artículos científicos, revistas científicas, etc.); 2 corresponde a parcialmente divulgativo (clínicas, asociaciones, etc.) o parcialmente especializado (manuales, portales web, tesis doctorales, etc.); 3 corresponde a poco divulgativo (guías, informes, etc.) o poco especializado (carta el editor, guía, etc.).
- *Destinatarios:* este campo corresponde a etiquetar los textos según su destinatario. Por tanto, tenemos al público general, los y las pacientes interesados en leer y comprender este contenido y los y las estudiantes universitarios. Además, si encontramos que en algunos de los textos se hace referencia al lector y lo tratan como "mujer" u "hombre", también lo anotamos.
- *Géneros:* géneros textuales recogidos en las Figuras 1 y 2.
- *Palabras clave:* palabras o frases breves que remiten a los contenidos más importantes de los textos. En su mayoría, se han extraído gracias a que aparecen acompañando al texto. Para el resto de textos que no aparecen, se extraen manualmente.
- *URL*: link de acceso al texto.

A continuación, en la Tabla 4, podemos ver un ejemplo real de los campos para el subcorpus divulgativo con la muestra textual *RAD00001* y otro ejemplo para el subcorpus especializado con la muestra textual *RAD00010:*

Campos Metadatos	Subcorpus Divulgativo	Subcorpus Especializado
Nombre de Codificación	RAD00001	RAE00010
Tema	Reproducción Asistida	Reproducción Asistida
Subtema	Donación	Técnicas/Tratamientos Anatomía, fisiología y embriología
Origen	Sociedad Española de Fertilidad	RUA
Nº de palabras	1752	47907
Nivel de especialización	Divulgativo (1)	Especializado (2)
Destinatarios	Pacientes Lego	Profesional de la salud
Géneros	Folleto	Tesis doctoral
Recursos audiovisuales/ Información gráfica	No	Sí
Palabras clave	Donación, legislación, gametos	Reproducción asistida, Fecundación in vitro, Biomarcadores espermáticos, Fosforilación de proteínas, Fragmentación del DNA, Reacción acrosómica, ICSI, Cultivo embrionario
Link	https://cnrha.sanidad.gob.es/registros/pdf/SIRHA_DONANTES.pdf	*http://rua.ua.es/dspace/handle/10045/69569*

Tabla 4: Ejemplo con los metadatos para los subcorpus

5. Resultados

Una vez que hemos compilado nuestros dos subcorpus representativos de la temática de la RA, decidimos analizar la representatividad global de nuestro corpus; esto es, conocer si ya contábamos con dos subcorpus que abarcasen de manera significativa esa parte real y especializada del lenguaje, así como el tipo de muestras textuales que componían cada subcorpus en términos de cantidad de palabras y su distribución en las muestras textuales. Para ello, hicimos dos estudios cuantitativos que veremos detallados en los siguientes epígrafes.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

5.1. *Representatividad del corpus*

Como comentamos previamente en la descripción del corpus, consideramos, primeramente, las propiedades de los textos–calidad, representatividad, relevancia y diversidad–, para establecer una aproximación a los tipos de textos que constituyen nuestro corpus *ad hoc*. Sin embargo, en cuanto al parámetro de representatividad, determinar una cantidad exacta de textos, palabras o *tokens* que constituirán el corpus previamente es complejo. Unido a esto, varios autores han estudiado el concepto de representatividad. Por ejemplo, Biber puntualiza que la representatividad es "the extent to which a sample includes the full range of variability in a population" (1993: 243). Precisamente, sugiere que se pueden representar casi todos los elementos de un registro con apenas pocos ejemplos, unas mil palabras, y un número reducido de textos, específicamente diez.

Dentro de este paradigma, en el que se han hecho aproximaciones para medir la representatividad *a priori*, sin resultados objetivos ni concretos, nace la herramienta *ReCor* (Corpas, Seghiri 2007), el cual estudia el tamaño mínimo de un corpus o colección textual para que pueda considerarse representativo en términos cuantitativos. De ahí que nos planteáramos usar *ReCor*, ya que decidimos estudiar si ambos subcorpus podrían aportar novedades significativas si seguíamos aumentándolos. Además, nos llamó la atención la sencillez de su interfaz ya que no mostraba al usuario el algoritmo matemático ni la interfaz de línea de comandos (CLI), algo que también describen sus creadoras, Corpas y Seghiri (2007:167).

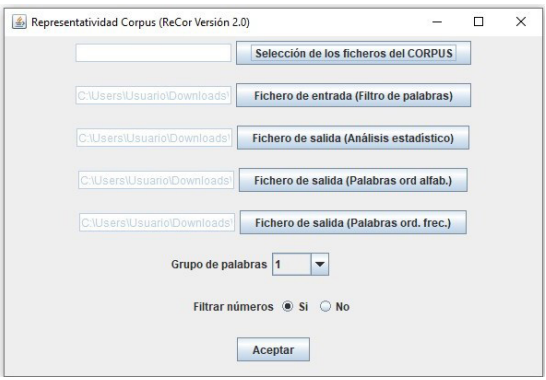

Figura 4: Interfaz *ReCor*

A grandes rasgos, ReCor usa el algoritmo N-Cor, que permite ilustrar gráficamente el punto en el que un corpus o una colección textual empieza a ser representativo en términos cuantitativos. El método del algoritmo N-Cor, de acuerdo con Corpas y Seghiri, "calcula el tamaño mínimo de un corpus mediante el análisis de la densidad léxica (d) en relación a los aumentos incrementales del corpus (C) documento a documento" (2007:166). Una vez realizada la prueba, se generan las representaciones gráficas y ficheros de salida en .txt con datos estadísticos sobre las palabras y types/tokens del corpus.

Para nuestro corpus *ad hoc*, hicimos una prueba con la versión 2.0 de *ReCor*. En primer lugar, estudiamos la colección de muestras textuales divulgativas. A continuación, puede verse la Figura 5 que arroja los resultados:

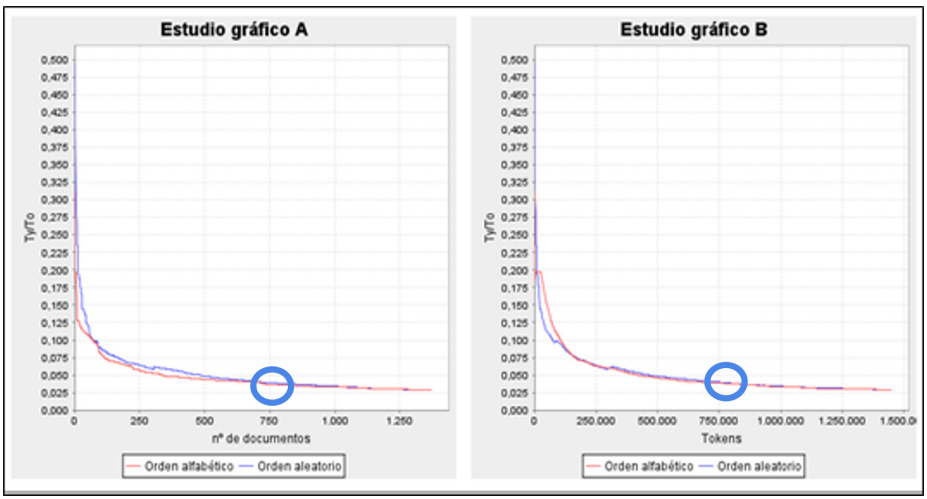

Figura 5: Gráficos con los resultados de *ReCor* para las muestras divulgativas

La representación gráfica, a partir de las dos líneas que corresponden a documentos incluidos alfabéticamente, línea roja, y aleatoriamente, línea azul, que se unen a medida que se acercan al valor cero, muestra el tamaño mínimo del corpus especializado para ser considerada representativa. Por tanto, el subcorpus divulgativo es representativo a partir de 750 documentos y 750.000 palabras. Mientras que para el subcorpus especializado, el número de documento mínimo para poder ser representativo es 130 y 1 800 000 palabras, según la Figura 6:

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

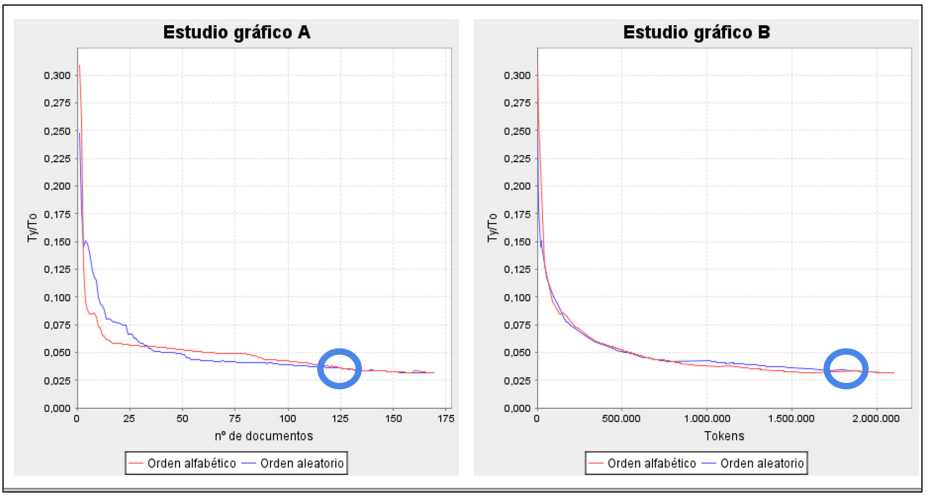

Figura 6: Gráficos con los resultados de *ReCor* para las muestras especializadas

De este modo, se comprueba que el corpus ya es verdaderamente representativo para este campo de especialidad en cuestión. Además, se evidencia que la inclusión de nuevos textos apenas incorporará novedades significativas al corpus, consolidando así su validez y exhaustividad en términos de contenido.

5.2. *Estadística descriptiva aplicada a la lingüística de corpus*

La lingüística de corpus se enfoca en el estudio del lenguaje a través del análisis de su uso real. Esto implica la necesidad de realizar un análisis cuantitativo, lo que nos llevó a utilizar la estadística descriptiva en nuestro corpus. Mediante este enfoque, podemos comparar muestras textuales y obtener conclusiones sobre la distribución de las palabras, proporcionando así una comprensión más profunda de nuestras muestras textuales.

Por lo tanto, dentro de la estadística descriptiva, en cuanto a las medidas de tendencia central, calculamos la media y la mediana y para las medidas de dispersión, calculamos el rango y la desviación típica. Antes de mostrar los resultados, es importante mencionar que, para el subcorpus divulgativo, contamos con 1370 muestras textuales concentradas en 1.456.239 palabras. Por tanto, para los cálculos se usaron todas las muestras. Para el subcorpus especializado, contamos con

175 muestras textuales, agrupadas en 2.001.280 palabras. Esta información ya nos informa de que tenemos muestras textuales mayores en cuanto a cantidad de palabras en el subcorpus especializado.

Resultados Subcorpus Divulgativo		Resultados Subcorpus Especializado	
Nº documentos	**Palabras**	**Nº documentos**	**Palabras**
1370	1.456.239	175	2.001.280
Media	1.093,77	Media	1.2083,96
Mediana	566	Mediana	5624
Rango	47.514	Rango	97.731
σ	2958	σ	16.402

Tabla 5: Resultados estadística descriptiva

Como es de esperar, los valores para el subcorpus especializado son mayores, ya que la cantidad de palabras por muestra es mayor y las muestras textuales están más concentradas. La extensión de palabras entre una muestra y otra varía más para el subcorpus especializado que para el subcorpus divulgativo. Esto se puede apreciar claramente con los valores de la desviación típica (σ), ya que para el subcorpus divulgativo tenemos una desviación de 2958, esto quiere decir que en promedio las muestras textuales están a una distancia de 2958 palabras arriba o abajo, mientras que en el especializado hay 16.402 palabras. Por tanto, tenemos mucha más separación de palabras entre las muestras textuales para el subcorpus especializado. Esto puede verse ejemplificado de la siguiente manera: si cogemos tres muestras textuales aleatorias del subcorpus especializado, *RAE00008* (19 358 palabras), *RAE000020* (608 palabras) y *RAE00039* (8709 palabras), observamos prontamente la distancia que existe de una muestra a otra.

5.3. *Aplicaciones del corpus*

En este apartado queremos destacar las amplias posibilidades que brinda el uso de un corpus digital especializado en investigaciones lingüísticas. Este tipo de corpus se presenta como una valiosa herramienta para el análisis y estudio del lenguaje, permitiendo un acceso rápido y eficiente a una gran cantidad de datos lingüísticos auténticos y representativos.

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

En primer lugar, la creación de recursos terminológicos a partir de las extracciones terminológicas es una de las ventajas más destacadas. El corpus digital especializado facilita la identificación y recopilación de términos clave en un dominio específico, lo que resulta fundamental para la comprensión y el avance en áreas técnicas y científicas. Es por eso que gracias a la herramienta *Sketch Engine* hemos podido hacer una extracción terminológica semiautomática a partir de nuestro subcorpus especializado para la creación de un recurso terminológico de Reproducción Asistida. En la actualidad, este recurso terminológico se encuentra en proceso de creación, ya que debemos recordar que hay diferentes etapas para su creación, desde la realización de cuestionarios para dichos estudiantes para poder conocer sus necesidades y perfilar la ficha terminológica hasta la creación de la base terminológica.

En segundo lugar, el uso de la inteligencia artificial en el análisis de este tipo de corpus representa un avance significativo en el campo de la lingüística. La aplicación de técnicas de procesamiento del lenguaje natural y aprendizaje automático permite el desarrollo de algoritmos y modelos capaces de identificar patrones, tendencias y regularidades en los datos lingüísticos. En nuestro caso, gracias a sistemas de aprendizaje automático, estamos creando un modelo de entrenamiento con WEKA para estudiar la clasificación del nivel de especialización del corpus. Próximamente, mostraremos el desarrollo y los resultados.

6. Conclusiones

En este trabajo se ha presentado el proyecto de investigación *NEOTERMED*, el cual se enfoca en la variación y el análisis multidimensional del discurso biomédico en el contexto de la neología y la terminología en ciencias de la salud, concretamente en el área de la Reproducción asistida. Desde el contexto en que surge el proyecto, hemos observado la necesidad de crear recursos atendiendo a dos grupos de receptores: a) los y las estudiantes universitarios del campo de la Biomedicina y b) los y las pacientes que buscan tratamientos y técnicas de reproducción asistida.

Hemos descrito el corpus representativo que hemos diseñado y construido en base a su finalidad. Además, nos hemos aproximado a la metodología empleada para crear un corpus *ad hoc* a partir de dos subcorpus: uno de textos divulgativos y otro de textos especializados. Del mismo modo, hemos observado que existen diferencias notorias en los procesos de diseño y compilación para ambos subcorpus, a pesar de que se constatan también muchas semejanzas.

Con respecto a los resultados del corpus, concluimos lo siguiente. Por un lado, el corpus es representativo para la finalidad que tiene, y esto es algo que hemos comprobado *a posteriori* con la aplicación informática *ReCor.* Por otro lado, gracias al análisis cuantitativo, aplicando la estadística descriptiva, observamos que se necesita un número mayor de muestras textuales del subcorpus divulgativo para poder tener un corpus equilibrado con el especializado y que, en el subcorpus especializado, las muestras textuales tienen un número mayor de palabras y hay una separación mayor de palabras entre muestra y muestra.

Creemos que el uso de un corpus digital, en este caso especializado en investigaciones lingüísticas, ofrece una amplia gama de oportunidades. Desde la creación de recursos terminológicos hasta la aplicación de la inteligencia artificial, lo que brinda resultados fundamentales para el avance en el conocimiento del lenguaje y su uso en el ámbito médico. Esto es algo que hemos estudiado a partir de la elaboración de un recurso terminológico de reproducción asistida para estudiantes universitarios todavía en proceso de creación, y la aplicación de la inteligencia artificial a través del aprendizaje automático, para el estudio de la terminología en las muestras textuales.

Bibliografía citada

Acebes de la Arada, Desireé (2018), "Diseño y compilación de un corpus especializado en fisioterapia", *Léxico y cultura en LE/L2: corpus y diccionarios*, eds. María Bargalló; Esther Forgas; Antoni Nomdedeu. Tarragona, Rull: 15-23.

Atkins, Sue; Clear, Jeremy; Ostler, Nicholas (1992), "Corpus Design Criteria", *Literary and Linguistic Computing*, 7/1: 1-16.

Barona, Josep Lluís (2004), "Hacer ciencia de la salud: los diagnósticos y el conocimiento científico de las enfermedades", *Panace@. Revista de Medicina, Lenguaje y Traducción*, 15: 37-44.

Biber, Douglas (1993), "Representativeness in Corpus Design", *Literary and Linguistic Computing*. 8/4: 243-57.

Bowker, Lynne; Pearson, Jennifer (2002), *Working with Specialized Language: A practical guide to using corpora*, London, Routledge.

Cabré, Teresa (1993), *La terminología: teoría, metodología y aplicaciones,* Barcelona, Empúries.

Cabré, Teresa (1999), *La terminología*: *representación y comunicación*, Barcelona, Institut

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X

universitari de lingüística aplicada.

Cabré, Teresa (2003), "El lenguaje científico desde la terminología", *Aproximaciones al lenguaje de la ciencia*, ed. B. M. Rodríguez Rodillo. Madrid, Fundación Instituto Castellano y Leonés de la Lengua: 19-52.

Cabré, Teresa; Estopà, Rosa (2002), "El conocimiento especializado y sus unidades de representación: diversidad cognitiva", *Sendebar. Revista de la Facultad de Traducción e Interpretación*, 13: 141-53.

Collins, Luke (2019), *Corpus Linguistics for Online Communication: A Guide for Research*, CRC Press. https://www.crcpress.com/Corpus-Linguistics-for-Online-Communication-A-Guide-for-Research/Collins/p/book/9781138718968

Combe, Christelle (2022), "Alfabetización digital, géneros digitales y enseñanza a distancia", *Tecnología versus/para el aprendizaje de lenguas,* ed. Fernando Trujillo. Barcelona, Difusión: 41-46.

Corpas, Gloria; Seghiri, Miriam (2007), "Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor", *Procesamiento del lenguaje natural*, 39: 165-72.

Cremades, Francesc (2003), "Medicina: canvi social i canvi lingüístic", *Llengua, societat i ensenyament*, ed. Vicent Martines. Edició electrónica Espagràfic, vol. II: 125-60 [20/08/2023] https://rua.ua.es/dspace/bitstream/10045/90988/2/Llengua-societat-i-ensenyament_vol-II.pdf

Domènech, Ona; Estopà, Rosa; Santamaría Pérez, Isabel (2002), "La terminología de la reproducción asistida en los diccionarios", *La terminologia, espejo de la evolución del conocimiento científico: El caso de la reproducción asistida*, eds. Rosa Estopà; Mercè Lorente. Barcelona, Institut de Lingüística Aplicada de la Universitat Pompeu Fabra y Documenta Universitària: 137-56.

Domènech, Ona; Santamaría Pérez, Isabel (2023), "La evolución de la terminología sobre reproducción asistida en los diccionarios y corpus de lengua española", *Cultura, Lenguaje y Representación, 30:* 57-78.

Estopà, Rosa (2012), "Medicina i Llenguatge: les paraules de la salut", *Llengua, Societat i Comunicació*, 10: 1-5.

Estopà, Rosa (2022), "Neología general y especializada", *La neología del español. Del uso al diccionario*, eds. Elisenca Bernal; Judit Freixà; Sergi Torner. Madrid, Iberoamericana: 151-74.

Estopà, Rosa (coord.) (2019), *Comunicación, lenguaje y salud*, Barcelona, Institut de Lingüística Aplicada de la Universitat Pompeu Fabra & Documenta Universitària, Sèrie Activitats 25.

Estopà, Rosa; Lorente, Mercè (eds.) (2022), *La terminologia, espejo de la evolución del conocimiento científico. El caso de la reproducción asistida*, Barcelona, Institut de Lingüística Aplicada de la Universitat Pompeu Fabra & Documenta Universitària, Sèrie Monografies 15.

Faya, Goretti (2016), "Propuesta de tipología textual para el campo médico", *Revista*

Española de Lingüística Aplicada, 29/1: 64-87.

Goźdź-Roszkowski, Stanislaw (2021), "Corpus Linguistics in Legal Discourse", *International Journal for the Semiotics Law*, 34: 1515–40.

Guardiola, Elena; Baños, Josep-Eladi (2011), "Eponímia mèdica. Els altres epònims", *Annals de Medicina,* 94: 130-32.

Guerrero Ramos, Gloria (2016), "Nuevas orientaciones en la percepción de los neologismos: neologismos de emisor y neologismos de receptor o neologismos de receptor", *La neología en las lenguas románicas. Recursos, estrategias y nuevas orientaciones*, eds. Joaquín García Palacios *et al.* Bern, Peter Lang: 57-68.

Guerrero Ramos, Gloria (2017), "Nuevas orientaciones de la terminología y de la neología en el ámbito de la semántica léxica", *RILCE. Revista de Filología Hispánica*, 33/3: 1385-1415.

López Gálvez, José Jesús; Moreno García, Juan Manuel (2015), "¿'Industria de la fertilidad' o respuesta a la búsqueda del hijo biológico?", *Treinta años de reproducción asistida en España: una mirada interdisciplinaria a un fenómeno global y actual*, eds. Pilar Benavente Moreda; Esther Farnós Amorós. *Boletín del Ministerio de Justicia*, LXIX/2179: 239-66.

Martín Peris, Ernesto (2008), *Diccionario de términos clave de ELE*, Madrid, SGEL.

Martínez, Luis Javier (2016), *Cómo buscar y usar información científica. Guía para estudiantes universitarios,* Santander, Universidad de Cantabria.

Mayor, Blanca (2008), *Cómo elaborar folletos de salud destinados a los pacientes*, Barcelona, Fundación Dr. Antonio Esteve.

Pearson, Jennifer (1998), *Terms in context*, Amsterdam, John Benjamins.

Reppen, Randi (2022), "Building a corpus: what are key considerations?", *The Routledge Handbook of Corpus Linguistics*, eds. A. O'Keefe; M. McCarthy. London, Routledge: 13-20.

Real Academia Española (2014), *Diccionario de la lengua española* (DLE) [05/05/2023] https://www.rae.es/

Sánchez Manzanares, Carmen; Santamaría Pérez, Isabel (2021), "Neology and terminology in health sciences. An approach to terminological metaphor in the discourse of Assisted Reproduction", *Metaphor in Economics and Specialised Discourse*, eds. José Mateo; Francisco Yus. Bern, Peter Lang: 321-60.

Sánchez Ramos, María del Mar (2017), "Compilación y análisis de un corpus *ad hoc* como herramienta de documentación electrónica en Traducción e Interpretación en los Servicios Públicos, (TISP)", *Estudios de Traducción*, 7: 177-90.

Santamaría Pérez, María Isabel (2021), "Español e inglés de la medicina: diseño e implementación de una experiencia docente en un contexto biosanitario", *Memorias del Programa de Redes-I3CE de calidad, innovación e investigación en docencia universitaria: Convocatoria 2020-2021*, eds. Asunción Menargues Marcilla; Rocío Díez Ros; Neus Pellín Buades; Rosana Satorre Cuerda. Alicante, Universidad de Alicante: 229-53.

Santamaría Pérez, María Isabel (2023), "Salud y comunicación: análisis lingüístico

de las páginas web sanitarias. El caso de la Reproducción Asistida", *La traducción en la encrucijada interdisciplinar. Temas actuales de traducción especializada, docencia, transcreación y terminología, eds.* Antonio Sánchez Fajardo; Chelo Vargas Sierra. Valencia, Tirant lo Blanch: 337-72.

Santamaría Pérez, María Isabel; Congost Maestre, Nereida (2021), "Español e inglés de la medicina: diseño e implementación de una experiencia docente en un contexto biosanitario", *Memorias del Programa de Redes-I3CE de calidad, innovación e investigación en docencia universitaria*, eds. Asunción Menargues Marcilla; Rocío Díez Ros; Neus Pellín Buades; Rosana Satorre Cuerda. Alicante, Universidad de Alicante: 233-57.

Santamaría Pérez, María Isabel; Congost Maestre, Nereida (2022), "Una experiencia educativa innovadora. El lingüista en la sociedad actual: diseño e implementación de una experiencia docente en un contexto biosanitario", *Actas I International Congress: Education and Knowledge,* eds. Vicent Martines Peres; Jordi M. Antolí Martínez; Rosabel Roig Vila. Barcelona, Octaedro: 258.

Santamaría Pérez, María Isabel; Congost Maestre, Nereida; Gómez Torres, María José *et al*. (2022), "El lingüista como mediador en la comunicación médico-paciente: experiencias en innovación educativa en un contexto profesional", *Memoria del Programa de Redes de Investigación en Docencia Universitaria*, eds. Rosana Satorre Cuerda; Asunción Menargues Marcillas; Rocío Díez Ros. Alicante, Universidad de Alicante: 451-66.

Santamaría Pérez, María Isabel; María José Gómez Torres (eds.) (2022), *Glosario de Reproducción Asistida (español-inglés)* [08/09/2023] <https://terms.iulma.ua.es/es/glosario-de-fertilidad-humana>

Sef (Sociedad Española de Fertilidad) (2018), *Saber más sobre fertilidad y reproducción asistida*, Madrid, SEF.

Sef (Sociedad Española de Fertilidad); Ministerio de Sanidad (2021), *Informe estadístico de Técnicas de Reproducción Asistida 2019*, Departamento de Estadística, 53-54.

Seghiri, Miriam (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.

Sinclair, John (2004), *Developing Linguistic Corpora: a Guide to Good Practice*, Tuscan Word Centre, AHDS.

Vargas Sierra, Chelo (2005), "A pragmatic model of text classification for the compilation of special-purpose corpora", *Thistles. A homage to Brian Hughes. Essays in Memoriam*, eds. José Mateo; Francisco Yus. Alicante, Universidad de Alicante, vol. II: 295-315.

Vargas Sierra, Chelo (2008), "La sistematización terminográfica: una propuesta metodológica para la elaboración de diccionarios traductológicos", *Actas del X Simposio Iberoamericano de Terminología*, Montevideo, Uruguay, 7-10 de noviembre

de 2006 [Recurso electrónico].
Zegers-Hochschild, Fernando *et al*. (2017), "The International Glossary on Infertility and Fertility Care, 2017", *Human Reproduction*, 32/9: 1786-1801.

Ovidia Martínez Sánchez se ha graduado en Estudios ingleses (2017-2021) por la Universidad de Alicante. Es Máster en Inglés y Español para Fines Específicos (2021-2022) de la misma Universidad, donde obtuvo el 2º Premio en su Trabajo Final de Máster en *Premios de Investigación para Trabajos Fin de Grado y Máster en Materia de Transparencia, Acceso a la Información Pública, Buen Gobierno, Datos Abiertos e Integridad Institucional*. Actualmente desempeña el cargo de Técnico Superior en el Instituto de Lenguas Modernas y Aplicadas, adscrita al proyecto de investigación *NEOTERMED* (CIAICO/2021/074). Es miembro del Grupo de Investigación del Español Profesional y Académico (EPA) y recientemente ha iniciado su tesis doctoral en el campo del lenguaje médico y procesamiento de lenguaje natural. Además, cursa su segundo máster en Traducción Médico-Sanitaria de la Universitat Jaume I.
ovidia.martinez@ua.es

Isabel Santamaría Pérez actualmente desarrolla su labor como docente e investigadora en el Departamento de Filología española de la Universidad de Alicante. Es directora del Grupo de Investigación del Español Profesional y Académico, directora de la Sede Universitaria de La Marina y forma parte del Instituto Interuniversitario de Lenguas Modernas y Aplicadas (IULMA) y el Instituto de Investigación de Estudios de Género (IUIEG). Ha participado en numerosos proyectos de investigación nacionales e internacionales financiados por diversas entidades públicas: algunos vigentes son el proyecto NEOTERMED (CIAICO/2021/074), donde es IP1 responsable, METAPRES-COLING (PID2019-107265GB-I00), donde es IP2 responsable, y DIGITENDER (TED2021-130040B-C21), donde es IP2 responsable. Es directora y coautora del *Diccionario del turrón* (español, catalán, inglés, ruso, árabe y chino) (2015), coautora del *Diccionario de neologismos del español actual* (Neoma) (2016) y directora del *Glosario de Reproducción asistida* (2023). Entre sus líneas de investigación destacan la lexicología, la lexicografía monolingüe y bilingüe, las lenguas de especialidad, la terminología y la neología.
mi.santamaria@ua.es

CUADERNOS AISPI 22 (2023): 197-226
ISSN 2283-981X