UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Análisis y Tipificación de Errores Lingüísticos para una Propuesta de Mejora de Informes Médicos en Español

D.ª Jésica López Hernández

2022

UNIVERSIDAD DE MURCIA

Facultad de Letras - Facultad de Informática

Tesis Doctoral

Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español

D.ª Jésica López Hernández

Directores:

Dra. Dña. Ángela Almela Sánchez-Lafuente
Dr. D. Fernando Molina Molina
Dr. D. Rafael Valencia García

2022

Esta tesis doctoral ha sido financiada por el Ministerio de Educación, Cultura y Deporte de España a través de las Ayudas para la formación de profesorado universitario (FPU), del Programa Estatal de Promoción del Talento y su Empleabilidad, con referencia FPU16/03324.

AGRADECIMIENTOS

Quiero aprovechar estas líneas para transmitir mi eterna gratitud a aquellas personas que han sido fundamentales durante el desarrollo de este proyecto.

En primer lugar, a los directores de esta tesis, Ángela Almela, Fernando Molina y Rafael Valencia. Gracias por vuestra inestimable ayuda, implicación y confianza durante estos años de investigación. También a Sonia Madrid, por su labor como tutora a lo largo de mi etapa investigadora.

A la empresa Vócali, por su colaboración y generosidad. Allí nació la idea de esta tesis, mientras descubría con ilusión el procesamiento del lenguaje natural y me acercaba por primera vez a los corpus de informes médicos.

A Ruslan Mitkov y al Research Group in Computational Linguistics de la Universidad de Wolverhampton, por darme la oportunidad de realizar una estancia de investigación, conocer otros métodos de trabajo y proyectos, y ampliar el campo de visión.

A los compañeros del laboratorio, por recibirme y tratarme siempre como a una más del equipo.

Por último, especialmente, a mi familia, a Javi y a mis amigos. No tengo palabras suficientes para agradeceros el apoyo incondicional durante estos años. Gracias por los ánimos, la paciencia y el cariño, soy muy afortunada de teneros en mi vida.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTADO DE ABREVIATURAS

ASALE: Asociación de Academias de la Lengua Española

CDSS: Clinical decision support system (sistema de apoyo a la decisión clínica)

CNN: Convolutional neural network (red neuronal convolucional)

DCI: Denominación común internacional

DLE: Diccionario de la lengua española

DL: Deep learning (aprendizaje profundo)

DPD: Diccionario panhispánico de dudas

DTM: Diccionario de términos médicos

IA: Inteligencia artificial

ISMP: Instituto para el Uso Seguro de los Medicamentos

ML: Machine learning (aprendizaje automático)

NLP: Natural language processing (procesamiento del lenguaje natural)

NER: Named entity recognition (reconocimiento de entidades nombradas)

OCR: Optical character recognition (reconocimiento óptico de caracteres)

OLE: Ortografía de la lengua española

OMS: Organización Mundial de la Salud

PLN: Procesamiento del lenguaje natural

POS: Part-of-speech (categoría gramatical)

RAE: Real Academia Española

RAMN: Real Academia Nacional de Medicina de España

SI: Sistema internacional

RESUMEN

El objetivo principal de esta investigación es la detección, análisis y clasificación de errores lingüísticos presentes en informes médicos en español.

Los sistemas de corrección automática más actuales y potentes, como las arquitecturas basadas en redes neuronales, requieren grandes conjuntos de datos de entrenamiento para un rendimiento óptimo. Por tanto, debido a la ausencia de corpus de dominio biomédico disponibles, en el procesamiento del lenguaje natural ha ganado importancia la recopilación y generación artificial de errores para el entrenamiento de los sistemas. El desarrollo de una tipología de errores a partir del estudio empírico de un corpus de informes médicos va a permitir añadir nuevos patrones a la generación de errores de forma más exhaustiva y, con ello, la creación de modelos más robustos para el procesamiento de datos en medicina.

Para la detección y clasificación de errores se ha analizado un corpus formado por informes médicos reales pertenecientes a cuatro especialidades (urgencias, UCI, cirugía general y psiquiatría), con más de dos millones de *tokens*. El enfoque metodológico desarrollado ha incluido distintas técnicas de detección y corrección automática, entre las que se encuentran la generación de un modelo lingüístico basado en n-gramas, la representación vectorial de las palabras del corpus a partir de *Word2Vec* y el etiquetado gramatical del corpus. Se ha desarrollado una herramienta de cómputo y clasificación de errores, y se ha realizado un análisis cuantitativo y cualitativo de los resultados obtenidos.

Los resultados han permitido identificar semejanzas y diferencias entre estas especialidades y han reflejado que la especialidad cuyos informes médicos presentan una mayor tasa de errores lingüísticos es urgencias. La mayoría de las palabras con errores están a distancia de edición 1 de la palabra correcta correspondiente, gran parte de los errores detectados se concentran en un número específico de caracteres y el tipo de error más cometido con una alta incidencia es el de omisión. Muchos de los errores presentan patrones de reproducción consistentes que es posible sistematizar, como la sustitución de caracteres con similitudes fonéticas, los errores provocados por desconocimiento de la norma ortográfica actual y los errores derivados del uso del teclado.

En síntesis, esta tesis doctoral pretende ser una contribución al estudio de errores lingüísticos en informes médicos para aportar una base de conocimiento lingüístico a los métodos de detección y corrección existentes para este dominio.

Palabras clave: detección automática de errores, análisis de errores, errores *real-word*, informes médicos, procesamiento del lenguaje natural

SUMMARY

The main purpose of this research is the detection, analysis and classification of linguistic errors in medical reports in Spanish.

The most current and powerful automatic correction systems, such as neural network-based architectures, require large training data sets for optimal performance. Therefore, artificial error collection and generation for training systems have gained importance in natural language processing, due to the scarcity of available biomedical domain corpora. The development of an error typology from the empirical study of a corpus of medical reports will make it possible to add new patterns to the generation of errors in a more exhaustive way and the creation of more robust models for data processing in medicine.

A corpus made up of real medical reports from four specialties (emergency medicine, ICU, general surgery and psychiatry), with more than two million tokens, has been analyzed for error detection and classification. The methodological approach developed has included different detection and automatic correction techniques, including the implementation of a linguistic model based on n-grams, the vector representation of the corpus words from Word2Vec and the grammatical labeling of the corpus. An error calculation and classification method has been developed, and a quantitative and qualitative analysis of the results obtained has been carried out.

The results have made it possible to identify similarities and differences between these specialties and have shown that the specialty with the highest rate of errors in medical reports is emergency medicine. Most of the erroneous words are within one edit distance of the corresponding correct word, and a large part of the errors detected are concentrated in a small number of characters and the most common type of error is omission. Many of the errors have consistent reproduction patterns that can be systematized, such as the substitution of characters with phonetic similarities, errors caused by ignorance of the current orthographic norm, and errors derived from the use of the keyboard.

To summarize, this doctoral thesis aims to be a contribution to the study of linguistic errors in medical reports in order to provide a base of linguistic knowledge to the existing detection and correction methods for this domain.

Keywords: automatic error detection, error analysis, real-word errors, medical reports, natural language processing

1. INTRODUCCIÓN

1.1. Justificación del estudio

La disciplina conocida como procesamiento del lenguaje natural[1] es una rama de la inteligencia artificial encargada de estudiar y desarrollar sistemas capaces de procesar el lenguaje humano. Está sumamente presente en nuestras vidas, aunque no seamos conscientes, a través de herramientas comunes que utilizamos a diario, como buscadores de información, traductores automáticos, asistentes virtuales, sistemas de GPS o correctores ortográficos subyacentes en aplicaciones de mensajería instantánea.

La finalidad esencial del procesamiento del lenguaje natural es la construcción de sistemas que sean capaces de analizar, comprender y extraer información expresada en lenguaje natural. El concepto de lenguaje natural alude a cualquier manifestación de una lengua que ha surgido de forma natural, a diferencia de lenguajes formales, como los lenguajes de programación, que han sido diseñados artificialmente para aplicaciones específicas. Por tanto, estas tecnologías buscan extraer conocimiento de datos no estructurados, mediante la construcción de recursos lingüísticos y el desarrollo de soluciones y técnicas para automatizar el tratamiento de la información contenida en textos.

Una de las principales áreas de investigación y de actuación del procesamiento del lenguaje natural es el dominio biomédico. La digitalización de los registros clínicos en el entorno sanitario ha conllevado un crecimiento exponencial de los datos en los últimos años. La documentación clínica contiene información sumamente valiosa para la investigación y la práctica sanitaria, por tanto, su tratamiento automatizado tiene un gran potencial que está siendo explorado. Resulta fundamental poder emplear tecnologías basadas en procesamiento automático de datos que permitan la extracción y clasificación de información clínica, la codificación de diagnósticos, la anonimización de documentos, la interoperabilidad semántica o el soporte a la decisión clínica, entre otras muchas posibilidades. Así, en los últimos años se están desarrollado múltiples proyectos para el

[1] Habitualmente se utilizan las siglas PLN y NLP (*natural language processing*). A lo largo de este trabajo en ocasiones se empleará esta forma abreviada.

español desde el sector de las tecnologías del lenguaje. No obstante, actualmente las infraestructuras lingüísticas y los recursos para el español continúan siendo limitados en comparación con los desarrollados para el inglés y existen dificultades para la explotación de datos lingüísticos, como la escasez de corpus públicos para entrenar los sistemas.

El lenguaje médico o biosanitario, al tratarse de un lenguaje científico, tiene como objetivo expresar con exactitud, precisión y claridad conceptos e ideas para facilitar el intercambio de información (Bello, 2016). No obstante, presenta particularidades lingüísticas que dificultan su procesamiento automático, como su riqueza y complejidad léxica, pues se trata de un lenguaje especializado sumamente fértil, resultado de veinticinco siglos de historia (Navarro, 2015). Además del empleo de terminología específica del dominio y de la elevada presencia de términos formados por derivación y composición, se caracteriza por el uso de mecanismos especiales de formación, como epónimos, abreviaturas o siglas (Gutiérrez, 2005). Por otro lado, los avances de la medicina se reflejan en la lengua como vehículo de comunicación e implican el surgimiento constante de neologismos. Todo ello da lugar a una falta de asentamiento de formas y a una variabilidad que provoca incorrecciones y dudas entre los especialistas.

Si nos centramos en los informes médicos, las dificultades aumentan, pues a las características inherentes anteriores hay que sumar un formato más desestructurado, la presencia de expresiones no gramaticales y un estilo telegráfico, la ausencia de puntuación y abundantes errores en la escritura (Meystre, 2006). Estos textos suelen contener errores lingüísticos debido a las restricciones de tiempo de los facultativos, que se ven obligados a redactarlos con rapidez y no suelen disponer de tiempo para atender a la forma o para llevar a cabo una revisión posterior.

Los errores dificultan el procesamiento informático de estos textos y generalmente es necesario un proceso de corrección previa. Son numerosos los trabajos (Ruch et al., 2003; Wong y Glance, 2011; Lai et al., 2015; entre otros) que constatan esta realidad y proponen métodos de corrección automática adaptados al dominio. Por tanto, la corrección automática se ha convertido en un eslabón fundamental para el procesamiento de datos en informes clínicos.

Es en esta realidad donde surgió la pregunta que motivó el comienzo de esta tesis doctoral, situada en la intersección entre la lingüística y la informática: ¿qué puede aportar la lingüística a los sistemas de detección y corrección automática de errores en el dominio biomédico?

Las técnicas más actuales de corrección, basadas en modelos de aprendizaje profundo (*deep learning*) requieren grandes conjuntos de datos para ser entrenados y obtener un rendimiento eficaz. Sin embargo, no siempre es posible contar con corpus disponibles y adecuados para esa labor, especialmente en el caso de los textos de dominios especializados. Esta escasez de corpus con errores disponibles influye en el rendimiento de las herramientas y es uno de los grandes desafíos técnicos. Por tanto, debido a la ausencia de corpus de dominio accesibles para el entrenamiento de modelos, en la producción actual para corrección automática ha ganado importancia la generación artificial de errores para el entrenamiento de los sistemas (Dziadek et al., 2017).

El método tradicional de generación artificial de errores se ha basado en reemplazar palabras por variantes ortográficamente correctas que están a distancia de edición cercana. Esos reemplazos de palabras se realizan de forma aleatoria mediante el empleo de operaciones de inserción, sustitución, omisión o transposición de caracteres, por lo que en muchas ocasiones no representan la realidad de forma exhaustiva y esta circunstancia repercute en el rendimiento de los sistemas. Por este motivo, autores como Davidson et al. (2020) consideran fundamental desarrollar métodos de generación de errores tras el análisis lingüístico de textos reales.

Hasta el momento no se ha realizado un análisis sistemático o una descripción exhaustiva sobre la naturaleza de los errores presentes en informes médicos en español. Además, tampoco existen corpus del dominio médico en español anotados con errores que puedan servir de base para generar ejemplos sintéticos de entrenamiento. Díaz (2005) defiende que los tipos de errores y sus distribuciones varían significativamente entre dominios y conjuntos de datos, por tanto, considera que el desarrollo de una tipología de errores universal no es posible, sino que depende de «los ejes en torno a los cuales se defina la clasificación propuesta» (Díaz, 2005: 409), de ahí que sea necesario contar con información específica del dominio.

En los informes médicos se observan patrones y estructuras que se repiten, lo que puede ayudar a sistematizar y aportar información para todo aquel que desee generar errores sintéticos para la corrección de este tipo de textos clínicos. Con los datos que se obtengan tras el estudio es posible ampliar el repertorio de reglas lingüísticas utilizadas para generar errores sintéticos específicos que se incluyan en el corpus de entrenamiento (Felice et al., 2014). La incorporación de nuevos tipos de errores no contemplados en el

sistema va a ayudar a la creación de conjuntos de datos de entrenamiento más exhaustivos, que incorporen casuísticas reales de lo que ocurre en el dominio.

Por consiguiente, el objetivo de esta investigación es la detección de errores lingüísticos presentes en informes médicos en español para su posterior análisis, estudio y clasificación. Al tratarse de textos clínicos, no es fácil contar con grandes cantidades de datos disponibles debido a las restricciones para acceder a este tipo de documentos, de ahí que sea especialmente relevante poder estudiar la tipología de errores en un corpus de estas características.

En suma, esta investigación surge con la intención de contribuir al conocimiento sobre los errores lingüísticos presentes en informes médicos, a través de un estudio exploratorio con carácter descriptivo, que pueda añadir otra capa de información a los métodos de detección y corrección automática disponibles y, por consiguiente, pueda contribuir a la mejora de corpus pertenecientes al dominio biomédico.

1.2. Estructura de la tesis

La tesis se estructura en dos partes claramente delimitadas: una primera fase dedicada a la presentación de la tesis doctoral y a la investigación teórica sobre el estado de la cuestión, y una segunda fase práctica, de carácter eminentemente descriptivo, centrada en el desarrollo metodológico y en el análisis de los resultados

En el primer capítulo, que sirve de preámbulo, se da a conocer la finalidad de la investigación y se define el marco de referencia en el que se inserta. Por un lado, se exponen de forma sucinta las razones que han motivado la realización de este trabajo y, por otro, se detalla la distribución de los distintos apartados que lo componen.

En el segundo capítulo se abordan los fundamentos teóricos y se documenta el estado de la cuestión en lo que respecta a los dos pilares que sustentan esta investigación: la corrección automática de errores y el lenguaje médico. La primera sección de este capítulo tiene una naturaleza preliminar, pues en él se introducen algunos conceptos básicos sobre procesamiento del lenguaje natural, como los distintos niveles de procesamiento, la distinción habitual entre procesamiento del lenguaje natural y lingüística computacional, o los principales recursos desarrollados que se encuentran disponibles para el español. A continuación, se ahonda en la detección y corrección

automática de errores, y se describen los principales enfoques y trabajos desarrollados en esta área de investigación.

Las páginas siguientes de este capítulo están centradas plenamente en el dominio médico. Por un lado, se analiza el lenguaje médico en español, prestando especial atención a las características que presentan los informes clínicos al ser los textos que componen el corpus de estudio. Por otro, se presenta el panorama actual del estado de las tecnologías del lenguaje en el ámbito de la salud, sus principales aplicaciones y los diferentes desafíos a los que se enfrenta. Finalmente, se profundiza en la corrección y detección automática de errores en este dominio y se realiza una revisión de los trabajos más destacables en el campo, identificando los tipos de corpus empleados, métodos y recursos.

En el tercer capítulo se explica la propuesta metodológica empleada y los experimentos desarrollados. Se define el objetivo principal de la tesis y se formulan los objetivos específicos para dar respuesta al problema de investigación planteado. En segundo lugar, se presenta el corpus objeto de estudio recopilado, se proponen los criterios de análisis que se van a tener en cuenta y las distintas convenciones en cuanto al tratamiento de los datos. En la sección dedicada al procedimiento se describen las distintas fases del enfoque metodológico llevado a cabo, que incluye el preprocesamiento del corpus, la detección y corrección de errores *non-word* y *real-word* respectivamente, y el cómputo y clasificación de los errores detectados.

El cuarto capítulo comprende la ejecución del análisis de datos a partir de los resultados obtenidos. Se realiza un análisis cuantitativo teniendo en cuenta la frecuencia, la distancia de edición, el tipo y subtipo de error, y la posición del error. Además, se efectúan pruebas estadísticas para comprobar si se producen diferencias significativas entre las especialidades analizadas. Por su parte, en el análisis cualitativo se realiza un desglose pormenorizado de los distintos tipos de patrones de error detectados, mencionando otros aspectos lingüísticos que pueden ser de utilidad para la finalidad del estudio.

La siguiente sección está destinada a la discusión de los resultados, donde se relacionan los contenidos teóricos previamente expuestos con los resultados obtenidos. Reflexionamos, además, sobre las diferencias detectadas a nivel de error entre el corpus de dominio médico y un corpus de español no especializado. Finalmente, el capítulo

culmina con la elaboración de una propuesta de módulo basado en conocimiento lingüístico a partir de los resultados obtenidos.

El quinto capítulo incluye las conclusiones obtenidas, así como las limitaciones y desafíos presentes en la investigación y, por último, las sugerencias de líneas de trabajo futuras que servirán para mejorar y ampliar los resultados.

El sexto capítulo incorpora las principales conclusiones de la investigación traducidas al inglés, lengua habitual de la comunicación científica, porque la tesis posee la Mención de Doctorado Internacional.

Finalmente, el séptimo capítulo presenta las principales aportaciones científicas derivadas de esta tesis doctoral, incluyendo los artículos de investigación, los capítulos de libro y las comunicaciones en congresos.

2. FUNDAMENTACIÓN TEÓRICA

2.1. Procesamiento del lenguaje natural

Antes de ahondar en la metodología y los experimentos propuestos es necesario conocer cuál es la situación actual en torno a los principales núcleos de actuación de esta investigación. En las páginas siguientes, se llevará a cabo un estudio sobre los sistemas de detección y corrección ortográfica. Presentaremos las características del lenguaje médico y, más concretamente, de los informes médicos. Posteriormente, analizaremos el estado del procesamiento del lenguaje natural en medicina y el estudio se centrará en la detección y corrección ortográfica en el lenguaje médico.

El procesamiento del lenguaje natural (PLN) es el campo interdisciplinar de la inteligencia artificial que tiene como finalidad proporcionar mecanismos para conseguir la comunicación entre humanos y sistemas (Névéol et al., 2018). Distintos algoritmos informáticos son capaces de extraer significado de entradas no estructuradas habladas o escritas, es decir, llevan a cabo la explotación de conjuntos de datos en lenguaje natural. Como mencionábamos en el capítulo de Introducción, por lenguaje natural se entiende el lenguaje hablado o escrito que es utilizado por las personas para comunicarse. Se diferencia, por tanto, de los conocidos como lenguajes formales, que se diseñan artificialmente y son empleados en ramas como la programación.

Otros términos que se emplean habitualmente para referirse al PLN son «lingüística computacional» y «tecnologías del lenguaje». Suelen remitir a la misma realidad, únicamente cambia la perspectiva de enfoque. El término PLN suele ser utilizado habitualmente en el área de ingeniería informática y tiene una connotación más aplicada y enfocada en las técnicas, el término «lingüística computacional» suele ser de uso más común por lingüistas y tiene un componente más descriptivo, mientras que el concepto de «tecnologías del lenguaje» es el habitual en entornos empresariales.

Existen distintos niveles de procesamiento del lenguaje. Estos se corresponden con los establecidos por la lingüística como sus áreas de estudio y pueden dividirse en análisis fonético-fonológico, morfosintáctico, léxico-semántico y pragmático (Bender, 2013).

El nivel fonético-fonológico analiza la estructura de los fonemas y atiende a la forma en la que se organizan los sonidos del habla de una lengua y sus características. En PLN se desarrollan herramientas de análisis científico del habla, sistemas de reconocimiento de voz y asistentes virtuales.

El nivel morfosintáctico estudia la estructura o formas de las palabras en un idioma y cómo se relacionan entre sí dentro de la oración a nivel sintáctico. En el análisis a nivel de palabra se lleva a cabo la tarea de tokenización, que comprende la división del texto en unidades de análisis o cadenas que normalmente se corresponden con palabras. La división del texto en sus *tokens* constituyentes suele ser uno de los primeros procesos en cualquier tarea de PLN. El procesamiento morfológico tiene en cuenta los distintos morfemas que componen las palabras y sus mecanismos de formación. También se utiliza la lematización, que consiste en asignar el lema o forma base a una palabra flexionada; y la identificación de la categoría gramatical o *POS-tagging*, cuyo fin es el etiquetado de partes del discurso de una palabra. Por su parte, el análisis sintáctico o *parsing* determina la estructura sintáctica de una oración y puede ser realizado mediante árboles de constituyentes o dependencias.

El nivel léxico-semántico se encarga de estudiar el conjunto de palabras de una lengua, el significado que poseen y las relaciones de sentido que se establecen entre ellas. El análisis semántico requiere, entre otras tareas, la desambiguación del sentido de la palabra, esto es, ser capaz automáticamente de asignar el sentido adecuado a una palabra polisémica según el contexto donde aparece. También es importante el reconocimiento de entidades nombradas, que consiste en la identificación de nombres, ubicaciones, organizaciones u otros conceptos relevantes dependiendo de la finalidad de la tarea.

El nivel pragmático explora la forma en la que las circunstancias contextuales de la comunicación influyen en el lenguaje. Analiza el lenguaje situándose en el plano del texto, en lugar de en el nivel de oración. Las tareas de PLN en este plano incluyen la detección de la negación, la ironía o el sarcasmo, el análisis de sentimientos o la resolución de la anáfora y la correferencia (Mitkov, 2005).

Entre las herramientas más populares para PLN en español destacan las librerías de NLTK[2] (*Natural Language Toolkit*) o Spacy[3], disponibles en Python. También

[2] <https://www.nltk.org/>

[3] <https://spacy.io/>

podemos mencionar FreeLing[4], que es una librería de C++, e Ixa pipes[5], una librería formada por un conjunto de herramientas desarrolladas en Java. Permiten la personalización de distintas funcionalidades según nuestro objetivo, e incluyen múltiples posibilidades para la explotación de datos lingüísticos, entre las que se encuentran la tokenización, la lematización, la segmentación de frases, el *POS-tagging*, el recuento de n-gramas, el reconocimiento y clasificación de entidades nombradas, el análisis sintáctico de dependencias, la generación de resúmenes automáticos o la creación de chatbots. Los componentes de cada paquete ofrecen rendimientos distintos dependiendo del corpus procesado y de la tarea específica.

2.1.1. Detección y corrección automática de errores

En esta sección se realiza una presentación sobre la detección y corrección automática de errores lingüísticos. Se describen las principales técnicas y recursos, y los correctores disponibles actualmente en el mercado. Se dedica una sección específica para investigar y profundizar en la problemática existente en torno a la detección y corrección de errores *real-word*, debido a los mayores desafíos que plantea.

La detección y corrección automática de errores fue una de las primeras tareas de procesamiento del lenguaje en comenzar a abordarse, debido a que es uno de los componentes clave de los sistemas de PLN. Los primeros sistemas surgieron en torno a 1960 y desde esa fecha se han ido desarrollando y perfeccionando (Kukich, 1992).

En primer lugar, es fundamental señalar que el proceso de corrección automática consta de tres fases: una primera fase de detección, una segunda fase de generación de candidatos de corrección y una tercera fase de clasificación de los candidatos para la corrección. La fase de detección consiste en descubrir si una palabra ha sido escrita de forma errónea o es incorrecta en ese contexto. Los correctores suelen usar diccionarios[6], que son grandes listados de palabras correctas, para detectar los errores. Si una palabra no es reconocida en el diccionario se considera incorrecta.

[4] <https://nlp.lsi.upc.edu/freeling/node/1>

[5] <https://ixa2.si.ehu.eus/ixa-pipes/>

[6] En la literatura consultada se utilizan los términos «diccionario», «lista de palabras» y «lexicón» de forma indistinta.

En el momento en el que una palabra es identificada como errónea entra en juego la fase de corrección, que se divide a su vez en la generación de correcciones candidatas y la clasificación de estas correcciones. Cuando se detecta el error, se necesita identificar la forma correcta de la palabra. Para ello, el sistema debe ser capaz de generar una lista de posibles sugerencias que puedan reemplazar la palabra mal escrita y elegir la adecuada entre ellas. Para la generación de ese listado de candidatos de corrección se hace uso de la llamada distancia de edición. Este concepto se refiere al número de operaciones que es necesario llevar a cabo para convertir una palabra o cadena de caracteres en otra. Habitualmente se conoce como distancia de Damerau-Levenshtein, en honor a dos de los primeros científicos que comenzaron a estudiarla (Damerau, 1964; Levenshtein, 1966). Las operaciones básicas de edición de errores pueden ser de cuatro tipos: inserción, omisión, sustitución y transposición. En la operación de inserción se agrega un carácter adicional a la palabra, en la operación de omisión se elimina un carácter de la palabra, en la de sustitución se utiliza un carácter equivocado en lugar del que corresponde, y en la de transposición se intercambia la posición de dos caracteres adyacentes. Por ejemplo, en la palabra «fármaco», un error de omisión sería «fámaco», un error de inserción daría lugar a «fárrmaco», un error de sustitución desembocaría en «fárnaco» y un error de transposición generaría la forma «fármcao». A diferencia de la distancia de Damerau-Levenshtein, la distancia clásica de Levenshtein (1966) considera el error de transposición o intercambio de caracteres como dos operaciones de edición en lugar de una.

Por tanto, la lista de sugerencias se produce a partir de la distancia de edición, se calcula el número mínimo y el tipo de operaciones que se van aplicando sobre la palabra errónea hasta obtener una coincidencia correcta con la palabra correspondiente en el diccionario. Damerau (1964) determinó en sus investigaciones que más del 80% de los errores suelen estar a distancia 1 de la palabra correcta, y que la distancia de edición rara vez excede dos operaciones.

Para poder elegir la sugerencia o candidato de corrección adecuado es necesario establecer métodos de ordenación o *ranking*, como estimaciones probabilísticas o medidas de similitud léxica, que limiten el número de candidatos y que asignen el primer lugar a la sugerencia correcta en ese contexto. Para conseguirlo es necesario hacer uso de distintas técnicas que ayuden a restringir el espacio de búsqueda y a tomar la decisión idónea para la corrección, las cuales veremos con más detalle en la siguiente sección.

Kukich (1992) realiza un exhaustivo recorrido sobre la corrección automática y afirma que los errores pueden clasificarse en dos importantes grupos: errores *non-word* y errores *real-word*. Cuando se comete un error *non-word* se obtiene un grupo de caracteres que no dan lugar a una palabra válida en esa lengua (por ejemplo, el error de inserción de 'n' da lugar a «transtorno», pero la palabra correcta es «trastorno»). Por su parte, un error *real-word* da origen de forma accidental a una palabra existente y correcta idiomáticamente, pero errónea en ese contexto, es decir, errónea en el plano semántico o sintáctico (por ejemplo, el error de omisión de 'd' provoca la aparición de «olor» en lugar de «dolor»). El primer caso es fácilmente identificable por un corrector común al comparar el resultado con el diccionario que incorpore, sin embargo, en el segundo aumenta en gran medida la complejidad de detección y requiere el análisis semántico del contexto circundante. Por tanto, en la primera fase de detección de un corrector automático, conocida como la fase de búsqueda en diccionario (y basada en la comparación del corpus con un listado de palabras validadas), los errores *real-word* pasan desapercibidos al ocasionar palabras ortográficamente correctas que están recopiladas en el lexicón previamente validado. Por consiguiente, son palabras que están enmascarando errores en el corpus y se necesitan otros niveles de análisis y técnicas basadas en contexto que permitan detectarlos.

En la clasificación y análisis de errores se pueden tener en cuenta otras dimensiones, como la causa del error, el tipo de error, la posición del error o el contexto en el que aparece (Rambell, 2000). Las tipologías de errores pueden presentar distinta granularidad según el fin para el que vayan a ser empleadas. Algunos autores distinguen entre errores de competencia o de actuación según sea la causa que provoca los errores (Naber, 2003). Los errores de competencia están relacionados con el desconocimiento de la norma ortográfica de la lengua y tienen una motivación cognitiva, mientras que los de actuación se consideran errores que han ocurrido de forma fortuita y están sujetos a factores no lingüísticos, como distracciones o cuestiones mecánicas (Díaz, 2005). Corder (1967) propuso la distinción *error* para el primer concepto y *mistake* para el segundo. La creación de tipologías de errores ha sido especialmente productiva en el campo de aprendizaje de lenguas y son numerosos los trabajos de análisis y clasificación de errores desarrollados a partir de corpus de aprendices de segundas lenguas (Nagata et al., 2017).

Kukich (1992) estableció una serie de hallazgos generales sobre patrones de errores que han permanecido en la literatura posterior. Entre ellos destaca que la mayoría

de los errores tienden a estar motivados por un único cambio de inserción, omisión, sustitución o transposición en la palabra, por consiguiente, la mayoría de los errores tienden a estar a distancia 1 de longitud de la palabra correcta. También constata que se producen pocos errores en la primera letra de una palabra y estos mismos resultados son proporcionados por Damerau (1964), Yannakoudakis y Fawthrop (1983), o Pollock y Zamora (1983), entre otros. Yannakoudakis y Fawthrop (1983) recopilaron 1377 errores en un corpus de más de 60 000 palabras y concluyeron que la mayoría de los errores siguen reglas específicas a partir de consideraciones fonológicas y ortográficas. Por ejemplo, establecieron que estos podían tipificarse a partir de diecisiete reglas heurísticas, entre las que se encuentran la omisión de la consonante 'h' en los conjuntos 'ch', 'ph' y 'rh', o la no duplicación de consonantes. También determinaron que las palabras incorrectamente escritas no suelen contener más de un error, tesis que defienden otros autores —como ya hemos mencionado—, entre los que se encuentran Pollock y Zamora (1983), que también hablan de un porcentaje de errores múltiples muy bajo, en torno al 6 % en una recopilación de 50 000 errores. Otros hallazgos que señalan son la mayor dificultad de corrección de errores fonéticos porque provocan una mayor distorsión de la cadena escrita y los fuertes efectos de adyacencia del teclado. Consideran que la mayor parte de los errores de escritura en el teclado están causados al presionar una tecla adyacente en el teclado o al presionar dos teclas que están juntas.

Estos hallazgos se han utilizado a lo largo de los años en la implementación de algoritmos de corrección para reducir el tiempo de búsqueda. Los resultados recopilados en estos trabajos a partir de corpus en inglés se han comprobado en otras lenguas, que presentan diferencias motivadas por la propia idiosincrasia de cada idioma a nivel fonético y ortográfico y que deben ser tenidas en cuenta. Han sido realizados estudios sobre patrones de error en lenguas como el francés (Ren y Perrault, 1992), el portugués (Gimenes et al., 2014), el español (Ramírez y López, 2006), el húngaro (Siklósi et al., 2016), el japonés (Baba y Suzuki, 2012), el danés (Paggio, 2000), el euskera (Aduriz et al. 1997), o el persa (Miangah, 2014). Autores como Gimenes et al. (2014), que realizan un estudio de errores para el portugués de Brasil, destacan que un porcentaje importante de los errores lingüísticos estarían relacionados con el uso de signos diacríticos o el uso de la cedilla, patrones que deberían tenerse en cuenta en los cálculos de probabilidad. A partir de las estadísticas presentadas en ese estudio, los autores agregan un nuevo módulo al procesador de textos OpenOffice Writer para reordenar la lista de sugerencias.

En cuanto a trabajos sobre tipologías de errores enfocadas a tareas de corrección automática para el español pueden destacarse dos estudios. El trabajo de Ramírez y López (2006) forma parte del desarrollo de un corrector para el español en Microsoft Corporation y en él se ofrece un análisis cuantitativo sobre patrones de errores en español. En él se refleja que los errores están estrechamente relacionados con las reglas ortográficas de cada idioma, se señala que el tipo de error con mayor presencia es el de omisión, que el 89 % de las palabras erróneas está a distancia 1 de la forma correcta, y que gran parte de los errores ortográficos (54,9 %) están relacionados con el uso incorrecto de los signos diacríticos, con especial presencia del error de omisión de tilde. Por su parte, en Díaz (2005) se aborda el tratamiento de errores gramaticales y de motivación cognitiva, y en él se defiende la relevancia de la creación de una tipología de errores para crear un corrector gramatical y de estilo. Profundizaremos en los resultados de estos trabajos en las secciones de Corrección automática de errores *real-word* (2.1.1.2.) y Discusión (4.3.).

En lo que respecta a correctores disponibles en código abierto, los más comunes son Aspell[7] y Hunspell[8], que poseen diccionarios para diferentes idiomas. Para el español encontramos también correctores ortográficos, gramaticales y de estilo, entre los que pueden mencionarse Stilus[9] o LanguageTool[10]. Sin embargo, son sistemas desarrollados para detectar errores en el dominio general, no están preparados para textos de dominios especializados como el sanitario, lo que afecta al rendimiento. La mayor parte de los correctores suelen ser sistemas comerciales cuyo uso no está disponible en abierto. Para el dominio médico se encuentran CorrectM[11] y Spellex Medicina[12]. El primero es un editor de texto plano con un corrector ortográfico que contiene terminología médica y el segundo es un corrector ortográfico que se integra en las herramientas de Microsoft Office. Por tanto, reconocen términos médicos y farmacológicos y proveen sugerencias para reemplazar aquellas palabras incorrectas. De igual forma, debemos mencionar un trabajo final de grado (Merino Torre, 2015) desarrollado en la Escuela Universitaria de

[7] <http://aspell.net/>

[8] <http://hunspell.github.io/>

[9] <http://www.mystilus.com/>

[10] <https://www.languagetool.org/>

[11] <https://www.cpimario.com/correctm.html>

[12] <https://tudiccionariomedico.com/>

Ingeniería Técnica Industrial de Bilbao, cuyo objetivo fue la implementación de un editor de texto plano con un corrector ortográfico para textos médicos. El autor defiende la necesidad de cubrir la mayor cantidad de palabras técnicas y se apoya en la creación de un modelo de lenguaje para suministrar posibles palabras correctas ante el error ortográfico detectado. No obstante, los tres funcionan con errores *non-word*, verificando la ortografía y reconociendo como válidos términos específicos que los correctores comunes no contienen, pero no detectan errores *real-word*, errores gramaticales, semánticos o de estilo.

2.1.1.1. Técnicas

Los métodos tradicionales de corrección de errores se han basado principalmente en la consulta de diccionarios y en la distancia de edición mínima entre un error y sus candidatos de corrección. Con el transcurso del tiempo a estos métodos se fueron incorporando nuevas técnicas, como las basadas en similitud fonética, en reglas y heurísticas; métodos basados en aprendizaje automático y técnicas probabilísticas, como el análisis de n-gramas y los modelos de lenguaje; o las más innovadoras basadas en aprendizaje profundo y redes neuronales. A continuación, se presentan de forma sucinta las técnicas más comunes empleadas en los distintos sistemas de corrección automática, aunque se profundizará en ellas en la sección dedicada a las técnicas para detectar errores *real-word*.

Los sistemas más tradicionales basados en reglas (Yannakoudakis y Fawthrop, 1983; Naber, 2003; Patrick et al., 2010) se apoyan en la coincidencia de cadenas y patrones, y en el uso de expresiones regulares y listas de palabras. El análisis de errores ha sido empleado para aportar información lingüística y estadística que ayude al diseño de técnicas basadas en reglas. Los algoritmos utilizan el conocimiento sobre patrones de errores mediante reglas para transformar estos en palabras correctas. Suelen tener alta precisión en conjuntos de datos específicos, pero dependen de la efectividad de otras herramientas y es necesario conocimiento especializado del idioma. Una de las principales desventajas es que no tienen en cuenta el contexto, por lo que no son adecuados para tipos de errores que implican una mayor complejidad. No obstante, sí es posible trascender el nivel de palabra a pequeña escala si se usan reglas que involucren varias palabras o que trabajen a nivel sintáctico con categorías gramaticales. Gran parte

de los productos de comprobación y corrección gramatical utilizan reglas lingüísticas, reglas de análisis sintáctico y de estructura de frases, como LanguageTool o Microsoft Office Word. Los enfoques tradicionales pueden implementar reglas para la detección de errores basadas en la similitud de palabras y calcular frecuencias de n-gramas para determinar la palabra correcta.

Asimismo, también pueden utilizarse técnicas basadas en similitud fonética. Al igual que existen métodos que se basan en la similitud léxica entre palabras o la distancia de edición, también encontramos técnicas que trabajan con la distancia de edición o similitud a nivel fonético, especialmente útil en palabras cuya pronunciación es similar pero su grafía es distinta, y que consiste en mapear a cada cadena de caracteres una clave. El algoritmo fonético genera cadenas de caracteres válidas que son fonéticamente similares al error ortográfico. El conjunto de cadenas generadas se verifica con la base de conocimiento y todas aquellas cadenas que son válidas pasan al algoritmo de clasificación. En este método debemos señalar el algoritmo fonético Soundex desarrollado por Odell y Russell, o sistemas mejorados como Metaphone y Double-Metaphone (Kilicoglu et al., 2015), que están incluidos en el corrector ortográfico Aspell.

En tercer lugar, un set o conjunto de confusión es un grupo de palabras que pueden confundirse entre sí y son almacenadas en una lista (Pedler, 2007; Pedler y Mitton, 2010). Las palabras pueden confundirse entre sí por similitud fonológica, gráfica o gramatical. Jurafsky y Martin (2014) señalan algunos ejemplos prototípicos como *peace/piece, among/between, affect/effect* o *there/their*. En este caso se aborda la tarea de corrección como un problema de resolución de ambigüedad y mediante el uso de sets de confusión se intenta guiar al algoritmo para predecir el candidato más probable en ese contexto, focalizando en errores determinados y reduciendo el tiempo en el proceso de clasificación (Rozovskaya y Roth, 2010). Durante el proceso de detección, si el sistema encuentra una palabra del conjunto de confusión, se utilizan distintas reglas y clasificadores supervisados para decidir si esa palabra es adecuada o lo es otra del conjunto. Esas reglas y clasificadores pueden basarse en diferentes características del texto, como la relación sintáctica o semántica, o la frecuencia de aparición.

Por otra parte, el método basado en contexto más habitual suele ser el modelo de lenguaje de n-gramas. Un modelo de lenguaje es una distribución de probabilidades que se aprenden de un corpus y permite obtener la probabilidad de aparición de secuencias de unidades lingüísticas, como bigramas o trigramas (Golding y Schabes, 1996; Wilcox-

O'Hearn, 2008; Samanta y Chaudhuri, 2013; Faili et al., 2016; Al-Jefri y Mahmoud, 2013). Se fundamenta en que las combinaciones o secuencias de palabras que son correctas reflejarán un valor de probabilidad mayor que las incorrectas. Por tanto, los enfoques basados en estadística dependen en gran medida del tamaño y características del corpus. No es necesario texto anotado y son versátiles, pero no pueden manejar bien las dependencias de largo alcance. También encontramos trabajos que combinan enfoques estadísticos, el uso de sets de confusión y sistemas basados en reglas (Faili et al., 2016). En los enfoques de corrección más vanguardistas se siguen utilizando ampliamente los modelos de lenguaje, especialmente para clasificar sugerencias de corrección, en combinación con técnicas más potentes a nivel computacional (Yuan et al., 2016).

Más recientemente, se han explorado enfoques basados en aprendizaje profundo. Las arquitecturas de redes neuronales llevan a cabo una representación vectorial de las palabras (*word embeddings*) según sus características. Esta representación de las palabras como objetos matemáticos logra capturar las propiedades semánticas y sintácticas de las palabras de manera que los sistemas sean capaces de interpretarlas. Por tanto, estos modelos de aprendizaje profundo aprenden a partir de grandes cantidades de texto mediante un proceso de entrenamiento y son capaces de predecir qué palabras encajan mejor en ese contexto calculando la similitud entre los vectores que representan cada palabra (Wu et al., 2020). Un ejemplo es la traducción automática neuronal seq2seq, cuyo objetivo inicial era la traducción de lenguas, pero que se ha convertido en uno de los métodos más probado en los últimos años para corrección automática. Su funcionamiento se basa en el uso de un codificador, que mapea una frase de una lengua a un formato vectorial intermedio, y un decodificador, que la decodifica a un segundo idioma (Beloki et al., 2020). En el caso de la corrección automática, se trata de una tarea monolingüe en la que se transforma el idioma de origen (texto con errores) en un nuevo idioma de destino o meta (texto sin errores). Asimismo, modelos como ELMo («Embeddings from Language Models») o BERT («Bidirectional Encoder Representations from Transformers») se consideran algunos de los codificadores bidireccionales preentrenados más avanzados en la actualidad para el modelado del lenguaje, y pueden ajustarse y entrenarse para tareas de corrección automática. Sin embargo, el mayor obstáculo de estos codificadores lingüísticos es la necesidad de contar con cantidades muy grandes de datos para el entrenamiento y la optimización de los parámetros del modelo. Como solución, se generan datos sintéticos introduciendo errores en las oraciones correctas (Grundkiewicz

et al., 2019, White y Rozovskaya, 2020). Para ello, se modifican los corpus y se introducen errores mediante sustitución léxica, el uso de la distancia de edición, de patrones de error o el uso de reglas inversas (Choe et al., 2019).

2.1.1.2. Corrección automática de errores *real-word*

Los errores *real-word* o errores sensibles al contexto dan lugar a palabras reales, que están verificadas en el diccionario, por lo que es necesario hacer uso de técnicas basadas en contexto para detectar estos errores que suelen pasar desapercibidos. Se han desarrollado numerosos métodos para abordar la corrección de los errores non-*word*, pero los errores dependientes del contexto aún siguen siendo un desafío para los investigadores. Kukich (1992), que estableció una de las jerarquías de tipos de errores para corrección automática más importantes, situó los errores *real-word* en el grado de mayor dificultad para ser procesados.

Los errores *real-word* afectan al plano sintáctico y semántico y las causas que los desencadenan pueden ser cognitivas, tipográficas o fonéticas, entre otras. Así, estos errores pueden clasificarse a su vez en errores de homófonos, errores tipográficos, errores gramaticales o errores de límite entre palabras (Kim et al., 2013). En los últimos años, las investigaciones sobre estos errores se han centrado principalmente en el subgrupo de los errores gramaticales, motivo por el que tienen una gran presencia en la revisión del estado de la cuestión. No obstante, aunque muchos errores *real-word* dan lugar a errores sintácticos, hemos de precisar que no todos los errores gramaticales están constituidos por errores *real-word*, pues pueden darse errores sintácticos por cuestiones de puntuación o de omisión de constituyentes, por ejemplo.

Los errores gramaticales han sido estudiados principalmente con fines educativos y con especial relevancia en la enseñanza de segundas lenguas (Yannakoudakis, 2013; Lawley, 2015; Hernández, 2012; Carlini et al., 2014). El objetivo del análisis de errores en este campo es ser un indicador del proceso de enseñanza y aprendizaje en el contexto educativo. Los estudios se han centrado en las producciones de estudiantes de inglés como segunda lengua (ESL), por lo que la mayoría de corpus anotados con errores están escritos en inglés. Este campo de investigación cuenta con pocos recursos para otros idiomas (Rozovskaya y Roth, 2019).

Concretamente, los corpus disponibles en español que contienen errores *real-word* son escasos y el número de trabajos es limitado. Los corpus disponibles para su uso son recopilaciones de textos escritos por aprendices de español como segunda lengua (Davidson et al., 2020). Los ejemplos más destacados son CEDEL2 (*Corpus Escrito del Español como L2*) desarrollado por Lozano y Mendikoetxea (2013) y CAES (*Corpus de Aprendices de Español*) de Rojo y Palacios (2016), aunque no incluyen etiquetado que facilite el entrenamiento de modelos. Más recientemente, Davidson et al. (2020) han creado COWS-L2H, un corpus de estudiantes de español como segunda lengua que contiene anotaciones de error y texto corregido en paralelo. Sin embargo, al tratarse de errores cometidos por estudiantes de español como segunda lengua, los tipos de errores recopilados difieren en gran medida de los cometidos por hablantes nativos.

Una característica común a la mayoría de sistemas que corrigen los errores de palabras dependientes del contexto es el uso de conjuntos de confusión y de modelos de lenguaje basados en n-gramas (Sharma y Gupta, 2015; Azmi et al., 2019). Posteriormente, se han desarrollado métodos basados en aprendizaje automático, como los sistemas basados en clasificación supervisada estadística, que emplean un clasificador para cada tipo de error gramatical a partir de corpus anotados y características extraídas (Gamon, 2010; Tetreault et al., 2010). Estos métodos han ido evolucionando hasta arquitecturas de aprendizaje automático más avanzadas usadas para detectar y corregir errores mediante el análisis de su contexto.

En los últimos años, el enfoque que ha ganado más relevancia debido a su efectividad es el de abordar este tipo de errores como una tarea de traducción automática (Napoles y Callison-Burch, 2017). Estos sistemas son más eficientes que los métodos basados en clasificación supervisada, pues tienen capacidad para abordar patrones de errores más complejos (Beloki et al., 2020). Los primeros métodos de traducción automática estaban basados en traducción automática estadística (Brockett et al., 2006; Junczys-Dowmunt y Grundkiewicz, 2014). Recientemente se están desarrollando modelos más potentes basados en traducción automática neuronal (Yuan y Briscoe, 2016; Grundkiewicz et al., 2019; Chen et al., 2020), pues tienen una gran capacidad para generalizar patrones. Funcionan a nivel de frase, no de palabras individuales, y pueden ser entrenados para otros idiomas siempre que los datos estén disponibles. No obstante, los modelos son difíciles de interpretar y modificar, el costo computacional para procesar estos candidatos es muy alto y requieren muchos datos de entrenamiento paralelo. Así,

como solución ante la escasez de corpus con errores disponibles se está llevando a cabo la generación de datos de forma artificial, como ya hemos mencionado anteriormente.

Estos enfoques más recientes basados en *deep learning* y redes neuronales han sido aplicados mayoritariamente al inglés, no son comunes los trabajos para el español. En español únicamente encontramos el trabajo de Davidson et al. (2020), en el que los autores entrenan un modelo de traducción automática neuronal para la corrección de errores gramaticales de estudiantes de español. A excepción del trabajo anterior, la detección y corrección de errores *real-word* o gramaticales en español se ha realizado mayoritariamente mediante enfoques basados en reglas, etiquetado y modelos estadísticos basados en n-gramas. A continuación, mencionaremos algunos de estos trabajos.

En primer lugar, Ramírez y Sánchez (1996) desarrollaron una propuesta de corrector gramatical (Gramcheck) para el español y el griego; para ello emplearon un enfoque lingüístico basado en reglas mediante el uso de extensiones de Prolog. Posteriormente se han creado distintos algoritmos apoyados en el análisis estadístico para detectar errores gramaticales. Estos algoritmos complementan a los correctores automáticos que requieren el etiquetado y el análisis morfosintáctico previo para desempeñar su función (San Mateo, 2016). Nazar y Renau (2012) emplean un método basado en estadísticas de coocurrencias tomando como referencia un extenso corpus de n-gramas, el corpus Google Books N-gram. Los n-gramas son usados para verificar y sugerir correcciones en un texto de entre las posibilidades léxicas existentes. Estos autores consideran que, a pesar de ser una de las primeras tareas de procesamiento del lenguaje natural en comenzar a abordarse, la corrección automática gramatical y los verificadores comerciales desarrollados siguen siendo deficientes, pues muchos de los errores no son detectados. Otros investigadores utilizan un enfoque similar basado en métodos estadísticos y bigramas de palabras (López-Ferrero et al., 2014; Blázquez, 2019). También encontramos trabajos para detectar errores específicos, como los relacionados con el uso o ausencia de tilde que da lugar a errores *real-word* (Atserias et al., 2012) o los errores de coocurrencias léxicas (Ferraro et al., 2014).

San Mateo (2016) desarrolló un corrector ortográfico y gramatical a partir del análisis estadístico de la frecuencia de ocurrencia de bigramas en un corpus. En este caso, se trata de un corpus de hablantes nativos, pues es una herramienta enfocada a personas que realizan corrección y revisión de textos. Además, Bustamante-Rodríguez et al. (2018) desarrollaron una herramienta para el ámbito educativo que fuese de utilidad como apoyo

a las tareas de corrección. El modelo tiene dos fases, una primera en la que se emplea la técnica basada en etiquetado sintáctico mediante la librería NLTK y una segunda fase de detección sintáctica de errores mediante reglas gramaticales y léxicas.

Para el español se han realizado también diversos estudios sobre análisis de errores y diseño de tipologías. Díaz Villa (2005) presentó una propuesta de tipología de errores gramaticales para un corrector automático gramatical y de estilo, centrada en los errores de motivación cognitiva. Los errores se obtuvieron tras la revisión de textos con un grado medio-alto de revisión que fueron escritos por hablantes nativos de español. En esta tipología encontramos una combinación entre errores *non-word* y *real-word*. No obstante, algunos de los tipos de errores señalados caerían en la categoría de recomendaciones o cuestiones de estilo. Entre los errores *real-word* que señala se encuentran los errores que superan al nivel de palabra, como los relacionados con paronimia sintáctica («si que creo» en lugar de «sí que creo») y paronimia semántica («surgir» en lugar de «surtir»), construcciones sintácticas incorrectas o errores de concordancia intrasintagmática e intersintagmática. No incorpora datos cuantitativos que permitan saber la presencia real de este tipo de errores en corpus de español no especializado. Los errores señalados en los anteriores trabajos tienen que ver mayoritariamente con concordancia de género y número, posición de clíticos, uso de preposiciones, acentuación o paronimia, entre otros.

Los errores *real-word* también han sido abordados para estudiar trastornos específicos de aprendizaje. Pedler (2007) compiló un listado de errores *real-word* para el inglés, compuesto por 833 parejas de palabras que tienden a ser confundidas, extraídas de textos escritos por personas con dislexia. Para el español destaca *DysList* (Rello et al. 2016), un recurso lingüístico formado por un listado de errores cometidos por personas con dislexia. En este trabajo concluyen que solo el 8,97 % del total de los errores cometidos fueron errores *real-word* (Rello et al., 2014).

Por último, se han llevado a la práctica investigaciones centradas en otras lenguas de la península. Es reseñable el trabajo de Beloki et al. (2020), en el que se hace uso de modelos neuronales secuencia a secuencia, basados en la arquitectura *transformer*, para la corrección de errores gramaticales en euskera. Además, debido a la inexistencia de datos de entrenamiento para esta lengua, los autores desarrollan un enfoque basado en reglas lingüísticas para producir de forma artificial oraciones incorrectas a nivel gramatical. Para el gallego encontramos el trabajo de Gamallo et al. (2015), en el que se desarrolla una herramienta de corrección gramatical que detecta y analiza errores

comunes en aprendices de gallego. Los autores afirman tener versiones básicas para portugués y español, pero es necesario el desarrollo de recursos lingüísticos como listas con tipos de errores o reglas sintácticas para la identificación de estos.

2.2. El lenguaje médico

El lenguaje médico es un lenguaje científico y en él prima la función representativa del lenguaje, es decir, se trata de un tipo de discurso que principalmente busca informar y transmitir conceptos. Gutiérrez (2006) considera que en él debe imperar la precisión terminológica, la neutralidad y la economía.

Uno de los rasgos distintivos del lenguaje científico es el uso de léxico especializado (Cabré, 1993). El léxico especializado del lenguaje médico está presente en numerosos entornos, desde contextos académicos y profesionales, como la actividad diaria de los profesionales sanitarios en hospitales y publicaciones en revistas científicas biomédicas, hasta contextos más informales, como la conversación con un familiar sobre la enfermedad que padece. Si nos situamos en el contexto profesional sanitario, la producción de textos médicos es muy fecunda, día a día se generan informes médicos, protocolos de actuación, ensayos clínicos o documentos de alta, entre otros.

Son diversos los trabajos existentes en la literatura que estudian y analizan el lenguaje médico en español desde una perspectiva prescriptivista. Entre ellos, debemos destacar el *Diccionario de Términos Médicos* (2012) elaborado por la Real Academia Nacional de Medicina, cuyo fin es la normalización del lenguaje médico, y que ha sido desarrollado gracias al trabajo colectivo de especialistas médicos y académicos. Además de aportar definiciones actualizadas, en él se dan indicaciones para el buen uso del lenguaje médico: se aclaran conceptos que generan duda, se señalan errores frecuentes, acompañados de su correspondiente corrección, y también se proponen términos en español para evitar el uso innecesario de anglicismos, o la mejor forma de adaptación de estos. También se incluye información sobre la procedencia etimológica de los términos, sinónimos, abreviaciones y distintas observaciones.

Resulta también especialmente destacable la labor de divulgación realizada por diversas asociaciones y entidades. Entre ellas debemos destacar TREMÉDICA[13]

[13] <https://www.tremedica.org/>

(Asociación Internacional de Traductores y Redactores de Medicina y Ciencias Afines), que ofrece gran variedad de recursos terminológicos, lingüísticos y normativos, y se encarga de la publicación de la revista *Panace@. Revista de Medicina, Lenguaje y Traducción*, para la difusión del conocimiento lingüístico y traductológico especializado en medicina y ciencias afines. También podemos mencionar la labor de la Fundación Dr. Antoni Esteve[14], que busca favorecer la investigación biomédica y contribuir a la formación de investigadores. Concretamente, resulta relevante la publicación de una de sus monografías, titulada «La importancia del lenguaje en el entorno biosanitario» (2014) y coordinada por el traductor médico Fernando Navarro y por la profesora de la Universidad de Salamanca Bertha Gutiérrez. En esta publicación se aborda la importancia del lenguaje en la labor asistencial médica, en el ámbito de la investigación biomédica y la divulgación científica, en los planes de estudio de las titulaciones biosanitarias, y el tratamiento del lenguaje médico en los medios de comunicación y los problemas que plantea.

Asimismo, es habitual encontrar publicaciones en revistas especializadas de medicina que ofrecen recomendaciones y hojas de estilo para mejorar la corrección lingüística en el ámbito médico (Hernández, 1992; Aguilar, 2013b). Entre ellas se encuentran revistas como *Medicina Clínica*[15] y *Educación Médica*[16], que incorporan artículos sobre distintos aspectos del lenguaje médico en español. Un ejemplo es el trabajo de Aleixandre-Benavent et al. (2017), en el que se realiza una caracterización del lenguaje médico en los artículos científicos y se describen algunos de los problemas y defectos más habituales que presenta. Entre ellos los autores destacan el abuso de abreviaciones (abreviaturas, siglas y acrónimos), el uso de extranjerismos innecesarios, la utilización de títulos efectistas y metafóricos en los artículos, la presencia de pleonasmos y solecismos, y el abuso de las mayúsculas y de las formas en gerundio.

Por su parte, Hernández y Bustabad (2015) analizan las características lingüísticas de los trabajos científicos en el ámbito de urgencias y destacan la presencia de tecnicismos y una marcada influencia anglicista. Consideran que la incorporación de nuevas palabras no es un problema, pero sí lo es el empobrecimiento injustificado del léxico, y lo ilustran con las voces «patología», «evidencia» y «severo». También critican el empleo excesivo

[14] < https://www.esteve.org/ >

[15] <https://www.elsevier.es/es-revista-medicina-clinica-2>

[16] <https://www.elsevier.es/es-revista-educacion-medica-71>

de la voz pasiva y del gerundio, el uso anafórico de «el mismo», la puntuación incorrecta, como la presencia de coma entre el sujeto y el predicado, y los errores de concordancia.

Aguilar (2013a) lleva a cabo un listado de recomendaciones ortográficas y ortotipográficas a partir de las principales novedades incorporadas en la nueva *Ortografía de la lengua española* (2010) que pueden aplicarse a publicaciones biomédicas en español para evitar errores de escritura. Por otra parte, Mayor (2010) analiza textos médicos destinados a pacientes, concretamente folletos de salud, páginas web y prospectos de medicamentos. Muestra algunos ejemplos de los errores más habituales de estos géneros, como la falta de precisión y claridad expositiva de los folletos; el lenguaje demasiado técnico y difícilmente comprensible de los prospectos; y la falta de rigor en el empleo del lenguaje médico de algunas páginas web.

Por tanto, observamos que en los últimos años se han publicado diversos estudios sobre las características del lenguaje médico desde una perspectiva cualitativa y descriptiva, sin embargo, la literatura sobre errores lingüísticos en informes médicos escritos en español es más limitada.

2.2.1. Informes médicos

La información médica en los sistemas clínicos se almacena en registros de salud electrónicos, donde quedan recogidos los distintos informes médicos y las interacciones entre médico y paciente. Un informe médico es un documento escrito por un facultativo en el que se indica el proceso asistencial prestado a un paciente. Es un texto expositivo y descriptivo en el que se describen síntomas, procesos, pruebas, procedimientos y observaciones para llegar a un diagnóstico y tratamiento adecuado (Estopà, 2020). La Comisión de Deontología del Colegio de Médicos de Vizcaya (*apud* Llopart-Saumell y Da Cunha, 2020: 14) lo define de la siguiente forma:

> Es el documento mediante el cual la médica o el médico responsable de un paciente, o el que lo ha atendido en un determinado episodio asistencial, da a conocer aspectos médicos relacionados con los trastornos que sufre, los métodos diagnósticos y terapéuticos aplicados, y, si procede, las limitaciones funcionales que se puedan derivar. Sirve para dejar constancia de un estado de salud incluso anterior al de la fecha de petición; por lo tanto, su vigencia no está limitada a un periodo de tiempo. Su petición puede estar vinculada a motivos de interés particular o de orden legal o público.

Aunque depende de la especialidad médica, los informes médicos suelen tener una estructura prototípica y ciertas partes distinguibles que siempre aparecen: la identidad del paciente, los antecedentes, la anamnesis, los resultados de la exploración física, la sintomatología, la descripción de las pruebas y análisis, el tratamiento prescrito, la evolución, y los datos del médico que ha redactado el informe (Dalianis, 2018).

Según Domènech-Bagaria et al. (2020), los grupos de usuarios receptores de la información contenida en los informes pueden ser diversos. En primer lugar, el propio paciente o los familiares que le acompañan. En segundo lugar, otros profesionales de la salud, entre los que se encuentran otros especialistas, enfermeros, investigadores clínicos o personal administrativo. En tercer lugar, científicos de datos y empresas tecnológicas que los utilizan para el desarrollo de sistemas de procesamiento y gestión. Por último, se convierten en documentos legales que pueden cobrar mucha importancia en procesos judiciales.

Los informes médicos son una fuente de información de incalculable valor porque pueden ser utilizados para desarrollar estudios clínicos que ayuden a la investigación de cohortes de pacientes y enfermedades y, por consiguiente, pueden mejorar la atención médica de los pacientes. Por esta razón, es tan importante la calidad de estos textos como fuente de entrada para poder utilizar sistemas que los procesen para fines analíticos, como abordaremos con detalle en la sección de medicina y PLN.

Recientemente se están desarrollando importantes proyectos centrados en mejorar los procesos de comunicación entre los profesionales sanitarios y los pacientes. Uno de ellos es el proyecto de investigación «Junts»[17] dirigido por Rosa Estopà y desarrollado por investigadores de la Universitat Pompeu Fabra y la Universitat Oberta de Catalunya, cuyo objetivo es contribuir a la alfabetización en el ámbito de la salud para derribar las barreras de comunicación que se producen entre los pacientes y sus familias. Concretamente, se centran en el entorno de la enfermedad rara pediátrica y consideran un problema comunicativo las interferencias lingüísticas y cognitivas de los textos médicos.

En el marco de este proyecto de investigación, recientemente Estopà (2020) ha coordinado la obra *L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió*, que representa una importante contribución para el análisis del lenguaje en los informes clínicos. Cada uno de los capítulos está dedicado a analizar distintas

[17] <https://www.upf.edu/web/medicina_comunicacio/junts>

vertientes del lenguaje, entre las que se encuentran la terminología, la comprensión textual, la sintaxis, la pragmática o la ortografía. Concretamente, uno de los capítulos, «Els errors ortotipogràfics, ortogràfics i de puntuació» (Vivaldi, 2020) está dedicado a la presencia de errores ortográficos en los informes médicos y profundizaremos en él en la siguiente sección. Estopà (2020) considera necesario concienciar a los facultativos sobre aquellos aspectos lingüísticos que pueden ser un impedimento para la comprensión de los informes que van a ser leídos por los pacientes y aboga por una medicina colaborativa. La autora también defiende la necesidad de dotar de mayor relevancia cuestiones comunicativas y de redacción en la formación médica.

Por su parte, Terroba (2016) analiza cuatrocientos informes médicos del sistema sanitario de La Rioja. El objetivo principal de este trabajo es estudiar los rasgos que caracterizan los informes clínicos de esta comunidad y aportar una visión general sobre la situación del lenguaje escrito en la sanidad riojana. La autora se centró principalmente en los sistemas de acortamiento, pues señala que suponen la principal causa de incomprensión de estos textos y repercuten en la calidad de la asistencia sanitaria.

2.2.1.1. Características lingüísticas

Los informes médicos presentan unas características lingüísticas especiales, conocerlas es importante para comprender mejor el corpus de estudio. El informe médico es un texto que difiere del texto estándar especializado de los documentos científicos en cuanto al estilo y corrección, pues presenta unas convenciones particulares. Al tratarse de lenguaje natural, la información se presenta desestructurada y con formas no normalizadas que no siempre se ajustan a las normas lingüísticas del español. También se caracteriza por el uso de siglas y abreviaturas específicas del dominio, formas compuestas, errores ortográficos y la presencia de extranjerismos (Aleixandre-Benavent et al., 2015). El estilo de escritura presenta diferencias dependiendo de la unidad clínica y la especialidad (Dalianis, 2018), hay especialidades que utilizan un estilo más telegráfico y otras más narrativo. A continuación, se presentan sus principales rasgos lingüísticos atendiendo a distintos aspectos.

En primer lugar, los informes médicos se caracterizan por su elevada densidad terminológica en comparación con otros textos. Constantemente se utilizan términos específicos del dominio, incluso cada especialidad presenta su propio lexicón

terminológico que puede plantear dudas entre facultativos de otras especialidades. Esto es debido a la elevada cifra de especialidades médicas y al vasto repertorio de términos existentes para cada una. En los informes se nombran enfermedades, signos, síntomas, hallazgos, medicamentos, procesos, marcas y principios activos, entre otros.

El origen de gran parte de la terminología médica se encuentra en lenguas clásicas, como el griego, el latín y el árabe; en lenguas modernas europeas, como el francés, el alemán y el inglés; e incluso otras lenguas del mundo, como el japonés (Pérez Castro, 1997; Gutiérrez, 2006). Especialmente importante ha sido el vocabulario latino en el origen de muchos de los términos anatómicos que se usan en la actualidad, mientras que los términos patológicos suelen proceder del griego clásico (Dalianis, 2018). Es también destacable el considerable uso de epónimos (Martín, 2008), es decir, términos vinculados con el nombre propio de una persona, lugar o época. Por otro lado, no debemos obviar la constante creación de neologismos para nombrar los nuevos avances y realidades que van surgiendo. Como señala Gutiérrez (2006:279):

> Cabe señalar que ni el mejor de los repertorios especializados contiene todos los términos de una zona del conocimiento, ni está permanentemente actualizado, tal es la velocidad con la que se suceden los descubrimientos y las diferentes interpretaciones que se dan a los mismos fenómenos, con las repercusiones terminológicas que ello conlleva. Como tampoco pueden dar cuenta con exhaustividad de la sinonimia, polisemia y otros fenómenos semánticos que, a pesar de los pesares, existen en el lenguaje de la ciencia.

Debido a su importancia como lengua de la comunicación científica, la lengua inglesa posee una notable presencia en el lenguaje médico. La mayor parte de las investigaciones biomédicas son publicadas en esa lengua, lo que conlleva la incorporación de nuevos términos en inglés en el vocabulario de los profesionales médicos. No obstante, algunos autores, como Aleixandre-Benavent y Amador-Iscla (2001) advierten que en ocasiones se produce el uso incorrecto de expresiones médicas en español por traducciones erróneas o falsos amigos que se propagan y acaban impregnándose en el léxico de la comunidad médica.

Por otro lado, uno de los rasgos más característicos de los informes médicos es la abundante presencia de abreviaturas y siglas, como constatan los estudios de Piñero et al., (2006) o Plasencia y Moliner (2012). Las abreviaturas y siglas se utilizan para representar de forma reducida y compacta expresiones mayores, lo que ahorra tiempo y espacio

durante la escritura. Concretamente, una abreviatura es «la representación gráfica reducida de una palabra o grupo de palabras, obtenida por eliminación de algunas de las letras o sílabas de su escritura completa» (OLE, 2010: 568). Por tanto, son palabras que son acortadas mediante la eliminación de alguna de sus letras. Cualquier hablante de una lengua puede crear las abreviaturas que considere en su día a día en busca de la economía lingüística, de ahí que un entorno como el médico, con las particularidades que lo caracterizan, se preste de forma especial a la creación de abreviaturas de forma muy prolífica.

Piñero et al. (2006) detectaron un promedio de 14,7 abreviaciones en informes de urgencias e informes de servicios sanitarios especializados. Por su parte, Plasencia y Moliner (2012) hallan más de veinte abreviaturas por informe en una muestra de sesenta informes de alta hospitalaria y enfermería. Las abreviaturas son también muy abundantes en informes médicos escritos en otras lenguas, como el inglés (Wu et al., 2012) o el húngaro (Siklósi et al., 2014). Estos autores también señalan la ambigüedad que se produce con siglas que tienen más de un significado. Un ejemplo es la sigla «AP», cuya búsqueda en el diccionario de siglas SEDOM[18] devuelve doce posibles resultados. La interpretación adecuada de las abreviaciones supone un obstáculo en muchas ocasiones incluso para los propios facultativos al tratarse de casos muy ambiguos, aunque tengan información contextual. Liu et al. (2017) analizan textos clínicos en inglés e informan de que el 33 % de las abreviaturas detectadas plantearon dificultades y fueron consideradas ambiguas. La ambigüedad también se produce al utilizar abreviaciones que no están estandarizadas en la comunidad médica y son creadas de forma particular (Aleixandre-Benavent et al., 2015).

A nivel sintáctico los informes se caracterizan por la búsqueda de la brevedad a través de la elipsis y el uso de estructuras oracionales muy simples. Se suele presentar el sujeto omitido, al igual que se omiten verbos auxiliares, determinantes y otros elementos discursivos como conectores. Los textos suelen estar formados por sintagmas nominales breves y se produce una mayor nominalización y presencia de adjetivos especificativos (Terroba, 2016).

Por último, destaca la presencia de errores ortográficos. Vivaldi (2020) expone las razones que considera causantes de la presencia de errores en los informes, el autor

[18] <http://diccionario.sedom.es/>

considera que los profesionales médicos no suelen otorgar importancia al uso de caracteres especiales y signos de puntuación, que redactan limitados por el tiempo y que poseen desconocimiento de diferentes aspectos de la lengua, como ortografía, sintaxis y técnicas de comunicación. Vivaldi (2020) analiza un corpus en español de 1090 informes de urgencias emitidos por el Hospital Italiano de Buenos Aires entre los años 2014 y 2015, con un total de 420 000 *tokens*. Tras la revisión establece que se producen errores de acentuación, la unión de dos o más palabras y la inclusión de un espacio que provoca la separación de una palabra en dos o más fragmentos. También menciona la omisión o inserción indebida de una o más letras en una palabra, la alteración del orden de las letras o el uso de abreviaturas no estándares o poco transparentes.

Es también pertinente mencionar la labor del Instituto para el Uso Seguro de los Medicamentos (ISMP)[19], cuyo objetivo es prevenir los errores de medicación y fomentar la seguridad en el uso de medicamentos. Para tal fin, han creado el «Programa de Notificación de Errores de Medicación» y desarrollan documentos y recursos de interés, entre los que se encuentra un listado[20] de pares de nombres de medicamentos que se prestan a equivocación por similitud ortográfica o fonética.

Terroba (2016) en su estudio sobre informes médicos en español refiere la presencia de errores fonéticos y ortográficos, así como errores provocados por la confusión en el uso de mayúsculas y minúsculas, fluctuaciones acentuales, ambigüedades semánticas, repeticiones de lexemas que empobrecen el estilo, abundantes elipsis, y errores sintácticos con oraciones incorrectamente construidas, que no siguen el orden lógico o con errores de puntuación.

Los autores no incorporan datos cuantitativos sobre la presencia de errores, pero señalan que es elevada. Por tanto, estos trabajos suponen un buen sustento teórico y punto de partida, pero es esencial continuar profundizando en los tipos de errores mediante el

[19] <http://www.ismp-espana.org/>

[20] «Lista de los pares de nombres de medicamentos con mayor riesgo de causar errores por similitud fonética u ortográfica» disponible en <http://www.ismp-espana.org/ficheros/relaciondenombres.pdf> y una lista complementaria a la anterior publicada en 2010 con nuevos pares de nombres <http://www.ismp-espana.org/ficheros/Actualizaci%C3%B3n%20nuevos%20pares%20nombres%202005-2010.pdf>.
También se puede consultar un documento con «Recomendaciones para prevenir los errores causados por confusión en los nombres de los medicamentos», disponible en <http://www.ismp-espana.org/ficheros/Recomendaciones%20nombres%20ISMP-Espa%C3%B1a.pdf>.

análisis de un corpus real de informes médicos de varias especialidades. Compararemos estos trabajos con los resultados obtenidos en el presente estudio en la sección de Discusión (4.3).

2.3. Medicina y procesamiento del lenguaje natural

La presente sección pretende aportar una fuente de conocimiento sobre el estado actual del procesamiento del lenguaje natural en el dominio de la medicina. Se van a presentar las principales tareas para el procesamiento de textos clínicos médicos, los desafíos existentes y las aplicaciones desarrolladas, profundizando en los proyectos y corpus disponibles para el español.

En el campo del PLN y la lingüística computacional encontramos numerosos proyectos que se centran en la creación de recursos lingüísticos y el desarrollo de técnicas que puedan ser útiles para el procesamiento de documentación clínica en formato digital y el desarrollo de soluciones de PLN orientadas a sanidad. La documentación clínica incluye información sobre síntomas, diagnósticos, tratamientos, uso de medicamentos —y sus efectos adversos—, que puede ser utilizada para mejorar la investigación y la atención médica de los pacientes. Sin embargo, durante décadas esa información no ha podido ser reutilizada al no disponer de los medios necesarios para su correcto procesamiento. Esa realidad ha cambiado, los métodos computacionales desarrollados en los últimos años están permitiendo procesar información en lenguaje natural a gran escala, teniendo un importante impacto en la investigación biomédica y la industria de la salud. Estos métodos son capaces de analizar y extraer información tanto de los datos estructurados como de los no estructurados procedentes de diversas fuentes, como historias médicas, ensayos clínicos, resultados de laboratorio o publicaciones científicas, mediante distintas técnicas.

Por tanto, los sistemas de PLN tienen el objetivo de ayudar a convertir textos clínicos escritos en formato narrativo en representaciones estructuradas de datos clínicos fácilmente interpretables. Para conseguirlo se han puesto en práctica un amplio número de tareas, entre las que se encuentra la minería de textos, la codificación clínica automatizada, la anonimización, la corrección automática, la detección de la negación, la desambiguación de siglas, o el reconocimiento de entidades clínicas, como veremos con más detalle en las siguientes secciones.

Los grandes avances en procesamiento del lenguaje natural en este dominio se han producido especialmente para el inglés. Esto es debido a varias razones: su consolidación como lengua principal de la ciencia, el mayor número de recursos específicos del dominio desarrollados para esa lengua y la mayor disponibilidad de corpus de gran tamaño. Asimismo, las grandes corporaciones (como *Google* o *Facebook*), que lideran el desarrollo de nuevos sistemas en la industria tecnológica y realizan las contribuciones más relevantes a nivel de software, trabajan principalmente con esta lengua.

Aunque el procesamiento de textos en otros idiomas ha experimentado menos atención, en los últimos años han surgido múltiples proyectos para otras lenguas en el área de procesamiento del lenguaje natural que tienen como objetivo el desarrollo y adaptación de recursos para sanidad. Este es el caso del español, debido al aumento exponencial de documentación clínica y a la necesidad de poder acceder y reutilizar esos datos. Concretamente, tiene un papel fundamental el Plan de Impulso de las Tecnologías del Lenguaje[21], cuyo objetivo es ofrecer acceso a recursos y herramientas de procesamiento del lenguaje natural para el español, con una sección específica enfocada a sanidad. Otras líneas de actuación son la formulación de instrumentos de evaluación y la difusión de proyectos. Además, también tienen como finalidad proporcionar a la comunidad investigadora y a la industria corpus en los que se puedan probar y evaluar herramientas, contribuir a la reproducibilidad de experimentos y a la transferencia de resultados de investigación (Villegas et al., 2017). Muchos de los recursos se encuentran alojados en repositorios abiertos como Zenodo[22] o GitHub.

Asimismo, es habitual en el área organizar talleres y competiciones comunitarias sobre determinadas tareas de PLN, donde participan grupos de investigación académicos y de la industria, para experimentar con técnicas de vanguardia, evaluar su rendimiento, generar nuevos recursos, y promover el desarrollo y mejora de sistemas de procesamiento de textos médicos (Huang y Lu, 2015).

[21] <https://plantl.mineco.gob.es/sanidad/Paginas/sanidad.aspx>

[22] «Medical NLP – language technology resources for clinical and biomedical documents in multiple languages». <https://zenodo.org/communities/medicalnlp>

2.3.1. Desafíos

Las particularidades inherentes a los textos clínicos, comentadas anteriormente, causan complicaciones importantes para su procesamiento automático e intentan ser superadas. A pesar de los importantes avances realizados en el área, el procesamiento de informes clínicos aún continúa presentando sustanciales desafíos (Sun et al., 2018; Iroju y Olaleke, 2015). Entre ellos pueden destacarse:

2.3.1.1. Errores lingüísticos

Como hemos abordado anteriormente con más detalle, una proporción significativa de la información clínica se encuentra en textos en lenguaje natural y presentan una serie de características que dificultan su procesamiento. Los repositorios clínicos están integrados por grandes volúmenes de datos de naturaleza muy heterogénea. Perera et al. (2013) afirman que los textos clínicos contienen alrededor del 80 % de datos no estructurados. Es imprescindible que los documentos sean sometidos a un preprocesamiento preliminar efectivo para que tengan el formato de entrada de datos idóneo y sean más fáciles de interpretar. El preprocesamiento se convierte, por tanto, en una etapa del proceso sumamente importante, con una carga de trabajo superior al 60% (Sun et al., 2018). El preprocesamiento de datos incluye la limpieza y normalización de estos, mediante la conversión de determinados caracteres, la expansión de abreviaturas, la tokenización, el reconocimiento de límites de oraciones o la eliminación de duplicados, entre otros.

La presencia de errores lingüísticos es una de las dificultades más destacadas, debido a los problemas de reconocimiento de entidades y de confusión semántica que puede plantear. Un error de sustitución del carácter *o* por *i*, adyacentes en el teclado, puede conllevar que la palabra *hemitórax*[23] se convierta en *hemotórax*[24]. Por tanto, la calidad de los resultados del modelado depende de manera crucial de la calidad de los datos de entrada. Profundizaremos en esta cuestión en la sección dedicada a la detección y corrección automática de errores en el dominio médico.

[23] El *DTME* define *hemitórax* como «cada una de las dos mitades laterales del tórax».

[24] El *DTME* define *hemotórax* como «presencia de sangre en la cavidad pleural».

2.3.1.2. Datos confidenciales y anonimización

Las distintas estrategias de preprocesamiento están estrechamente relacionadas con las características que presente el corpus. En el caso de los registros clínicos es imprescindible tener en cuenta que contienen información confidencial sobre el paciente, por lo que es fundamental la anonimización[25] de esos datos (Gkoulalas-Divanis, 2014). Por responsabilidad ética y legal hay establecidos protocolos de protección de datos y métodos de control de acceso que deben cumplirse antes de que los informes puedan ser utilizados para la investigación o para el desarrollo de aplicaciones.

Para que la identidad de las personas involucradas no pueda ser revelada se eliminan todo tipo de datos identificativos, como nombres y apellidos de pacientes y facultativos, direcciones, fechas, números de teléfono o correos electrónicos. Según Medlock (2006) la anonimización puede verse como un proceso de dos etapas, una primera en la que se deben identificar las referencias o ítems sensibles y una segunda etapa en la que se deben neutralizar. La neutralización de la información sensible se puede llevar a cabo de tres formas, mediante la eliminación, el reemplazo y la pseudoanonimización. La eliminación consiste en reemplazar la entidad con un espacio en blanco; la categorización es el reemplazo de la entidad con una etiqueta que representa la categoría en la que se inserta («nombre», «apellidos», «hospital», etc.); y la pseudoanonimización es la sustitución de una referencia por una variante aleatoria del mismo tipo o categoría. Por tanto, los pasos para llevar a cabo la anonimización incluirían la identificación de los ítems como información confidencial, su posterior clasificación en la categoría que le corresponda y la sustitución de esos datos por una categoría general, por espacio en blanco o por ítems similares generados de forma aleatoria.

Llevar a cabo la anonimización de esos documentos de forma manual es un proceso sumamente tedioso y costoso cuando se trata de conjuntos de datos de grandes dimensiones, pero aún se convierte en un proceso más complejo e inviable cuando se trata de lenguaje natural no estructurado (Dalianis, 2018). Por este motivo, en los últimos años han surgido un importante número de estudios que investigan sobre el desarrollo de herramientas para la anonimización automática de registros clínicos (Vovk et al., 2021).

[25] La anonimización es el proceso llevado a cabo para eliminar toda la información de un documento que puede identificar a una persona determinada.

Son múltiples los esfuerzos para desarrollar tecnologías eficaces para la anonimización automática y a lo largo de los años se han propuesto múltiples sistemas automáticos de anonimización (Mamede et al., 2016; Dernoncourt et al., 2016). En el marco del Plan de Tecnologías del Lenguaje se han organizado distintos retos comunitarios sobre anonimización de documentación médica en español. El primero de ellos fue la tarea MEDDOCAN[26] («Medical Document Anonymization»), que se estructuró en dos subtareas: la detección del texto confidencial y la clasificación de la información en categorías. Marimon et al. (2019) resumen los datos y resultados obtenidos tras la organización de esta tarea sobre anonimización de documentos médicos en español y concluyen que «advanced deep learning approaches in combination with rule based systems and gazetteer resources can provide very competitive results when a high quality manually labeled dataset is available»[27] (Marimon et al, 2019: 639).

Para esta competición se creó el corpus MEDOCCAN[28], un corpus sintético que contiene mil casos clínicos recopilados de SciELO[29], con alrededor de 495 000 palabras en total y un promedio de 494 palabras por caso clínico. Los casos clínicos se enriquecieron con información sanitaria protegida procedente de resúmenes de alta y registros clínicos de genética médica. El corpus se dividió en conjuntos de entrenamiento, desarrollo y prueba. El conjunto de entrenamiento contiene 500 casos, mientras que los otros dos conjuntos contienen 250 casos cada uno.

Por último, uno de los grupos participantes en la edición 2019 de MEDDOCAN desarrolló HitzalMed[30], una herramienta web disponible que realiza la detección automática de información confidencial en textos clínicos para el español utilizando algoritmos de aprendizaje automático.

---

[26] «Campañas de Evaluación Plan TL: Anonimización de documentación médica» <https://plantl.mineco.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/Paginas/anonimizacion-doc-medicos.aspx>

[27] «Los enfoques avanzados de aprendizaje profundo en combinación con sistemas basados en reglas y lexicones pueden proporcionar resultados muy competitivos cuando se dispone de un conjunto de datos etiquetados manualmente de alta calidad» (Marimon et al, 2019: 639).

[28] <https://github.com/PlanTL-GOB-ES/SPACCC_MEDDOCAN>

[29] Biblioteca electrónica que reúne publicaciones de revistas científicas de Hispanoamérica, Brasil, Sudáfrica, Portugal y España (<http://www.scielo.org>).

[30] <https://snlt.vicomtech.org/hitzalmed>

2.3.1.3. Escasez de corpus disponibles y formatos heterogéneos

En general, es difícil acceder a corpus clínicos para la investigación. Esto se debe principalmente a la información de carácter sensible que contienen, como hemos mencionado anteriormente. Muchos de los corpus utilizados han sido desarrollados por grupos de investigación y se requieren permisos especiales para poder acceder a ellos, otros han sido compilados por empresas privadas y no se encuentran a disposición pública. Santiso et al. (2018: 10970) afirman que «the data sparsity and the lexical variability tend to decrease the performance of models inferred from data»[31].

Según Wen et al. (2020), hay pocos conjuntos de datos de texto médico a gran escala disponibles públicamente que sean adecuados para los modelos de preentrenamiento, y consideran que los corpus disponibles con datos reales a menudo son de pequeña escala y están desequilibrados. Como consecuencia, uno de los mayores desafíos en la construcción de sistemas de PLN basados en aprendizaje profundo para corpus biomédicos es la disponibilidad de conjuntos de datos públicos.

La ausencia de herramientas que puedan resolver problemas específicos de cada lengua y la falta de corpus del dominio disponibles para entrenamiento y evaluación de modelos supone un obstáculo que intenta ser superado por parte de los investigadores, contribuyendo con proyectos y el desarrollo de recursos (Névéol et al., 2018).

Otro inconveniente tiene que ver con el formato heterogéneo que poseen los informes dependiendo de cuál sea su procedencia. No suelen presentar una estructura uniforme que esté estandarizada y se observan notables diferencias entre los registros de cada hospital, con distintos títulos, códigos y tratamiento de los datos. Como consecuencia, no siempre es posible utilizar una misma herramienta en distintos entornos y deben someterse a adaptaciones.

La codificación clínica es una tarea clave que consiste en la transformación de textos médicos clínicos a un formato estructurado o codificado utilizando conceptos reconocidos internacionalmente, como CIE-10 o SNOMED-CT, con formato numérico o alfanumérico.

[31] «La escasez de datos y la variabilidad léxica tienden a disminuir el rendimiento de los modelos inferidos a partir de los datos» (Santiso et al., 2018: 10970).

En 2020 se organizó CodiEsp[32] («Clinical Case Coding in Spanish Shared Task») en el marco del eHealth CLEF 2020. Consistió en la primera tarea comunitaria específicamente dedicada a la codificación automática de casos clínicos en español. Los equipos participantes tenían que asignar automáticamente códigos ICD-10 (CIE-10) a los documentos de casos clínicos, que serían evaluados con codificaciones ICD-10 generadas manualmente (Miranda-Escalada et al., 2020). Para el desarrollo de la tarea se creó un corpus sintético de 1000 casos clínicos que abarcan diversas especialidades médicas y que está formado por 411 067 *tokens*. Además, se puso a disposición un conjunto de 2751 documentos con antecedentes. Fue seleccionado manualmente por facultativos y documentalistas clínicos y anotado por profesionales de la codificación clínica (Miranda-Escalada et al., 2020). El conjunto de entrenamiento está compuesto por 500 documentos y los conjuntos de desarrollo y prueba están compuestos por 250 cada uno.

2.3.1.4. Detección de la negación y la incertidumbre

En estos textos es común hacer uso de la negación para indicar la presencia o ausencia de determinadas entidades, como síntomas y enfermedades, o para descartar diagnósticos. La negación puede aparecer de forma explícita o implícita, y cuando se da de forma implícita implica relaciones semánticas entre las expresiones que pueden resultar complejas de analizar computacionalmente. Palabras como «no», «sin» o «ninguno» alteran todo el sentido de una oración. De igual manera ocurre con prefijos como «in-» o «a-» que reflejan negación morfológica, como en «afebril», que indica la ausencia de fiebre o en «insuficiencia», que indica la escasez de una cantidad necesaria. La importancia de esta tarea radica en la influencia que la negación puede tener en la comprensión automática de la información dado que invierte el valor de una cláusula.

Asimismo, resulta fundamental la detección de la expresión de la incertidumbre. Son múltiples las palabras que pueden transmitir incertidumbre según el contexto en el que se encuentren (por ejemplo, «sospecha de», o «sin aparente»), pues su presencia marca que el contenido no puede ser plenamente verificado y debe ser interpretado con cautela. En el dominio biomédico el significado implícito de una oración puede ser crucial debido al tipo de información que está en juego. Es imprescindible procesar y extraer con

[32] <https://temu.bsc.es/codiesp/>

total precisión el conocimiento de los registros clínicos para que el uso de estos sistemas se pueda estandarizar.

En los corpus anotados con fenómenos de negación e incertidumbre se suelen etiquetar, por un lado, los marcadores que desencadenan el cambio de significado y, por otro, el alcance o las palabras que se ven afectadas por la negación.

Uno de los primeros trabajos sobre detección de entidades negadas en textos clínicos fue llevado a cabo por Chapman et al. (2001), que desarrollan NegEx, un algoritmo basado en reglas con el que buscan determinar si un hallazgo o una enfermedad mencionada en informes médicos está presente o ausente. Este trabajo ha sido tomado como punto de referencia para el avance de investigaciones posteriores, como la de Deléger y Grouin (2012) para textos clínicos en francés; Kang et al. (2017) detectan la negación y su alcance en notas clínicas en chino mediante *word embeddings*; Skeppstedt (2011) desarrolla una adaptación de NegEx para textos clínicos en sueco; y Morante et al. (2008) investigan sobre el alcance de la negación en los textos biomédicos.

Para el español encontramos disponible NegEx-MES[33], un sistema de detección de negaciones en textos clínicos en español basado en el algoritmo NegEx de Chapman et al. (2001). Otros trabajos que también abordan la detección de negaciones en el dominio médico son los siguientes:

- Santiso et al. (2018) crean un sistema basado en redes neuronales para detectar entidades médicas negadas, como enfermedades o medicamentos, en historias clínicas electrónicas escritas en español.

- Lima et al. (2020) desarrollan el corpus NUBES («Negation and Uncertainty annotations in Biomedical texts in Spanish»), formado por extractos de informes clínicos que han sido enriquecidos con anotaciones de negación e incertidumbre.

- Cruz et al. (2017) presentan un trabajo de anotación con marcadores de negación y eventos negados realizado en un corpus de informes de anamnesis y radiología.

- Marimon et al. (2017) recopilan el corpus IULA «Spanish Clinical Record Corpus», formado por 3194 frases extraídas de historias clínicas anonimizadas y anotadas manualmente con marcadores de negación y su alcance.

- Cotik et al. (2017) generan un corpus de informes radiológicos anotados con entidades, eventos y relaciones de negación e incertidumbre. El corpus fue concebido

[33] <https://github.com/PlanTL-GOB-ES/NegEx-MES>

como un recurso de evaluación para el reconocimiento de entidades nombradas y extracción de relaciones, y como entrada para el uso de métodos supervisados.

- Oronoz et al. (2015) crean un corpus *gold standard*[34] formado por historias clínicas en español anotadas manualmente por expertos en farmacología y farmacovigilancia Anotaron entidades relacionadas con patologías y medicamentos, y también relaciones entre entidades que indican eventos de reacciones adversas a medicamentos.

2.3.1.5. Reconocimiento y desambiguación de siglas y abreviaturas

Los profesionales de la salud se enfrentan a serias limitaciones de tiempo, por lo que es habitual que con frecuencia utilicen abreviaturas y siglas específicas de dominio para ahorrar tiempo de escritura y espacio. Su presencia en los informes médicos ha proliferado y, a pesar de su utilidad, también presentan desafíos para la comprensión del texto, especialmente si su uso no está estandarizado y la sigla no aparece definida. Cuando no poseen una difusión generalizada y estandarizada puede no ser fácilmente interpretable por la mayoría de los receptores.

Además, en ocasiones se produce ambigüedad debido a que a una misma abreviatura o sigla se le pueden asociar distintos sentidos. Es habitual encontrar siglas que remiten a distintos conceptos y resultan difíciles de analizar computacionalmente.

La desambiguación del sentido de la abreviación es la lógica que determina cuál de los sentidos (forma expandida) de una abreviación es el más relevante en el contexto en el que aparece. Para paliar este problema, tanto la comunidad investigadora como los desarrolladores de software realizan esfuerzos considerables para crear sistemas que identifiquen siglas y encuentren sus significados correctos en el texto.

Schwartz y Hearst (2002) proponen uno de los métodos que ha sido más ampliamente utilizado en la investigación de identificación de siglas. Desarrollan un algoritmo basado en reglas que utiliza la coincidencia de caracteres entre las letras de la sigla, determinados patrones y su contexto para detectar la sigla y su forma expandida en el texto. También se han desarrollado modelos basados en características para la

[34] En PLN se utiliza el término *gold standard* para referirse al corpus o conjunto de datos con anotaciones validado por los investigadores y que se utiliza para el entrenamiento y evaluación de los diferentes sistemas de aprendizaje automático desarrollados.

identificación y desambiguación de siglas (Liu et al., 2017). Además, parte del software existente emplea modelos basados en reglas heurísticas para la identificación de siglas en el dominio biomédico junto con métodos de aprendizaje profundo más avanzados para la desambiguación (Xu et al., 2007; Ciosici y Assent, 2018). Joopudi et al. (2018) aplican modelos de red neuronal convolucional (CNN) de codificación de características de notas clínicas y contextuales para la desambiguación de abreviaturas. Uno de los trabajos más recientes es el de Wen et al. (2020), que generan un gran conjunto de datos para la desambiguación de siglas en el dominio médico a partir de más de 18 millones de resúmenes de PubMed para un entrenamiento previo. Usan sustitución inversa para generar muestras sin etiquetado humano, es decir, identifican términos completos en el texto que tienen abreviaturas conocidas y los reemplazan con estas. Ellos proponen que la desambiguación de abreviaturas sea parte de una tarea de preentrenamiento que sirva para transferir el aprendizaje adquirido a otras tareas posteriores con arquitecturas de aprendizaje profundo.

Intxaurrondo et al. (2017) consideran que se han publicado un número marginal de estudios sobre reconocimiento y resolución de abreviaturas para el español, además de que existen pocos recursos léxicos para interpretar abreviaturas en esta lengua que puedan ser usados en formato legible para sistemas. Por este motivo, desarrollaron una competición de identificación y resolución de abreviaturas llamada «Biomedical Abbreviation Recognition and Resolution» (BARR), que formaba parte del «Workshop on Evaluation of Human Language Technologies for Iberian Languages» (IberEval 2017). Tuvo como principal objetivo impulsar el desarrollo y la evaluación de sistemas de identificación de abreviaturas. La tarea consistió en identificar las abreviaturas y sus correspondientes formas expandidas a partir de resúmenes de literatura biomédica en español. Además, los organizadores proporcionaron el corpus BARR, con resúmenes médicos etiquetados manualmente por expertos en la materia, y una plataforma de evaluación llamada BARR-Markyt. Mientras que esta primera competición se centró en literatura biomédica, en 2018 se organizó una segunda edición, «Biomedical Abbreviation Recognition and Resolution 2nd Edition»[35] (BARR2) cuyo objetivo fue el reconocimiento de abreviaturas en los resúmenes de textos clínicos y estudios de casos

[35] <https://temu.bsc.es/BARR2/>

clínicos escritos en español procedentes de SciELO. El corpus BARR2[36] está formado por 3343 casos clínicos de SciELO, que fueron clasificados por un facultativo para incluir los que eran similares a textos clínicos reales en cuanto a estructura y contenido, y descartar los que no eran adecuados para esta tarea. El corpus cubre un amplio abanico de especialidades, entre las que se encuentran oftalmología, urología, cirugía, atención primaria, pediatría y oncología, entre otros.

Por último, Pomares-Quimbaya et al. (2020) contribuyen a la resolución de siglas en español a través de la creación de un conjunto de inventarios de sentido organizados por especialidad clínica que contiene siglas, sus expansiones y características basadas en corpus y está compuesto por 3603 siglas en total.

2.3.2. Aplicaciones

En la industria de la salud, el procesamiento del lenguaje natural tiene muchas aplicaciones sumamente beneficiosas y que merece la pena reseñar. Se han dedicado notables esfuerzos de investigación en el desarrollo de software para etiquetar texto, clasificar información, extraer entidades, detectar la negación, resolver la anáfora, desambiguar siglas y abreviaturas, entre otras (Wong y Glance, 2011). Todas estas tareas son concebidas como un recurso de apoyo para los sistemas de minería de textos clínico y otras aplicaciones, actuando como distintos engranajes que posibilitan el procesamiento de grandes volúmenes de datos.

La extracción de información o minería de texto es la tecnología que tiene como objetivo encontrar información específica a partir de textos en lenguaje natural y extraer conocimiento implícito en texto no estructurado. Al convertir el texto libre en datos estandarizados se puede acceder a la información de forma mucho más eficiente mediante interfaces de consulta. Por ejemplo, se puede extraer información determinada —relativa a medicación, enfermedades, síntomas o fechas— de la historia clínica de un paciente con el objetivo de ver su evolución o estudiar una variable específica.

Otra importante aplicación está relacionada con el apoyo a la decisión clínica y la predicción de enfermedades. Según Sun et al. (2018), los sistemas de apoyo a las decisiones médicas permiten a los expertos médicos obtener asesoramiento sobre el

[36] El corpus BARR2 está disponible para su descarga en <https://temu.bsc.es/BARR2/datasets.html>.

diagnóstico y tratamiento de los síntomas, basándose en datos reales previos de otras historias clínicas electrónicas. Un ejemplo es DXplain[37], un sistema de soporte de decisiones desarrollado en el Laboratorio de Ciencias de la Computación en el Hospital General de Massachusetts, que utiliza un algoritmo de pseudoprobabilidad para generar una secuencia de enfermedades ingresando datos del paciente, los síntomas, los resultados de laboratorio y el tratamiento clínico (Sun et al., 2018). Asimismo, Iroju y Olaleke (2015) mencionan sistemas de apoyo a la decisión clínica (CDSS) como cTAKES[38] (*Clinical Text Analysis and Knowledge Extraction System*) y MedLEE (*Medical Language Extraction and Encoding System*), que han sido desarrollados para extraer información significativa de textos clínicos no estructurados al identificar y etiquetar las entidades de dominio en el corpus con codificación clínica estándar, y para comprender las diferentes relaciones lingüísticas y semánticas entre los distintos elementos del texto. De esta forma se facilita el análisis automático significativo de la información sanitaria.

Otros ejemplos son herramientas que pueden ayudar a detectar síntomas tempranos de enfermedades, reacciones adversas provocadas por medicamentos o diagnósticos gracias a modelos inferidos a partir de observaciones a gran escala. En particular, al tener acceso a un gran conjunto de casos de cáncer con patrones de datos identificados, podría ser posible detectar y pronosticar casos futuros y realizar análisis predictivos (Jensen et al., 2017). De igual forma, autores como Jacobson y Dalianis (2016) aplican aprendizaje profundo en registros de salud electrónicos en sueco para predecir infecciones asociadas a la atención médica. También es posible la detección de efectos adversos provocados por el uso de determinados medicamentos, como en las investigaciones realizadas por Henriksson et al. (2015) y Casillas et al. (2016).

Los hallazgos sobre enfermedades y fármacos tradicionalmente han conllevado un enorme coste humano y temporal. Sin embargo, la tecnología de minería de datos médicos puede descubrir con mayor rapidez la trayectoria médica de una enfermedad a lo largo del tiempo (Sun et al., 2018). Esta realidad es posible mediante la detección automática de valores atípicos, el reconocimiento de patrones y la clasificación de estos.

Además de para la toma de decisiones, otra aplicación del procesamiento del lenguaje natural en el dominio médico tiene que ver con la simplificación y síntesis de

[37] <http://www.mghlcs.org/projects/dxplain>

[38] <https://ctakes.apache.org/>

información. La síntesis de información supone la extracción automática de la información más importante, como puede ser el resumen del registro clínico de un paciente. A su vez, el resumen automático es la representación narrativa simplificada del documento original. La simplificación de informes médicos puede resultar especialmente útil para aquellos pacientes que tienen acceso a estos desde aplicaciones, como el Portal del Paciente, y que puede ayudar a una mayor comprensión de estos.

También es posible llevar a cabo la codificación y categorización de documentos para organizar y clasificar en categorías relevantes la información, de manera que se facilite la búsqueda y navegación de los usuarios mediante palabras clave (Asker et al., 2016). Un ejemplo es la implementación de sistemas de indización automática y la asignación de vocabularios controlados como CIE-10 o de la terminología de SNOMED-CT. Investigadores del grupo de Innovación Tecnológica y la Unidad de Gestión Clínica de Hematología del Hospital Universitario Virgen del Rocío de Sevilla desarrollaron un codificador automático CIE-O-3 y CIE-10 dentro del proyecto COCO, cuyo objetivo era diseñar y evaluar un sistema de extracción de conocimiento y codificación automática de diagnósticos de oncohematología (Sánchez et al., 2018).

En el sector industrial español se están desarrollando importantes proyectos que buscan implementar soluciones que ayuden a acelerar la investigación en salud en beneficio de los pacientes. Savana[39], empresa fundada en 2014, ha desarrollado *EHRead* para procesar historias clínicas electrónicas en tiempo real y extraer automáticamente información médica. Por su parte, Iomed[40] crea herramientas de PLN para transformar el texto de las notas médicas en bases de datos que puedan ser consultadas para la investigación clínica. Han desarrollado el software *Compass,* con el que es posible crear cohortes, saber cuántos pacientes cumplen con unos criterios específicos y llevar a cabo estudios de viabilidad e investigación clínica. Mediante estas herramientas es posible extraer todas las variables clínicas de relevancia de registros electrónicos de salud para estudios observacionales y ensayos clínicos. Otro ejemplo destacable es Medbravo[41], que ofrece soporte para implementar arquitecturas en procesamiento de datos en salud. Genera modelos predictivos de aprendizaje profundo, y algunos de los proyectos que se

[39] <https://savanamed.com/>

[40] <https://iomed.es/>

[41] <https://www.medbravo.org/>

mencionan en su página web son la predicción de biomarcadores en el carcinoma colorrectal, la predicción de la respuesta a fármacos y la sensibilidad a partir de datos genómicos, o el multietiquetado a gran escala de informes clínicos en lenguaje natural en español.

Por último, otra aplicación estaría relacionada con las interfaces de usuario, los sistemas conversacionales y la comunicación eficiente con los sistemas informáticos para optimizar el trabajo de los facultativos. Podemos mencionar el uso de tecnologías de reconocimiento de voz que permiten a los usuarios comunicarse mediante la voz, o hacer uso de los sistemas de reconocimiento de voz para el dictado automático de informes médicos, como en el caso de INVOX Medical[42], desarrollado por la empresa Vócali[43]. Esta aplicación permite a los médicos transcribir automáticamente la información en lugar de tener que utilizar el teclado, lo que optimiza el proceso y les permite realizar otras tareas de forma simultánea.

2.4. Detección y corrección automática en el dominio médico

El procesamiento automático de textos médicos es una materia emergente y de relevancia en PLN. Los avances en los métodos y tecnologías de ciencia de datos han impulsado la utilización de los documentos electrónicos como fuente inestimable de información para la investigación en el ámbito sanitario. A continuación, abordaremos la literatura existente sobre detección y corrección automática de errores en el dominio médico.

En primer lugar, es necesario manifestar que, si bien es cierto que existen multitud de trabajos realizados en el área de PLN para el dominio médico —como hemos constatado en la sección anterior—, la mayoría están enfocados actualmente en tareas de extracción de información médica, codificación de documentos o reconocimiento de entidades nombradas (Wu et al., 2020; Casey et al., 2021), y no existen tantos estudios centrados específicamente en el proceso de detección y corrección automática. En muchos casos, el proceso de corrección se presenta como un eslabón más en investigaciones que incluyen otras técnicas y tienen un propósito más amplio. Por

[42] <https://invoxmedical.com/>

[43] <https://vocali.net/>

ejemplo, Dziadek et al. (2017) pretenden mapear texto clínico no estructurado a una ontología médica para facilitar la interoperabilidad semántica, pero primero deben lidiar con la abundancia de errores ortográficos presentes en las notas clínicas; es el mismo caso del trabajo de Sayle et al. (2011), cuyo objetivo principal es llevar a cabo minería de datos para poder identificar eficientemente nombres químicos; o el de Wong y Glance (2011), que trabajan con técnicas de desambiguación automática.

Por tanto, los estudios centrados específicamente en detección y corrección automática en el dominio médico son limitados y, además, se caracterizan por su heterogeneidad, pues se han desarrollado en diferentes entornos, idiomas y problemáticas. Como se observa en la Tabla 1, se han aplicado técnicas de corrección en diversos tipos de textos médicos. Entre ellos, mayoritariamente destacan los trabajos dedicados a la corrección de informes médicos electrónicos, pertenecientes a múltiples especialidades, como emergencias (Patrick et al., 2010), patología quirúrgica (Workman et al, 2019), fetopatología (D'hondt et al., 2016), radiología (Zech et al., 2019), repositorios de alergias (Lai et al., 2015), oncología (Dzieciątko, 2019), notas de progreso (Wong y Glance, 2011) o registros de salud (Fivez et al., 2017), entre otros. Asimismo, también se han realizado estudios sobre los errores contenidos en consultas realizadas por pacientes o consumidores para mejorar sistemas de búsqueda sobre información de salud y sugerencias de aplicaciones (Senger et al., 2010; Kilicoglu et al., 2015; o Lu et al., 2019), o sobre patentes de interés farmacéutico (Sayle et al., 2011) para mejorar el proceso de minería de datos. Puede destacarse que, a pesar de la variabilidad, las contribuciones tienen en común que los corpus presentan un formato desestructurado y contienen terminología médica.

También encontramos trabajos que analizan corpus con errores provocados por sistemas de dictado. La utilización de software de reconocimiento de voz en entornos clínicos ha ganado protagonismo en los últimos años, debido a las facilidades que otorga su uso a los facultativos. No obstante, en las transcripciones de voz a texto también se producen errores, de ahí la necesidad de desarrollar métodos automatizados para identificar y corregir estos errores en el texto generado (Ringler et al., 2017; Zhou et al., 2018).

Asimismo, en los últimos años se ha normalizado la conversión de documentación clínica conservada en formato físico al formato digital para construir bases de conocimiento y recursos que puedan ser consultados por el personal médico. Si bien la

mayoría de los documentos médicos contemporáneos se crean de forma electrónica, muchos informes más antiguos se mantienen solo en versión impresa. Por tanto, se escanean y se les aplica software de reconocimiento óptico de caracteres (OCR). Sin embargo, este procesamiento de caracteres también provoca errores, en muchas ocasiones por similitud gráfica entre las letras o por falta de resolución y calidad en los documentos, tanto en los escaneados como en los originales en papel. Se han desarrollado numerosos sistemas de corrección para detectar y corregir automáticamente este tipo de errores en entornos clínicos (Thompson et al., 2015; D'hondt et al., 2016; Romero et al., 2011). Tanto los errores provocados por reconocimiento de voz, como los provocados por reconocimiento óptico de caracteres presentan unas particularidades directamente relacionadas con el medio en el que se producen y, por tanto, presentan diferencias con aquellos errores cometidos por escribir en un teclado o los provocados por motivación cognitiva.

La mayoría de investigaciones en el área sobre detección y corrección automática han sido desarrolladas con corpus en inglés (Lai et al., 2015; Zech et al., 2019; Fivez et al., 2017; Workman et al., 2019; entre otros), pero en los últimos años han surgido numerosas propuestas para otras lenguas. Se han realizado experimentos con corpus en sueco (Dziadek et al., 2017), en holandés (Fivez et al., 2017), en francés (Ruch et al., 2003; D'hondt et al., 2016), en húngaro (Siklósi et al., 2016), en persa (Yazdani et al., 2019); en ruso (Balabaeva et al., 2020), en polaco (Mykowiecka y Marciniak, 2006) o en español (Bravo Candel et al., 2021; Lima-López et al., 2021).

Los trabajos consultados que han utilizado corpus formados por informes clínicos coinciden en destacar el considerable número de errores lingüísticos que presentan estos y la dificultad de su procesamiento, tanto por la elevada presencia de abreviaturas que contienen, como por la compleja terminología, la ausencia de estandarización de las formas y la inexistencia de revisión posterior (Lai et al., 2015; Patrick et al., 2010; Siklósi et al., 2016). Las tasas de errores detectadas varían dependiendo del trabajo consultado. En el caso de Ruch et al. (2003) el porcentaje de error alcanza el 10 % en los informes clínicos en francés analizados; Nizamuddin y Dalianis (2014) señalan una tasa de error del 7,6 % en un corpus de informes médicos en sueco; Lai et al. (2015) detectan un 4 % en informes en inglés; Ringler et al. (2017) mencionan un 3,2 % en informes de radiología en inglés; y Liu et al. (2012) estiman que hay alrededor de un 0,4 % de errores ortográficos en documentos médicos en inglés. Por tanto, observamos fluctuaciones en

los resultados dependiendo del idioma y entorno. Además, estos autores señalan que los patrones de errores más comunes tienen que ver con la adición, omisión y transposición de caracteres, junto con el uso incorrecto de la puntuación, los errores gramaticales, el abuso de abreviaturas y la presencia de terminología con errores ortográficos (Siklósi et al., 2016). La mayoría de estos trabajos investigan sobre el tratamiento de errores *non-word*, que suelen ocurrir con un porcentaje mayor a los *real-word,* y apenas se abordan los errores semánticos y gramaticales.

Además, los sistemas de registro clínico no incorporan correctores ortográficos, lo que sería una herramienta de apoyo, especialmente en la escritura de terminología como medicamentos y epónimos (Dzieciątko, 2019). Los profesionales médicos a menudo soportan exceso de trabajo y disponen de poco tiempo para la redacción de estos registros, que generalmente no tienen revisión posterior. Debido a estas circunstancias, el número de errores en la documentación clínica es generalmente alto, lo que supone un desafío para el procesamiento de este tipo de documentos. Aunque la presencia de esos errores ortográficos no suponga una dificultad cognitiva significativa en la comprensión de las personas, sí afecta a la eficacia de los sistemas automatizados que se emplean para la extracción de información, la traducción automática, la síntesis de información o la codificación de documentos. Por tanto, estos sistemas se consideran una fase esencial en el preprocesamiento para mejorar la minería de textos (Yazdani, 2020). En la Tabla 1 aparece recopilada la información disponible sobre los distintos corpus utilizados en los principales trabajos publicados sobre detección y corrección automática en el dominio médico:

Estudio	Corpus
«Misspellings in drug information system queries: Characteristics of drug name spelling errors and strategies for their prevention»	Consultas de sistemas electrónicos de información sobre medicamentos del Hospital Universitario de Heidelberg (Alemania). Sistema basado en la web que contiene más de 95 000 marcas y más de 10 000 principios activos. 221 437 usuarios del Hospital Universitario de Heidelberg consultaron el DIS 575 142 veces.
«Spelling correction in clinical notes with emphasis on first suggestion accuracy»	Corpus de historias clínicas del Servicio de Urgencias del Hospital Concord de Sídney (Australia). 57 523 palabras únicas y 7442 errores ortográficos (datos de entrenamiento). 164 302 palabras únicas y 65 429 errores ortográficos (datos de prueba).
«Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes»	Notas de evolución clínica en inglés. Conjunto de pruebas (test set) de 30 muestras de un corpus de 2433 notas de evolución reales y 961 palabras, cada nota con una media de 32 palabras.
«Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction»	Minería de textos químicos recuperados de la base de datos de IBM de alrededor de 12 millones de patentes estadounidenses, europeas y mundiales. Los resúmenes que no estaban en inglés se tradujeron utilizando la funcionalidad de traducción Lexichem de OpenEye.
«Context-aware correction of spelling errors in Hungarian medical documents»	Corpus clínico en húngaro. Documentos clínicos anonimizados de varios departamentos de un hospital húngaro. *Gold standard*: 50 394 *tokens*. El tamaño del conjunto de prueba fue de 3722 tokens. El conjunto de prueba (*test set*) tiene 89 palabras diferentes mal escritas. 2000 oraciones (17 243 *tokens* y 6234 *types*) seleccionadas aleatoriamente de todo el corpus clínico de varios departamentos.
«An ensemble method for spelling correction in consumer health questions»	Preguntas de salud del consumidor (*consumer health questions*) en inglés recopiladas por la Biblioteca Nacional de Medicina de Estados Unidos. 372 preguntas para entrenamiento y 100 preguntas para prueba.

«Automated misspelling detection and correction in clinical free-text records»	Registros clínicos de texto libre en inglés. Primer conjunto de entrenamiento de 275 notas, 106 668 palabras, y un conjunto de prueba de 40 notas, con 15 247 palabras. Formado por notas clínicas seleccionadas al azar de las clínicas de atención primaria del *Brigham and Women's Hospital* (Boston). El segundo conjunto de datos se construyó a partir de entradas de texto libre sobre alergia seleccionadas al azar tomadas del Repositorio de alergias de *Partners Enterprise* (PEAR). 6460 palabras en el conjunto de entrenamiento. El conjunto de pruebas estaba formado por 442 entradas con 1380 palabras. El tercer conjunto de datos estaba compuesto por órdenes de medicación en texto libre seleccionadas al azar e introducidas por los médicos a través del sistema EHR ambulatorio de Partners. 5069 palabras formaron el conjunto de entrenamiento, mientras que 392 entradas (872 palabras) formaron el conjunto de prueba.
«Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care»	Contenido generado por el consumidor sobre atención médica, como publicaciones en sitios web de redes sociales en inglés. 150 publicaciones (21 358 palabras) del sistema de tablón de anuncios (BBS) de MedHelp. Este conjunto de publicaciones está relacionado con un medicamento llamado Zoloft y contiene descripciones de los consumidores sobre sus síntomas y sugerencias de otros.
«Customised OCR Correction for Historical Medical Text»	Textos médicos históricos en inglés. Reconocimiento de caracteres (OCR) en imágenes de páginas escaneadas. Colección de 24 documentos corregidos a mano, compuesta por tres documentos tomados de cada una de las siguientes décadas: 1840, 1860, 1880, 1900, 1920, 1940, 1960 y 1980.
«Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records (EMR)»	Historias clínicas electrónicas en inglés. 250 historias clínicas, que incluyen apartados como resultados, medicamentos e instrucciones para pacientes.

«Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings»	Texto libre clínico en inglés. Base de datos MIMIC-III (notas de progreso generadas por el médico, notas de enfermería, etc.). 873 instancias contextualmente diferentes.
«Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction»	Dos corpus en sueco. Un corpus de literatura formado por artículos de revistas editados (1,3 millones de *tokens*), y un corpus clínico, que comprende notas de EHR, texto clínico no estructurado (4,4 millones de *tokens*). Evaluación controlada con texto de literatura médica con errores inducidos. Evaluación parcial sobre notas clínicas.
«Improving Spelling Correction with Consumer Health Terminology»	Corpus en inglés con preguntas de salud del consumidor. Datos de salud del consumidor. 471 preguntas de salud del consumidor con 24 837 *tokens*.
«An efficient prototype method to identify and correct misspellings in clinical text»	Dos corpus diferentes: informes de patología quirúrgica, y notas de visita y progreso del departamento de emergencias, extraídos de los recursos de la Administración de Salud de Veteranos. 76 786 notas clínicas.
«Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models»	Informes clínicos en inglés. 91 867 informes de TC[44] de cabeza y 1 013 287 de radiografías de tórax del Sistema de Salud de Mount Sinai (2006-2017). Mount Sinai Hospital (61 722 TC de cabeza, 818 978 radiografías de tórax) y Mount Sinai Queens (30 145 TC de cabeza, 194 309 radiografías de tórax). Además, se obtuvo de la base de datos MIMIC-III un total de 32 259 TC de cabeza y 54 685 informes de radiografías de tórax del centro médico Beth Israel Deaconess.
«Using lexical disambiguation and named-entity recognition to improve	Informes clínicos electrónicos en francés. 424 informes. Conjunto inicial se divide en dos subconjuntos iguales (212 registros), conjunto A se utiliza para ajustar el sistema, mientras que el conjunto B se utiliza como conjunto de prueba final.

[44] Tomografía computarizada.

spelling correction in the electronic patient record»	
«Low-resource OCR error detection and correction in French Clinical Texts»	Corpus de informes clínicos franceses. Notas de pacientes del dominio de la fetopatología, digitalizados (OCR) en el contexto del proyecto Accordys, Expedientes de 2476 pacientes individuales, lo que equivale a 16 573 documentos en papel.
«Automated Misspelling Detection and Correction in Persian Clinical Text»	Corpus de informes clínicos en persa. Tres fuentes de datos proporcionadas por el Hospital Imam Khomeini en Teherán. El primer conjunto de datos incluye informes de pacientes que se han realizado una ecografía mamaria para comprobar la presencia de ganglios linfáticos. El conjunto de entrenamiento (*training set*) con 15 639 informes y 871 275 palabras. El conjunto de prueba (*test set*) con 249 informes y 101 863 palabras. Un segundo conjunto de datos consistió en informes de ecografía de cabeza y cuello, 15 472 informes con 168 055 palabras. El conjunto de prueba contenía 75 informes con 26 856 palabras. El tercer conjunto de datos consistió en informes de ultrasonido abdominal y pélvico. Además, se consideraron 3531 informes como un conjunto de entrenamiento que contenía 106 084 palabras. El conjunto de prueba contiene 428 informes con 19 264 palabras.
«Correcting Polish Bigrams and Diacritical Marks»	Historias clínicas electrónicas en polaco del campo de la oncología del Instituto de Oncología Maria Sklodowska-Curie (Varsovia). 130 000 registros.
«Detection of spelling errors in Swedish clinical text»	Corpus en sueco que contiene 100 registros de pacientes, notas de los médicos y enfermeros de cinco unidades clínicas diferentes: neurología, ortopedia, infecciones, cirugía dental y nutrición, con un total de 151 924 palabras

«Grammatical error correction for Spanish health records»	Dos corpus en español. Una colección de historias clínicas en español corregidas (10 007 oraciones) llamada IMEC. Y el corpus TMAE, un corpus sintético formado a partir de la recopilación de textos procedentes de IBECS, SciELO, Pubmed y SPACCC con errores artificiales. El corpus resultante tiene un tamaño de más de 2,3 millones de oraciones paralelas y 51 millones de *tokens*.

Tabla 1. Información sobre los corpus utilizados en estudios de detección y corrección automática de errores en el dominio médico

2.4.1 Técnicas

En las últimas dos décadas, las técnicas de corrección ortográfica han sido ampliamente estudiadas. Si nos centramos específicamente en el ámbito médico, la abundancia de trabajos es menor pero también destacable. La mayoría de estos estudios se han enfocado en informes clínicos de distintas especialidades, como ya hemos mencionado anteriormente, aunque también se localizan otros dedicados a textos generados por el consumidor en el área de la atención médica o en foros de consulta. Por tanto, intervienen diferentes especialidades y tipos de textos, circunstancia que influye en el tipo de enfoques elegidos.

Son diversos los métodos empleados en los procesos de detección y corrección automática en este dominio, como puede examinarse en el Tabla 2, en la que hemos recogido las distintas técnicas y recursos empleados en cada uno de los trabajos revisados. Es destacable mencionar que hay técnicas que son indispensables y se usan en todos los trabajos analizados. De igual forma, se observa una tendencia, conforme avanzamos hacia estudios más recientes el uso de técnicas basadas en contexto aumenta. El punto de partida suelen ser los sistemas de búsqueda en diccionario, que son usados en todos los trabajos consultados. Además de los diccionarios generales, contenidos en correctores como Aspell, Hunspell o Google Spell Checker, es necesario incorporar terminología específica del dominio para que estos sean efectivos. Estos lexicones se crean a partir de la combinación de distintas fuentes y en ellos se suele incluir listas de abreviaturas y acrónimos, afijos específicos, *gazetteers*, nomenclaturas, taxonomías, tesauros, listas con

enfermedades, medicamentos, síntomas, principios activos y otros términos médicos. Son diversos los diccionarios y terminologías especializados utilizados en la literatura consultada: Diccionario médico de Webster, Diccionario médico de Dorland, Diccionario médico de Stedman, el Sistema de Lenguaje Médico Unificado UMLS[45] (*The Unified Medical Language System*), la terminología clínica SNOMED-CT[46] (*Systematised Nomenclature of Medicine-Clinical Terms*), The Moby Lexicon[47], MedlinePlus[48], PubMed[49], SPECIALIST Lexicon[50], entre otros. El procedimiento de consulta en diccionario es sencillo, si una palabra o conjunto de caracteres no aparece en la búsqueda en el diccionario empleado es susceptible de ser una palabra mal escrita.

También tiene una presencia sumamente alta en la literatura el uso del método basado en la distancia mínima de edición. La distancia de edición de Damerau-Levenshtein representa el número mínimo de operaciones obligatorias para transformar una cadena de caracteres en otra. También encontramos técnicas que trabajan con la distancia de edición o similitud a nivel fonético. En este método debemos señalar el algoritmo fonético Soundex o sistemas mejorados como Metaphone y Double-Metaphone (Kilicoglu et al., 2015).

También se hace uso de sistemas basados en reglas y heurísticas (Lai et al., 2015; Thompson, 2016), como aquellos que usan expresiones regulares. En estos casos las búsquedas en los diccionarios se pueden realizar de acuerdo a unas reglas previamente definidas. Estos sistemas son especialmente útiles para expandir abreviaturas, para corregir de forma automática casos que están claramente definidos, o palabras que habitualmente se escriben mal y su corrección no plantea ambigüedades. También se utilizan los métodos basados en reglas para la generación de errores sintéticos que se emplean para el aumento de datos de entrenamiento, como ya hemos referido anteriormente. Otros también incorporan etiquetado de partes del discurso e información

[45] <https://www.nlm.nih.gov/research/umls/index.html>

[46] International Health Terminology Standards Development Organisation, SNOMED CT. <http://www.ihtsdo.org/snomed-ct/>

[47] Moby Project. <https://mobyproject.org>

[48] Medline Plus. <https://medlineplus.gov/>

[49] PubMed. <https://www.ncbi.nlm.nih.gov/pubmed/>

[50] SPECIALIST Lexicon. <http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicon.html>

morfosintáctica mediante *POS-tagging* y análisis sintáctico (Zhou et al., 2015; Hussain y Qamar, 2016).

Asimismo, son mayoritarios los estudios que incorporan métodos estadísticos (Wong y Glance, 2011), como datos de frecuencia de ocurrencia de palabras en corpus, para la ordenación y elección de candidatos. Dentro de los métodos estadísticos encontramos el modelo de error, que es utilizado en la técnica conocida como *noisy channel model* o modelo de canal ruidoso (Kernighan, 1990; Jurafsky y Martin, 2014). Esta técnica parte de la teoría de la comunicación (Shannon, 1948), en ella un emisor envía una secuencia de símbolos a un receptor, sin embargo, durante la transferencia ciertos símbolos de la secuencia transmitida se confunden debido a las deficiencias del canal de transmisión. El objetivo del receptor es reconstruir la secuencia original utilizando el conocimiento de la fuente y las propiedades del canal de transmisión. Lai et al. (2015) utilizan esta técnica en su estudio y consiguen una precisión de corrección en torno al 80 % en un conjunto de notas clínicas. No obstante, este modelo no tiene en cuenta información contextual.

En aquellos casos en los que las palabras no pueden ser corregidas de manera aislada, entran en juego otras técnicas estadísticas como el uso de modelos de lenguaje, el análisis de n-gramas (normalmente de bigramas o trigramas) u otras técnicas de aprendizaje automático (Patrick et al., 2010; Wong y Glance, 2011; Lu et al., 2018). En todos estos casos se pretende añadir al sistema información contextual. Es relevante mencionar también el uso de ontologías (Zhou et al., 2015), como Wordnet[51], RxNorm[52] o SNOMED-CT, que pueden ayudar a identificar la distancia semántica de una palabra con respecto a las palabras adyacentes. La corrección de errores dependientes del contexto es fundamental en casos donde la palabra escrita correctamente se reemplaza por otra palabra existente o real, lo que dificulta su identificación. En estos casos ya no nos situamos a nivel de palabra, sino a nivel de frase o texto, pues entran en juego cuestiones gramaticales y semánticas.

Más recientemente han surgido nuevas técnicas y enfoques basados en aprendizaje profundo, como el uso de redes neuronales (Shickel et al., 2018; Fivez et al., 2017; Bravo et al., 2021) que pueden explotarse con éxito en aquellos casos en los que se requiere

[51] WordNet. A Lexical Database for English. <https://wordnet.princeton.edu/>

[52] National Library of Medicine, RxNorm. <http://www.nlm.nih.gov/research/umls/rxnorm/>

contexto. En ellos se mide la similitud entre vectores de palabras y se representan los contextos en los que aparece una palabra. No obstante, el uso de estos métodos no habría sido posible sin un aumento significativo de la potencia computacional, que ha permitido utilizar conjuntos de datos más grandes y algoritmos computacionales más complejos. En el caso del español, el trabajo de Bravo-Candel et al. (2021) propone el uso de un modelo seq2seq de traducción automática neuronal para la corrección de estos errores. El corpus usado para entrenar y evaluar el modelo fue recopilado a partir de Wikicorpus y de un corpus médico formado por casos clínicos de CodiEsp, MEDDOCAN y SPACC. No obstante, estos corpus estaban revisados y no contenían errores reales, por lo que se generaron errores sintéticos aplicando un conjunto de reglas. Entre los tipos de errores introducidos se encuentran errores de discordancia de género y número, y de confusión entre palabras homófonas. Otro trabajo destacable realizado con un corpus en español es el de Lima-López et al. (2021), que presentan una aproximación a la corrección de errores gramaticales de historias clínicas en español y para ello llevan a cabo un experimento con un codificador-decodificador convolucional multicapa y también recurren a la generación de errores sintéticos para entrenar el modelo.

Aunque son notables los avances llevados a cabo en los últimos años, siguen existiendo limitaciones, pues muchas de las técnicas resultan complejas de configurar y los correctores no son aún totalmente exitosos. En los trabajos consultados se obtienen tasas de acierto prometedoras, aunque todavía queda trabajo por hacer y margen de mejora para que los sistemas sean totalmente exhaustivos y precisos. Algunos experimentos que lograron resultados fructíferos y que podemos destacar son los desarrollados por Siklósi et al. (2016), que implementaron un sistema sensible al contexto basado en traducción automática estadística con una precisión del 87,23 % para corregir errores en texto clínico húngaro; Zhou et al. (2015), que desarrollaron un sistema de corrección ortográfica con una precisión de corrección de los errores del 86 %; o Lai et al. (2015), que lograron una precisión de corrección de errores ortográficos del 80 % en textos clínicos. No obstante, es necesario precisar que gran parte de los trabajos consultados, especialmente los menos recientes, miden la eficacia y precisión de sus métodos de corrección ortográfica aplicados solo en casos de errores *non-word,* debido a la mayor complejidad que implican los casos de errores *real-word.* Las mayores dificultades y desafíos aparecen en los errores gramaticales y semánticos (Kilicoglu et al., 2015) y, a pesar de que en los últimos años se han incrementado los estudios que se

adentran en esta problemática, aún es necesario un mayor desarrollo de técnicas que involucren aspectos semánticos.

Otra de las limitaciones observadas se encuentra en los métodos de evaluación, se suele medir el rendimiento en relación al grado de acierto en la corrección de palabras mal escritas, pero también puede resultar útil medir aquellos casos de falsos positivos en los que se corrigen erróneamente palabras que son correctas. Es también relevante poder contar con conjuntos de datos comunes a disposición de la comunidad científica para poder comparar el rendimiento de los distintos sistemas desarrollados, contribuyendo así a la mejora y validación de estos.

La calidad del diccionario utilizado tiene una influencia determinante en la calidad de los resultados obtenidos. Sin embargo, es costoso recopilar diccionarios que estén actualizados porque la ciencia avanza cada día y continuamente se crean neologismos para aludir a nuevas realidades que van surgiendo. De igual forma, a medida que aumenta el tamaño del diccionario, el número de errores que se incluyen en él de forma involuntaria puede incrementarse. Es, por tanto, esencial la calidad de los diccionarios usados y que estos sean exhaustivos.

En estrecha relación con esta última cuestión se encuentra el problema de la disponibilidad de datos. En el caso de las arquitecturas basadas en redes neuronales, son necesarios volúmenes ingentes de textos para que se puedan entrenar y den lugar a sistemas competitivos. Con corpus de menor tamaño se reduce la efectividad y este es uno de los principales problemas a los que se enfrentan estos métodos contextuales, ya que cuanto más pequeño es el conjunto de datos, menos información se puede extraer a partir del contexto, y peor será la elección del candidato para la corrección. A esto se suma que los textos suelen contener errores y abreviaturas que pueden distorsionar los resultados y minimizar la precisión. Por tanto, es necesario que haya una mayor disponibilidad de corpus de especialidad para el entrenamiento de modelos, pero el dominio clínico se enfrenta a un mayor desafío debido a cuestiones de privacidad en el intercambio de datos.

Los mayores avances se han conseguido en corpus en inglés, hay un número mayor de herramientas disponibles para este idioma, de ahí la necesidad de analizar, adaptar y generar nuevos recursos para otras lenguas. Son numerosos los autores que manifiestan la necesidad de otorgar un mayor grado de universalidad en las soluciones desarrolladas, para poder ser aplicadas en otros entornos e idiomas. La reproducibilidad

y la comprensión del funcionamiento interno de los modelos resulta fundamental para el avance científico.

En suma, durante la revisión de la literatura hemos podido constatar el crecimiento exponencial acaecido en el área en los últimos años. Como muchos otros campos que utilizan procesamiento del lenguaje natural, ha sido particularmente destacable en el ámbito clínico el reciente desarrollo de los métodos basados en aprendizaje profundo, siendo las arquitecturas basadas en redes neuronales recurrentes o *transformers* las más populares. Sin embargo, aunque estos métodos cada vez ganan más peso, su rendimiento todavía no se encuentra en una fase plenamente estable y sigue enfrentándose a dificultades debido a la escasez de datos disponibles y a la complejidad que plantea la interpretabilidad de estos modelos computacionales. Como sostienen Shickel et al. (2018), los avances en la interpretabilidad de los algoritmos de aprendizaje profundo son fundamentales para su adopción plena en la práctica clínica. Así, los métodos tradicionales de aprendizaje automático y basados en reglas todavía se continúan utilizando ampliamente en numerosos entornos, y la elección de enfoques híbridos y la combinación de distintas técnicas que se complementen parece la mejor solución para reforzar el rendimiento de los sistemas. En la Tabla 2 se reúnen todos los métodos y recursos empleados en los estudios analizados sobre detección y corrección automática en el lenguaje médico, de manera que resulten fácilmente comparables:

Estudio	Método	Recursos
«Misspellings in drug information system queries: Characteristics of drug name spelling errors and strategies for their prevention»	Búsqueda en el diccionario. Distancia de edición. Prueba de Friedman. Clave de similitud. Enfoque basado en la frecuencia de ocurrencia.	Aspell. Codificación Metaphone y Double Metaphone. Algoritmo Needleman-Wunsch. Paquete estadístico R.
«Spelling correction in clinical notes with emphasis on first suggestion accuracy»	Búsqueda en el diccionario. Distancia de edición. Sistema de generación de sugerencias basado en reglas.	Aspell. Herramientas de modelado de lenguaje

	Algoritmo de clasificación sensible al contexto. Modelo de lenguaje estadístico-Modelo *trigrams*.	estadístico CMU-Cambridge. Snomed-CT.
«Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes»	Búsqueda en el diccionario. Distancia de edición. Distancia de Hamming. Análisis semántico estadístico basado en datos Web.	Aspell. Servicio de sugerencias ortográficas de Yahoo!
«Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction»	Búsqueda en el diccionario. Distancia de edición. Minería de textos (*text mining*) Clasificación de nombres de entidades químicas. Entrada de diccionarios infinitos (gramáticas) para software de nombre a estructura. Distancia de Hamming. Expresiones regulares.	Aspell. CaffeineFix. Diccionarios de máquinas de estados finitos.
«Context-aware correction of spelling errors in Hungarian medical documents»	Búsqueda en el diccionario. Distancia de edición. Decodificador de traducción automática estadística (SMT). Modelado del lenguaje.	Aspell. *Moses*, kit de herramientas de traducción automática estadística (SMT).
«An ensemble method for spelling correction in consumer health questions»	Búsqueda en el diccionario. Distancia de edición. Método contextual basado en similitudes.	Word2vec. ESpell. Double Metaphone.

	Enfoque basado en la frecuencia. Edición de sugerencias basada en la distancia ortográfica y fonética.	
«Automated misspelling detection and correction in clinical free-text records»	Búsqueda en el diccionario. Distancia de edición. Modelo de canal ruidoso de Shannon. Reconocimiento de entidades nombradas (NER). Expresiones regulares. Sistema basado en reglas. Clave de similitud (*similarity key*). Etiquetado *part-of-speech.* Lista de sugerencias basada en sufijos y prefijos.	Aspell. Stanford NER. UMLS metatesauro. Versión simplificada del algoritmo Double Metaphone.
«Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care»	Búsqueda en el diccionario. Distancia de edición. Ontologías. Etiquetado *part-of-speech*.	Google Spell Checker. MedHelp. Snomed CT. RxNorm. Anotador del National Center for Biomedical Ontology (NCBO).
«Customised OCR Correction for Historical Medical Text»	Búsqueda en el diccionario. Distancia de edición. Reconocimiento óptico de caracteres (OCR). Corrección de errores basado en reglas. Enfoque basado en la frecuencia.	Diccionario Hunspell Diccionario OpenMedSpel
«Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records (EMR)»	Búsqueda en el diccionario. Distancia de edición. Etiquetado *part-of-speech*. *Stemming*. Lematización.	Aspell. Gspell. Etiquetado *part-of-speech* del corpus Brown.

	Expresiones regulares. Cálculo de frecuencia de término (TF). Semejanza de coseno. Recuperación de información. Enfoque basado en la frecuencia.	
«Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings»	Búsqueda en el diccionario. Distancia de edición. Distancia de edición Damerau-Levenshtein de 2 a partir de un léxico de referencia. Incrustaciones neuronales (*neural embeddings*) Tokenización. Clave de similitud.	Aspell. Modelo de skipgram de FastText. SPECIALIST Lexicon. Tokenizador de patrones. Diccionario de Jazzy, un corrector ortográfico de código abierto en Java. Algoritmo Double Metaphone.
«Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction»	Búsqueda en el diccionario. Distancia de edición. Ontologías. Frecuencias de trigramas.	Aspell. Snomed CT.
«Improving Spelling Correction with Consumer Health Terminology»	Búsqueda en el diccionario. Distancia de edición. Enfoque basado en la frecuencia. Análisis de n-gramas.	Diccionario CSpell. SPECIALIST Lexicon. UMLS metatesauro and MEDLINE.
«An efficient prototype method to identify and correct misspellings in clinical text»	Búsqueda en el diccionario. Distancia de edición. Frecuencias de los términos del corpus. Incrustaciones neurales (*neural embeddings*)	Word2Vec. SPECIALIST Lexicon.

«Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models»	Modelo neuronal de secuencia a secuencia (S*eq2seq*) Generación artificial de errores para entrenar el modelo *Seq2seq*. Revisión manual de una muestra. Redes de memoria a largo plazo (LSTM).	PyTorch OpenNMT-py
«Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record»	Búsqueda en el diccionario. Distancia de edición. Reconocedores de entidades nombradas y herramientas de desambiguación léxica. Reglas. Modelo de lenguaje. *Part-of-speech* (POS). Análisis de n-gramas.	UMLS Semantic Network MULTEXT Extractor de entidades nombradas.
«Low-resource OCR error detection and correction in French Clinical Texts»	Búsqueda en el diccionario. Distancia de edición. Reconocimiento óptico de caracteres (OCR). Generación artificial de errores. Redes de memoria a largo plazo (LSTM). Redes neuronales recurrentes (RNN). Generación de errores de forma artificial para aumento de datos.	Keras[53]
«Automated Misspelling Detection	Búsqueda en el diccionario. Distancia de edición	Vafa spell-checker dictionary.

[53] <https://keras.io/>

and Correction in Persian Clinical Text»	Modelo de lenguaje estadístico basado en n-gramas. Análisis de n-gramas.	Radiological Sciences Dictionary (David J. Dowsett)
«Correcting Polish Bigrams and Diacritical Marks»	Búsqueda en el diccionario. Distancia de edición. Modelo de lenguaje estadístico basado en n-gramas.	SAS 4GL. ONKO.SYS.
«Detection of spelling errors in Swedish clinical text»	Tokenización, Lematización Búsqueda en diccionario. Algoritmo basado en reglas. Revisión manual.	Swedish Parole dictionary. CST lematizador. ICD-10.
«Grammatical error correction for Spanish health records»	Redes neuronales. Experimentación con un codificador-decodificador convolucional multicapa. Generación artificial de errores para aumento de datos. Método basado en reglas para introducir errores. Revisión manual. Búsqueda en diccionario. Distancia de edición.	MeSpEn. SPACCC. Errant. fastText. Aspell. KenLM toolkit. MLConv.

Tabla 2. Métodos y recursos utilizados en detección y corrección automática en el dominio médico

3. METODOLOGÍA

3.1. Naturaleza del estudio

La orientación metodológica de este estudio se enmarca en la rama de la lingüística de corpus y la lingüística computacional. Se ha tratado y analizado un corpus sincrónico de especialidad mediante herramientas computacionales para investigar sobre los errores lingüísticos que contienen los informes médicos de diversas especialidades médicas. El método adoptado presenta la descripción formal y el análisis de los errores presentes en los informes para poder tipificarlos, encontrar las posibles causas de estos y contribuir al desarrollo de distintas soluciones para su corrección. Se trata, por tanto, de una investigación descriptiva que se fundamenta en la formulación de objetivos que permiten describir el objeto de estudio (Van Peer et al., 2012). Además, el fenómeno se estudia tal como ocurre de forma natural, sin manipulación, pues los informes están en lenguaje natural y no han sido revisados o editados previamente. En este estudio se combina la metodología cuantitativa y cualitativa para contabilizar y categorizar los errores detectados en los informes clínicos. Se van a tomar en consideración distintas características, como la frecuencia de aparición, la distancia de edición, el tipo de error, si da lugar a palabras existentes y ortográficamente correctas o no (error *real-word* o error *non-word*), la posición del error dentro de la palabra, la existencia de multierror en la palabra, la causa del error y el contexto en el que aparece.

Por tanto, el propósito principal de esta tesis doctoral es la recopilación, clasificación y análisis de errores lingüísticos presentes en informes médicos en español para contribuir al desarrollo de un módulo basado en conocimiento lingüístico, que pueda ser utilizado en sistemas de corrección automática en el dominio médico. En los corpus de especialidad es de singular importancia para el desarrollo de herramientas de corrección conocer qué tipos de errores aparecen más frecuentemente, para que sean sistematizados y abordados de la manera más apropiada, contribuyendo así al desarrollo de sistemas más robustos —mediante la generación de errores sintéticos para corpus de entrenamiento—, a la ponderación de sugerencias y a la elección de candidatos en el proceso de detección y corrección automática.

Este objetivo principal se divide a su vez en una serie de objetivos específicos entre los que se encuentran:

- Analizar el estado actual del procesamiento del lenguaje natural en el dominio médico, así como la corrección automática, tanto en el dominio general como en el dominio específico de la medicina.
- Compilar y preprocesar el corpus de estudio a partir de la recolección de informes médicos digitalizados pertenecientes a varias especialidades médicas.
- Revisar y analizar informes médicos desde el punto de vista lingüístico.
- Definir los criterios de análisis de error.
- Estudiar las principales técnicas de detección y corrección de errores *non-word* y *real-word*.
- Desarrollar una herramienta de cómputo y clasificación de errores.
- Identificar, analizar y clasificar de forma sistemática los errores contenidos en informes médicos desde un enfoque cuantitativo y cualitativo.
- Recopilar los tipos de errores más frecuentes y diseñar una tipología de errores.
- Comprobar si hay diferencias significativas entre las distintas especialidades y entre los errores en el dominio médico y el español general.
- Contribuir a la creación de conjuntos de datos de entrenamiento más exhaustivos, que incorporen casuísticas reales de lo que ocurre en los informes médicos.

Los objetivos son fundamentales en las decisiones metodológicas en cuanto a técnicas e instrumentos de investigación. Por consiguiente, en esta fase se van a decidir los instrumentos de detección y recolección de datos y el tipo de análisis. Se van a desarrollar distintas herramientas para el análisis cuantitativo y cualitativo del lenguaje, y se va a llevar a cabo un enfoque estadístico dependiente del corpus, como veremos posteriormente con más profundidad.

3.2. Descripción del corpus

El corpus objeto de estudio está constituido por una recopilación de informes médicos electrónicos en español pertenecientes a las especialidades médicas de urgencias, unidad de cuidados intensivos (UCI), psiquiatría y cirugía general.

El informe médico es el documento emitido por el facultativo médico con la información sobre un proceso asistencial prestado al paciente. En él encontramos descripción de pruebas, procesos y observaciones realizadas al paciente para obtener un diagnóstico y tratamiento apropiado. En términos de estructura y contenido, los informes presentan concreciones específicas según la especialidad, pero todos incluyen apartados como los siguientes: antecedentes, motivo de consulta, exploraciones, diagnóstico, tratamiento y pautas de actuación. Los informes presentan una estructura prototípica, muy formalizada y propia de este campo de especialidad, que se repite a lo largo del corpus. Para ello, se toma como referencia la Ley 41/2002 del Boletín Oficial del Estado, de 14 de noviembre, en el que se establecen los derechos y obligaciones de pacientes y profesionales de la salud en materia de información y documentación clínica.

Los informes de urgencias del corpus incluyen los siguientes apartados: antecedentes, diagnóstico principal, enfermedad actual, exploraciones complementarias, exploraciones radiológicas, exploración física, tratamiento de alta, y tratamiento crónico. Los informes de UCI tienen la siguiente estructura: motivo de ingreso, antecedentes, enfermedad actual, exploración física, exploraciones complementarias, juicio clínico, evolución en UCI, juicio clínico final y procedimientos realizados. El informe clínico de la especialidad de psiquiatría presenta mayor variabilidad en su estructura, pero los apartados que suelen repetirse son los siguientes: motivo de consulta, antecedentes, enfermedad actual, exploración neurológica, exploraciones complementarias, evolución, juicio diagnóstico, tratamiento y revisiones. Por último, el informe de cirugía general contiene: motivo de ingreso, antecedentes médicos y antecedentes quirúrgicos, historia actual, exploración física, resumen de pruebas complementarias, evolución y comentarios, diagnóstico principal, y tratamiento. En la Figura 1 se incorpora una muestra del corpus, extraída de la sección de exploración física de un informe de urgencias. Los informes fueron redactados a ordenador directamente por los médicos, no fueron transcritos ni revisados de forma posterior, lo que supone un referente real de errores en un contexto profesional.

```
AC ritmico sin soplos ni extratonos destacables. AP: MVC
en ambos campos, no ruidos agregados. CyO en las tres
esferas con BEG. ABD blnado y depresible doloros en
epigastrio. No irradiación. Blumber negativo, sin signos
de irritacion peritoneal. Rigidez mandibular. RHA
conservados. PPRB negativa.  MMII edemas con fovea
++.TA:130/66. FC:77 Dolor a la abdducción de la cadera,
Lassegue y Bragard neg. Pulsos conservados y simétricos.
No puntos herniarios.
E.NEUROLOGICA: sin focalidad
```

Figura 1. Muestra del corpus: sección de exploración física de urgencias[54]

Los informes clínicos están completamente anonimizados y no incluyen información sobre el centro, los pacientes y los facultativos, fechas o lugares, según establece el Reglamento General de Protección de Datos (RGPD)[55]. Debido a la confidencialidad de los datos analizados, el corpus no ofrece información específica desde el punto de vista sociolingüístico y se desconoce la procedencia, edad o sexo de los facultativos que han escrito los informes, lo que supone una de las limitaciones del estudio.

Es un corpus monolingüe y privado, cedido a la empresa Vócali[56], que lo emplea para la generación de modelos lingüísticos que se utilizan en el desarrollo de sistemas de reconocimiento de voz en el dominio biosanitario. Está formado por un conjunto de ficheros de texto plano no estructurado y no está etiquetado. Ha sido sometido a un preprocesamiento para uniformar su formato y eliminar etiquetas HTML y XML en aquellos que las contuviesen.

El corpus, compuesto por los cuatro subcorpus, contiene un total de 2 321 826 *tokens*. En la Tabla 3 se muestran los datos estadísticos del corpus desglosados según

[54] Muestra del corpus sin alterar, la versión corregida sería: «AC rítmico sin soplos ni extratonos destacables. AP: MVC en ambos campos, no ruidos agregados. CyO en las tres esferas con BEG. ABD blando y depresible doloroso en epigastrio. No irradiación. Blumberg negativo, sin signos de irritación peritoneal. Rigidez mandibular. RHA conservados. PPRB negativa. MMII edemas con fóvea ++. TA: 130/66. FC:77 Dolor a la abducción de la cadera, Lasègue y Bragard neg. Pulsos conservados y simétricos. No puntos herniarios. E. NEURÓLOGICA: sin focalidad».

[55] *Reglamento General de Protección de Datos* (RGPD): <https://gdpr-info.eu/>

[56] <https://vocali.net/>

especialidad. El término *token*[57] se suele utilizar en el campo de la lingüística de corpus para hacer referencia al número total de formas léxicas o palabras en el corpus, mientras que el concepto *type* se utiliza para aludir al número de palabras distintas que contiene el corpus. La relación entre el número de *types* y el número de *tokens* se conoce como ratio *type/token* (TTR, por sus siglas en inglés) y se utiliza para medir la riqueza léxica de una muestra. Debido a la diferencia de tamaño entre los subcorpus de urgencias y UCI con respecto a psiquiatría y cirugía general, se ha calculado con *WordSmith Tools*[58] la proporción estandarizada correspondiente para poder cotejar apropiadamente corpus con diferente longitud. En la ratio *type/token* estandarizada (STTR) se calcula cada 'n' palabras del corpus (de forma predeterminada, n = 1000) y se calcula la media de esos extractos, para lograr una proporción promedio de *type/token* a partir de fragmentos consecutivos de 1000 palabras (Chipere et al., 2004). De acuerdo con la medida, la especialidad médica que dispone de una mayor riqueza léxica es psiquiatría, debido a que los informes tienen una estructura más narrativa y un estilo más libre; en cambio, las especialidades de urgencias y cirugía general presentan una estructura mucho más telegráfica y formas análogas (López-Hernández y Almela, 2021).

Subcorpus	***Tokens***	***Types***	**Relación *Type/Token* Estandarizada (STTR)**
Urgencias	730 468	25 870	40,91
UCI	725 690	22 897	43,63
Psiquiatría	424 775	19 489	45,25
Cirugía general	440 893	14 697	42,04

Tabla 3. Datos estadísticos del corpus

Es relevante mencionar que el corpus es representativo desde un punto de vista cualitativo y cuantitativo, basándonos en criterios externos e internos. Las muestras de

[57] En este trabajo usaremos *token* como sinónimo de palabra, que es la forma más habitual de tratamiento en los estudios de procesamiento automático de la información. Por tanto, se entiende por palabra o *token* «cualquier cadena de caracteres limitada por espacios en blanco o por cualquier signo de puntuación» (Capsada Blanch y Torruella Casañas, 2017: 349).

[58] WordSmith Tools: <https://www.lexically.net/wordsmith/>

cada especialidad son representativas a nivel cualitativo teniendo en cuenta los criterios de compilación con los que fueron recopiladas y las características específicas del corpus. Asimismo, para confirmar la representatividad del corpus se llevó a cabo un análisis con el programa ReCor 1 (Figura 2), creado por Gloria Corpas Pastor, Miriam Seghiri Domínguez y Romano Maggi.

Este programa trabaja a partir del algoritmo N-Cor, que analiza la densidad léxica en relación al aumento incremental del corpus y extrae información sobre la frecuencia de los *types* y *tokens* en los distintos ficheros que lo componen (Corpas y Seghiri, 2007). En el eje horizontal se muestra, por un lado, el número de ficheros y, por otro, el número de *tokens*, y en el eje vertical el cociente, que se obtiene tras dividir el número de n-gramas distintos entre el número de n-gramas totales (Corpas y Seghiri, 2007). Se puede considerar que un corpus es una muestra representativa de la población a nivel estadístico cuando la incorporación de nuevos documentos no aporta un número significativo de *types* o palabras distintas. Gráficamente se representaría mediante el descenso exponencial de las líneas y la estabilización en valores cercanos a cero, pues gradualmente deben aparecer menos palabras nuevas.

Como se puede apreciar en los resultados reflejados en la Figura 2, el corpus, que está formado por las cuatro muestras, es representativo a nivel cuantitativo para el objetivo de este estudio. La tendencia es descendente y la incorporación de nuevas palabras o ficheros no evidencia cambios significativos y se mantiene constante a partir de 1,5 millones de tokens, por tanto, el corpus crece en tamaño, pero no en riqueza terminológica. Cada uno de los documentos o ficheros contiene un elevado número de informes clínicos, de ahí que el número de documentos sea limitado.

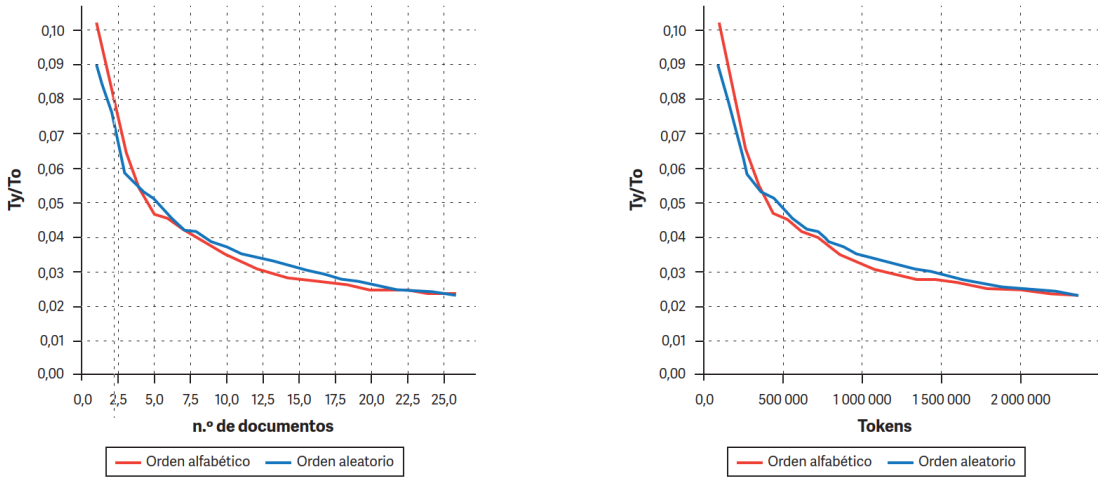

Figura 2. Análisis cuantitativo de la representatividad del corpus mediante ReCor según el número de documentos y el número de *tokens*

3.3. Criterios de análisis

Una decisión fundamental en el planteamiento metodológico fue establecer qué requisitos debían reunir las palabras del corpus para ser consideradas erróneas. Fueron contabilizadas como palabras incorrectas y que contenían un error aquellas que no cumplían las normas vigentes establecidas en fuentes de autoridad y, como consecuencia, dificultaban el procesamiento automático del corpus.

Para tal cometido tomamos como referencia los requisitos y criterios ortográficos y ortotipográficos recogidos en la normativa académica confeccionada por la Real Academia española (RAE) y la Asociación de Academias de la Lengua española (ASALE), a través de diccionarios normativos, como el *Diccionario de la lengua española*[59] en formato electrónico, el *Diccionario panhispánico de dudas*[60] (2005) y el *Diccionario de dudas y dificultades de la lengua española* (Seco, 2004); obras académicas como la *Ortografía de la lengua española*[61] (2010) y la *Nueva gramática de la lengua española*[62] (2010); manuales de estilo (Martínez de Sousa, 2008; 2012); trabajos que abordan el tratamiento de los errores y dudas más habituales del español

[59] En adelante nos referiremos al *Diccionario de la lengua española* como *DRAE*.

[60] En adelante nos referiremos al *Diccionario panhispánico de dudas* con las siglas *DPD*.

[61] En adelante nos referiremos a la *Ortografía de la lengua española* (2010) con las siglas *OLE*.

[62] En adelante nos referiremos a la *Nueva gramática de la lengua española* con las siglas *NGLE*.

(Instituto Cervantes, 2013); así como la sección «preguntas frecuentes»[63] de la página web de la Real Academia Española. Asimismo, se consultaron glosarios y obras especializadas que tienen por objetivo contribuir a la normalización del lenguaje médico, como el *Diccionario de términos médico*s (2011), desarrollado por la Real Academia Nacional de Medicina, además de revistas y recursos profesionales para la traducción y la redacción médicas (Navarro, 2015).

De igual forma, se definieron una serie de convenciones en cuanto al tratamiento de siglas, abreviaciones, errores de puntuación, uso de mayúsculas y minúsculas, unión y separación de vocablos, términos procedentes de otras lenguas y expresiones numéricas, entre otros. El proceso de detección y corrección de errores cubre aspectos ortográficos, ortotipográficos y gramaticales, pero no se tuvieron en cuenta cuestiones relativas al estilo o a la semántica textual y discursiva. Se decidió no incluir en el proceso de corrección y análisis cuantitativo la presencia de siglas no estandarizadas y abreviaturas debido a la amplitud y complejidad del problema. La desambiguación de abreviaturas en sí misma es un tópico de investigación que requiere de conocimientos específicos de medicina y que trasciende los objetivos de este trabajo. En el *Diccionario de términos médicos* (2015: XIX) se constata esta complejidad:

> La presencia de las siglas en la literatura médica es, en muchos casos, abrumadora. Su gran diversificación, unida a su uso restringido por comunidades de profesionales agrupados las más de las veces por especialidades, campos de estudio o ámbitos de trabajo, hace que resulte imposible recoger, ni siquiera de forma aproximada, todas las siglas, acrónimos y abreviaturas con sus distintos usos y significados en los límites de este diccionario.

Como veremos con más detalle posteriormente, se realizará el reconocimiento de patrones de errores mediante el desarrollo de una herramienta de extracción, cómputo y clasificación (Sección 3.4.3.). Los errores van a ser analizados sistemáticamente teniendo en cuenta los siguientes criterios:

- Frecuencia de aparición: contabiliza el número de veces que aparece el error en el corpus.

[63] Preguntas frecuentes en la sección «Español al día» de la Real Academia de la Lengua: <https://www.rae.es/espanol-al-dia/preguntas-frecuentes>.

▪ Distancia de edición: alude al número de operaciones de edición necesarias para convertir una cadena de caracteres en otra.

▪ Tipo de error: se define a partir de la operación de edición que convierte una palabra o cadena de caracteres correcta en una incorrecta. Las cuatro operaciones de edición básicas son omisión, sustitución, inserción y transposición.

▪ Subtipo de error: es un grado de especificidad mayor con respecto al anterior, pues se señala en qué tipo de carácter ocurre el error de omisión, sustitución, inserción o transposición. Se detecta si el tipo de error afecta a un carácter, signo diacrítico o espacio. En el grupo de signo diacrítico se tiene en cuenta la tilde y también otros signos como el acento grave o la diéresis.

▪ Posición del error en la palabra: lugar que ocupa en la palabra el carácter erróneo.

▪ Multierror en la palabra: cuantifica aquellas palabras que contienen más de un error, es decir, que requieren más de una operación de edición para convertirlas en correctas. Asimismo, calcula el número de errores que contienen.

▪ Error *non-word* o *real-word*: si el error da lugar a una palabra existente (error *real-word*) o idiomáticamente incorrecta (error *non-word*).

▪ Contexto de la palabra errónea: en el caso de los errores *real-word* o dependientes del contexto, resulta imprescindible conocer el contexto inmediato de la palabra para poder establecer que se trata de un error.

3.4. Procedimiento

Las distintas fases del enfoque metodológico adoptado para esta investigación se muestran en la Figura 3. En primer lugar, se llevó a cabo la compilación de un corpus de estudio formado por informes médicos de varias especialidades. Este corpus fue sometido a un preprocesamiento que abarcó la eliminación de etiquetas, la tokenización y la normalización del corpus. El proceso de normalización, que busca la homogeneización del corpus para facilitar su procesamiento, consistió principalmente en la conversión a letra de caracteres numéricos y signos de puntuación, y la expansión de acrónimos, entre otros.

Seguidamente se desarrolló el sistema para la detección y corrección de errores ortográficos (errores *non-word*), que incluyó la comparación con un lexicón, la técnica de distancia de edición mínima para la generación de candidatos y la revisión manual

asistida. Una vez detectados y corregidos los errores *non-word*, se procedió con los experimentos para la detección de errores *real-word*. De esta forma, se pretendió reducir el número de errores ortográficos presentes en el corpus que pudiesen alterar los resultados en el experimento posterior para detectar errores *real-word*. Durante este proceso, se llevó a cabo la generación de un modelo lingüístico, la representación vectorial de las palabras del corpus a partir de *Word2Vec* y el etiquetado gramatical del corpus.

Posteriormente, se desarrolló una herramienta de cómputo y clasificación con la que se realizó la clasificación sistemática de los errores detectados, junto con la creación de categorías adaptadas para estos errores y el diseño de una tipología de error específica. Finalmente, se efectuó el análisis de los resultados obtenidos. A continuación, explicaremos con más detalle cada una de las fases del procedimiento metodológico llevado a cabo para detectar y recopilar errores.

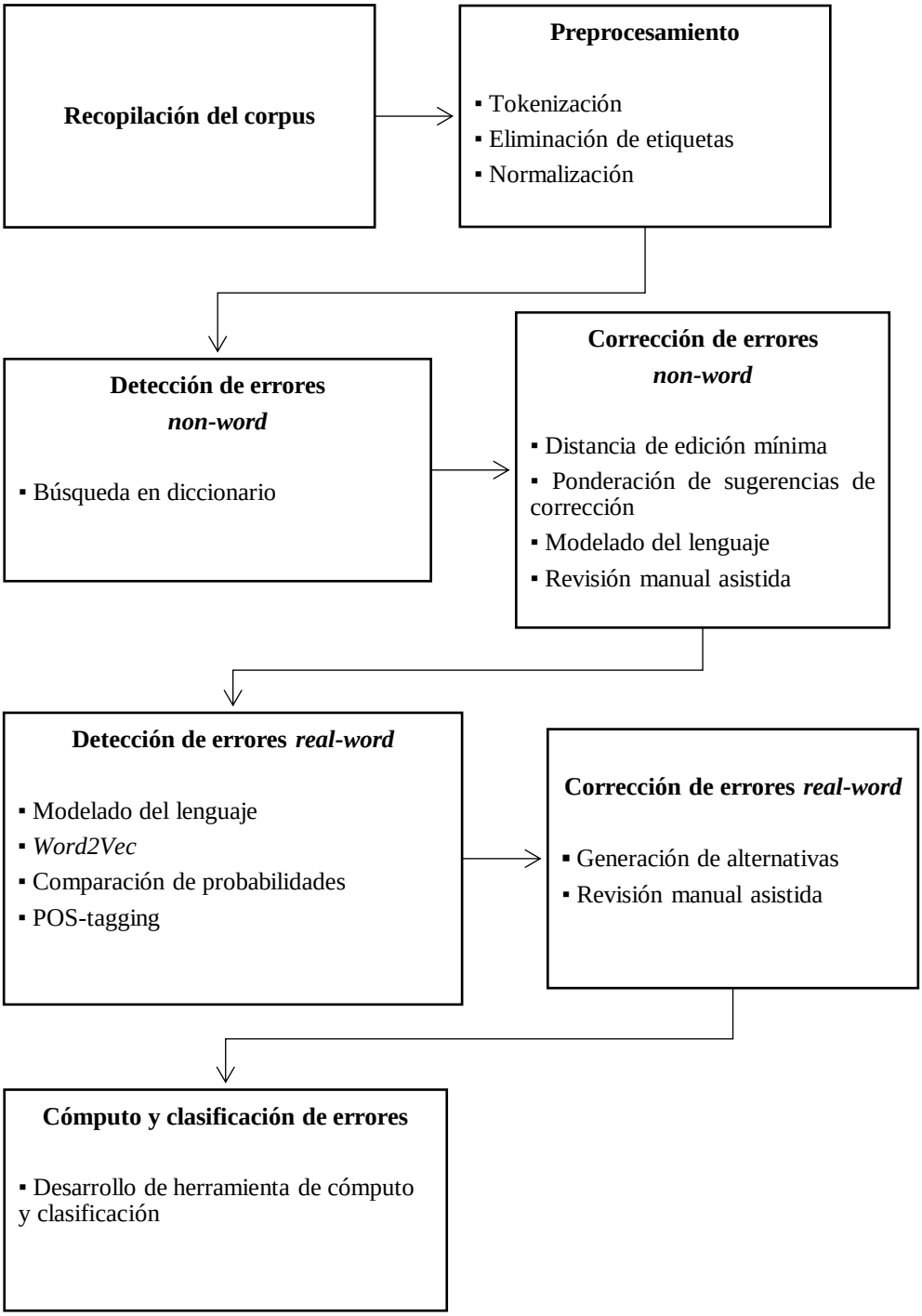

Figura 3. Fases del enfoque metodológico

3.4.1. Detección y corrección de errores *non-word*

Para la detección de errores *non-word* se utilizó la técnica de búsqueda en diccionario, que se basa en la comparación automática del corpus con un lexicón compilado por palabras previamente validadas. Las palabras del corpus que aparecen en ese listado fueron consideradas correctas, mientras que las que no aparecen en él fueron consideradas candidatas a palabra errónea. Entre las candidatas a palabra errónea se hallaron falsos positivos, es decir, palabras que eran correctas y que posteriormente se excluyeron de la lista de errores. Resulta fundamental que el lexicón tenga una gran cobertura léxica para evitar la detección de falsos positivos, como neologismos, de ahí la importancia de su continuo mantenimiento y actualización. Una de las características más distintivas del dominio médico es la abundante presencia de terminología, es decir, de voces que poseen un valor especializado y se utilizan para representar conocimientos sobre medicina, como hemos mencionado anteriormente.

En este caso el lexicón utilizado se compiló a partir del diccionario para el español de Hunspell[64], un corrector ortográfico multiplataforma de código abierto, en el que encontramos formas flexionadas y derivadas de las palabras. Como el corpus objeto de estudio pertenece al dominio médico, el diccionario de español común que incorpora Hunspell resulta insuficiente al no contener léxico especializado. Por consiguiente, es necesario añadir terminología biomédica recopilada a partir de diversas fuentes. Principalmente pueden mencionarse las nomenclaturas sistematizadas de Snomed-CT[65] y CIE-10[66], glosarios especializados y recursos léxicos obtenidos de fuentes como la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)[67], la Sociedad Española de Documentación Médica (SEDOM)[68] o Vademecum[69], artículos

[64] <http://hunspell.github.io/>

[65] Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) es una terminología clínica multilingüe distribuida por la International Health Terminology Standards Development Organisation (IHTSDO): <http://www.ihtsdo.org/snomed-ct/>

[66] Clasificación Internacional de Enfermedades (CIE-10), décima edición, correspondiente a la versión en español de la International Statistical Classification of Diseases and Related Health Problems (ICD): <https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html>

[67] Agencia Española de Medicamentos y Productos Sanitarios (AEMPS): <https://www.aemps.gob.es/>

[68] Sociedad Española de Documentación Médica (SEDOM): <http://www.sedom.es/>

[69] Guía farmacológica en español: <https://www.vademecum.es/>

especializados, así como la incorporación de léxico menos frecuente o neologismos recopilados tras la revisión y corrección de documentación clínica. Gracias a estos recursos se añadieron listados de medicamentos, siglas, principios activos, enfermedades, síntomas, procesos, técnicas, estructuras anatómicas y protocolos, entre otros.

Además, fue añadido un módulo específico con expresiones regulares[70] para la detección de errores en el empleo de símbolos y palabras compuestas con guiones.

A continuación, se llevó a cabo la fase de corrección y clasificación de sugerencias. En esta fase se asignó la corrección correspondiente a cada palabra incorrecta. Para ello, se generaron automáticamente candidatos de corrección con el objetivo de hallar una palabra correcta que estuviese recopilada en nuestro lexicón. Para la generación de sugerencias, se utilizó la técnica de distancia de edición mínima, conocida como la distancia de Damerau-Levenshtein, que alude al número mínimo de ediciones necesarias para transformar una cadena de caracteres en otra. Como ya hemos mencionado en el capítulo teórico de esta tesis, son cuatro las operaciones básicas de edición: inserción, omisión, sustitución y transposición.

Para ordenar la lista de sugerencias y que la sugerencia correcta apareciese en primera posición también se tuvo en cuenta la frecuencia de aparición de cada sugerencia en el corpus y técnicas probabilísticas, como el análisis de n-gramas. Finalmente, se validaron las correcciones obtenidas mediante la revisión manual semisupervisada y la comprobación del contexto del error en aquellos casos que lo requerían.

Resulta fundamental una primera fase de corrección de errores ortográficos (*non-word*) presentes en el corpus antes de comenzar con el proceso de detección de errores en contexto.

3.4.2. Detección y corrección de errores *real-word*

3.4.2.1. Generación del modelo de n-gramas

Una vez detectados y corregidos los errores ortográficos del corpus, llevamos a cabo la fase del experimento para detectar errores *real-word*. Para detectar este tipo de errores la búsqueda en diccionarios resulta ineficaz, pues se trata de palabras que aparecen

[70] Las expresiones regulares (también conocidas como «regex») son patrones de búsqueda que se utilizan para encontrar un determinado conjunto de caracteres en cadenas de texto (Jurafsky y Martin, 2014).

en el lexicón al ser correctas y se convierten en falsos negativos. Para la identificación de los errores en contexto hemos desarrollado una propuesta basada en el modelado estadístico del lenguaje mediante n-gramas. Un n-grama es una subsecuencia de *n* elementos de una secuencia dada (Manning y Schütze, 1999). Los modelos de lenguaje son modelos de aprendizaje automático utilizados en lingüística computacional y procesamiento del lenguaje natural, con ellos se obtiene la distribución de probabilidad de los elementos de un corpus y se fundamenta, por tanto, en una descripción probabilística de los fenómenos del lenguaje. Se pueden modelar distintas secuencias de caracteres, palabras o fonemas dependiendo del objetivo de aplicación, en este caso son *tokens*. Un modelo de lenguaje va a resultar de mayor utilidad que el cómputo de n-gramas puros obtenido con herramientas de análisis de corpus, debido a que los modelos de lenguaje tienen una noción de probabilidad condicional que hace que la evaluación de las probabilidades de ocurrencias de palabras en un texto sea más precisa.

Se utilizó la herramienta *SRI Language Modeling* (SRILM)[71], con la que es posible crear modelos estadísticos de lenguaje para diversas tareas de PLN, como reconocimiento de voz, etiquetado o traducción automática, entre otros. Se ejecutó el algoritmo para cada uno de los subcorpus, dando como resultado cuatro modelos lingüísticos de orden 3, que fueron entrenados a partir de las cuatro muestras de cada especialidad respectivamente. Los modelos generados son de orden 3 al estar formados por secuencias de unigramas, bigramas y, por último, trigramas. A pesar de que órdenes superiores, como el 4 y 5 aportan más contexto para analizar errores, se decidió trabajar con trigramas al ser la medida que mejores resultados había aportado en anteriores experimentos para reconocimiento de voz, porque una mayor amplitud conlleva una mayor posibilidad de combinación y menor viabilidad de manejo de los datos.

Los datos obtenidos tras el modelado se pueden almacenar en diversos formatos; en este experimento se ha utilizado el formato de archivo conocido como ARPA, el estándar para la representación de n-gramas en modelos de lenguaje. El fichero de salida en formato ARPA se divide en secciones de n-gramas según la longitud. En él se observa un encabezado, introducido por la palabra clave \data\, que indica el número de n-gramas que contiene el corpus analizado según la longitud (unigrama, bigrama y trigrama).

[71] *SRI Language Modeling Toolkit* es una plataforma que ofrece herramientas para crear y aplicar modelos estadísticos de lenguaje: <https://www.sri.com/case-studies/srilm/>

Seguidamente, el fichero se divide en secciones de n-gramas según la longitud y cada n-grama aparece en una línea acompañado a su izquierda de un logaritmo (en base 10) de la probabilidad condicional de ese n-grama en el corpus. Se utiliza el logaritmo en base 10 para representar las probabilidades al dar lugar a un resultado más compacto, ya que pueden darse probabilidades de aparición sumamente pequeñas. En el caso de los unigramas, se obtienen probabilidades incondicionales, mientras que los n-gramas con dos o más elementos muestran probabilidades condicionales, pues un elemento depende del otro. Al n-grama también le puede seguir opcionalmente el logaritmo en base 10 del peso del retroceso asociado. La representación del modelo concluye con la palabra clave \end\. La Tabla 4 muestra un extracto con bigramas y trigramas de un modelo de lenguaje almacenado en el formato de texto ARPA:

```
\data\
ngram 1= 42113
ngram 2= 458119
ngram 3= 1331927

\2-grams:
-2.034176   CONTRACTURA   PARACERVICAL    -0.060702
-3.211859   CONTRACTURA   PARACERVICALES  0.278013
-2.716428   CONTRACTURA   PARAESCAPULAR   -0.060173
-3.211859   CONTRACTURA   PARAESPINAL     -0.149138
-2.716428   CONTRACTURA   PARALUMBAR      0.007389
-1.467361   CONTRACTURA   PARAVERTEBRAL   0.238207

\3-grams:
-0.991717   MÚLTIPLES   INFARTOS      ANTIGUOS
-0.991717   MÚLTIPLES   INFARTOS      CEREBRALES
-0.991717   MÚLTIPLES   INFARTOS      LACUNARES
-1.116656   MÚLTIPLES   INFECCIONES   NOSOCOMIALES
-0.556790   MÚLTIPLES   INFECCIONES   RESPIRATORIAS
-1.116656   MÚLTIPLES   INFECCIONES   URINARIAS
-0.514596   MÚLTIPLES   INFORMES      DE
-2.075304   MÚLTIPLES   INGRESO       EN
```

Tabla 4. Extracto de un modelo del lenguaje

Los números más cercanos a cero indican mayor probabilidad y, por tanto, mayor número de apariciones en el corpus. Así, esta representación del lenguaje da como resultado probabilidades que reflejan con precisión la frecuencia de combinación de las palabras que integran el corpus, lo que nos va a permitir detectar combinaciones anómalas o sospechosas que deben ser analizadas para hallar posibles errores *real-word*. Una baja frecuencia de aparición del bigrama o trigrama en el modelo o una frecuencia menor de lo que sería esperable puede ser un claro indicio de que se trata de una combinación de palabras anómala y hay un error. El modelado del lenguaje nos permite hacer pruebas y generar listas con distintos umbrales, valores y ajustes.

3.4.2.2. Enfoque con *word embeddings*

En segundo lugar, se añadió un enfoque basado en *word embeddings* (incrustaciones de palabras) para llevar a cabo una representación vectorial de las palabras del corpus, contextualizarlas más allá de los trigramas y de esta forma detectar más casos de posibles errores *real-word* analizando la similitud del coseno. Los *word embeddings* se convierten en una solución efectiva para guardar información semántica sobre las palabras y sus relaciones entre sí. Para generar los *word embeddings* se utilizó *Word2Vec*[72]. Concretamente, se utilizó el algoritmo basado en la arquitectura de *skip-grams*. Los *word embeddings* fueron entrenados durante 30 iteraciones y los parámetros de configuración más relevantes fueron el tamaño de los *embeddings*, que se fijó en 200, y el tamaño de la ventana de contexto alrededor de cada palabra, que osciló en nuestros experimentos entre 3 y 6.

En la Tabla 5, se observa un fragmento de los resultados generados con *Word2Vec* a partir de la palabra «radioterapia». Las palabras que aparecen en los mismos contextos tienen una distancia más cercana a 1. De esta forma, es posible detectar combinaciones de palabras cuya distancia de coseno es alejada y son susceptibles de ser errores.

[72] *Word2Vec*: <https://www.tensorflow.org/tutorials/text/word2vec>

Tokens	**Distancia del coseno**
RT	0.651676
QT	0.544842
QUIMIOTERAPIA	0.528390
TERAPIA	0.483247
RADIOQUIMIOTERAPIA	0.480375
ADYUVANTE	0.472046
ESTEROIDES	0.454999
HISTERECTOMÍA	0.451256
QUIMIO	0.446215
TEMOZOLOMIDA	0.439546
RADIOTERÁPICO	0.438578
RDT	0.431811
MENISCECTOMÍA	0.424197
CIRUGÍA	0.422718

Tabla 5. Extracto de resultados de *Word2Vec*

Mediante las distintas pruebas con modelos lingüísticos, los *word embeddings* y el análisis de los bigramas y trigramas obtenidos detectamos posibles errores a nivel sintáctico y semántico en los informes médicos que habían pasado desapercibidos en la primera fase de corrección automática centrada en los errores ortográficos. Los casos susceptibles de ser errores fueron comprobados directamente en el corpus para asegurarnos de que no se trataran de falsos positivos.

3.4.2.3. Generación de alternativas y comparación de resultados

El siguiente paso del proceso, tras obtener los valores del modelo lingüístico y los *word embeddings*, fue generar automáticamente alternativas mediante distancia de edición mínima, y comparar los valores que estas alternativas tenían en el modelo lingüístico y en la representación vectorial. A cada n-grama se le aplicaron las distancias de edición mínima 1 y 2 (con operaciones de inserción, sustitución, omisión y transposición) para obtener bigramas y trigramas alternativos. Los nuevos bigramas y trigramas tenían que estar formados por palabras existentes, por lo que se pidió al sistema que solo devolviese los candidatos formados por palabras correctas que estaban en el

diccionario y en el corpus. De cada bigrama y trigrama original se obtuvo distinto número de alternativas, según la cantidad de palabras existentes cercanas a su grafía.

En el caso del modelo lingüístico, el objetivo de esta fase fue comprobar las puntuaciones de cada bigrama y trigrama original y de sus alternativas cercanas para ver cuáles eran más probables y si existía una diferencia significativa entre ellas. Se comparó el valor que aparecía en el ARPALM para cada uno de ellos y si la diferencia estaba por encima del umbral establecido se detectó como bigrama o trigrama susceptible de contener un error. Este umbral, que supone el valor de confianza estimado, se estableció a partir de varias pruebas y la observación de los respectivos resultados. Se probó sistemáticamente con diversos valores de diferencia entre los originales y sus alternativas y se revisaron los resultados obtenidos, de esta forma fue posible ajustar los parámetros más adecuados para detectar errores. Se extrajeron aquellos conjuntos a distancia de edición 1 y 2 con una diferencia en el valor entre 0,75 y un máximo de 3 en el caso de los bigramas y un umbral de diferencia entre 0,1 y 0,5 para los trigramas. En la Tabla 6 se muestra un ejemplo del ARPALM que refleja que aquellos bigramas con el número negativo en escala logarítmica más alejado de cero son los que indican una relación anormal y en la mayoría de casos resultaron ser palabras que contenían errores *real-word*.

ALTERACIONES RADIOLÓGICA	-3,560761
ALTERACIONES RADIOLÓGICAS	-0,020771
LEVES CAMBIO	-3,503668
LEVES CAMBIOS	-0,021522
ALTERACIONES RADIOLÓGICOS	-3,316706
ALTERACIONES RADIOLÓGICAS	-0,020771
ALTERACIONES SIGNIFICATIVOS	-4,025419
ALTERACIONES SIGNIFICATIVAS	-1,748291
CAMBIOS DEGENERATIVAS	-3,455923
CAMBIOS DEGENERATIVOS	-0,033766

Tabla 6. Casos detectados en el modelo lingüístico susceptibles de ser errores

Asimismo, se utilizaron los *word embeddings* creados para determinar qué palabras solían aparecer cerca de estas según la medida de similitud del coseno. En el espacio vectorial, aquellas palabras que suelen aparecer en los mismos contextos tienen representaciones vectoriales similares y una medida de similitud del coseno alejada de

cero. Los vectores de las palabras que componían los n-gramas originales y los n-gramas alternativos generados mediante distancia de edición se compararon con los del contexto inmediato utilizando la medida de similitud del coseno que es próxima a cero para vectores muy diferentes. De esta forma, aquellas palabras que no suelen aparecer habitualmente en ese contexto reflejaron una distancia del coseno cercana a cero, por lo que fueron señaladas como posibles errores a partir del umbral establecido. Un ejemplo es la combinación aparentemente correcta de «radioterapia ayudante», al consultar si ambas palabras suelen aparecer en los mismos contextos, el resultado fue un valor próximo a cero, pero al calcular las palabras a distancia 1 y 2 de «ayudante» y comprobar la distancia de los *embeddings* se detectó que «adyuvante» tenía una distancia del coseno mucho más cercana a 1 que «ayudante». Posteriormente, todos los casos extraídos y recopilados como susceptibles de ser errores fueron validados sobre el corpus.

3.4.2.4. Enfoque con etiquetado o *pos-tagging*

Por último, se añadió de forma complementaria al experimento el etiquetado del corpus o *POS-tagging*. Etiquetar el corpus posibilita hacer búsquedas de categorías de palabras, no de palabras concretas, lo que permite subir un nivel de abstracción en el proceso de detección y análisis. Se utilizó el etiquetador gramatical *Spanish Clinical Case Corpus Part-of-Speech Tagger* (SPACCC_POS-TAGGER)[73] para corpus del dominio médico desarrollado a través del Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)[74], que busca facilitar el acceso a recursos y herramientas de PLN e infraestructuras lingüísticas para el área de biomedicina. Concretamente, este recurso es una versión implementada de la herramienta Freeling3.1[75] adaptada para casos clínicos en español. No obstante, encontramos algunas limitaciones en su uso porque el etiquetado no fue del todo preciso en el corpus y muchas de las palabras específicas del dominio y morfológicamente más complejas no fueron correctamente etiquetadas. Por tanto, esta técnica fue utilizada únicamente de forma complementaria para detectar formas verbales que enmascararan errores o discordancias de género y número. En la Tabla 7 observamos un fragmento del corpus etiquetado:

[73] <https://github.com/PlanTL/SPACCC_POS-TAGGER>

[74] <https://plantl.mineco.gob.es/sanidad/Paginas/sanidad.aspx>

[75] Conjunto de etiquetas: <https://freeling-user-manual.readthedocs.io/en/latest/tagsets/tagset-es/>

Token	**Lema**	**Etiqueta**
estenosis	estenosis	NCFN000
aórtica	aórtico	AQ0FS0
severa	severo	AQ0FS0
hernia	hernia	NCFS000
de	de	SPS00
hiato	hiato	NCMS000
desde	desde	SPS00
hace	hacer	VMIP3S0
30	30	Z
años	año	NCMP000

Tabla 7. Fragmento del corpus etiquetado con SPACCC_POS-TAGGER

A partir de la revisión de los resultados, se establecieron patrones y mediante reglas formuladas con expresiones regulares llevamos a cabo las búsquedas de esos patrones. Entre los casos detectados con esta técnica se encuentran formas verbales de primera persona de singular que estaban enmascarando errores y realmente no eran esa forma verbal. En la mayoría de casos se había omitido la tilde y esa circunstancia había conllevado su incorrecta clasificación a nivel gramatical. Algunos ejemplos de los patrones de búsqueda utilizados son:

▪ Verbo de indicativo del presente en primera persona del singular: VMIP1S0

- Fiebre en contesto de paciente oncológico.
- Cierre de shunt porto sistémico
- Ultimo ingreso hospitalario

▪ Verbo de indicativo del presente en segunda persona del singular: VMIP2S0

- No otras enfermedades medicas conocidas.
- Epilepsia vas
- HTA desde hace 12 años con regulas control domiciliario

▪ Sustantivo masculino singular + adjetivo femenino singular: NCMS000 + AQ0FS0

- Dolor a nivel de la musculatura paracervical y en el hombro izquierda.
- Solo se toma un comprimido diaria.
- Estreñimiento crónica, edemas crónicos en MMII

▪ Sustantivo masculino singular + adjetivo masculino plural: NCMS000 + AQ0MP0

- MVC en ambos campo pulmonares sin ruidos patológicos
- DM tipo 2 en tratamiento farmacológicos con ADOs
- No hábito tóxicos conocidos.

▪ Sustantivo femenino singular + adjetivo femenino plural: NCFS000 + AQ0FP0

- No alergia medicamentosas.
- Qx: Cirugía previas de cadera, fémur y meseta tibial
- Hernia discales C5 y C4

3.4.3. Cómputo y clasificación de errores

En esta fase se desarrollaron dos herramientas. Una primera herramienta de cómputo que se basó en la búsqueda en el corpus de cadenas de texto y expresiones regulares previamente definidas para cuantificar cuántas veces ocurrían en el corpus los errores detectados anteriormente mediante las distintas técnicas. Además, los resultados fueron validados manualmente. De esta forma, se obtuvo la frecuencia de aparición de errores en cada una de las especialidades y se recopiló una lista de resultados que fue utilizada como lista de entrada para la clasificación de estos.

En segundo lugar, para la clasificación de errores se generó una herramienta cuyo funcionamiento se fundamenta en la comparación entre la cadena de caracteres errónea y la respectiva corrección otorgada para recopilar conocimiento sobre el tipo de operación de edición que ha convertido una cadena de caracteres en otra, el subtipo de operación y la distancia de edición, entre otros. A la herramienta se le pasa como input una lista de entrada con los errores y sus respectivas correcciones, con el formato «error;corrección», y esta devuelve varios ficheros de resultados en formato CSV.

Error	Corrección	Distancia de edición	Tipo de error	Subtipo de error	Carácter	Sustitución
ambos brazo	ambos brazos	1	omisión	omisión de letra	s	-
antecedentes ginecológicas	antecedentes ginecológicos	1	sustitución	sustitución de letra	a	o
días presentas	días presenta	1	inserción	inserción de letra	s	-
esta asintomático	está asintomático	1	omisión	omisión de tilde	a	á

Tabla 8. Extracto de resultados de la herramienta de clasificación de errores

La Tabla 8 muestra un extracto de uno de los ficheros de resultados tras la configuración de parámetros y ejecución de la herramienta. En este fichero aparece en cada línea la palabra errónea o combinación de palabras errónea, seguidas de la solución, la frecuencia, la distancia de edición, el tipo y subtipo de error, el carácter o caracteres afectados y su posible sustitución. También analiza la posición del carácter de la palabra afectada por el cambio. Al generar ficheros en formato CSV es posible trabajar con hojas de cálculo y filtrar los resultados dependiendo del tipo de análisis estadístico que se quiera hacer, como investigar sobre un tipo de error específico o categoría determinada.

La herramienta ha sido diseñada para que tenga capacidad de analizar errores de cualquier distancia. Uno de los grandes retos tuvo que ver con el procesamiento de multierrores, porque una mayor distancia de edición conlleva una mayor explosión combinatoria. Fue necesario plantear una solución generalizable e ideas para optimizarla, como definir el orden más natural de operaciones que llevan desde la palabra errónea a la palabra correcta para ser clasificados correctamente.

De esta forma, obtuvimos la frecuencia y descripción de errores en cada una de las especialidades. Además, se crearon matrices de confusión a partir de los resultados obtenidos en cada especialidad. Finalmente, la estructura de la herramienta ha sido desarrollada para poder ser ampliada con otras categorías en el futuro, en función del tipo de error o fenómeno que se investigue, con dimensiones extensibles y configurables.

Tiene, por tanto, posibilidad de conexión con otras herramientas de análisis de corpus, pudiendo formar parte de *pipelines*[76] de procesamiento más complejos.

[76] En informática se utiliza el término *pipeline* (tubería) para referirse a una cadena de procesos conectados de manera que la salida de cada elemento de la cadena es la entrada del próximo.

4. ANÁLISIS DE DATOS Y DISCUSIÓN

4.1. Análisis cuantitativo

A continuación, se presentan y comentan los resultados cuantitativos obtenidos tras la detección, corrección y clasificación de errores. Se han detectado un total de 76 711 errores en un corpus formado por 2 321 826 *tokens*, lo que supone una tasa de error del 3,3 %. En la Tabla 9 se constata que la presencia de errores *real-word* es notablemente inferior a la de errores *non-word* en el corpus, se han detectado un total de 76 711 errores, de los cuales 2556 (3,33 %) son errores *real-word*. Es un resultado esperable, pues los errores *real-word* son un subtipo que requiere de unas condiciones específicas, esto es, que el error provocado dé lugar a una palabra existente. Sin embargo, es un tipo de error que tiene una importancia significativa debido a los problemas de detección y, como consecuencia, de procesamiento automático que plantea.

Tipo	**Errores**	**Porcentaje de error**
Non-word	74 155	96,668
Real-word	2556	3,332
Total	76 711	100

Tabla 9. Errores totales detectados en el corpus

Los resultados de la Tabla 10, que incluye la frecuencia absoluta y relativa, reflejan los resultados obtenidos en cada especialidad. La especialidad con mayor porcentaje de errores de ambos tipos es urgencias. Este resultado puede deberse a las circunstancias que suelen darse en esta especialidad, como la limitación de tiempo para redactar los informes y la necesidad de dar respuesta inmediata al paciente. Además, es una especialidad que cubre un extenso repertorio del lenguaje médico, con una gran variabilidad de etiologías y enfermedades, lo que puede conllevar más dificultades a nivel terminológico.

Especialidad	***Tokens***	**Errores *non-word***		**Errores *real-word***	
		Frec. Abs.	**Frec. Rel.**	**Frec. Abs.**	**Frec. Rel.**
Urgencias	730 468	35 819	4,90	1271	0,174
UCI	725 690	19 890	2,74	521	0,071
Psiquiatría	424 775	7271	1,71	491	0,115
Cirugía General	440 893	11 175	2,53	273	0,062
Total	2 321 826	74 155	3,19	2556	0,110

Tabla 10. Errores *non-word* y *real-word* sobre el total de *tokens* del corpus según la especialidad médica

4.1.1. Errores *non-word*

En esta sección vamos a plasmar los resultados obtenidos específicamente en el grupo de errores *non-word* y los datos cuantitativos de cada especialidad atendiendo a los criterios definidos. Anteriormente hemos indicado que la especialidad con una mayor tasa de errores es urgencias, seguida de UCI, cirugía y, por último, psiquiatría. Los resultados muestran que el tipo de error que ocurre con un porcentaje mayor es el de omisión de tilde y la mayor parte de los errores se producen a distancia de edición 1, entre pares de caracteres con similitudes fonéticas y pares de caracteres que se encuentran adyacentes en el teclado, como abordaremos con más detalle a continuación.

4.1.1.1. Distancia de edición

En cuanto a la distancia de edición, los resultados de la Tabla 11 corroboran que la mayor parte de las palabras erróneas del corpus están a distancia 1, es decir, tienen un único error y la distancia de edición entre la palabra incorrecta (palabra origen) y la palabra correcta (palabra de destino) conlleva una única operación de edición. En la categoría de multierror se incluyen las palabras incorrectas que necesitan más de un paso u operación para llegar a la palabra meta, esto es, las palabras a distancia 2 o un número superior de su correspondiente corrección. En el total del corpus se detectan 71 735 errores (96,74 %) a distancia 1 y 2420 errores (3,26 %) a una distancia de edición superior.

	Urgencias		UCI		Psiquiatría		Cirugía General	
	Frec. abs.	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Distancia 1	34 544	96,44	19 387	97,47	6981	96,01	10 823	96,85
Multierror	1275	3,56	503	2,53	290	3,99	352	3,15

Tabla 11. Palabras con errores *non-word* según la distancia de edición

El porcentaje de errores con distancia 1 es superior al 95 % en las cuatro especialidades analizadas, por tanto, el porcentaje de multierrores en las palabras es muy limitado, por debajo del 4 % en la cuantificación total de errores e inferior al 0,2 % en el total de palabras que conforman el corpus (López Hernández y Almela, 2021). Estudios consultados avalan que, aunque el porcentaje de errores cometidos al escribir en el teclado puede ser alto en determinados entornos, cada palabra errónea no se suele desviar demasiado de la palabra correcta. Durante la escritura, las posibilidades de que el facultativo omita, inserte, sustituya o transponga más de dos caracteres son escasas. Así, la mayor parte de las palabras erróneas contabilizadas como multierror contienen dos errores.

4.1.1.2. Tipo y subtipo de error

	Urgencias		UCI		Psiquiatría		Cirugía General	
Tipo de error	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Omisión	29 283	84,77	17 957	92,62	5762	82,54	9816	90,70
Inserción	2471	7,15	733	3,78	553	7,92	620	5,73
Sustitución	1741	5,04	336	1,73	543	7,78	235	2,17
Transposición	1049	3,04	361	1,86	123	1,76	152	1,40

Tabla 12. Errores *non-word* según el tipo de operación de edición

En la Tabla 12 se reflejan los resultados de cada especialidad según el tipo de error en distancia 1. Hay un evidente predominio de los errores de omisión sobre los demás,

siendo el tipo de error *non-word* más frecuente en todas las especialidades. Urgencias, psiquiatría y cirugía general coinciden en la jerarquía (omisión > inserción > sustitución > transposición), mientras que en UCI el error de transposición se sitúa en la tercera posición y el de sustitución en cuarta posición, aunque no presentan una diferencia significativa.

Si se atiende a un nivel de especificidad superior en la clasificación de errores, en la Tabla 13 se aprecia que el subtipo de error con más presencia en el corpus es el de omisión de signo diacrítico. El segundo lugar en la jerarquía es ocupado por la omisión de letra, excepto en UCI, cuya segunda posición es ocupada por la omisión de espacio. El tercer tipo de error más frecuente en urgencias y cirugía general es el de inserción de letra, en UCI es omisión de letra y en psiquiatría es sustitución de letra.

	Urgencias		**UCI**		**Psiquiatría**		**Cirugía General**	
Subtipo de error	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Inserción de signo diacrítico	602	1,74	230	1,19	204	2,92	109	1,00
Inserción de letra	1869	5,41	503	2,59	349	4,99	511	4,72
Omisión de signo diacrítico	26 494	76,67	15 668	80,82	5049	72,33	6846	63,25
Omisión de letra	2631	7,62	1122	5,79	585	8,38	2847	26,30
Omisión de espacio	158	0,46	1167	6,02	128	1,83	123	1,14
Sustitución de letra	1726	5,0	333	1,71	529	7,59	231	2,13
Sustitución de signo diacrítico	15	0,04	3	0,02	14	0,20	4	0,04
Transposición de letra	1049	3,04	361	1,86	123	1,76	152	1,40

Tabla 13. Errores *non-word* según el subtipo de operación de edición

En la Figura 4 se muestra un gráfico de barras apiladas que permite obtener una mejor visión del conjunto de datos, debido a que la escala ha sido ajustada al 100 %.

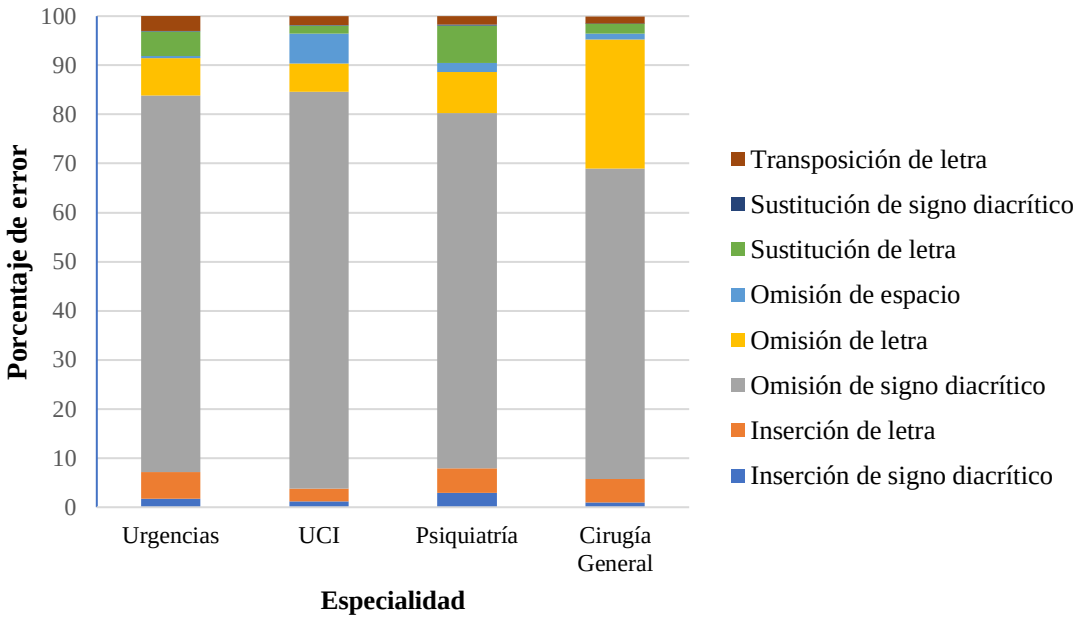

Figura 4. Errores según el subtipo de operación de edición y especialidad

Para poner a prueba la independencia de los datos obtenidos se ha utilizado el estadístico chi-cuadrado. Entre las distintas pruebas no paramétricas, la prueba chi-cuadrado es probablemente una de las más usadas, debido a sus numerosas ventajas para propósitos lingüísticos (Cantos, 2013). Se genera una tabla cruzada en la que las dos variables son de tipo cualitativo (Tabla 14), y se clasifican en especialidad médica y subtipo de error, con la intención de averiguar si ambas variables se encuentran relacionadas. Por consiguiente, con ello se pretende comprobar la hipótesis alternativa que implicaría que existe una relación entre la especialidad médica contenida en cada corpus y el subtipo de error cometido. Se codificaron los datos en el paquete de software estadístico *IBM SPSS* y se generó la tabla cruzada que incluía una tercera variable denominada frecuencia, la cual se corresponde con los valores observados. Debido a que la prueba chi-cuadrado fundamentalmente se basa en detectar posibles diferencias entre dichos valores observados y los valores esperados, la Tabla 14 incluye ambos tipos de valor.

			Especialidad				Total
			Urgencias	UCI	Psiquiatría	Cirugía	
Subtipo Error	Inserción de signo diacrítico	Recuento	602	230	204	109	1145
		Recuento esperado	551,4	309,4	111,4	172,8	1145,0
	Inserción de letra	Recuento	1869	503	349	511	3232
		Recuento esperado	1556,4	873,5	314,5	487,6	3232,0
	Omisión de signo diacrítico	Recuento	26 494	15 668	5049	6846	54 057
		Recuento esperado	26031,2	14609,4	5260,6	8155,8	54057,0
	Omisión de letra	Recuento	2631	1122	585	2847	7185
		Recuento esperado	3459,9	1941,8	699,2	1084,0	7185,0
	Omisión de espacio	Recuento	158	1167	128	123	1576
		Recuento esperado	758,9	425,9	153,4	237,8	1576,0
	Sustitución de letra	Recuento	1726	333	529	231	2819
		Recuento esperado	1357,5	761,9	274,3	425,3	2819,0
	Sustitución de signo diacrítico	Recuento	15	3	14	4	36
		Recuento esperado	17,3	9,7	3,5	5,4	36,0
	Transposición de letra	Recuento	1049	361	123	152	1685
		Recuento esperado	811,4	455,4	164,0	254,2	1685,0
Total		Recuento	34 544	19 387	6981	10 823	71 735
		Recuento esperado	34 544,0	19 387,0	6981,0	10 823,0	71 735,0

Tabla 14. Tabla cruzada de subtipo de error y especialidad

En Tabla 15 se presentan los resultados de la prueba chi-cuadrado, en la que puede observarse que su valor es de 6753,31 con 21 grados de libertad, cuyo valor de significación asintótica asociado es claramente inferior a 0,05, el nivel umbral considerado habitualmente en lingüística y ciencias sociales en general. Por tanto, puede rechazarse la hipótesis nula y concluirse que al 95 % de confianza existe relación entre las variables de especialidad médica y subtipo de error (López-Hernández y Almela, 2021).

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	6753,317[a]	21	,000
Razón de verosimilitud	5804,407	21	,000
Asociación lineal por lineal	21,696	1	,000
N de casos válidos	71 735		
a. 1 casillas (3.1 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 3.50.			

Tabla 15. Prueba de chi-cuadrado

En la Tabla 16 presentamos algunos ejemplos extraídos del corpus según el subtipo de operación. Consta de cinco bloques generales, los cuatro primeros engloban errores cuya solución necesita un único paso, es decir, existe una distancia de edición 1. El último bloque recoge algunas casuísticas detectadas en los llamados multierrores, esto es, palabras incorrectas que necesitan más de un paso u operación para llegar a la palabra meta.

Omisión	
Omisión de un carácter	mostar [mostrar][77]
Omisión de un signo diacrítico	
Tilde	perdida [pérdida]
Acento grave	Lasegue [Lasègue]
Diéresis	linguísticos [lingüísticos]
Omisión de espacio	
Espacio entre palabras independientes	esque [es que]
Espacio entre número y unidad	20cm [20 cm]
Inserción	
Inserción de un carácter	
Misma letra	dirrección [dirección]
Distinta letra	transtorno [trastorno]
Inserción de un signo diacrítico	
Tilde	fué [fue]
Diéresis	ambigüo [ambiguo]
Guion entre prefijo y raíz de la palabra	ex-fumador [exfumador]
Sustitución	
Sustitución de un carácter	arañaso [arañazo]

[77] Entre corchetes se ofrece la versión correcta de la palabra.

Sustitución de signo diacrítico	
Acento grave por acento agudo	informaciòn [información]
Diéresis por acento agudo	ülcera [úlcera]
Trasposición	
Trasposición de un carácter	sobregaregados [sobreagregados]
Trasposición de espacio	hayq ue [hay que]
Multierror	
Omisión + adición	paguina [página]
Omisión + omisión	tnel [túnel]
Omisión + sustitución	balido [válido]
Omisión + trasposición	anlagesico [analgésico]
Sustitución + inserción	hepaticoyiyunostommía [hepaticoyeyunostomía]
Sustitución + sustitución	sinbastatina [simvastatina]
Sustitución + trasposición	antihieprtebsivo [antihipertensivo]
Inserción + inserción	hepatoniesplenomegalia [hepatoesplenomegalia]
Inserción + trasposición	sobregaregadoos [sobreagregados]
Trasposición + trasposición	cerivcodosrolumbalgia [cervicodorsolumbalgia]

Tabla 16. Tipología de errores non-word según el subtipo de operación

4.1.1.3. Posición del error

En este apartado quisimos obtener información sobre las posiciones de las palabras a distancia 1 en las que con mayor frecuencia aparecen errores, para saber si era posible establecer patrones de errores según la distribución de la posición de los mismos. A diferencia de otros estudios y dominios, como abordaremos con más detalle en el apartado de Discusión (4.3.), la variabilidad y amplitud de la ventana de posiciones de caracteres con errores es mayor en nuestro análisis (Tabla 17). Esto es debido a que el léxico del lenguaje especializado médico contiene términos de mayor longitud que las palabras usadas en el lenguaje común. En las cuatro especialidades observamos ciertas similitudes (Tabla 18), el mayor porcentaje de errores se concentra entre la posición segunda y sexta de las palabras, acumulando la segunda posición el número más alto de errores en las especialidades de urgencias, UCI y psiquiatría, mientras que en cirugía es la quinta posición. Urgencias, UCI y cirugía coinciden además en tener un alto número de errores en la sexta y tercera posición. Por tanto, las posiciones que presentan más errores son 2-6-3 en urgencias y UCI, 2-3-5 en psiquiatría y 5-6-2 en cirugía.

Posición	Frecuencia absoluta	Frecuencia relativa
1	1179	1,644
2	14 755	20,569
3	9063	12,634
4	6830	9,521
5	9768	13,617
6	11 096	15,468
7	5528	7,706
8	6306	8,791
9	3710	5,172
10	1976	2,755
11	601	0,838
12	324	0,452
13	270	0,376
14	199	0,277
15	49	0,068
16	40	0,056
17	20	0,028
18	14	0,02
19	7	0,01

Tabla 17. Errores *non-word* según la posición en el total del corpus

	Urgencias		UCI		Psiquiatría		Cirugía General	
Posición	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
1	550	1,592	245	1,264	155	2,22	229	2,116
2	7954	23,026	4637	23,918	1014	14,525	1150	10,626
3	4821	13,956	2392	12,338	981	14,052	869	8,029
4	3306	9,57	1593	8,217	823	11,789	1108	10,237
5	3708	10,734	1979	10,208	928	13,293	3153	29,132
6	4998	14,469	3435	17,718	840	12,033	1823	16,844
7	2563	7,42	1629	8,403	682	9,769	654	6,043
8	3250	9,408	1961	10,115	428	6,131	667	6,163
9	1979	5,729	973	5,019	316	4,527	442	4,084
10	954	2,762	214	1,104	418	5,988	390	3,603
11	206	0,596	135	0,696	102	1,461	158	1,46

12	97	0,281	75	0,387	74	1,06	78	0,721
13	84	0,243	49	0,253	98	1,404	39	0,36
14	38	0,11	32	0,165	94	1,347	35	0,323
15	14	0,041	16	0,083	11	0,158	8	0,074
16	12	0,035	11	0,057	8	0,115	9	0,083
17	3	0,009	7	0,036	5	0,072	5	0,046
18	4	0,012	3	0,015	3	0,043	4	0,037
19	3	0,009	1	0,005	1	0,014	2	0,018

Tabla 18. Errores *non-word* según la posición en cada especialidad

Algunos ejemplos son:

Error	**Corrección**	**Tipo**	**Subtipo**	**Posición**	**Carácter**
hepaticoyeyunostomia	hepaticoyeyunostomía	Omisión	Tilde	19	´ (I → Í)
tetrahidrocannabionoides	tetrahidrocannabinoides	Inserción	Letra	18	O
espondilolistesi	espondilolistesis	Omisión	Letra	17	S
hiperparatiroidosmo	hiperparatiroidismo	Sustitución	Letra	16	O → I
hidroclorotiazuda	hidroclorotiazida	Sustitución	Letra	15	U → I
laterocervicaes	laterocervicales	Omisión	Letra	14	L
lumbociatalgía	lumbociatalgia	Inserción	Tilde	13	´ (Í →I)
histerectomia	histerectomía	Omisión	Tilde	12	´ (I →Í)
intervenidda	intervenida	Inserción	Letra	11	D
intrapareqnuimatosas	intraparenquimatosas	Transposición	Letra	10	QN → NQ
Micardisplus	Micardis Plus	Omisión	Espacio	9	
tratamineto	tratamiento	Transposición	Letra	8	NE → EN
acidopeptica	acidopéptica	Omisión	Tilde	7	´ (E →É)
adminsitración	administración	Transposición	Letra	6	SI → IS
anemía	anemia	Inserción	Tilde	5	´ (Í →I)
aparacen	aparecen	Sustitución	Tilde	5	A→ E
apedicectomizado	apendicectomizado	Omisión	Letra	4	N
epístaxis	epistaxis	Inserción	Tilde	3	´ (Í→I)
calcico	cálcico	Omisión	Tilde	2	´ (A→Á)
area	área	Omisión	Tilde	1	´ (A→Á)

Tabla 19. Ejemplos de errores *non-word* según la posición del error

4.1.1.4. Matriz de confusión

El desarrollo de la herramienta de clasificación de errores posibilitó la generación de matrices de confusión, que permiten visualizar qué carácter es sustituido por otro y con qué frecuencia. Como resultado, se identificaron los errores de sustitución más comunes y las combinaciones de caracteres involucradas. Fueron incluidos también los errores de inserción y omisión de signos diacríticos para obtener más información sobre parejas de caracteres, pero estos no han sido cuantificados como errores de sustitución. A continuación, incorporamos una matriz de confusión con el cómputo total de este tipo de errores del corpus (Tabla 20).

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	ñ	à	á	â	ä	è	é	ê	ë	ì	í	ò	ó	ö	ú	ü
a	0	0	2	0	87	0	0	1	11	0	0	2	0	0	36	0	1	1	4	0	4	0	0	0	0	0	0	1	7372	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	3	4	1	0	1	0	0	0	0	0	1	8	0	3	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	1	0	15	1	2	1	0	0	0	6	1	0	4	0	2	4	1	59	3	0	15	0	1	0	278	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	1	3	8	0	0	5	3	0	0	0	0	4	0	6	0	2	0	3	14	15	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	79	0	0	0	0	0	0	1	33	0	0	0	0	0	18	0	0	2	3	1	0	0	0	0	0	0	0	0	2	0	0	1	4149	0	1	0	0	0	0	0	0	0	0
f	0	2	0	3	0	0	4	1	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	3	5	2	0	9	0	3	0	7	0	1	0	1	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	5	0	2	0	28	0	0	0	0	0	0	0	0	0	50	0	0	2	0	0	30	1	0	0	169	0	0	0	0	0	0	0	0	0	0	1	21066	0	0	0	0	0	0
j	0	0	0	0	0	0	8	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	2	0	0	1	0	0	0	2	0	4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	3	1	0	1	1	0	0	0	0	0	2	8	0	0	0	10	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	2	1	0	0	1	0	1	0	0	0	0	0	368	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	4	7	14	0	0	1	1	0	0	0	4	471	0	0	1	0	11	12	0	0	1	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	22	0	3	0	19	0	0	0	61	0	0	1	0	0	0	9	0	2	1	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18945	5	0	0
p	1	2	6	0	0	0	1	0	0	0	0	1	1	0	16	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
q	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	1	2	1	1	4	1	0	0	0	0	0	14	1	5	1	0	0	0	15	18	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s	13	0	35	17	1	1	0	0	0	0	0	4	0	8	0	0	0	16	0	1	0	2	0	58	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	3	20	0	3	2	0	0	1	0	3	1	1	0	1	0	26	4	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

u	1	0	0	0	0	0	0	0	15	0	0	0	0	1	20	0	0	0	1	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2507	6
v	0	26	8	2	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	0	0	7	0	0	0	0	0	0	0	1	0	0	3	0	0	0	1	40	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	37	2	0	1	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z	0	0	17	2	0	0	0	0	0	0	0	1	0	0	0	0	0	1	26	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	0	0	0	0	0	0	0	0	0	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
á	229	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
â	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ä	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
è	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
é	0	0	0	0	196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ê	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ë	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ì	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
í	0	0	0	0	0	0	0	0	471	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0
ò	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0
ó	0	0	0	0	0	0	0	0	0	0	0	0	0	0	212	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ö	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	
ú	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ù	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
ü	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0

Tabla 20. Matriz de confusión para errores en informes médicos [X(error), Y(corrección)]

Los resultados obtenidos en la matriz de confusión junto con el análisis cuantitativo de los tipos de errores reflejan que la mayoría de los errores se concentran en un número limitado de pares de caracteres. Primero, los relacionados con el uso o ausencia de tilde. Seguidamente, aquellos pares de caracteres que generan confusión por su similitud fonética o por desconocimiento de las normas académicas que regulan su uso. Por último, los casos cuyo error es motivado por las posiciones adyacentes de estas letras en el teclado y que evidencian que son errores de actuación o de tipo mecánico. Los resultados de la matriz de confusión serán utilizados para el desarrollo del módulo basado en conocimiento lingüístico. La Tabla 21 muestra los patrones de error más frecuentes y algunos ejemplos:

Patrón	Ejemplos	Tipo
E → A	altereciones [alteraciones] creneoencefálico [craneoencefálico]	Confusión fonética
O → A	Citolopram [Citalopram] Alprozolam [Alprazolam]	Confusión fonética
A → E	anaxectomía [anexectomía] atelactasia [atelectasia]	Confusión fonética
I → E	Seritide [Seretide] craniotomía [craneotomía]	Confusión fonética
E → O	estonosis [estenosis] goserolina [goserelina]	Confusión fonética
O → I	tóxocos [tóxicos] cofosis [cifosis]	Adyacencia en el teclado
I → O	autilimitada [autolimitada] neutrifilia [neutrofilia]	Adyacencia en el teclado
S → C	leucositosis [leucocitosis] residivante [recidivante]	Confusión fonética
C → S	doxazocina [doxazosina] distención [distensión]	Confusión fonética
S → X	estrasistolia [extrasistolia] hallus [hallux]	Confusión fonética
X → S	herpex [herpes] extrabismo [estrabismo]	Confusión fonética
C → Z	Cyprexa [Zyprexa] Becocyme [Becozyme]	Confusión fonética
Z → C	cabezera [cabecera] enzías [encías]	Confusión fonética
I → Y	Enantium [Enantyum] Misoline [Mysoline]	Confusión fonética
Y → I	Betadyne [Betadine] Secalyp [Secalip]	Confusión fonética
V → B	fivrinógeno [fibrinógeno] varicoflevitis [varicoflebitis]	Confusión fonética / Adyacencia en el teclado[78]
B → V	bancomicina [vancomicina] cabidad [cavidad]	Confusión fonética / Adyacencia en el teclado

[78] Encontramos casos en los que no es posible saber con certeza si el error es provocado por confusión fonética al ser fonemas con similitudes o por estar en posición adyacente en el teclado. Los ejemplos prototípicos tienen que ver con el uso de B/V y M/N.

M → N	parémquima [parénquima] preferemte [preferente]	Confusión fonética / Adyacencia en el teclado
N → M	abdoninal [abdominal] desconpensación [descompensación]	Confusión fonética / Adyacencia en el teclado
L → R	dolsalgia [dorsalgia] Omeplazol [Omeprazol]	Confusión fonética
R → RR	bilirubina [bilirrubina] hipereactividad [hiperreactividad]	Confusión fonética
RR → R	aérreo [aéreo] contracturra [contractura]	Confusión fonética
C → V	cognitico [cognitivo] cercical [cervical]	Adyacencia en el teclado
G → F	ecogragía [ecografía] diclogenaco [diclofenaco]	Adyacencia en el teclado
S → A	hallszgos [hallazgos] orsles [orales]	Adyacencia en el teclado
S → D	calensario [calendario] absomen [abdomen]	Adyacencia en el teclado
T → R	dolot [dolor] femotal [femoral]	Adyacencia en el teclado
R → T	duranre [durante] gastroenteriris [gastroenteritis]	Adyacencia en el teclado
R → S	verpertino [vespertino] cuerpor [cuerpos]	Adyacencia en el teclado

Tabla 21. Patrones de sustituciones más frecuentes

4.1.2. Errores *real-word*

4.1.2.1. Distancia de edición

Al igual que ocurría con los errores *non-word*, casi la totalidad de las palabras erróneas detectadas contienen un único error y están a distancia 1 de la palabra correcta correspondiente (Tabla 22). En ninguna de las especialidades analizadas el porcentaje de multierrores es superior al 5 % en el total de errores.

	Urgencias		UCI		Psiquiatría		Cirugía General	
	Frec. abs.	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Distancia 1	1 222	96,144	511	98,080	477	97,149	264	96,703
Multierror	49	3,856	10	1,920	14	2,851	9	3,297

Tabla 22. Errores *real-word* según la distancia de edición

4.1.2.2. Tipo y subtipo de error

En la Tabla 23 aparecen cuantificados los tipos de errores que dan lugar a palabras existentes en español clasificados según la operación de edición en distancia 1. El tipo de error que se comete con más frecuencia en todas las especialidades es el de omisión, con una diferencia significativa sobre el resto. En urgencias, UCI y psiquiatría el siguiente tipo de error más cometido es el de inserción y en tercer lugar el de sustitución, mientras que en cirugía general el segundo tipo de error más frecuente es el de sustitución y posteriormente el de inserción. Por último, el error de transposición no tiene apenas presencia en el corpus, lo que nos indica que no es habitual que este tipo de error dé lugar a palabras existentes en español.

	Urgencias		UCI		Psiquiatría		Cirugía General	
Tipo de error	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Omisión	936	76,60	404	79,06	388	81,34	214	81,06
Inserción	188	15,38	67	13,11	58	12,16	17	6,44
Sustitución	95	7,77	39	7,63	31	6,5	32	12,12
Transposición	3	0,25	1	0,20	0	0	1	0,38

Tabla 23. Errores *real-word* según el tipo de operación de edición

En un nivel de especificación mayor (Tabla 24) detectamos que el subtipo de error más ampliamente cometido es el de omisión de signo diacrítico (*diagnostico principal* [*diagnóstico principal*]) y, en segundo lugar, el de omisión de letra (*ambos brazo* [*ambos*

brazos]). En urgencias, UCI y psiquiatría el tercer subtipo de error es el de inserción de letra (*cardiofrénicos liberes* [*cardiofrénicos libres*]), mientras que en cirugía general es el de sustitución de letra (*color local* [*calor local*]). En la Figura 5 se muestra una representación de barras apiladas con una escala ajustada al 100 % de los resultados de la Tabla 24, que permite una mejor visualización de los datos.

	Urgencias		**UCI**		**Psiquiatría**		**Cirugía General**	
Subtipo de error	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**	**Frec. abs.**	**Frec. rel.**
Inserción de signo diacrítico	14	1,15	4	0,78	5	1,05	1	0,38
Inserción de letra	142	11,62	53	10,37	39	8,18	13	4,92
Inserción de espacio	32	2,62	10	1,96	14	2,94	3	1,14
Omisión de signo diacrítico	623	50,98	275	53,82	292	61,22	142	53,79
Omisión de letra	285	23,32	85	16,05	80	16,77	52	19,70
Omisión de espacio	28	2,29	44	8,22	16	3,35	20	7,58
Sustitución de letra	95	7,77	39	7,63	31	6,50	32	12,12
Transposición de letra	3	0,25	1	0,20	0	0	1	0,38

Tabla 24. Errores *real-word* según el subtipo de operación de edición

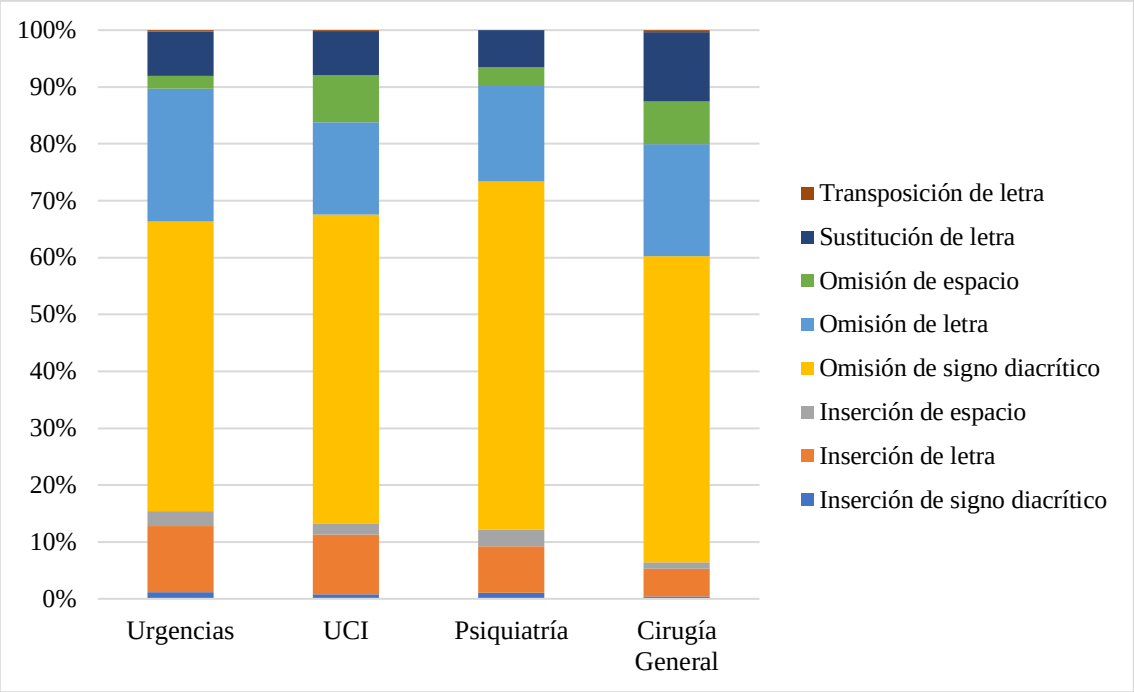

Figura 5. Errores *real-word* según el subtipo de operación de edición

También debemos hacer mención a la detección de falsos positivos durante el proceso, es decir, bigramas y trigramas cuya puntuación en el modelo lingüístico y los *word embeddings* indicaban que podía haber un error, pero en realidad eran casos correctos con poca presencia en el corpus, por lo que fueron suprimidos del listado. Algunos de los casos en los que se produjeron falsos positivos fueron aquellos en los que aparentemente había un error de ausencia de concordancia entre sustantivo y adjetivo, pero realmente se trataba de un sintagma mayor y el adjetivo concordaba con un elemento anterior. Algunos ejemplos son:

- *tratamiento específica* [traslado a unidad de tratamiento específica]
- *fractura costales* [varias líneas de fractura costales]
- *herida limpio* [exudado de herida limpio]
- *imagen compatibles* [datos de imagen compatibles con tep.]
- *masa significativos* [efectos de masa significativos sobre la línea media].
- *modo controlada* [ventilación mecánica en modo controlada]
- *múltiples diagnosticada* [encefalopatía vascular con infartos lacunares múltiples diagnosticada en abril de 2008]

- *paciente deciden* [ante la mala evolución del paciente deciden apertura del trayecto]
- *paciente están* [dado que el resto de problemas del paciente están controlados.]

Asimismo, en la Tabla 25 se incorporan algunos ejemplos de los tipos de errores detectados según la operación de edición:

Omisión	
Omisión de letra	- Poliartosis con lumbalgia cónica [crónica] - Refiere 2 ingreso [ingresos] previos - Con diuresis mantenida, con buena situación cínica [clínica]
Omisión de tilde	- Fractura de vertebra [vértebra] lumbar - Exéresis de deposito [depósito] de grasa en esclera lateral - Habito [hábito] enólico ocasional
Omisión de espacio	- Tensiones arteriales entorno [en torno] a 100/40 mmHg - Doloroso en hemiabdomen inferior sobretodo [sobre todo] FII - Con saturaciones entorno [en torno] al 95%
Inserción	
Inserción de letra	- Traslado a planta para controlo [control] y tratamiento postquirúrgico - Buena [buen] intercambio gaseoso con adecuada dinamica - Paciente muestras [muestra] rasgos destacados en esta escala
Inserción de tilde	- Última cita está [esta] mañana - Antecedentes de heteroagresividad hacía [hacia] familiares - Monitorización pulsioximétrica continúa [continua]
Inserción de espacio	- No ingresos hospitalarios a parte [aparte] de las Qx - No enfermedades médico quirúrgicas [medicoquirúrgicas] de interés - Índice cardio torácico [cardiotorácico] normal
Sustitución	
Sustitución de letra	- STOP-COLD dada [cada] ocho horas - Presenta crisis mioclónicas sobre todo en hombre [hombro] derecho - Último [última] valoración en consulta de Neumología
Transposición	
Trasposición de letra	- Al lata [alta] consciente y orientado - Se decide lata [alta] a planta - Fue dado de lata [alta] con hematoma hepático

Tabla 25. Ejemplos de errores *real-word* según el subtipo de operación

4.2. Análisis cualitativo

Además del análisis cuantitativo, es pertinente recopilar toda la información disponible a nivel cualitativo sobre la presencia de los errores en el corpus, junto con otros aspectos lingüísticos que puedan ser de utilidad para el desarrollo del módulo basado en conocimiento lingüístico y el tratamiento automatizado de estos. Por esta razón, en esta sección vamos a llevar a cabo una catalogación y descripción cualitativa de los errores detectados, además de una explicación de las posibles causas de su aparición.

En primer lugar, no podemos obviar los distintos factores que influyen en el proceso comunicativo, entre los que se encuentran el canal, el emisor, el receptor, el código o el contexto (Alcaraz y Martínez, 1997; Martinet, 1984; Moreno, 2000). Es determinante tener en cuenta el contexto situacional o conjunto de circunstancias que rodean la situación de escritura de los informes médicos y que condicionan el proceso comunicativo. Los profesionales de la salud sufren escasez de tiempo para la redacción de informes, debido a la presión asistencial provocada en muchas ocasiones por la saturación de los centros hospitalarios. Como consecuencia, el facultativo debe llevar a cabo una escritura rápida y normalmente no dispone de tiempo para una revisión posterior. Además, no suele contar con herramientas de corrección integradas en el sistema de gestión para redactar los informes —y que podrían ayudar a la detección de algunos errores ortotipográficos—, como los incluidos en la mayoría de procesadores de textos. De igual modo, es importante conocer el emisor o, al menos, el grupo en el que puede insertarse. Debido a la confidencialidad de los datos analizados, no es posible obtener más información específica, como procedencia, edad o sexo, únicamente grupo profesional. Se trata de facultativos médicos, poseen estudios superiores y un nivel formativo que se supone por encima de la población media, por lo que tienen un conocimiento lingüístico sólido. Los informes médicos han sido escritos por múltiples profesionales, por lo que la forma de redactar varía ampliamente a lo largo del corpus, con informes que presentan una escritura más depurada y que siguen en mayor grado los estándares normativos, y con otros que poseen más errores lingüísticos y rasgos idiosincráticos más marcados.

El canal utilizado es el ordenador y la comunicación escrita se realiza de forma digital mediante la redacción a través de un teclado, influyendo en gran medida en los

tipos de errores detectados. A su vez, el código es el español y el mensaje se corresponde con la información recopilada tras el proceso asistencial sanitario prestado, con unas características lingüísticas intrínsecas, como el uso de terminología médica o abreviaturas. El contexto temático, que constituye el tema fundamental en torno al cual gira el mensaje y que sustenta el desarrollo comunicativo, se enmarca en textos especializados del dominio médico, en los que se lleva a cabo el registro de antecedentes, exploraciones, pruebas, diagnósticos o tratamientos, entre otros. Por otra parte, este mensaje le será entregado a distintos receptores encargados de interpretarlo, entre los que se encuentran mayoritariamente otros profesionales sanitarios, el propio paciente, familiares o tutores de este, o incluso responsables en procesos judiciales en casos excepcionales de investigación sobre el procedimiento médico realizado.

Tras el análisis de resultados y teniendo en cuenta los factores mencionados, podemos constatar que muchos de los errores parecen estar provocados por la rapidez en el uso del teclado del ordenador al escribir, lo que implica deslizamientos y fallos de la coordinación motora. Como resultado, se pulsan dos teclas en lugar de una, se presionan accidentalmente teclas adyacentes en el teclado en lugar de las correspondientes, se omiten letras o se escriben en el orden equivocado, provocando los errores de omisión, inserción, sustitución y transposición anteriormente analizados cuantitativamente. Algunos ejemplos prototípicos son: *palapción* (palpación), *bilñateral* (bilateral), *hroas* (horas), *comrpimido* (comprimido), o *evolcuion* (evolución). Estos son, por tanto, errores de actuación que ocurren accidentalmente y están ocasionados por factores no lingüísticos, como distracciones o problemas mecanográficos.

No obstante, también han sido detectados errores susceptibles de ser considerados de motivación cognitiva o errores de competencia, como veremos posteriormente en detalle. Los facultativos se enfrentan a dudas específicas desde el punto de vista lingüístico a la hora de redactar documentación clínica en contextos académicos y profesionales. Bello (2016: 391) considera que «la formación lingüística es materia olvidada en la mayor parte de los planes de estudio y en las carreras de ciencias». Tras analizar el corpus se ha detectado que en ocasiones hay desconocimiento de la normativa lingüística ortográfica y gramatical y de las últimas reformas establecidas que rigen el funcionamiento general del español. También observamos el caso contrario, es decir, la producción de errores por un fenómeno de hipercorrección. Asimismo, se producen errores motivados por cuestiones fonéticas, por discrepancias entre norma y habla, por

analogía con palabras semánticamente relacionadas, o por confusiones lingüísticas debido a la complejidad terminológica del dominio.

Por otro lado, es destacable el uso intencional de una escritura no estándar por diversas razones. La principal suele ser la necesidad de agilizar la escritura, por lo que no consideran relevante prestar atención a la forma y a la corrección lingüística, sino únicamente al contenido.

Debido a la ingente cantidad de errores recopilados, se ha hecho especial énfasis en recopilar aquellos casos que pueden resultar más problemáticos para el procesamiento informático. A lo largo de este estudio hemos recogido información relevante sobre los distintos cotextos en los que se producen los errores, es decir, sobre el entorno lingüístico que rodea al patrón de error. También hemos detectado errores que pueden ser directamente vinculados con el dominio médico o patrones de error que pueden identificarse especialmente en ese sublenguaje. A partir del análisis de los distintos niveles de error detectados, que incluyen el nivel léxico, sintáctico, u ortotipográfico, entre otros, se organiza la casuística que describimos. Se incluyen numerosos ejemplos, además de la explicación normativa sobre la causa del error, los distintos fenómenos y la formalización de los patrones de errores.

Dicho todo lo anterior, presentamos una colección de patrones de errores que va a resultar beneficiosa tener en cuenta en el proceso de generación de errores y entrenamiento de modelos.

4.2.1. Errores *non-word*

Son numerosos los errores en el plano ortográfico que pueden ser abordados, como ya hemos constatado durante el análisis cuantitativo. Encontramos errores diversos (Tabla 26), que hemos intentado sistematizar y que iremos desglosando a lo largo de esta sección. Entre los patrones detectados, debemos mencionar los siguientes: uso erróneo de tildes, formación errónea de palabras mediante derivación y composición, escritura errónea de extranjerismos y nombres propios, simplificación de grupos consonánticos, representación gráfica de fonemas errónea, analogía con otras formas, uso equivocado de minúsculas y mayúsculas, creación y uso incorrecto de abreviaturas, y tratamiento erróneo de siglas y símbolos.

Uso de tildes	
Omisión de tilde	dias [días] ubicacion [ubicación]
Palabras monosílabas	vió [vio] fué [fue]
Demostrativos	éste [este] aquéllos [aquellos]
Grupo vocálico *-ui-*	incluído [incluido] influído [influido]
Diptongos decrecientes	terapéuta [terapeuta] hemoptóico [hemoptoico]
Acentuación de palabras parónimas	líbido [libido o lívido] éstasis [éxtasis o estasis]
Analogía con formas en plural	volúmen [volumen o volúmenes] gérmen [germen o gérmenes]
Acentuación de letras mayúsculas	QUIRURGICOS [QUIRÚRGICOS] C. ISQUEMICA [C. ISQUÉMICA]
Formación de palabras mediante derivación y composición	
Prefijos unidos con guion a la base	ex-fumador [exfumador] sub-condral [subcondral]
Prefijos separados de la raíz mediante espacio en blanco	pre menstrual [premenstrual] anti hipertensivo [antihipertensivo]
Combinación del guion con la adición de espacio	post- quirúrgica [postquirúrgica] pre- síncope [presíncope]
Alternancia de formas	postparto-posparto seudoaneurisma-pseudoaneurisma
Errores en el uso de *trans-* y *tras-*	transplante [trasplante] transtorno [trastorno]
Coordinación de prefijos	infra y supratentorial [infra- y supratentorial] pre y postraqueal [pre- y postraqueal]
Prefijos unidos a la base cuando comienza con mayúscula	anti VHD [anti-VHD] anti HTA [anti-HTA]
Inserción de guion en la construcción de compuestos univerbales	cardio-vascular [cardiovascular] colo-rectal [colorrectal]
Yuxtaposición de adjetivos con espacio	abdomino pélvica [abdominopélvica] coxo femoral [coxofemoral]
Compuestos univerbales con dos acentos	médicoquirúrgicas [medicoquirúrgicas] abdóminopélvica [abdominopélvica]
Compuestos sintagmáticos con espacio	inflamatorio infeccioso [inflamatorio-infeccioso] grisácea amarillenta [grisácea-amarillenta]
Combinación híbrida de guion y espacio	infero- posterior [inferoposterior] córtico- subcortical [córtico-subcortical]
Escritura de extranjerismos y nombres propios	
Incorrecta grafía inglesa	screaning [screening] distres [distress]

Ausencia de concordancia numeral	5 stent [5 stents]
Confusiones en la escritura de epónimos	Canal de Schlem [Schlemm] Síndrome de Schmit [Schmidt]
Errores en la escritura de medicamentos	Inhaladuo [Inaladuo] Lovibon [Lobivon]
Simplificación de grupos consonánticos	
Grupo consonántico -*bs*-	abceso [absceso] astinencia [abstinencia]
Grupo -*cc*-	infeciosa [infecciosa] micional [miccional]
Analogía con voces con la secuencia gráfica -*cc*-	delección [deleción] insercción [inserción]
Grupo consonántico -*ns*-	conciente [consciente] transplantar [trasplantar]
Grupo consonántico -*ct*-	apendicetomía [apendicectomía] laringetomía [laringectomía]
Representación gráfica de fonemas	
Uso de *h*	homalgia [omalgia] inhapetencia [inapetencia]
Fonema /rr/	alrrededor [alrededor] microrotura [microrrotura]
Uso de *b* y *v*	fotofovia [fotofobia] absorver [absorber]
Uso de *g* y *j*	vegiga [vejiga] sujestivas [sugestivas]
Uso entre *y* y *ll*	apollarse [apoyarse] maya [malla]
Fonema vibrante simple /r/ como fonema lateral alveolar /l/	dolsalgia [dorsalgia] Calvedilol [Carvedilol]
Fonemas /s/ y /z/	doxazocina [doxazosina] Balsak [Balzak]
Uso de *x*	espectoración [expectoración] epixtasis [epistaxis]
Sustitución de *e* por *a*	amigdelectomizada [amigdalectomizada] amigdelectomía [amigdalectomía]
Sustitución de *i* por *e*	alimenticeo [alimenticio] diverger [divergir]
Monoptongación	cratinina [creatinina] peritonal [peritoneal]
Cierre vocálico	cuagulación [coagulación] cuagulado [coagulado]
Analogía con otras formas	

Inserción de *l*	Borrellia [Borrelia]
Interferencia con otros paradigmas verbales	apreta [aprieta]
Inserción de *r*	discursión [discusión]
Inserción de *s*	disgresiones [digresiones]
Sustitución de *d* por *t*	curvadura [curvatura]
Sustitución de *c* por *t*	absortiometría [absorciometría]
Formación de singular en palabras invariables	carie [caries]
Uso de minúsculas y mayúsculas	
Mayúscula inicial en palabras comunes	Hipotiroidismo [hipotiroidismo] Esclerosis múltiple [esclerosis múltiple]
Mayúscula inicial en los nombres de los meses y días de la semana	Julio [julio] Lunes [lunes]
Mayúscula inicial en nombres de principios activos	Metamizol [metamizol] Adenosina [adenosina]
Nombres comerciales de medicamentos en minúscula	prozac [Prozac] nolotil [Nolotil]
Tratamiento inadecuado de mayúsculas y minúsculas después de un signo de puntuación	- Fumador de 20 cigarrillos al día. asma en tratamiento. hipotiroidismo [Asma en tratamiento. Hipotiroidismo]. - Cirugías: Apendicectomía. [Cirugías: apendicectomía]
Creación y uso de abreviaturas	
Falta de uniformidad en la creación	glu, gl, glc, gluc. [glucosa] creat., cr., CR [creatinina]
Misma abreviatura para distintos términos	bil [bilateral, bilirrubina] aprox. [aproximado, aproximadamente]
Ausencia de punto abreviativo	sd [sd., síndrome] carc [carcinoma]
Tratamiento de siglas y acrónimos	
Siglas en letras minúsculas	got [GOT] gea [GEA]
Inserción de punto abreviativo	H.T.A [HTA] A.R.N [ARN]
Añadir *-s* para marcar el plural	AINEs [AINE] TAs [TA]
Tratamiento de símbolos	
Inserción de punto abreviativo	gr. [g] kg. [kg]

Añadir *-s* para marcar el plural	kgs [kg] mls [ml]
Omisión de espacio de separación entre el símbolo y la cifra que acompaña	50mg [50 mg] 20cm [20 cm]

Tabla 26. Clasificación cualitativa de errores *non-word*

4.2.1.1. Uso de tildes

Como ya hemos constatado anteriormente, el error más ampliamente cometido a lo largo de todo el corpus es la omisión de tilde y las circunstancias de escritura nos llevan a pensar que en la mayoría de ocasiones la omisión de la tilde se produce de forma consciente. La supresión de tilde no solo se produce en palabras de mayor complejidad, sino también en aquellas de uso común y sobradamente conocidas por un hablante nativo de español (*dias** [*días*], *ubicacion** [*ubicación*]). La escritura de tilde implica pulsar una tecla más en el teclado y, dado que en muchas ocasiones se infravalora su función en el lenguaje, su omisión consciente resulta un ejemplo ilustrativo de ortografía no estandarizada motivada por esa necesidad de llevar a cabo una redacción rápida.

Asimismo, también identificamos el caso contrario, es decir, la inserción de tildes en sílabas de palabras en las que no les corresponde. Estos casos no han sido provocados por la búsqueda de una mayor rapidez en la escritura, sino por desconocimiento de la norma académica. En primer lugar, plantean dificultades aquellas voces que siguen normas especiales de acentuación, como las palabras monosílabas y aquellas que contienen tilde diacrítica. Las palabras monosílabas, es decir, de una sola sílaba, no se acentúan gráficamente, al no ser necesario marcar en qué sílaba recae la mayor intensidad prosódica. Además, a partir de la última reforma ortográfica (OLE, 2010), tampoco se deben tildar aquellas palabras monosílabas en las que todas sus vocales forman un diptongo o triptongo, y que anteriormente sí se aceptaban como excepción. Sin embargo, en el corpus aparecen con frecuencia formas como: *vió** (*vio*), *dió** (*dio*), *fué** (*fue*), *ión** (*ion*), *fuí** (*fui*), *bién** (*bien*), *pié** (*pie*), o *més** (*mes*). Únicamente deben acentuarse gráficamente aquellos monosílabos que contienen tilde diacrítica, pues en ellos se emplea la tilde diacrítica para diferenciar inequívocamente palabras con la misma grafía, pero con categorías gramaticales distintas y diferente significado (*dé/de*, *té/te*, *sí/si*, *más/mas*). Por tanto, la omisión de la tilde diacrítica en estos casos da lugar a palabras existentes

pero incorrectas en ese contexto, por lo que serán abordados con mayor profundidad en el apartado de errores *real-word*.

En segundo lugar, identificamos la inserción de tilde de forma errónea en los demostrativos y el adverbio *solo*. La RAE recomienda en su última versión de la *OLE* (2010) que el adverbio *solo* y los pronombres demostrativos *este, ese* y *aquel*, con sus respectivas formas en femenino y plural, no sean tildados, debido a que no cumplen el requisito que justifica el uso de tilde diacrítica, que es diferenciar vocablos tónicos de los vocablos átonos cuya forma era la misma. Si bien es cierto que anteriores versiones de la normativa ortográfica sí prescribían el uso de la tilde diacrítica y, a pesar de que esta última reforma data de 2010, este cambio aún no ha calado plenamente entre la población y se mantiene su uso entre muchos hablantes.

También plantean dificultades en la grafía aquellas palabras formadas por diptongos e hiatos, constituyendo su escritura una fuente común de errores en textos en español, no solo en el lenguaje médico. El grupo vocálico *-ui-* es considerado un diptongo a nivel acentual, por tanto, no deben escribirse con tilde los participios de verbos acabados en *-uir* al tratarse de palabras llanas. Como consecuencia de esta confusión algunos de los errores detectados son: *disminuído** (*disminuido*), *incluído** (*incluido*), *influído** (*influido*), *constituído** (*constituido*) o *construído** (*construido*). De igual forma ocurre con los diptongos a final de palabra, cuya pronunciación incorrecta en hiato se ha generalizado. Así localizamos también la forma *estadío*[79]* (*estadio*), cuyo uso según el DTM está muy extendido en prensa y en algunas especialidades como oncología. Sin embargo, la única pronunciación considerada correcta actualmente es la etimológica, en la que el acento prosódico se sitúa en la segunda sílaba. Igualmente sucede en formas con diptongos decrecientes como *terapéuta** (*terapeuta*), *protéico** (*proteico*), *hemoptóico** (*hemoptoico*) o *nucléico** (*nucleico*), que son tildados de forma incorrecta en palabras llanas.

Hemos detectado también voces terminadas en *-ia* (como *-fagia, -falia, -scopia, -opia, -plejia,* o *-plastia*) con acentuación alternante en el corpus. Según la *OLE* (2010:210): «esas variantes acentuales responden a la divergencia existente entre la acentuación del étimo griego y la del étimo latino». Es por esto que se encuentran

[79] <https://www.ranm.es/recomendaciones-de-la-ranm-sobre-el-lenguaje-medico/1560-estadio-no-estadio.html>

vocablos constituidos con elementos compositivos como *-plejia/-plejía*[80] (*hemiplejia/hemiplejía, paraplejia/paraplejía*) o *-scopia/-scopía*[81] (*artroscopia/artroscopía, laringoscopia/laringoscopía*), «en los que la terminación *-ía* refleja la prosodia griega y la terminación *-ia*, la latina» (*OLE*, 2010: 210). Además, la *OLE* (2010:210) señala que «existe una clara preferencia en el ámbito hispánico por la acentuación *-plejia*, mientras que en las voces que incluyen la terminación *-scopia/-scopía* las preferencias pueden variar según los términos e incluso los países». En los países americanos de habla hispana la forma con hiato (*-ía*) es la más usada, mientras que la forma con diptongo (*-ia*) es la que predomina en España. Así en el corpus la tendencia mayoritaria es la acentuación en diptongo, en consonancia con la articulación latina. Algunos ejemplos son: *hemiplejia* (38)[82]/*hemiplejía* (9), *paraplejia* (16)/*paraplejía* (2), *endoscopia* (198)/*endoscopía* (2), *colonoscopia* (263)/*colonoscopía* (3). Hallamos también voces como *epilepsia* (145)/*epilepsía* (3), tildadas posiblemente por analogía con otras voces cultas procedentes del griego como *anomalía, apoplejía* o *cirugía*.

Un caso similar en cuanto a la alternancia mencionada lo encontramos en *síndrome/sindrome**, cuya pronunciación llana parece generalizarse en países hispanohablantes de América, aunque en la actualidad aún no es una forma aceptada. En el corpus encontramos *síndrome* (1156) y *sindrome** (928), aunque en este caso al enmarcarse en la variedad peninsular, la omisión de tilde parece desencadenada de forma mayoritaria por la necesidad de agilizar la escritura, no por factores diatópicos.

Los principales diccionarios académicos y de uso consultados se inclinan por la forma diptongada en los casos anteriormente mencionados, pero es pertinente tener en cuenta la variedad diatópica a la hora de llevar a cabo el procesamiento informático de los textos, para aplicar las preferencias y cambios de forma adecuada al contexto de actuación. Es recomendable también optar por una elección y no alternar las formas con tilde y sin tilde en el mismo texto, para dotarlo de mayor homogeneidad y coherencia ortográfica. Este corpus está compuesto por informes que han sido redactados por múltiples facultativos, entre los que pueden encontrarse profesionales de otras regiones,

[80] Terminación que designa diferentes tipos de parálisis.

[81] Elemento compositivo que significa «examen, exploración, vista».

[82] Entre paréntesis se ofrece la frecuencia absoluta de aparición de la palabra en el corpus.

motivo por el que posiblemente se encuentra una pequeña alternancia en la acentuación de algunas de estas formas.

Por otra parte, detectamos palabras cuya grafía es errónea por influencia e interferencia de la acentuación de otras palabras parónimas, como *líbido** (*lívido* o *libido*), *éstasis**[83] (*éxtasis* o *estasis*), *contínuo** (*continuo* si es la forma adjetiva y *continúo* si es la forma verbal), o *colón** (*colon* o *Colón*). Un proceso similar ocurre en *psiquíatra** o *foníatra**, por analogía con *psiquiatría* y *foniatría*. Otros casos que presentan errores debido a la analogía con sus formas en plural —que sí deben llevar tilde al ser esdrújulas— son *volúmen** (*volúmenes*), *exámen** (*exámenes*), *jóven** (*jóvenes*), *resúmen** (*resúmenes*), *gérmen** (*gérmenes*) y *líquen** (*líquenes*).

Aunque la acentuación de letras mayúsculas es obligatoria —únicamente no se acentúan las mayúsculas que forman parte de siglas— en el corpus encontramos fragmentos redactados en su totalidad con letras mayúsculas que no contienen tildes, influidos quizá por la falsa creencia de que no es necesario su uso. La idea de que la escritura en mayúsculas exime del uso de la tilde ha estado extendida durante años y aún continúa presente entre algunos hispanohablantes, aunque se trata de un falso mito porque nunca se ha establecido esa norma en la ortografía académica[84].

4.2.1.2. Formación de palabras mediante derivación y composición

Son constantes los fenómenos de creación de nuevas voces, especialmente en el dominio médico. Sin embargo, la creación de nuevas palabras a partir de afijos y ciertos elementos compositivos plantea algunas dudas y vacilaciones a los hablantes. En las líneas siguientes se analizan los errores detectados en la escritura de este tipo de formas o expresiones complejas.

[83] Según el DRAE, *estasis* es el 'estancamiento de sangre o de otro líquido en alguna parte del cuerpo'. No debe confundirse con *éxtasis*, cuya acepción principal en este caso es 'estado de exaltación'.

[84] El germen de esta creencia se asocia con el uso de las antiguas máquinas de escribir e imprentas en las que se solapaba la tilde con las propias letras al insertarla, afectando a la visibilidad del texto, por lo que se renunciaba a su uso.

Primeramente, la derivación[85] prefijada, que tiene una alta frecuencia como mecanismo de formación en los informes médicos, es la que plantea mayores problemas. Los prefijos, y también los sufijos, no tienen entidad de palabra, sino de elementos afijos, pues carecen de autonomía y deben unirse a una base léxica, aportando nuevos matices significativos. Por consiguiente, constituyen una unidad morfológica y prosódica y deben escribirse siempre unidos gráficamente a la base sobre la que actúan. Sin embargo, encontramos en el corpus prefijos unidos con guion a la base que acompañan (*ex-fumador** [*exfumador*], *trans-metatarsiana** [*transmetatarsiana*], *sub-condral** [*subcondral*], o *re-polarización** [*repolarización*]) o separados de la raíz mediante espacio en blanco (*post traumática** [*postraumática*], *pre menstrual** [*premenstrual*], o *anti hipertensivo** [*antihipertensivo*]). También se han detectado casos en los que combina el uso del guion con la adición de espacio (*post- quirúrgica** [*postquirúrgica*], *pre- síncope** [*presíncope*]). Este uso puede deberse a la influencia del inglés, que utiliza el guion para la formación de palabras con determinados prefijos, como *ex-* o *self-*.

Por otra parte, hay falta de uniformidad en el uso de prefijos como *pseudo-* y *seudo-*, así como de *post-* y *pos-*. Ambas formas son permitidas, por lo que su alternancia no ha sido cuantificada como error en el corpus, pero sí es recomendable optar por una única forma en la escritura para que el texto posea una mayor consistencia ortográfica. De igual forma ocurre con los afijos *trans-* y *tras-*. Son formas que proceden del prefijo de origen latino *trans-*, siendo *tras-* la forma simplificada debido a la relajación en la articulación. La mayor parte de las palabras con el prefijo *trans-* podrán escribirse opcionalmente con *tras-*, sin embargo, hay casos en los que solo es válida la forma con *tras-* y se producen errores por analogía: *transplante** (*trasplante*), *transplantado** (*trasplantado*), *transladar** (*trasladar*), o *transtorno** (*trastorno*).

Si los prefijos aparecen mencionados de forma coordinada junto a la voz que resulta de la unión de otro prefijo a esa misma raíz deben escribirse con un guion pospuesto (*anti-*, *multi-*, *pre-*, *super-*, *bio-*...) para señalar que no son palabras autónomas, sino elementos que deben ser interpretados semánticamente teniendo en cuenta la base léxica: *espacios pre** [*pre-*] *y postraqueal.* En este dominio la coordinación de prefijos

[85] Se denomina derivación al proceso de formación de nuevas palabras mediante la inserción de sufijos o prefijos.

suele ser frecuente para generar derivados con la misma base pero que indiquen un sentido opuesto, como *pre-* y *post-*, *infra-* y *supra-*, *anti-* y *pro-* o *extra-* e *intra-*.

Por otro lado, los prefijos deben unirse con guion a la base cuando esta empieza por letra mayúscula, por tanto, se utiliza este signo cuando la palabra es una sigla o un nombre propio porque tienen mayúscula inicial (*anti-NMDA*). De igual forma, también debe emplearse el guion cuando la base es un número. Algunos ejemplos son:

- *Se objetiva presencia de anticuerpos anti VHD** [*anti-VHD*] *positivo.*
- *Pastillas anti HTA** [*anti-HTA*].

El adverbio *no*[86] es usado de forma similar a los prefijos privativos cuando se usa antepuesto a ciertos sustantivos para expresar la realidad contraria. Según la *OLE* (2010: 542) el adverbio *no* en «este tipo de expresiones mantiene la tonicidad característica», por lo que debe escribirse separado de la base. En el corpus encontramos algún caso en el que aparece unido a la palabra que precede:

- *Fenómeno de no-reflujo* [*no reflujo*].
- *Linfoma no-Hodgkin* [*no Hodgkin*] *en progresión.*

En suma, los prefijos deben aparecer unidos gráficamente a la raíz cuando la base es univerbal; soldados con guion cuando la base es una sigla, una palabra con inicial mayúscula o un número; y separados con espacio cuando la base es pluriverbal.

Hemos detectado que, además de en la prefijación, los errores relacionados con la inserción del guion y espacio en blanco son un patrón presente en la construcción de palabras mediante composición[87]. El español rechaza la yuxtaposición de adjetivos con espacio y sin nexo (*abdomino pélvica** [*abdominopélvica*], *coxo femoral** [*coxofemoral*], *inguino escrotal** [*inguinoescrotal*]), de modo que en estos casos es necesario unir las

[86] Este tipo de uso del adverbio *no* se extendió a lo largo del siglo XX, especialmente en el lenguaje periodístico, muy probablemente por influjo del inglés y el francés, idiomas en que el elemento *non-* funciona como un prefijo muy productivo (OLE, 2010: 542).

[87] La composición es «el proceso morfológico por el que dos o más palabras o raíces léxicas (entre las que se incluyen también los elementos compositivos de origen grecolatino) se unen para formar conjuntamente un término nuevo, denominado palabra compuesta o compuesto» (*OLE*, 2010: 524).

dos bases o recurrir al guion (OLE, 2010: 413). De esta forma se pueden generar compuestos univerbales (*maxilofacial*) y compuestos sintagmáticos (*maxilar-facial*). No obstante, hay voces que funcionan a efectos prosódicos y morfológicos de la misma forma que las palabras simples y su grafía debe ser siempre unitaria, por tanto, no es correcto escribir sus componentes por separado ni unidos con guion (*cardio-vascular** [*cardiovascular*], *colo-rectal** [*colorrectal*]). El compuesto univerbal consta de un solo acento, mientras que el sintagmático consta de dos acentos y las voces mantienen su propia independencia gráfica y acentual. En el caso de los compuestos sintagmáticos, las palabras deben estar unidas al guion, sin dejar espacio entre ellas: *inflamatorio infeccioso** (*inflamatorio-infeccioso*), *grisácea amarillenta** (*grisácea-amarillenta*), *distímico depresivo** (*distímico-depresivo*).

También encontramos la combinación híbrida de guion y espacio en la formación de las palabras compuestas de forma errónea: *infero- posterior** (*inferoposterior*), *córtico- subcortical** (*córtico-subcortical*), *torácico- costal** (*torácico-costal*).

Otro error detectado en los compuestos univerbales es la acentuación gráfica de las dos raíces léxicas que se unen, a pesar de que formen una única voz y, por tanto, un solo grupo acentual (*médicoquirúrgicas** [*medicoquirúrgicas*]). Como indica la *OLE* (2010: 525): «a efectos prosódicos, constituyen un solo grupo acentual, es decir, poseen un único acento léxico o primario, que es el que corresponde al último de sus componentes».

4.2.1.3. Escritura de extranjerismos y nombres propios

Otro de los patrones detectados tiene que ver con la escritura de extranjerismos y nombres propios. Este dominio se caracteriza por su complejidad léxica, debido a la elevada densidad terminológica que posee, con un ingente repertorio de enfermedades, procesos, signos, técnicas, medicamentos y un largo etcétera de elementos especializados. Asimismo, de forma constante surgen nuevos términos para dar nombre a nuevos descubrimientos, tratamientos, instrumentos, productos y a todos los progresos que hacen avanzar la medicina.

Como ya hemos mencionado anteriormente, el lenguaje médico en español está muy influenciado por el inglés, que es la lengua dominante actualmente en la transmisión de conocimiento en el ámbito científico y de la medicina en particular (Navarro, 2001).

Se ha dado una creciente influencia de esta lengua sobre el léxico en los últimos años y en los informes encontramos apartados, como el de pruebas, que contienen gran cantidad de anglicismos. No obstante, el surgimiento de neologismos originados por el progreso de las distintas ramas de la medicina no siempre se produce de acuerdo con las normas y características de la lengua española. En el caso de los extranjerismos, el *DTM* suele recomendar la adaptación de estas voces al español siempre que sea posible. Por ejemplo, en el caso del anglicismo *stent*[88] (y *stents* en su forma en plural), aconseja la forma *estent* y el plural *estents*. En este trabajo no han sido cuantificados como error los anglicismos detectados, únicamente aquellos casos que presentaban una incorrecta grafía inglesa o una adaptación errónea a las normas ortográficas del español. Algunos ejemplos destacables son: *stres** (*stress*), *screaning** (*screening*), *standar** (*standard* o *estándar*) *fluter** (*flutter*), *distres** (*distress*), *cinnámico** (*cinámico* o *cinnamic*), *bypas** (*bypass* o *by-pass*), *sten** (*stent*), *flaping** (*flapping*), *pach** (*patch*), *scaner** o *sscaner** (*scanner* o *escáner*), entre otros. Algunas oraciones extraídas del corpus son:

- *Fluter** [*flutter*] *auricular atípico.*
- *Cirugía cardiaca con implante de bypas** [*bypass*] *de safena a DA.*
- *Se implanta sten** [*stent*] *Titan de 3.5 x 28 mm.*
- *Anastomosis arterial a pach** [*patch*] *de art. mesentérica.*

Constatamos que en ocasiones en la escritura de nombres propios por influencia del inglés se generan híbridos que son incorrectamente lexicalizados. Se imitan erróneamente las grafías de otros anglicismos, como la terminación *-ing* (*Hodking** [*Hodgkin*]) o el uso de doble *n* (*Ennantyum** [*Enantyum*]. Un ejemplo prototípico es la adición de doble *s* por influencia de la grafía inglesa: *Parapress** (*Parapres*), *Douglass** (*Douglas*), *Coropress** (*Coropres*), *Lassegue** (*Lasègue*) o *Capgrass** (*Capgras*).

[88] «Estent, mejor que stent». Artículo publicado en *Recomendaciones de la RANM sobre el lenguaje médico*.
<https://www.ranm.es/terminolog%C3%ADa-m%C3%A9dica/recomendaciones-de-la-ranm/4038-estent-mejor-que-stent.html#:~:text=Estent%2C%20con%20e%20inicial%20y,abierto%20un%20vaso%20previamente%20estenosado'>.

- *Drenaje: Penrose en Douglass** [*Douglas*].
- *Delirio de Capgrass** [*Capgras*].

Asimismo, observamos que en ocasiones en los sintagmas que contienen anglicismos se produce ausencia de concordancia numeral, utilizándose la forma en singular en lugar de la correspondiente en plural:

- *ACTP primaria con colocación de 5 stent** [*stents*] *en CD.*

Paralelamente, hemos podido reconocer que se producen numerosas confusiones en la escritura de epónimos, es decir, en términos médicos cuyo origen radica en el nombre propio de una persona. Suponen una compleja nomenclatura debido a la amplitud y la diversidad de este tipo de términos en ciencias de la salud, mayoritariamente suelen ser nombres procedentes de otras lenguas y en ocasiones son términos con formas muy parecidas, lo que plantea múltiples errores en su escritura. Algunos ejemplos son:

- *Células de Schwan** [*Schwann*]
- *Canal de Schlem** [*Schlemm*]
- *Síndrome de Schmit** [*Schmidt*]
- *Esófago de Barret** [*Barrett*]
- *Síndrome de Barter** [*Bartter*]
- *Síndrome de Guillain-Barre** [*Barré*]
- *Síndrome de Mallory-Weis** [*Weiss*]
- *Anillo de Schatzky** [*Schatzki*]
- *Signo de Lassegue** [*Lasègue*]

Las dificultades mencionadas en la escritura de anglicismos y epónimos también se detectan en los nombres de medicamentos y productos. Observamos errores de inserción de letra (como *h*, *r* o *m*), de sustitución (como *m*-*n*, *v*-*b*, *y*-*i* o *s*-*c*), o de omisión, por confusión de sonidos o por analogía con otras formas, entre otros. Entre los casos detectados encontramos:

- *Inhaladuo** [*Inaladuo*]
- *Diamben** [*Dianben*]
- *Lovibon** [*Lobivon*]
- *Lecardip** [*Lercadip*]
- *Spacmotyl** [*Spasmoctyl*]
- *Sandimun** [*Sandimmun*]
- *Omeoprazol** [*Omeprazol*]
- *Enconcor** [*Emconcor*]
- *Tranquimazin** [*Trankimazin*]
- *Lirica** [*Lyrica*]
- *Novorrapid** [*Novorapid*]

También es destacable la variabilidad en la acentuación de nombres de medicamentos, siendo una de las dificultades específicas de este dominio, como también confirman estudios anteriores (Navarro, 2001; 2015). Muchos de ellos no aparecen tildados (*Cemidon** [*Cemidón*]), y entre las causas que se pueden aducir se encuentran la mayor rapidez en la escritura, como ya hemos mencionado, pero también la complejidad que plantea su correcta acentuación al no existir consenso sobre la acentuación de fármacos en las fuentes oficiales y el amplio repertorio léxico existente. También encontramos casos con tilde que siguen las reglas de acentuación en español, pero el nombre oficial con el que aparecen registrados como fármaco no la lleva (*Primperán** [*Primperan*], *Utabón** [*Utabon*]). En páginas oficiales de consulta como Vademecum o AEMPS aparecen mayoritariamente en mayúsculas y sin tilde. Por último, también se detectan casos con tilde y sin tilde dependiendo de la farmacéutica que los suministra, así encontramos casos como *Irbesartan Sandoz-Irbesartán Normon* o *Urbasón-Urbason*.

4.2.1.4. Simplificación de grupos consonánticos

La articulación como grupo silábico de muchos grupos consonánticos plantea dificultades en su pronunciación, por lo que esta tiende a relajarse y se produce el surgimiento de variantes gráficas simplificadas por analogía con el habla. Esta simplificación considerada errónea la encontramos en varios grupos consonánticos, como *-bs-*, *-cc-* o *-ns-*.

En el grupo consonántico *-bs-*, que aparece en posición de final de sílaba en vocablos que provienen del latín, la articulación del sonido /b/ tiende a relajarse, provocando la reducción del grupo *-bs-* a *-s-*. Da lugar así a variantes gráficas simplificadas que son incorrectas, como *astinencia** (*abstinencia*), *abceso** (*absceso*) e incluso al erróneo *acceso* con el sentido de *absceso*.

En segundo lugar, observamos en el corpus la confusión que plantea la escritura de una o dos letras *c* en muchas palabras. Los fonemas que se representan con la secuencia gráfica *-cc-* suelen plantear dificultades, debido que el primer fonema tiende a debilitarse en la pronunciación rápida y relajada, lo que lo hace poco perceptible. Así encontramos casos en el corpus como: *obstrución** (*obstrucción*), *abdución** (*abducción*), *afeciones** (*afecciones*), *infeción** (*infección*), *infeciosa** (*infecciosa*), *micional** (*miccional*), *ocipital** (*occipital*), *reprodución** (*reproducción*), *reseción** (*resección*), *restrición** (*restricción*), *seción** (*sección*), o *sobreinfeción** (*sobreinfección*).

De forma paralela, observamos el caso contrario, con voces que deben ser escritas con una sola letra *c* pero por analogía con otras voces se escriben con la secuencia gráfica *-cc-*. Entre ellos hemos detectado: *replección** (*repleción*), *deplección** (*depleción*), *ablacción** (*ablación*), *delección** (*deleción*), *insercción** (*inserción*), *relaccionado** (*relacionado*), *secrección** (*secreción*), *sujección** (*sujeción*), *toráccico** (*torácico*), o *concrección** (*concreción*). Vemos que mayoritariamente los errores sobre el uso de *-cc-* se agrupan en las palabras terminadas en *-cción/-ción*.

Por su parte, el grupo consonántico *-ns-* cierra sílaba y también tiende a relajarse la articulación de una de sus consonantes en el plano oral. Podemos mencionar la confusión en formas como *conciente** (*consciente*). Además, como mencionábamos anteriormente, también detectamos el caso contrario, la adición de la consonante *n* de forma errónea por analogía con otras formas (*transtorno** [*trastorno*], o *transplantar** [*trasplantar*]). Esa relajación en la articulación de sonidos la observamos también en el grupo *-ct-* en casos como *apendicetomía** (*apendicectomía*) o *laringetomía** (*laringectomía*).

Por último, la pronunciación de ciertas secuencias con *r* también plantea dificultad, lo que se traduce en la supresión de ciertos sonidos y la utilización de formas simplificadas en casos como *postpandial** (*postprandial*), *protusiones** (*protrusiones*), o *protombina** (*protrombina*).

4.2.1.5. Representación gráfica de fonemas

Se detectan patrones de errores en el corpus que tienen que ver con la correcta representación gráfica de fonemas y, por tanto, la escritura de ciertas consonantes y vocales. Se producen confusiones por analogía con otras formas, por fenómenos de hipercorrección o por desconocimiento de la forma correcta. Podemos observar que en gran parte de estos casos no se trata de sustituciones de letras adyacentes en el teclado, como puede examinarse en la matriz de confusión.

En primer lugar, se identifican errores relativos al uso de *h.* Esta letra no posee correlato fónico en el español estándar, por lo que puede implicar una mayor complejidad de uso. En el corpus se encuentran grafías con *h* en palabras que no deben llevarla, especialmente al inicio de palabra: *homalgia** (*omalgia*), *habundante** (*abundante*) o *inhapetencia** (*inapetencia*).

El fonema /rr/ puede representarse de forma gráfica mediante dos formas, la letra *r* y con el dígrafo *rr,* y el uso de una u otra forma depende de la posición que el fonema tenga en la palabra. En posición de inicio de palabra y detrás de consonante que pertenece a la sílaba anterior se utiliza la letra *r* en representación del fonema /rr/. Sin embargo, en el corpus hay formas como *alrrededor** (*alrededor*) o *enrrojecimiento** (*enrojecimiento*). Por su parte, se utiliza el dígrafo *rr* para representar el sonido vibrante múltiple /rr/ en posición intervocálica, también en voces prefijadas y compuestas, pero es fuente de numerosas vacilaciones y dudas. Por ejemplo:

- *Normoreactivas** (*normorreactivas*)
- *Prerenal** (*prerrenal*)
- *Microrotura** (*microrrotura*)
- *Cardiorespiratoria** (*cardiorrespiratoria*)
- *Normorefléxico** (*normorrefléxico*)

Detectamos errores relacionados con el uso incorrecto de *v* y *b* en el corpus. Se producen abundantes confusiones entre una letra y otra en la escritura de muchas palabras, debido a que ambas representan el fonema bilabial sonoro /b/ en español. La presencia de una u otra en las palabras depende de criterios etimológicos, y también antietimológicos en algunos casos, por lo que la grafía de cada palabra depende del conocimiento implícito

sobre la misma que posee el hablante. Así, encontramos casos como *obnuvilación** (*obnubilación*), *fotofovia** (*fotofobia*), *absorver** (*absorber*), *bancomicina** (*vancomicina*), *provable** (*probable*), *fivrinógeno** (*fibrinógeno*), *parabertebral** (*paravertebral*), o *neoadyubante** (*neoadyuvante*).

La confusión entre el uso de *g* y *j* también tiene presencia en el corpus y conlleva errores al seleccionar el grafema correcto (*ingurjitación** [*ingurgitación*], *enrogecida** [*enrojecida*], *vegiga** [*vejiga*], *sujestivas** [*sugestivas*], o *vajinal** [*vaginal*]).

De igual forma ocurre con el uso entre *y* y *ll*, siendo más frecuente la sustitución de *ll* por *y* (*apollarse** [*apoyarse*], *maya** [*malla*]). También ocurren fenómenos de lambdacismo, que consiste en la articulación del fonema vibrante simple /r/ como el fonema lateral alveolar /l/ en posición implosiva y de final de palabra, en voces como *talsalgia** (*tarsalgia*), *dolsalgia** (*dorsalgia*), *Calvedilol** (*Carvedilol*) o *Glurenol** (*Glurenor*).

La representación gráfica de los fonemas /s/ y /z/ plantea numerosas dudas en el corpus, que se traducen en errores en el uso de las letras *s*, *c* y *z*. Estas dudas están en estrecha relación con factores fonéticos y su traslado a la escritura. El fonema /s/ posee en español varias formas de articulación fonética que son comunes a todo el ámbito hispánico y otras que son exclusivas de aquellas áreas en las que se dan los fenómenos del seseo o del ceceo. En el corpus encontramos errores en la representación gráfica de *s* influidos por ceceo: *doxazocina** (*doxazosina*), *ocaciones** (*ocasiones*), *nauceas** (*náuseas*), *tarzalgia** (*tarsalgia*), *supreción** (*supresión*). También detectamos el fenómeno contrario, el uso de *s* en lugar de *z* o *c*: *hallasgos** (*hallazgos*), *livides** (*lividez*), *leucositosis** (*leucocitosis*), *rigides** (*rigidez*), *pinsamientos** (*pinzamientos*), *Balsak** (*Balzak*).

La *x* suele articularse como /s/ cuando se produce una pronunciación relajada, lo que hace que los hablantes duden de cuándo escribir una u otra letra en la grafía de muchas palabras. Algunos ejemplos son: *espectoración** (*expectoración*), *Rexer Flax**, *Reser Flas** o *Reser Flash** (*Rexer Flas*), *Zypresa** (*Zyprexa*), *epixtasis** (*epistaxis*), *hallus valgus** (*hallux valgus*).

Los errores en la representación gráfica de fonemas también ocurren con vocales. Debido a procesos de asimilación fonética que se trasladan a la escritura se reemplazan vocales como *a* por *e* (*amigdelectomizada** [amigdalectomizada]) o *i* por *e* (*alimenticeo** [*alimenticio*]), (*diverger** [*divergir*]). En los diptongos también se produce

monoptongación en formas como *cratinina** (*creatinina*) o cierre vocálico (*cuagulacion** [*coagulación*]).

4.2.1.6. Analogía con otras formas

En este apartado comentamos algunos casos destacables de errores provocados por el fenómeno de analogía y que no pueden ser incluidos en las secciones anteriores.

En primer lugar, el género de bacterias llamada *Borrelia*, de la familia Borreliaceae, es erróneamente escrito como *Borrellia** por analogía con otros géneros de bacterias que poseen formas con *ll* y están relacionadas semánticamente, como *Brucella*, *Klebsiella* o *Shigella*.

También hay errores provocados por la interferencia con otros paradigmas verbales, como en la forma *apreta**, del verbo irregular *apretar*, cuya forma apropiada en tercera persona de presente debe mantener la diptongación (*aprieta*). En los verbos irregulares no se siguen los modelos comunes de conjugación y debe tenerse en cuenta que los lexemas presentan alteraciones, como alternancias fonéticas o heteróclitas.

Otros errores que parecen tener una motivación cognitiva son casos como *discursión** en lugar de *discusión*, debido posiblemente a la interferencia con la voz *discurso* (que a su vez procede del verbo *discurrir*); o la voz *disgresiones** en lugar de *digresiones*, provocado por la interferencia del término *digresión* con el prefijo *dis.* Es también digna de mención la forma híbrida *curvadura**, creada erróneamente por el cruce de las formas *curvatura* y *corvadura.* De igual manera ocurre con *absortiometría**, escrita erróneamente por analogía con formas como *densitometría*, siendo la forma correcta *absorciometría.* También hay fenómenos de transposición silábica, como *hidrosaluteril** en lugar de *hidrosaluretil.*

Por último, otro error presente en el corpus tiene que ver con la formación del plural en español, que suele fijarse añadiendo *-s* o *-es*, lo que hace pensar que todas las formas acabadas en *-es* están en plural. Sin embargo, la voz *caries*, del latín *caries*, es invariable en plural, por lo que no existe la forma en singular *carie**.

4.2.1.7. Uso de minúsculas y mayúsculas

Son diversos los errores relativos al tratamiento de minúsculas y mayúsculas que han sido detectados. En los informes clínicos detectamos principalmente un uso excesivo de mayúsculas. En primer lugar, la mayúscula inicial es utilizada en palabras y expresiones comunes que no deben llevarla para dotarlas de un mayor carácter enfático. Se trata de una peculiaridad en este dominio, con especial presencia en los informes médicos, y con la que se busca llamar la atención del receptor para que con un solo vistazo se percate de esa palabra, dotándola de mayor relevancia semántica con respecto al resto del informe. En la decisión de escribir la palabra con mayúscula inicial prevalece únicamente el criterio subjetivo del autor. Este fenómeno es conocido como mayúscula «de relevancia» o «enfática», como ya constatan otros autores en estudios previos (Aguilar, 2013a; Martínez de Sousa, 2008; Bello, 2016). En el corpus encontramos nombres de enfermedades, síntomas, pruebas o características escritas con mayúscula inicial. Sin embargo, según la *OLE* (2010) los nombres que designan enfermedades, síndromes y trastornos son sustantivos comunes, por esta razón deben ser escritos con minúscula inicial (*Cáncer** [*cáncer*], *Neumonía** [*neumonía*], *Hipotiroidismo** [*hipotiroidismo*], *Esclerosis múltiple** [*esclerosis múltiple*]).

- *NO* se irradia. NO* fiebre en casa, NO* diarrea, no náuseas ni vómitos.*
- *Presenta Hipertrofia* ventricular.*
- *Antecedente de Tos*.*

Por su parte, sí deben llevar mayúscula inicial los nombres propios que forman parte del nombre de la enfermedad (*síndrome de Down*), así como el nombre completo de la enfermedad cuando forma parte de la designación de una organización (*Asociación Española Contra el Cánce*r).

Al igual que ocurre con las enfermedades, también encontramos una clara tendencia a escribir con mayúscula inicial los nombres de los meses y días de la semana

(*consulta urgencias en Julio** [*julio*]), debido posiblemente a la influencia del inglés[89]. En la *OLE* (2010: 502) se especifica que, «al contrario que en inglés, los nombres de los días de la semana y de los meses del año deben escribirse con minúscula inicial». Por ejemplo:

- *Consulta urgencias en Julio*.*
- *Última consulta Febrero*.*
- *Ingreso en Junio*.*
- *En Octubre* de 2000 sustitución valvular*.

Algunos casos que presentan mayor dificultad tienen que ver con la escritura de nombres de principios activos. Las voces que aluden a principios activos deben tener minúscula inicial y los nombres comerciales de medicamentos deben escribirse con mayúscula inicial al tratarse de nombres propios registrados. Si consultamos de nuevo la *OLE* (2010: 502), en ella se afirma que «deben escribirse con letra minúscula inicial los nombres de los principios activos de las medicinas bajo los cuales se comercializan los medicamentos genéricos». A pesar de ello, encontramos una amplia variabilidad y confusión, con principios activos escritos en mayúscula (*Adenosina** [*adenosina*], *Metamizol** [*metamizol*]) y nombres comerciales de medicamentos en minúscula (*prozac** [*Prozac*], *nolotil** [*Nolotil*]). Además, en algunos casos el nombre del principio activo y de la marca registrada del medicamento genérico comercializado coinciden (*Aciclovir Sandoz, Fluoxetina Normon*), aunque suele ir acompañado del nombre del fabricante y las siglas EFG[90].

También se detecta el tratamiento inadecuado de mayúsculas y minúsculas después de un signo de puntuación, como la ausencia de mayúscula inicial al inicio de oración tras un punto.

Por último, también debemos mencionar que en el corpus analizado hay varios fragmentos de informes que han sido redactados en mayúsculas en su totalidad. Es una circunstancia que se observa con frecuencia en varias secciones a lo largo del corpus, lo

[89] A pesar de la creencia, nunca ha existido una norma académica expresa que estableciese que los nombres de los meses, días de la semana y estaciones deben escribirse con mayúscula, pero a partir de la *Ortografía* de 1969 se indica de forma explícita que deben escribirse con minúscula.

[90] Equivalente Farmacéutico Genérico.

que impide poder cuantificar de manera precisa la presencia real de estos errores. No obstante, el uso incorrecto de mayúsculas o minúsculas no afecta en gran medida a la legibilidad y procesamiento automático, por lo que en el proceso de corrección es de los errores ortográficos menos significativos.

4.2.1.8. Creación y uso de abreviaturas

En el corpus se produce una alta concentración de mecanismos de abreviación léxica, como siglas, abreviaturas, acrónimos y símbolos. Su uso es continuo a lo largo de los informes para ahorrar tiempo y espacio de escritura. Un gran número de las palabras del corpus están incompletas porque los facultativos acortan palabras para ir más rápido, ya sea mediante la creación de abreviaturas o siglas. No obstante, la creación de abreviaturas debe hacerse de acuerdo con las reglas de formación establecidas para este tipo de creaciones, y no es recomendable su uso indiscriminado, pues aquellas poco frecuentes o que no están estandarizadas pueden dar lugar a confusión entre los distintos receptores de la información y dificultar el procesamiento automático. Aunque en nuestro estudio no hemos analizado cuantitativamente los errores relativos a abreviaturas debido a que la magnitud del tema requeriría una investigación propia, vamos a mencionar algunos de los errores y características detectadas más destacables, pues son el punto de partida para investigaciones futuras.

En el corpus se usan procedimientos basados en truncamiento (*comp* por *comprimido*, *vasc* por *vascular*), cuyo funcionamiento consiste en suprimir caracteres finales de una palabra y conservar la parte inicial del término. El segundo procedimiento que detectamos para la formación de abreviaturas se realiza por contracción, y se basa en eliminar caracteres centrales y conservar solo los más representativos (*tto* por *tratamiento* o *hrs* por *horas*). Hay también casos de abreviación extrema, como *sd* para *síndrome* o *qx* para *quirúrgico*.

El error más destacable en el corpus en cuanto a creación de abreviaturas tiene que ver con la falta de uniformidad y la gran variabilidad en el tratamiento de estas, con diferentes abreviaturas para las mismas voces (*glucosa* [*glu, gl, glc, gluc*.]) o con la misma abreviatura para distintos términos (*aprox*. [*aproximado, aproximada, aproximadamente*], *bil* [*bilateral, bilirrubina*], *cef*. [*cefalea, cefálica*]), lo que provoca ambigüedad. Un ejemplo reseñable de opacidad semántica puede ser *neo*, que aparece en

el corpus en ejemplos como *neo mama*, *neo gástrico* y *neo faringo-laríngea.* En estos casos se trata de la abreviatura de *neoplasia*, pero al no llevar punto abreviativo y no ser una abreviatura estandarizada puede confundirse con el prefijo de origen griego *neo*, que significa 'nuevo o reciente'. Por tanto, entre otros errores detectados también se encuentra la ausencia de punto abreviativo, a pesar de que el cierre de la abreviatura con punto es un requisito obligatorio. Además, las abreviaturas deben mantener la tilde si están formadas por la vocal que la lleva en la palabra desarrollada, pero en muchos casos del corpus se omite. Un ejemplo representativo es *célula*, que aparece de forma abundante en el corpus como *cel.* y *cél*., de forma indistinta. Otros ejemplos son:

- glucosa: *glu, gl, glc, gluc.*
- actividad*: actv, act.*
- cefalea*: cef, cefale*
- carcinoma: *car., carc., CA*
- creatinina: *creat., cr., CR*
- síndrome: Sd, SD, sd, Sd., Sde, Sdme, Sdre

4.2.1.9. Tratamiento de siglas y acrónimos

Al igual que en las abreviaturas, también observamos variabilidad y ausencia de estandarización en la creación y uso de siglas y acrónimos. Según la *OLE* (2010), las letras que componen las siglas deben escribirse en mayúsculas, sin embargo, un número considerable de las presentes en el corpus han sido escritas en minúscula, lo que dificulta su detección de forma automática. Únicamente deben ser escritas en minúscula cuando se lexicalizan y se convierten en una palabra común (*sida*).

Asimismo, y a diferencia de las abreviaturas, las siglas no deben llevar tildes ni punto abreviativo, pero el punto aparece erróneamente entre las letras que forman algunas siglas en los informes analizados. Por último, aunque se indica que son invariables y, por tanto, tampoco cambian de forma en plural, es un error muy extendido añadir *-s* para marcar el plural de estas en el corpus. Al igual que en el caso de las abreviaturas, su uso indiscriminado puede provocar problemas de comprensión en los destinatarios y en los sistemas de procesamiento automático. Ejemplos:

- *itu** [ITU]: infección tracto urinario.
- *got** [GOT]: *glutamic oxaloacetic transaminase* (transaminasa glutámico-oxalacética).
- gea* [GEA]: gastroenteritis aguda.
- H.T.A* [HTA]: hipertensión arterial.
- No A.M.C.* [AMC], no H.T.A* [HTA], no D.M.* [DM].
- Biopsia de R.T.U* [RTU].
- AINEs* [AINE]: antiinflamatorios no esteroideos.
- IECAS* [IECA]: inhibidores del enzima conversor de la angiotensina.

4.2.1.10. Tratamiento de símbolos

También detectamos la presencia de un alto número de símbolos en el corpus, entre los que encuentran unidades de medida, especialmente en los apartados de los informes que hacen referencia a los tratamientos, a las exploraciones físicas y a las pruebas complementarias. Los símbolos son abreviaciones de carácter científico que han sido fijados convencionalmente por instituciones de normalización como la ISO (*International Organization for Standardization*) y tienen validez internacional. Los símbolos, como las unidades básicas y derivadas del sistema internacional, deben escribirse sin punto de abreviación, sin tilde y no variar su forma en plural. No obstante, un error muy extendido es convertir en plural las unidades de medida, mediante la adición de *s*: *grs** [*g*], *kgs** [*kg*], *cms** [*cm*] y *mls** [*ml*]. También debe haber un espacio de separación entre ellos y la cifra que acompañan, pero no suele cumplirse esta condición en el corpus. La variabilidad mencionada también se aprecia en el uso de mayúsculas y minúsculas en la escritura de unidades, como en el caso de *mEq*. Algunos ejemplos característicos son:

- mg: *25mg 1-1-1*
- mEq (miliequivalente): *meq, MEQ, meQ, Meq*
- m (metro): *mt*
- h (hora): *hr*
- g (gramo): *gr*
- h (horas): hs

- m (minuto): *min*
- s (segundo): *seg, sg*
- kg (kilogramo): *Kg, 100 kgs*

4.2.1.11. Diferencias geográficas o diatópicas

En esta investigación la variedad predominante se corresponde con el español peninsular, al ser un corpus compuesto por informes clínicos obtenidos de hospitales españoles. Como ya hemos mencionado anteriormente, no es posible tener más información sobre los distintos facultativos que han escrito los informes, debido a las características del corpus. No obstante, es relevante mencionar que también se han detectado rasgos lingüísticos aislados propios de distintas variedades diatópicas del español de América y hay patrones de error en el corpus que pueden asociarse con ciertas variedades de otras regiones hispanohablantes.

Entre estos errores detectados observamos el cambio de timbre de vocales átonas y tónicas. El cierre de la /e/ átona en /i/ es la representación gráfica del cierre oral llevado a cabo durante la pronunciación en algunas variedades del español americano (*dispués** [*después*]). Un ejemplo es el sufijo derivativo *-ear* que se convierte a menudo en la terminación *-iar* (*golpiar** [*golpear*]). Este proceso no se ha normativizado, por lo que se considera un error y no debe reflejarse en la escritura.

Son también destacables fenómenos fonéticos como el lambdacismo, o la articulación de *-r* implosiva y final como *-l*, pronunciación que se ve reflejada en la escritura en casos como *locomotol** (*locomotor*). A su vez, el seseo, o sustitución del sonido de /z/ y /c/, con sus distintas variantes polimórficas a nivel fonético en América y algunas zonas del sur de la península, tiene presencia en la escritura a través de errores como *trombolisado** (*trombolizado*) o *leucositos** (*leucocitos*).

Otro rasgo influenciado por factores diatópicos es el cambio acentual (*video-vídeo, gastroscopía-gastroscopia*), aunque en este caso ambas pronunciaciones son válidas. En el plano léxico observamos la variedad de formas como *peronero/peroneo, adenoflegmón/adenoflemón o aquileana/aquiliana,* con voces que se circunscriben a determinados dominios geográficos hispanohablantes. Otro rasgo revelador es la presencia de formas como *toráxico*. La forma recogida en los diccionarios académicos y

considerada de uso mayoritario es *torácico*, sin embargo, en el *DPD* se señala que la forma *toráxico* es usada en regiones de América como en el Cono Sur y es también válida.

Como se ha mencionado anteriormente, es fundamental tener conocimiento sobre la variedad del corpus y el destinatario final en el proceso de detección y corrección de errores.

4.2.2. Errores *real-word*

En esta sección llevamos a cabo una tipificación cualitativa de errores *real-word* o errores dependientes del contexto. Un error *real-word* en la escritura conlleva un cambio de significado y se materializa en una palabra correcta ortográficamente, por tanto, son errores que en muchas ocasiones pasan desapercibidos en los sistemas de corrección. Es mucha la variabilidad de palabras ortográficamente correctas, pero sintáctica y semánticamente erróneas en el corpus, debido a errores de omisión, inserción, sustitución y transposición de caracteres en el momento de la escritura, como se ha cuantificado en el apartado anterior. En la mayoría de estos casos el significado de la oración se ve comprometido y se dificulta el correcto procesamiento automático. Algunos ejemplos ilustrativos son: *cuatro* similar* (*cuadro similar*), *estenosis órtica** (*estenosis aórtica*), *evolución cínica** (*evolución clínica*), *hábito eólico** (*hábito enólico*), *hombre* derecho* (*hombro derecho*), *lumbalgia cónica** (*lumbalgia crónica*), *olor* torácico* (*dolor torácico*), *tacto recta** (*tacto rectal*) o *color* local seco (calor local seco).*

Tras el proceso de detección y análisis del corpus, estos errores han sido clasificados en grupos claramente delimitados que aportan información específica para el desarrollo de un módulo basado en conocimiento lingüístico. Estos grupos incluyen errores de paronimia, de concordancia gramatical, de formación de palabras por fenómenos de composición y prefijación, y formas verbales anómalas en el dominio. Como puede observarse en la Tabla 27, estos conjuntos designados se dividen a su vez en subgrupos más específicos. Además, se ha realizado una recopilación de casos para el diseño de un set de confusión, formado por parejas de palabras que tienden a ser confundidas de forma significativa en el corpus.

Paronimia	
Homófonos	vulvar-bulbar maya-malla
Tilde diacrítica	dé-de sí-si
Parónimos acentuales – sílaba tónica	liquido-líquido-liquidó vomito-vómito
Confusión entre formas verbales homófonas	echo-hecho halla-haya
Proximidad fonética	cociente-consciente consiente-consciente
Concordancia gramatical	
Discordancia nominal de género	niveles hidroaéreas [niveles hidroaéreos] buen función [buena función]
Discordancia nominal de número	cólico nefríticos [cólicos nefríticos] varios episodio [varios episodios]
Discordancia nominal de género y número	antecedentes traumática [antecedentes traumáticos] tratamiento farmacológicas [tratamiento farmacológico]
Discordancia verbal	se visualiza lesiones [se visualizan lesiones] se aprecia focos muy mal definidos [se aprecian focos muy mal definidos]
Unión y separación de palabras	
Formación de palabras por composición	tónico clónicas [tonicoclónicas] cardio torácico [cardiotorácico]
Formación de palabras por prefijación	post traumático [postraumático] pre auricular [preauricular]
Confusión entre pares de palabras	entorno-en torno aparte-a parte
Formas verbales anómalas en el dominio	
Primera persona de presente	trafico [tráfico] riño [riñón]
Formas en segunda persona	ultimas [últimas] limites [límites]
Imperativos en plural	adelgazad [adelgazado] colapsad [colapsado]
Formas en subjuntivo	señale [señales] completare [completar]

Tabla 27. Clasificación cualitativa de errores *real-word*

4.2.2.1. Usos erróneos de formas parónimas y homófonas

Los parónimos son aquellas palabras que se asemejan entre sí por su sonido o forma (proximidad gráfica o fonética), o incluso por su etimología, pero tienen significados diferentes. Gran parte de los errores se producen debido a similitudes en la pronunciación entre palabras. Es el caso de las voces homófonas, es decir, de aquellas palabras que se pronuncian de forma idéntica, pero se escriben de manera distinta y su significado es también distinto.

En el corpus localizamos parejas o grupos de palabras que tienen una naturaleza casi homófona, lo que provoca dudas e induce a errores en el proceso de escritura. Se producen confusiones entre verbos (*halla/haya, haber/a ver,* o *echa/hecha*) y en la escritura de fonemas con más de una posibilidad gráfica para ser representados (*maya/malla, bulbar/vulvar,* o *cabo/cavo*).

En ocasiones la confusión en la grafía está provocada por la relajación de sonidos en el habla, dando lugar a una articulación más laxa y desembocando en fenómenos de homofonía. Se neutraliza la oposición existente entre sonidos formados por *-x-*, *-cc-*, *-ns-* o *-bd-*, entre otros. Así, encontramos en el corpus ejemplos como *aducción-abducción, consciente-cociente-consiente,* o *pizza-pisa*.

- *Paciente <u>cociente</u>* [*consciente*] *y orientado en tiempo y espacio.*

Resulta revelador estudiar en qué entornos típicos suelen producirse los errores entre parejas parónimas u homófonas. Además de los errores por homofonía y paronimia, en el siguiente apartado vamos a profundizar en los errores motivados por la presencia o ausencia de la tilde diacrítica o por la elección incorrecta de la sílaba tónica en los parónimos acentuales.

4.2.2.2. Errores en el uso de tildes

En algunas lenguas, como el inglés, las palabras no se acentúan de forma gráfica, pero en español el acento gráfico o tilde debe añadirse de forma obligada a aquellas voces que lo requieran. Además, hay palabras cuyo significado se ve transformado si se omite la tilde. Por tanto, resulta fundamental acentuar gráficamente la vocal tónica de las

palabras siguiendo las reglas ortográficas generales del español, pues la ausencia de acentuación gráfica plantea problemas de ambigüedad.

En este grupo específico de errores por paronimia se incluyen las palabras cuya diferencia radica en la inserción de tilde, como en el caso de las palabras con tilde diacrítica. La tilde diacrítica sirve «para distinguir entre sí los miembros de ciertos pares de palabras grafemáticamente idénticas, pero de distinto significado y función, siendo uno de ellos tónico y el otro átono; la tilde se coloca, como es natural, sobre el elemento tónico del par» (OLE, 2010: 75).

En el corpus detectamos palabras que deben llevar tilde diacrítica para diferenciarse de palabras con idéntica forma, pero distinta categoría gramatical y significado, sin embargo, se omite su uso. También detectamos el caso contrario, es decir, el par al que le corresponde la pronunciación átona es tildado erróneamente. Algunos ejemplos de parejas átonas/tónicas incorrectamente tildadas que aparecen en el corpus según el contexto son: *de* (preposición) y *dé* (forma del verbo *dar*), *el* (artículo) y *él* (pronombre personal), *si* (conjunción) y *sí* (adverbio), *mas* (conjunción) y *más* (adverbio), *aún* (adverbio con valor temporal) y *aun* (adverbio que equivale a *incluso*), *que* (conjunción o pronombre relativo) y *qué* (pronombre interrogativo), *cual* (pronombre relativo) y *cuál* (pronombre interrogativo), *quien* (pronombre relativo) y *quién* (pronombre interrogativo), entre otros.

- *<u>aún</u>* [*aun*] *así aconsejan acompañamiento continuo*
- *función renal se deteriora <u>aun</u>* [*aún*] *mas*
- *convexidad frontal es <u>dónde</u>* [*donde*] *se observa mayor diámetro del hematoma*
- *Control de su médico de AP, <u>quién</u>* [*quien*] *decidirá*
- *no sabe determinar por <u>que</u>* [*qué*] *causa*

La mayoría de palabras con tilde diacrítica son monosílabas, pero la tilde diacrítica no solo afecta a palabras monosílabas, también incluye a ciertas palabras polisílabas, como los interrogativos y exclamativos: *cómo, cuándo, cuánto y (a)dónde*. Como adelantamos anteriormente, estos errores podrían ser considerados de motivación cognitiva o errores de competencia, pues implican el desconocimiento de algunas de las normas ortográficas que regulan el funcionamiento de la lengua.

De igual forma se incluyen en este apartado los llamados parónimos acentuales, palabras cuya forma es semejante, pero difieren en la sílaba tónica. En español encontramos gran número de palabras que comparten sonidos y grafía, pero su único rasgo distintivo es el lugar de la sílaba tónica (*deposito/depósito/depositó, vertebra/vertebra,* o *rotula/rótula*).

- *Presento* [*presentó*] *disfunción renal leve.*
- *Enterobacter aerogenes (liquido* [*líquido*] *biliar).*
- *Exéresis de deposito* [*depósito*] *de grasa en esclera lateral.*

Cuando la sílaba tónica varía conlleva un cambio en el significado. Este es el caso del sustantivo *médico,* que aparece con mucha frecuencia en el corpus sin tilde (*medico*) y que se correspondería con la primera persona de presente del verbo *medicar.*

- *Se le medico* [*medicó*] *con 50mg de Capoten.*
- *Seguimiento por medico* [médico] *de familia.*

Asimismo, la ausencia de la tilde puede dar lugar a múltiples verbos conjugados en persona, modo o tiempos distintos (*diagnostico/diagnosticó, comprobara/comprobará*). Por tanto, el corpus contiene tanto palabras tildadas que no deben llevar tilde como palabras que la llevan en la sílaba inadecuada para ese contexto de uso: *segmentarías* (*segmentarias*), *mamá* (*mama*), *broté* (*brote*), *liquidó* (*líquido*).

4.2.2.3. Secuencias que se escriben en una o más palabras con distinto valor

Además de errores relacionados con cuestiones de acentuación y tonicidad, proximidad fonética o discordancias gramaticales, localizamos errores relacionados con la unión y separación de palabras, estrechamente relacionados con los fenómenos de paronimia ya abordados. La elección de la escritura de una palabra en su forma unida o separada resulta especialmente problemática para los hablantes en algunas voces. Algunos pares de palabras empleados erróneamente en el corpus son: *por que/porque/por qué, si no/sino, acerca/a cerca, sobre todo/sobretodo, entorno/en torno,* o *aparte/a parte*.

En el caso de *porque*, sus distintas configuraciones y combinaciones posibles es causa recurrente de errores.

- *Traída por los servicios de urgencias por que* [*porque*] *esta mañana al levantarse comienza con mareos.*
- *Desconoce el por qué* [*porqué*].
- *Con buena dinámica ventilatoria y con saturaciones entorno* [*en torno*] *al 95%.*
- *Sino* [*si no*] *mejoría, subir hasta 200mg.*

4.2.2.4. Errores de concordancia gramatical

Es sumamente destacable la presencia de errores *real-word* relacionados con la concordancia gramatical. La concordancia es definida en la NGLE (2010: 583) como «la expresión formal de varias relaciones sintácticas» y se produce mediante la coincidencia obligatoria de determinados accidentes gramaticales en distintos elementos de la oración. Pueden ser distinguidos dos tipos de concordancia: nominal y verbal. La concordancia nominal alude a la coincidencia de género y número entre constituyentes, y se establece entre el sustantivo y el artículo o los adjetivos. Por su parte, la concordancia verbal se refiere a la coincidencia entre número y persona, y es la que tiene lugar entre verbo y sujeto. Los casos de discordancia detectados se producen frecuentemente en entornos intrasintagmáticos, entre sustantivos y adjetivos que los acompañan. Para poder analizar los posibles errores de discordancia verbal entre verbo y sujeto sería necesario ampliar la ventana de contexto, pues el análisis del contexto inmediato de la palabra mediante bigramas, trigramas y *word embeddings* resulta limitado para este fin.

En relación a la concordancia de género, esta debe implantarse entre determinantes, sustantivos y adjetivos, que se clasifican en masculinos o femeninos. Algunos de los errores provocados por discordancia de género son: *dolor torácica** (*dolor torácico*), *infección respiratorio** (*infección respiratoria*), *tratamiento crónica** (*tratamiento crónico*), *ambos* piernas* (*ambas piernas*), y *comprimido diaria** (*comprimido diario*).

NRL: consciente alerta orientado, normotonico, normoreflexico, no focalidad neurológico [*neurológica*].

Entre las causas de error pueden encontrarse causas mecánicas por la escritura rápida, o también la vacilación y confusión de género. También hay que tener en cuenta el caso de sustantivos femeninos que comienzan por /a/ tónica, en los que el artículo que los precede debe tomar la forma *el* para evitar cacofonía (*la* apéndice* [*el apéndice*], *la* área* [*el área*]).

Gran parte de las discordancias detectadas surgen a propósito de la concordancia de número, que expresa la referencia a una o a varias entidades mediante formas de plural y singular. Entre los errores detectados se encuentran: *consulta* externas* (*consultas externas*), *ambos campo** (*ambos campos*), *abundantes líquido** (*abundantes líquidos*), *hernia* discales* (*hernias discales*), *hábito* tóxicos* (*hábitos tóxicos*), *ruidos patológico** (*ruidos patológicos*).

También pueden darse discordancias en los planos de número y género simultáneamente (*soplos sistólica** [*soplos sistólicos*], *tratamiento farmacológicas** [*tratamiento farmacológico*].

4.2.2.5. Errores de formación de palabras

Como hemos mencionado en el apartado dedicado a la formación de palabras mediante derivación y composición de la sección de errores *non-word* (4.2.1.2), en el dominio médico es muy habitual la presencia de palabras formadas por mecanismos de prefijación y composición para la creación de terminología especializada. Además, es incorrecta la yuxtaposición de adjetivos con espacio y sin nexo para la formación de palabras compuestas.

En el plano de los errores *real-word*, el error se suele producir al separar con espacio los dos elementos que integran la unidad léxica univerbal. Si están correctamente escritas las dos palabras de forma independiente, pasan desapercibidas en el proceso de detección, pues ambas aparecen en el lexicón que se usa para validarlas. Así encontramos casos como *ansioso depresivo** (*ansiosodepresivo*), *sacro coccígeas** (*sacrococcígeas*), *dorso lumbar** (*dorsolumbar*), o *uretero vesical** (*ureterovesical*). Por su parte, los prefijos se deben escribir unidos sin guion a la base léxica, pero localizamos también casos en los que se deja un espacio entre el prefijo y la palabra (*pre menstrual* [*premenstrual*], *bi cameral** [*bicameral*] o *post quirúrgico** [*postquirúrgico*]).

4.2.2.6. Formas verbales anormales en el dominio

La búsqueda de la economía expresiva se traduce en una sintaxis simple con estructuras breves. Se suelen elidir elementos de forma generalizada, como conectores, preposiciones, determinantes o formas verbales, especialmente aquellos verbos que se corresponden con un significado existencial, como *haber*, *estar*, *mostrar* o *presentar*.

Por tanto, los informes médicos se caracterizan por contener pocos verbos en proporción a la extensión media de los informes y el peso recae en las unidades terminológicas, que se acompañan de adjetivos modificadores o adverbios.

En el corpus predomina el uso de determinados tiempos verbales, como el presente o el pretérito perfecto simple, o formas en tercera persona del singular, además de formas no personales, como el infinitivo. No obstante, el análisis de los resultados ha permitido recopilar formas verbales que deben ser marcadas como sospechosas en un corpus de estas características porque suelen enmascarar errores, al no ser su presencia común en textos especializados de estas características. Entre ellas podemos destacar:

4.2.2.6.1. Formas en primera persona

Como hemos adelantado, los informes médicos suelen estar redactados en un estilo impersonal y mayoritariamente se emplea la tercera persona. Hay una marcada tendencia a la despersonalización y a la impersonalidad, pues se busca ocultar la voz del sujeto como reflejo de objetividad científica. Así la mayoría de los verbos tienen la flexión de tercera persona.

Por consiguiente, no es habitual encontrar formas en primera persona, únicamente se observa su uso con los llamados verbos declarativos, verbos de comunicación o *verba dicendi*, que son empleados para designar acciones comunicativas: *solicito*, *informo*, *indico*, *ruego,* o *pido*. Con ellos el referente, que es el facultativo que redacta el informe y que suele estar elidido, expresa peticiones o decisiones médicas. Por tanto, otras formas en primera persona del presente de indicativo suelen ser impropias en los informes médicos y casos como *automedico** (*automedicó*) o *circulo** (*círculo*) ocultan errores. Estos casos de errores *real-word* suelen estar provocados mayoritariamente por la omisión de tilde, convirtiendo la tercera persona de pretérito perfecto simple en primera

persona del presente de indicativo (*integro** [*integró*]), o transformando sustantivos y adjetivos en formas de primera persona del presente (*transito** [*tránsito*], o *incomodo** [*incómodo*]). Aunque son mucho menos frecuentes, también hay casos provocados por fenómenos de multierror, como en *riño** [*riñón*], desencadenado por la omisión de tilde y la omisión del carácter final.

4.2.2.6.2. Formas en segunda persona

Al igual que las formas en primera persona, las formas verbales flexionadas en segunda persona resultan poco frecuente en informes médicos. A través de ellas se apela o se hace alusión a un receptor de forma directa, por tanto, la presencia de estas formas verbales en el corpus es impropia y suele enmascarar errores *real-word*. Algunos ejemplos son *orbitas** (*órbitas*), *ultimas** (*últimas*), o *limites** (*límites*), cuyo error es la omisión de tilde.

4.2.2.6.3. Imperativos en plural

La mayoría de las oraciones que componen los informes presentan verbos conjugados en modo indicativo. Este modo es el utilizado para marcar lo que se va a expresar como información real, y la principal función del informe médico es informar, exponer y describir. Se contrapone, por tanto, a las formas en subjuntivo para expresar conjeturas, información no verificada o que no ha sido experimentada; o a las formas de imperativo para emitir mandatos o exhortaciones.

En el corpus hay formas verbales de infinitivo utilizadas con valor exhortativo para dar indicaciones y recomendaciones dirigidas a un interlocutor indeterminado, como en los apartados de observaciones, evolución y comentarios (*ver informes previos, solicitar RMN, repetir analítica*). No obstante, en el corpus detectamos algunas formas en imperativo que resultan anómalas. Concretamente, la aparición de estas formas se da por varios tipos de errores. En primer lugar, por la omisión del carácter final de la palabra, convirtiendo el adjetivo o participio en una forma de imperativo en segunda persona del plural (*etiquetad** [*etiquetado*], o *engrosad** [*engrosado*]). En segundo lugar, por la sustitución de la consonante *s* por la consonante *d* al escribir las unidades léxicas en plural, debido a la adyacencia de ambas letras en el teclado y a la frecuencia con la que

suelen usarse durante la escritura (*dudad** [dudas], o *mamad** [mamas]). Un caso similar es provocado por la adyacencia de las vocales *i* y *u* en el teclado (*salid** [*salud*]). También se han detectado las formas *iníciales** (*iniciales*) o *cicatrízales** (*cicatrizales*), cuyo error ha sido la inserción de tilde en ambos casos.

4.2.2.6.4. Formas en subjuntivo

Como ya se ha anticipado, las formas en subjuntivo se utilizan para expresar suposiciones, deseos o posibilidades, situándose en una esfera de irrealidad. Por tanto, su presencia en el género de los informes clínicos es poco común. Algunos ejemplos detectados y que han resultado ser errores *real-word* son formas en presente de subjuntivo (*señale** [*señal* o *señales*]), o futuro de subjuntivo (*completare** [*completar*], *comprobare** [*comprobar*]). En el primer caso, el error puede ser de omisión del carácter *s* o inserción de la vocal *e*, mientras que en los casos de futuro de subjuntivo el error es por inserción de la vocal *e*.

4.3. Discusión de los resultados

Tras el análisis de los distintos resultados de forma cuantitativa y cualitativa abordamos la discusión de estos teniendo en cuenta los datos extraídos y la información sobre otros estudios recopilada en el capítulo de fundamentación teórica.

En primer lugar, el número de errores detectados en los informes clínicos con las técnicas aplicadas es significativo y la tasa de error se sitúa en un 3,3 %. Este resultado es coherente con los consultados en otros estudios previos, que coincidían en destacar la elevada presencia de errores de escritura en documentación clínica y se situaban en porcentajes de errores que van desde un 0,4 % (Liu et al., 2012), un 4 % (Lai et al., 2015), o incluso llegaban al 10 % (Ruch et al., 2003). Al igual que en los mencionados estudios, podemos afirmar que los informes médicos se caracterizan de forma general por contener un número elevado de errores lingüísticos y que este fenómeno representa un obstáculo para la comprensión del informe por parte del paciente y dificulta el procesamiento informático automatizado. Estas investigaciones aportan datos cuantitativos sobre la presencia de errores en documentación clínica en lenguas como el inglés, el francés, el sueco o el húngaro, sin embargo, los resultados obtenidos en esta investigación no pueden

contrastarse de forma directa con datos previos sobre el español, ya que no encontramos investigaciones que se hayan centrado de manera exhaustiva en la detección automática y análisis cuantitativo de errores en informes médicos en español.

No obstante, sí pueden tomarse como referencia algunos trabajos que se irán mencionando a continuación y que han investigado sobre las características del lenguaje médico desde una perspectiva cualitativa y descriptiva. Uno de ellos es la tesis doctoral de Terroba (2016), ya mencionada en el capítulo dedicado a la fundamentación teórica, que analiza cuatrocientos informes clínicos asistenciales del sistema sanitario de La Rioja. A lo largo de este trabajo la autora señala la presencia de errores fonéticos y ortográficos, errores en el uso de mayúsculas y minúsculas, vacilaciones acentuales, repeticiones de lexemas que empobrecen el estilo, abundantes elipsis, y errores sintácticos con oraciones incorrectamente construidas, que presentan errores de concordancia, de puntuación o que no siguen el orden lógico. Al no incorporarse datos cuantitativos sobre la presencia de estos errores en los informes no podemos comparar los resultados con mayor precisión, pero observamos que se producen similitudes entre los errores documentados con respecto a los detectados en nuestro estudio.

Asimismo, la autora dedica un apartado específico a la presencia de errores de medicación[91] debido a la trascendencia que tienen en los informes analizados, pues considera que son los causantes de muchos de los errores del tratamiento (Terroba, 2016: 145). Entre los errores en la escritura de fármacos que se enumeran en ese trabajo se encuentran la omisión de tilde y las alteraciones acentuales, las pérdidas de consonantes en posición implosiva, la adición de consonantes, la simplificación de geminadas, y un comportamiento anárquico en el uso de determinados pares de consonantes (*k-c, q-k, z-c*). Además, la autora considera que, aunque es difícil conocerlas, algunas de las causas de estos errores se pueden intuir y son: «lapsus y despistes, errores en el manejo del ordenador, influencia del inglés, celeridad en el trabajo, influencia de algunos términos farmacológicos sobre el resto, etc.» (Terroba, 2016: 212). En nuestro estudio constatamos la elevada presencia de errores fonéticos y ortográficos en la escritura de medicamentos

[91] Según el *CedimCat* (*Centro de Información de Medicamentos de Cataluña*), «un error de medicación es cualquier evento evitable que tiene lugar durante el proceso de prescripción, preparación, dispensación o administración de un medicamento» (Moreira, s.f.), recuperado de la página web <https://www.cedimcat.info/index.php?option=com_content&view=article&id=192:errores-de-prescripcion-ejemplos-de-errores-de-prescripcion-frecuentes-y-su-posible-prevencion&lang=es>.

(Sección 4.2.1.3), con semejanzas con los resultados presentados en el trabajo de Terroba (2016).

Por su parte, el Instituto para el Uso Seguro de los Medicamentos (ISMP) afirma que la confusión entre los nombres de medicamentos ocurre principalmente por la semejanza fonética y ortográfica entre nombres comerciales (*Aricept* y *Azilect*), entre nombres genéricos (*valaciclovir* y *valganciclovir*) o entre nombres comerciales y genéricos (*Rohipnol* y *ropinirol*). Hay correspondencia, por tanto, con algunos de los resultados obtenidos en nuestro trabajo, aunque en el listado[92] que publican se tienen también en cuenta errores que trascienden el plano lingüístico, al ser necesarios conocimientos de medicina para poder detectar la confusión en la prescripción errónea de medicamentos que han sido escritos correctamente (por ejemplo: *Adolonta-Atenolol*; *Alapryl-Enalapril*, o *Casenbiotic-Casenglicol*). Es relevante mencionar que en nuestro estudio hemos descubierto patrones y variantes ortográficas en nombres de medicamentos que no se habían documentado previamente y que pueden incorporarse a esta clasificación. Sería también interesante generar un listado similar al creado por el ISMP sobre los patrones de errores obtenidos a partir de nombres de enfermedades y otros epónimos.

Otra investigación sobre errores en la escritura de medicamentos fue realizada por Ferner y Aronson (2016), que llevaron a cabo un estudio cuantitativo de variaciones de nombres de fármacos en diferentes bases de datos del ámbito biosanitario, como PubMed, Medline o Cochrane. El objetivo fue examinar cómo los errores ortográficos de los nombres de los medicamentos pueden dificultar la búsqueda de literatura publicada. En este trabajo los errores de sustitución (45 %) fueron el tipo de error más frecuentemente cometido, seguido del error de omisión de letra (28 %). Se trata de un estudio realizado para la lengua inglesa, por lo que los resultados no son directamente comparables debido a las diferencias fonéticas y ortográficas entre ambas lenguas. El inglés cuenta con acento prosódico, pero no utiliza acentos gráficos, principal diferencia con nuestros resultados al no contener errores de omisión de tilde su clasificación. Las sustituciones de letras fueron principalmente entre *y-i*, además de *c-s, c-k, ch-k, f-ph, m-n, th-t,* y *x-ks*. También destaca la omisión de las letras *e* y *r* (*pednisolone* en lugar de *prednisolone*, o *popranolol*

[92] <http://www.ismp-espana.org/ficheros/relaciondenombres.pdf>

en lugar de *propranolol*). De forma menos notable detectan casos de transposición (*filgrastim* en lugar de *filgastrim*) o duplicación (*ll* en lugar de *l* y *nn* en lugar de *n*).

Vivaldi (2020) estudia la presencia de errores en un corpus de 1090 informes de urgencias en español (420 000 *tokens*) procedentes del Hospital Italiano de Buenos Aires. Clasifica los errores según afecten a la ortografía de las palabras o a los signos de puntuación que estructuran el texto. A nivel de palabra señala que los errores más frecuentes están relacionados con la acentuación y los caracteres especiales, y destaca los errores de omisión de tilde, de forma similar a nuestro estudio. También menciona errores provocados por la omisión de espacio que da lugar a palabras unidas, así como la separación de una palabra en dos fragmentos por la inserción de este. Además, añade la omisión o inserción indebida de una o más letras de una palabra, la alteración del orden de las letras o el uso de abreviaturas no estándares o poco transparentes. En cuanto al grupo de errores de puntuación, comenta la ausencia de signos de puntuación, el uso erróneo de estos, el tratamiento erróneo de las mayúsculas después de un punto o la puntuación entre sujeto y predicado.

4.3.1. Diferencias entre el dominio médico y el español común

Para averiguar si hay diferencias entre los errores cometidos en el dominio médico y el español común nos servimos del estudio de Ramírez y López (2006) para establecer la comparación, porque es el único trabajo que aporta datos cuantitativos de errores para el español general. No obstante, únicamente tienen en cuenta errores *non-word*, por lo que para esta comparación hemos utilizado exclusivamente los resultados *non-word* de nuestro estudio, con la intención de que el cotejo de resultados fuese lo más preciso posible.

Ramírez y López (2006) analizan un corpus que está compuesto por tres muestras de datos, la primera incluye textos editados y no editados, la segunda está formada por textos altamente editados y la última únicamente por textos no editados. Esta última muestra, con aproximadamente dos millones de palabras, es la que se escogió para la comparación de errores, pues es la que presenta unas condiciones más cercanas a los informes clínicos al no haber sido editada ni sometida a una revisión posterior. Además, está formada por un conjunto de español de México (OWC-México) y de Madrid (OWC-Madrid), por lo que hay participantes hablantes tanto de español de América como de

español peninsular. Esta misma circunstancia se da en nuestro corpus, como abordábamos en el apartado de diferencias geográficas (Sección 4.2.1.11.), pues el tipo de léxico y los errores detectados nos permiten afirmar que los informes han sido redactados por facultativos procedentes de España y también de Hispanoamérica, aunque no es posible determinar el porcentaje ni las regiones debido a la confidencialidad de los datos.

En primer lugar, debemos señalar que, al ser la misma lengua, se observan ciertas semejanzas entre el corpus de español general y el corpus de informes médicos. Sin embargo, si abordamos un nivel de concreción mayor son varias las diferencias a las que podemos aludir. Primeramente, debemos destacar que el porcentaje de errores es más elevado en nuestro corpus (3,19 %) con respecto a la muestra de español general, cuyo porcentaje se sitúa en 2,86 % (Tabla 28). A pesar de que ambos corpus están compuestos por textos no revisados, en el corpus de informes médicos el porcentaje de errores es superior, debido a las características lingüísticas y contextuales especiales que posee.

Especialidad	**Palabras**	**Errores *non-word***	**Porcentaje de error**
Dominio médico	2 321 826	74 155	3,19 %
Español general (Ramírez y López, 2006)	2 088 186	59 707	2,86 %

Tabla 28. Comparación de frecuencia de errores con corpus de español común

Los patrones de error de nuestro corpus parecen corroborar sólo parcialmente algunos de los hallazgos de patrones reportados por Ramírez y López (2006), como se puede observar en la Tabla 29 y la Figura 6. El error de omisión de tilde es el tipo de error más cometido en ambas investigaciones, aunque con un número significativamente más alto en el corpus de informes clínicos. En Ramírez y López (2006) se sostiene que la omisión es el error más frecuente, seguido de sustitución, inserción y transposición respectivamente. En cambio, en la muestra del dominio médico el orden es omisión, inserción, sustitución y transposición, por tanto, en los informes médicos los errores de inserción tienen una incidencia mayor que los de sustitución. La mayoría de los errores *non-word* son cometidos por la omisión de tilde en ambos corpus.

Es importante precisar que en el apartado de análisis de los subtipos de errores del trabajo de Ramírez y López (2006), los autores tienen en cuenta los resultados de la suma

de los tres conjuntos de datos, incluyendo también el conjunto de textos editados, por tanto, esa parte de los resultados no es directamente comparable debido a su naturaleza, pero se ha tomado como referencia al ser el único estudio disponible sobre tipificación de errores en un corpus de español común.

Tipo de error	Español general (Ramírez y López, 2006)	Dominio médico
Omisión	60,3 %	84,71 %
Inserción	9,1 %	5,91 %
Sustitución	16,8 %	3,85 %
Transposición	1,84 %	2,27 %
Multierror	11,96 %	3,26 %

Tabla 29. Comparación de tipos de errores con corpus de español común

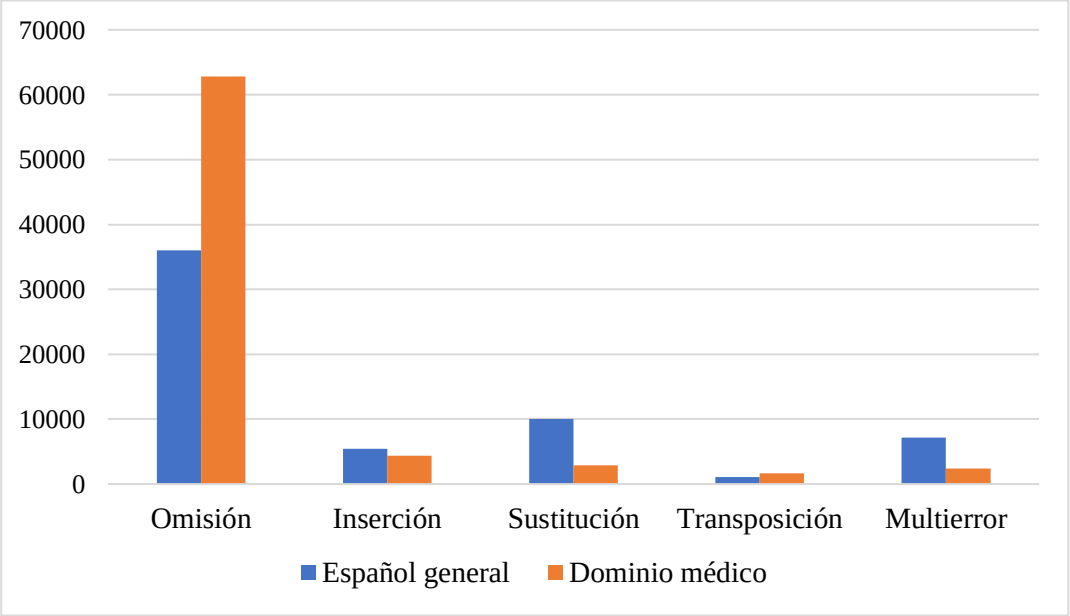

Figura 6. Errores según el tipo de operación de edición y el dominio

Ramírez y López (2006) sitúan los errores de omisión de tilde en torno al 52 %, mientras que en nuestro corpus se sitúa por encima del 60 % en todas las especialidades (76,67 % en urgencias, 80,82 % en UCI, 72,33 % en psiquiatría y 63,25 % en cirugía general).

La gran mayoría de los errores detectados en el corpus de español general (Ramírez y López, 2006) son errores ortográficos únicos en la palabra (más del 89 %), es decir, están a distancia de edición 1, y los errores ortográficos múltiples son menos del 9 %, con un porcentaje restante insignificante que incluye errores de espacios y cadenas de caracteres indescifrables (Ramírez y Lopez, 2006). En nuestro corpus se detectan 71 735 errores (96,74 %) a distancia 1 y 2420 errores (3,26 %) a una distancia de edición superior. Por consiguiente, la mayor parte de los errores presentes en el corpus tienden a ser casos únicos de inserción, omisión, sustitución o transposición en la palabra, de acuerdo también con los resultados de Damerau (1964), Kukich (1992), y Gimenes et al. (2015).

Las diferencias porcentuales observadas entre ambos corpus en la tabla nos llevan a pensar en la influencia que ejerce el dominio sobre los tipos de errores que se cometen, pero es necesario emplear un test estadístico que ponga a prueba esta hipótesis sobre la distribución de las frecuencias. Para ello se ha utilizado el test estadístico chi-cuadrado (Tabla 30), que nos permite reconocer la relación entre dos variables categóricas. Las variables que se pretenden analizar se clasifican en dominio (español general y dominio médico respectivamente) y tipo de error. De nuevo, se utiliza el programa estadístico *IBM SPSS*, con el que se genera una tabla cruzada que incluye los valores observados y los valores esperados según cada variable. El rechazo de la hipótesis nula implicaría que existe relación entre la variable dependiente y la independiente, apuntando a una asociación entre el dominio y los tipos de errores cometidos.

			Corpus		**Total**
			Español general	**Dominio médico**	
Tipo de error	Omisión	Recuento	36 003	62 818	98 821
		Recuento esperado	44 076,3	54 744,7	98 821,0
	Inserción	Recuento	5433	4377	9810
		Recuento esperado	4375,5	5434,5	9810,0
	Sustitución	Recuento	10 030	2855	12 885
		Recuento esperado	5747,0	7138,0	12 885,0

	Transposición	Recuento	1098	1685	2783
		Recuento esperado	1241,3	1541,7	2783,0
	Multierror	Recuento	7140	2420	9560
		Recuento esperado	4264,0	5296,0	9560,0
Total		Recuento	59 704	74 155	133 859
		Recuento esperado	59 704,0	74 155,0	133 859,0

Tabla 30. Tabla cruzada de tipo de error y dominio (corpus)

La Tabla 31 refleja los resultados obtenidos en la prueba. El valor de chi-cuadrado es 12 424,198 con 4 grados de libertad y el nivel de significación asintótica o *p-valor* es visiblemente inferior a 0,05. Por tanto, hay evidencia estadísticamente significativa para rechazar la hipótesis nula y confirmar que la distribución de los tipos de errores cometidos está relacionada con el dominio en el que se producen. A pesar de que esta prueba es considerada especialmente potente en tablas 2x2, el hecho de que nuestros datos no incluyan ninguna casilla igual a 0 ni ninguna frecuencia esperada inferior a 5 compensa el mayor tamaño de la tabla.

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	12424,198[a]	4	0,000
Razón de verosimilitud	12712,266	4	0,000
Asociación lineal por lineal	9177,531	1	0,000
N de casos válidos	133859		
a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 1241,28.			

Tabla 31. Prueba de chi-cuadrado

Por otro lado, a diferencia de otros trabajos y ámbitos, la variabilidad en las posiciones de los errores es mayor en nuestro corpus. La causa es que el léxico del lenguaje especializado de la medicina contiene palabras de mayor longitud que las usadas

en la lengua común. En nuestro corpus el mayor porcentaje de errores se concentra entre la posición segunda y sexta de las palabras (Figura 7). Según Ramírez y López (2006), la posición más probable para los errores ortográficos en español es alrededor del tercer, cuarto o quinto carácter de la palabra, siendo el tercer carácter el que presenta un porcentaje mayor (Figura 8). Asimismo, establecen que en alemán el cuarto carácter es la posición crítica para un error ortográfico. Por su parte, para el inglés podemos destacar el trabajo de Pollock y Zamora (1984), cuyo corpus de estudio concentró el mayor número de errores en la tercera posición (23 %). En el corpus del presente trabajo, la posición que presenta más errores es la segunda (20,56 %), seguida de la sexta (15,46 %), la quinta (13,61 %) y la tercera (12,63 %). La ventana de posición del error es más amplia en el corpus de informes médicos con respecto a otros estudios, debido a la mayor longitud media de la terminología médica, como mencionábamos anteriormente. Al igual que en el resto de trabajos consultados, son pocos los errores que ocurren en la primera letra de la palabra (1,64 %). Los resultados de los estudios llevados a cabo por Pollock y Zamora (1984), Yannakoudakis y Fawthrop (1983), Ren y Perrault (1992), Kukich (1992), Ramírez y Lopez, (2006), Gimenes et al. (2015), o Rodríguez-Rubio y Fernández-Quesada (2020), entre otros, establecen que se producen pocos errores ortográficos en la primera letra de una palabra, afirmación que coincide con nuestros resultados. Según Rodríguez-Rubio y Fernández-Quesada (2020), que describen patrones de errores detectados en tres diccionarios especializados de inglés-español, la menor presencia de errores en posición inicial obedecería a aspectos psicolingüísticos. Estos autores defienden que las posiciones inicial y final son más sobresalientes que las posiciones intermedias y, por tanto, los elementos de los extremos se recuerdan mejor y esta circunstancia afecta al proceso de escritura. Así, en su estudio, el 88 % de los errores *non-word* se produjeron en posiciones intermedias, el 2,3 % corresponde a errores de posición de la primera letra y el 9,7 % a errores de posición de la letra final (Rodríguez-Rubio y Fernández-Quesada, 2020). No obstante, se trata de una investigación sobre un corpus que había sido revisado previamente, por lo que no reúne las mismas condiciones que el nuestro.

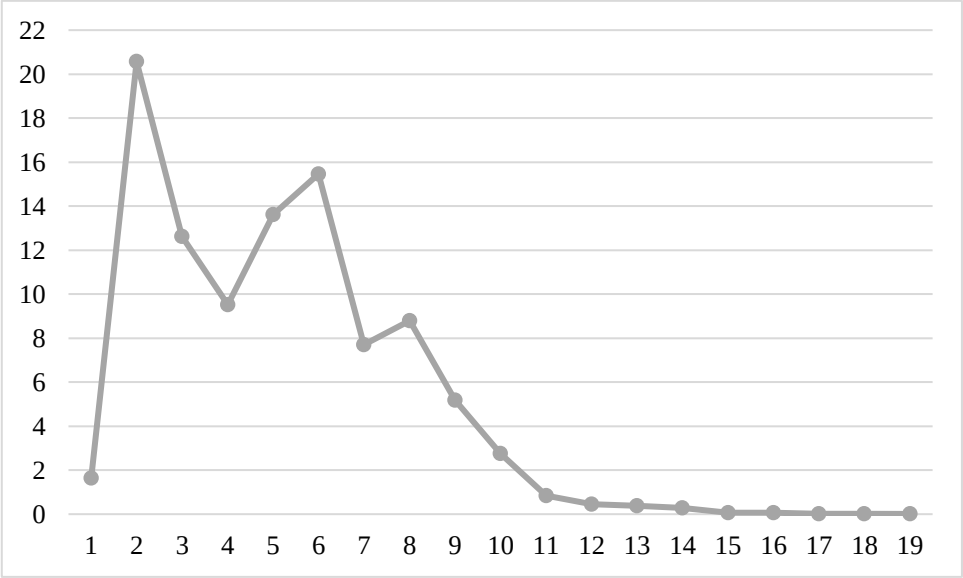

Figura 7. Posición de los errores *non-word* en las palabras de nuestro corpus según el porcentaje

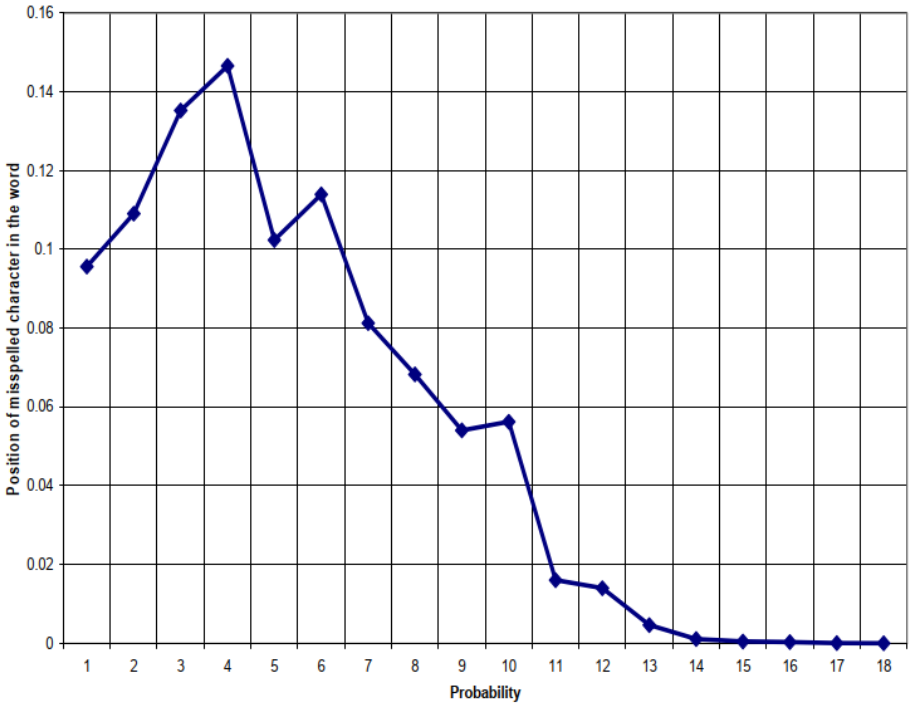

Figura 8. Probabilidad de que ocurra un error ortográfico en una posición determinada de una palabra en español (Ramírez y Lopez, 2006: 96)

En cuanto a los errores motivados por factores cognitivos, Ramírez y López (2006) afirman que en español este tipo de error consiste principalmente en sustituciones de una secuencia de letras fonéticamente correcta pero ortográficamente incorrecta, como

en el caso de los homófonos y los pares fonéticos similares (como *b-v*, *s-x*, *c-s*, *ll-y*). La mayoría de los errores cognitivos están provocados por fenómenos de sustitución, omisión e inserción. También consideran que los errores que involucran signos diacríticos, mayúsculas y minúsculas podrían considerarse errores cognitivos. Al incluir los errores de omisión de tilde como errores cognitivos, el porcentaje de este tipo de errores en su corpus sería muy alto, un 63 %, sin embargo, los resultados de nuestro corpus reflejan que las omisiones de tilde pueden tratarse de errores de actuación. En nuestro corpus, un importante porcentaje de los errores parecen provocados por la necesidad de agilizar el proceso de escritura, por causas mecánicas, o por el uso intencional de una ortografía no estándar, como hemos desarrollado con detalle en secciones anteriores. Como afirman Rodríguez-Rubio y Fernández-Quesada (2021), en ocasiones resulta imposible distinguir si se trata de un error cognitivo o de actuación al no conocer la causa. Debido a la naturaleza del corpus y la confidencialidad de la información, no se ha podido desarrollar un experimento controlado que pueda confirmar si se tratan de errores cognitivos o de actuación; no obstante, sí podemos mencionar ciertos casos susceptibles de ser errores cognitivos, como ha quedado recogido en la sección de análisis cualitativo.

Además de las diferencias a nivel cuantitativo sobre los tipos de errores *non-word* ya comentadas, podemos mencionar otras diferencias, como la presencia de errores relativos a la escritura de términos, extranjerismos, epónimos y nombres de fármacos, los errores en la formación de palabras mediante mecanismos de derivación y composición, la falta de consistencia en la creación de abreviaturas o el tratamiento de siglas, entre otros. La comparación de los resultados entre ambos corpus permite poner de manifiesto que se dan diferencias entre el dominio médico y el español general que deben ser tenidas en cuenta y es preciso, por tanto, profundizar en el análisis de corpus especializados.

4.4. Módulo basado en conocimiento lingüístico

El procesamiento del lenguaje natural se ayuda del aprendizaje automático para mejorar, acelerar y automatizar funciones de análisis de texto y de conversión de texto no estructurado en datos e información utilizables. Un modelo de aprendizaje automático es la suma del aprendizaje que se ha adquirido a partir de unos datos de entrenamiento, por tanto, resulta esencial crear un marco de aprendizaje propicio, mediante el suministro de

datos relevantes y con el formato adecuado. El objetivo de estos métodos es crear un sistema en el que el modelo mejore continuamente en la tarea que se le asigna, así el modelo irá cambiando a medida que adquiera más aprendizaje.

Existen diversos métodos de aprendizaje automático para PLN, según sea el problema a resolver. Es importante precisar la diferencia entre el aprendizaje supervisado y el no supervisado. El aprendizaje supervisado consiste en el uso de conjuntos de datos etiquetados o anotados con ejemplos para entrenar un modelo estadístico. Estos conjuntos de datos están diseñados para entrenar o supervisar algoritmos que sirven para clasificar datos o predecir resultados con precisión. Usando entradas y salidas etiquetadas, el modelo puede medir su precisión y aprender con el tiempo. Por su parte, el aprendizaje automático no supervisado emplea algoritmos de aprendizaje automático para analizar y agrupar muestras de datos sin etiquetar. Estos algoritmos son capaces de detectar patrones ocultos en los datos por sí solos, sin supervisión humana. Ambos tienen sus ventajas y limitaciones, por lo que habitualmente en PLN se suele ser optar por enfoques híbridos.

Nuestra tarea ha consistido en recopilar colecciones de ejemplos, patrones y datos a partir de un corpus real. De esta forma, preparamos un módulo basado en conocimiento lingüístico para que pueda implementarse, desarrollar el modelo y aplicar métodos de aprendizaje automático. La información recopilada a partir del análisis de errores en los informes clínicos reales puede ser utilizada en distintas partes del proceso de detección y corrección automática, como se observa en la Figura 9 y como explicaremos con detalle a continuación. Por esta razón, hablamos de una sección o módulo basado en conocimiento lingüístico, aunque realmente los resultados obtenidos se pueden utilizar en distintos estadios del proceso, como en la arquitectura de detección y corrección de errores (Sección 4.4.2) o en la de generación de errores para entrenar los sistemas (Sección 4.4.1.).

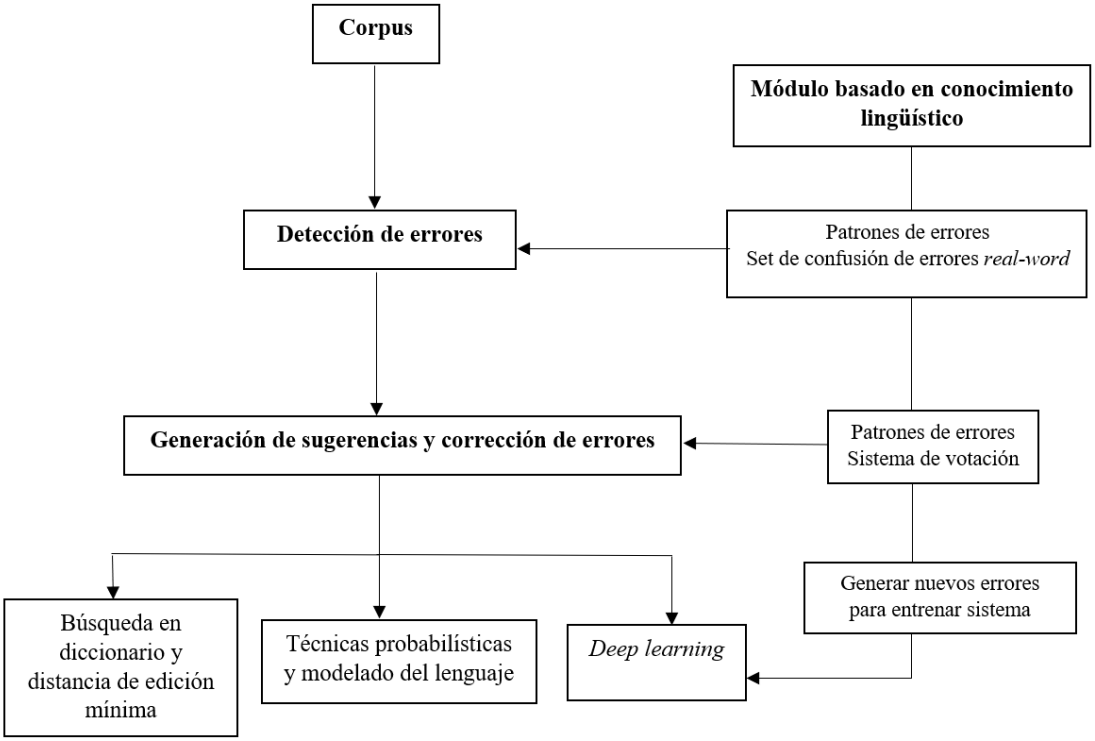

Figura 9. Aplicaciones del módulo basado en conocimiento lingüístico

4.4.1. Generación de errores sintéticos para conjuntos de entrenamiento

En el estado de la cuestión de los métodos de corrección automática se encuentran los modelos de redes neuronales profundas. El entrenamiento de estas arquitecturas requiere de grandes conjuntos de datos para la optimización de los parámetros del modelo. La falta de disponibilidad de estos grandes conjuntos supone un importante inconveniente, por lo que la solución adoptada más extendida consiste en introducir errores sintéticos en el corpus de entrenamiento. Los errores se pueden generar mediante métodos basados en reglas lingüísticas u operaciones de edición más genéricas y aleatorias.

En este caso, el análisis de un corpus real de dominio permite desarrollar nuevas reglas que imiten los errores detectados y, con ello, la creación de conjuntos de datos sintéticos. A partir de la información recopilada tras el análisis cuantitativo y cualitativo de resultados es posible establecer reglas formalizables por el sistema basadas en

conocimiento lingüístico e incorporarlas a herramientas de detección y corrección automática.

Hemos obtenido resultados respecto a la distancia de edición, los tipos de errores y los patrones que se repiten con más frecuencia en los informes analizados. Más del 95 % de los errores se producen a distancia 1 de la palabra correcta y el error más ampliamente cometido es el de omisión de tilde. Se pueden tener en cuenta los porcentajes y patrones detectados en los tipos y subtipos de errores, siendo los más habituales los de omisión de tilde, seguidos de omisión de letra, inserción, sustitución y transposición respectivamente. También se han recopilado los errores más comunes provocados por desconocimiento de la norma. Este tipo de datos proporciona pistas sobre qué patrones de error deben dotarse de mayor relevancia para mejorar la generación de listas de sugerencias para la corrección y la generación de datos sintéticos. Además, si se trata de errores dependientes del contexto o errores *real-word*, para establecer indicadores relacionados con los patrones de error más comunes que ayuden a detectarlos.

Asimismo, las matrices de confusión, generadas mediante la herramienta de cómputo y clasificación de errores, reflejan qué carácter se confunde por otro y con qué frecuencia. De esta manera, se identifican los errores de sustitución más frecuentes y las combinaciones de caracteres que están involucradas. Entre ellos se encuentran parejas de caracteres con similitudes fonéticas y parejas de caracteres adyacentes en el teclado (Tabla 32). En lugar de llevar a cabo cambios aleatorios para generar errores en el desarrollo de heurísticas, se pueden tener en cuenta estos patrones para que estos sean más representativos de los casos que se suelen dar.

E → A
O → A
A → E
I → E
E → O
O → I
I → O
S → C
C → S
S → X
X → S
C → Z
Z → C
I → Y
Y → I

V → B
B → V
M → N
N → M
L → R
R → RR
RR → R
C → V
G → F
S → A
S → D
T → R
R → T
R → S

Tabla 32. Patrones de sustituciones de caracteres más representativos

Además de los patrones sobre tipos y subtipos de errores, se han tenido en cuenta los errores detectados en el análisis cualitativo. A continuación, incorporamos de forma sucinta un listado de los patrones detectados, para obtener información más detallada sobre cada uno de ellos se puede consultar la sección 4.2. Entre los patrones podemos destacar:

- Inserción de tilde no diacrítica en palabras monosílabas.
- Demostrativos con tilde.
- Inserción de tilde en el grupo vocálico *-uí-*.
- Inserción de tilde en diptongos decrecientes (*-éu-*, *-éi-*, *-ói-*)
- Listado de palabras con alta frecuencia de aparición en los informes y con acentuación errónea por interferencia con palabras parónimas (ej.: *líbido**, *éstasis**, *contínuo*, psiquíatra**, etc.).
- Inserción de tilde en sustantivos en singular acabados en *-en* por analogía con formas en plural (ej.: *volúmen*, exámen*, jóven* resúmen*, gérmen*,* o *líquen**).
- Omisión de acentuación gráfica en letras mayúsculas.
- Formación errónea de palabras mediante derivación y composición:
    - Prefijos unidos con guion a la base que acompañan. Se utiliza listado de prefijos comunes en medicina acompañados de guion.
    - Prefijos separados de la raíz mediante espacio en blanco. Se utiliza listado de prefijos comunes en medicina separados por espacio.

- Confusión en el uso de los prefijos *trans-/tras-*. Intercambio de prefijo (ej.: *transplantar*, transladar**)

- Prefijos que aparecen de forma coordinada sin guion junto a otros prefijos (*pre** [*pre-*] *y postraqueal.*)

- Yuxtaposición de adjetivos con espacio y sin nexo para formar palabras compuestas (ej.: *inflamatorio infeccioso**).

- Compuesto univerbal con dos acentos. Palabras compuestas con tilde en las dos bases que la forman (ej.: *abdóminopélvica**).

- Combinación híbrida de guion y espacio en la formación de las palabras compuestas (ej.: *infero- posterior**).

- Listado de anglicismos escritos erróneamente:

  - Omisión de *s* en el grupo consonántico geminado *-ss-* (*bypas*, bypass*).

  - Omisión de *t* en el grupo consonántico geminado *-tt-* (*fluter*, flutter*).

  - Omisión de *p* en el grupo consonántico geminado *-pp-* (*flaping*, flapping*).

  - Omisión de *t* en el grupo consonántico *-tch* (*pach*, patch*).

  - Omisión de *t* en el grupo *-nt* posición final (*sten*, stent*).

  - Confusión *ea-ee* (*screaning*, screening*).

  - Omisión de *d* en grupos *-dt-* (*Schmit*, Schmidt*)

  - Imitación de la terminación *-ing* en voces terminadas en *-in* (*Hodking*, Hodgkin*).

  - Ausencia de concordancia numeral entre el anglicismo y el determinante o adjetivo que le acompaña.

- Listado de medicamentos y epónimos escritos erróneamente:

  - Inserción de letra *h, r* o *m.*

  - Inserción de doble *s* por influencia del inglés.

  - Sustitución de *m-n.*

  - Sustitución de *v-b.*

  - Sustitución de *y-i.*

  - Sustitución de *s-c.*

  - Sustitución de *q-k.*

- Simplificación de grupos consonánticos:

  - Grupo *-bs-*.

  - Grupo *-cc-*.

- Grupo -*ns*-.
- Grupo -*ct*-.
- Grupo -*pr*-.

- Representación gráfica de fonemas:
  - Confusión en el uso de *h* al inicio de palabra.
  - Confusión de *r* y *rr*.
  - Errores relativos al uso incorrecto de *b* y *v*.
  - Confusión en el uso de *j* y *g*.
  - Confusión entre *y* y *ll.*
  - Sustitución de *r* por *l* en posición implosiva o de final de palabra.
  - Representación gráfica de los fonemas /s/ y /z/. Errores en el uso de las letras *s, c* y *z.*
  - Errores en el uso de *x/s*
  - Confusión entre *c* y *k*
  - Reemplazo de vocal *a* por *e*.
  - Reemplazo de vocal *i* por *e*.
  - Monoptongación. Reducción de los elementos de un diptongo a una sola vocal (*ea-a, ei-i*).
  - Cierre vocálico. Cambio del timbre de la vocal que conlleva una subida en el trapecio vocálico. En español, las vocales más cerradas son *i* y *u*, la más abierta es *a*, mientras que *e* y *o* son intermedias en su grado de apertura (*oa-ua*).
- Listado de las formas erróneas por casos de analogía detectadas:
  - Errores provocados por la interferencia con otros paradigmas verbales.
  - Sustitución de *l* por *ll*.
  - Formación de singular en palabras invariables.
  - Inserción de *r.*
  - Inserción de *s.*
  - Sustitución de *d* por *t.*
  - Sustitución de *c* por *t.*
- Inserción de punto abreviativo en siglas y símbolos.
- Añadir -*s* para marcar el plural en siglas y símbolos.
- Omisión de espacio de separación entre el símbolo y la cifra que acompaña

En el caso de los errores *real-word*, es especialmente relevante contar con los patrones de error detectados para poder añadirlos a la tipología existente que ya ha sido empleada para crear corpus de entrenamiento. Son errores que suelen pasar desapercibidos en los procesos de detección, por esa razón, al entrenar el sistema para que aprenda de la casuística de errores que hemos recopilado, este será mucho más robusto.

En Bravo et al. (2021) implementamos un modelo de traducción automática neuronal *Seq2seq* para corregir errores *real-word* en textos clínicos. Para trabajar con modelos basados en *deep learning* es necesario un gran corpus de entrenamiento con errores, rico en el fenómeno que queremos modelar, por tanto, debe contener gran cantidad de ejemplos sobre los que queremos que el sistema aprenda a reconocer. No obstante, los corpus disponibles en abierto (Wikicorpus y una colección de tres conjuntos de datos clínicos) estaban revisados y no contenían errores reales, por lo que se generaron diferentes tipos de errores sintéticos mediante el uso de reglas. Entre los errores generados se encuentran: discordancia de género y número, sujeto-verbo, y ciertos homófonos (como *a/ha*, *b/v*, *ll/y* y *h*). Las reglas fueron diseñadas para identificar subcadenas específicas en la palabra de destino y reemplazarlas para generar una nueva palabra. Esta tipología ha sido ampliada y mejorada con los hallazgos sobre los tipos de errores detectados tras el análisis de los informes clínicos en la presente investigación. Se han recopilado nuevos patrones de errores que se introducen en el corpus de entrenamiento:

- Homófonos y palabras erróneas recopiladas con proximidad fonética. Se profundizó en la detección de homófonos, obteniendo una casuística mayor, como se especifica en la sección de análisis cualitativo de errores *real-word*.
- Tilde diacrítica, se inserta la pareja átona en las estructuras en las que debe ir la tónica y viceversa.
- Parónimos acentuales, formas erróneas según la posición de la sílaba tónica, como formas verbales.
- Secuencias que se escriben en una o más palabras con distinto valor.
- Errores de concordancia.
    - Discordancia nominal de género.
    - Discordancia nominal de número.
    - Discordancia nominal de género y número.

    - Discordancia verbal de número.
- Separación de las bases que forman palabras por composición.
- Formación de palabras por prefijación.
- Confusión entre pares de palabras.
- Secuencias que se escriben en una o más palabras con distinto valor.
- Introducción de formas verbales anómalas en el dominio.
    - Primera persona de presente en verbos no comunicativos.
    - Formas en segunda persona.
    - Imperativos en plural.
    - Formas en subjuntivo.
- Errores léxicos recopilados para el set de confusión.
- Información sobre los entornos típicos donde suelen producirse los errores.

Para el proceso de generación de errores sintéticos partimos de un corpus de entrada con frases correctas. Cada frase correcta pasa por un conjunto de heurísticas y, si sus elementos reúnen las condiciones establecidas, se aplican los cambios y se obtienen distintas frases con errores que servirían para entrenar el sistema. Para cada tipo de error se implementan distintas reglas, de forma que sea posible abarcar la mayoría de casos posibles. Los diferentes tipos de errores tienen características diferentes y requieren estrategias específicas para su corrección. Algunos ejemplos son:

- *Absceso preauricular izquierdo* → *Absceso pre auricular izquierdo*

Inserción de espacio entre el prefijo y la base a la que acompaña. Este cambio se aplicará si se detecta una palabra formada por un prefijo recogido en el lexicón de prefijos comunes en medicina y si su separación del resto de la raíz da lugar a una palabra correcta y recogida en el diccionario (*auricular*).

- *Exfumador con EPOC clínico* → *Ex-fumador con EPOC clínico*

Inserción de guion entre el prefijo y la base a la que acompaña. Este cambio se aplicará si se detecta una palabra formada por un prefijo recogido en el lexicón de prefijos comunes en medicina y si su separación del resto de la raíz da lugar a una palabra correcta y recogida en el diccionario (*fumador*).

- *Se halla lesión hepática segmento IV → Se haya lesión hepática segmento IV*

Sustitución de consonante en palabras homófonas (*ll-y*). El cambio se aplicará si se detecta un carácter o dígrafo codificado para ser sustituido al formar parte de la clasificación de parejas de caracteres con similitudes fonéticas y que habitualmente son confundidos. El cambio se llevará a cabo únicamente si la sustitución da lugar a una palabra correcta y recogida en el diccionario.

- *Ictus hemorrágico bulbar izq. → Ictus hemorrágico vulvar izq.*

Sustitución de consonante en palabras homófonas (*b-v*). Este cambio se aplicará si se detecta un carácter o dígrafo codificado para ser sustituido al formar parte de parejas de caracteres con similitudes fonéticas y que habitualmente son confundidos. El cambio se llevará a cabo únicamente si la sustitución da lugar a una palabra correcta y recogida en el diccionario.

- *No hábito enólico → No habito enólico*

Omisión o inserción de tilde en parónimos acentuales. Si la omisión de tilde da lugar a una palabra recogida en el diccionario, se realiza el cambio (*hábito-habito*).

- *Desconexión parcial del entorno → Desconexión parcial del en torno*
- *Glucemias en ayunas en torno a 110 → Glucemias en ayunas entorno a 110*

Confusión entre pares definidos de palabras. Si aparece en el corpus una palabra recogida en el set de confusión, se sustituye por su par (*entorno-en torno*).

- *Gases venosos sin alteraciones → Gases venoso sin alteraciones*

Discordancia nominal de número. Supresión de *-s* en la terminación de plural *-os* si da lugar a una palabra correcta. Este cambio no se aplicará en dos palabras seguidas para poder mantener la discordancia entre sustantivo y adjetivo.

- *Aplastamiento vértebras lumbares → Aplastamiento vertebras lumbares*
- *Fiebre en contexto de paciente oncológico → Fiebre en contesto de paciente oncológico*

Utilizar el set de confusión para introducir errores, como verbos en primera persona de presente del singular, segunda persona del singular y otras formas verbales

anómalas en el dominio. Se sustituirán las formas detectadas en el set de confusión por formas a distancia de edición cercana.

Estos son solo algunos ejemplos de cómo se pueden emplear las reglas mencionadas. Para evitar que se genere un corpus muy artificial, se deben limitar los cambios por oración, de manera que se generen distintas versiones según la palabra de la oración editada. Los conjuntos de datos de entrenamiento estarán compuestos por las oraciones con errores y las correspondientes versiones corregidas.

4.4.2. Detección de errores y generación de candidatos de corrección

Además de en la fase de generación de errores para el aumento de datos de entrenamiento, la recopilación de los fenómenos que más frecuentemente constituyen errores permite aportar información en la arquitectura de decisión y ponderación de alternativas de corrección. En este caso, la ayuda no se llevaría a la fase de entrenamiento, sino al momento de detección y decisión.

Los resultados permiten proporcionar indicaciones a los sistemas de corrección para optimizar el tiempo de búsqueda durante el proceso de detección. La especificación de ciertos patrones de error detectados en los informes puede facilitar la detección y tratamiento de estos mediante estrategias heurísticas. Un ejemplo es la detección de formas anómalas en el dominio, como algunas estructuras verbales (imperativos, subjuntivos, formas de segunda persona, etc.). También se puede tener en cuenta la información sobre criterios posicionales, léxicos o morfológicos. Resulta, por tanto, especialmente útil constituir un módulo dentro de la arquitectura que detecte automáticamente los patrones indicados en el caso de los errores *real-word*, pues habitualmente pasan desapercibidos. Para ello, nos podemos valer del uso de sets de confusión. En ellos se codifican específicamente las parejas de palabras que frecuentemente presentan error o son confundidas y su correspondiente corrección. Estos pueden estar formados por la recopilación de los errores detectados en el análisis, con los casos prototípicos que se han detectado, o ser creados a gran escala a partir de la aplicación de reglas sobre un lexicón. Los trabajos que abordan errores *real-word* o errores que necesitan contexto también suelen emplear modelos de lenguaje basados en

n-gramas, y muchos de ellos incluyen sets de confusión para focalizar en errores determinados y reducir el tiempo en el proceso de clasificación. En los sistemas más actuales cada aparición de una palabra en el texto se representa como un vector de características y un clasificador se entrena en un corpus previamente corregido. En el texto con errores, para cada palabra del texto que aparece en el set de confusión, el clasificador predice el candidato más probable del set de confusión teniendo en cuenta el contexto de la palabra.

Asimismo, es posible mejorar la precisión en la generación de sugerencias de candidatos para la fase de corrección. Esto es posible en arquitecturas que basan el proceso de corrección en un sistema de votación. Mediante las distintas operaciones de edición que se aplican en la palabra se generan distintas alternativas y para la ordenación de estas alternativas o candidatos de corrección se utilizan distintas técnicas de ponderación. Los sistemas con los que hemos trabajado habitualmente tienen en cuenta la frecuencia de aparición de la palabra, los resultados del modelo lingüístico y otras técnicas basadas en *deep learning*, y a cada criterio o técnica se le asigna un peso en la ponderación. La información lingüística recopilada puede utilizarse como un criterio más en el sistema. Una posibilidad está relacionada con las parejas de caracteres que plantean confusión entre sí con frecuencia, se les puede otorgar una puntuación mayor en el cálculo de probabilidades de la lista de candidatos de corrección, pudiendo así ayudar a resolver disyuntivas a la hora de ordenar sugerencias de candidatos en la ponderación de alternativas. También se pueden añadir reglas para corregir los errores más comunes provocados por desconocimiento de la norma, como los que tienen que ver con la formación de palabras. Mediante la suma de los resultados de cada técnica el sistema toma una decisión sobre el candidato más adecuado.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1. Consideraciones finales

Este trabajo ha aportado un análisis y clasificación de errores en un corpus formado por informes clínicos en español. Los resultados han sido recopilados a partir de cuatro especialidades médicas, por tanto, este enfoque comparativo ha permitido identificar semejanzas y diferencias entre estas especialidades. Asimismo, se ha contribuido a la tipificación de errores característicos en este dominio, mediante la detección de patrones lingüísticos, y a la mayor cobertura de casos para corrección automática, complementando las técnicas basadas en análisis estadístico y aprendizaje automático.

Para ello, se ha llevado a cabo una aproximación que ha incluido la implementación de un modelo lingüístico basado en n-gramas, la representación vectorial de las palabras del corpus a partir de *Word2Vec* y el etiquetado gramatical del corpus. Se ha estimado la verosimilitud de los diferentes bigramas y trigramas mediante una distribución de probabilidad, partiendo de que las secuencias correctas tienen una probabilidad notablemente más alta que las incorrectas. La utilización de estas distintas técnicas se ha llevado a cabo con el objetivo de que actuaran de forma complementaria para detectar el mayor número de errores, siendo especialmente fructífero para la detección de errores el modelo lingüístico basado en n-gramas y la representación vectorial de las palabras.

Los resultados han indicado que la especialidad cuyos informes médicos presentan una mayor tasa de errores es urgencias. Las frecuencias y tipos de errores detectados en el corpus han permitido conocer que muchos de los errores presentan patrones de reproducción consistentes que es posible sistematizar. La mayoría de las palabras que contienen errores están a distancia de edición 1 con respecto a la palabra correcta, gran parte de los errores *non-word* detectados se concentran en un número reducido de caracteres, y el tipo de error más común con una alta incidencia es el de omisión. Entre los patrones de error con más presencia en el corpus destacan la omisión de tilde, la inserción de tilde, la sustitución de caracteres con similitudes fonéticas (como *s*-*c*, *c*-*z*, *n*-*m*, o *y*-*i*), los errores provocados por desconocimiento de la norma ortográfica

actual, y los errores derivados del uso del teclado, como la sustitución y transposición de caracteres adyacentes. Además, se ha detectado la confusión y falta de consistencia en la formación de palabras mediante derivación y composición, errores en la escritura de extranjerismos y nombres propios, así como errores en la representación gráfica de determinados fonemas, grupos consonánticos y analogía con otras formas.

En cuanto a los errores *real-word* o errores dependientes del contexto, a los que también hemos dedicado una sección específica en este trabajo, los resultados han reflejado que la presencia de este tipo de errores es reducida en el corpus y considerablemente inferior a la de errores *non-word*. No obstante, para mejorar la precisión de los sistemas de corrección en el dominio médico es fundamental abordar y desarrollar técnicas para la detección de estos errores encubiertos. Los análisis realizados han permitido conocer con mayor exhaustividad los tipos de errores *real-word* que suelen aparecer en los informes clínicos y hemos detectado patrones que aparecen con frecuencia en el corpus. Al igual que en los errores *non-word*, la mayor parte de los errores detectados se han concentrado en un número reducido de caracteres, como los caracteres vocálicos en los que se ha producido la omisión o inserción de tilde, siendo la omisión de tilde el tipo de error que ha ocurrido con más frecuencia. También se han generado errores por la omisión de caracteres, especialmente al final de la palabra, o la sustitución de caracteres vocálicos por otros (*o-a*, *a-o*, *a-e* y *e-i*), dando lugar especialmente a errores de concordancia. Otros tipos de errores detectados han sido la inserción de letra, tilde o espacio, mientras que los errores de transposición no han tenido apenas presencia como desencadenantes de errores *real-word*. Asimismo, se han descubierto errores motivados por fenómenos de paronimia; errores de discordancia nominal y verbal; errores provocados por la incorrecta unión o separación de palabras; errores por desconocimiento de la norma ortográfica actual; formas verbales incorrectas en el contexto; y, por último, errores léxicos causados por errores de sustitución, omisión o inserción de un carácter. Este estudio ha permitido constatar la variabilidad de tipos de errores que es necesario tener en cuenta, por tanto, una tipología de error *real-word* centrada únicamente en errores de discordancia de género y número y homófonos, como se ha hecho habitualmente, resulta insuficiente para el proceso de detección y para el entrenamiento de modelos de corrección.

La tipología de errores desarrollada va a servir como guía para generación de errores de forma más exhaustiva y sistemática, dando lugar a mejores conjuntos de

entrenamiento. La definición de estos nuevos tipos de errores específicos va a contribuir al desarrollo de sistemas basados en reglas para la generación de errores artificiales que, a su vez, son utilizados para producir material de entrenamiento en los sistemas de corrección automática. Los datos de entrenamiento tienen una influencia determinante en el desempeño de los modelos basados en aprendizaje profundo. Además, esta tipología aporta conocimiento que puede ser empleado, por un lado, para mejorar la fase de detección mediante el uso de conjuntos de confusión y, por otro, para la ponderación de alternativas en la fase de corrección teniendo en cuenta los patrones y datos obtenidos sobre la frecuencia de los tipos de errores.

De esta forma, se ha pretendido contribuir al desarrollo y perfeccionamiento de sistemas de corrección automática que, a su vez, son utilizados para el procesamiento de datos en medicina.

5.2. Limitaciones

Aunque la aproximación empleada nos ha permitido detectar y analizar un gran número de errores y aportar una clasificación exhaustiva, presenta algunas limitaciones que deben ser tenidas en cuenta para investigaciones futuras.

El principal obstáculo de este trabajo se ha encontrado en la fase de detección. El núcleo de este ha sido clasificar los errores existentes en un corpus de grandes dimensiones, de ahí la importancia en la ejecución de la fase de detección. Si un error ha pasado desapercibido en nuestro experimento, no ha sido detectado y, como consecuencia, no ha sido posible analizarlo y cuantificarlo. Somos conscientes de que pueden haber quedado errores sin detectar en el corpus y es una circunstancia que debemos tener presente para el futuro. Pueden no haber sido detectados errores *real-word* relacionados con cuestiones semánticas y pragmáticas, especialmente aquellos casos cuyos elementos desencadenantes del error estaban alejados entre sí en la sentencia y han podido dar lugar a incoherencias gramaticales y semánticas. No obstante, a no ser que se llevara a cabo una lectura y revisión manual del corpus, una tarea inviable en corpus de grandes dimensiones como el de este trabajo, era inevitable que estas restricciones pudiesen producir. Es por ello por lo que continuamos trabajando para mejorar la fase de detección de errores, ampliando la ventana de contexto y experimentando con distintos

parámetros de ejecución, para que esta sea totalmente exhaustiva y sea posible la detección de todo tipo de error en los distintos planos del lenguaje.

La otra limitación de este trabajo tiene que ver con la evaluación. Esta tesis doctoral se ha situado en una esfera descriptiva, y ha aportado un análisis lingüístico y una tipificación sobre tipos de errores para ser utilizada en distintas fases del proceso de corrección automática. Sin embargo, consideramos relevante poder ir un paso más allá y evaluar el rendimiento que la incorporación del módulo basado en conocimiento lingüístico puede aportar a los sistemas. Un ejemplo puede ser comprobar si gracias a la incorporación de los nuevos patrones de error en el corpus de entrenamiento la red neuronal es capaz de decidir mejor a nivel global o en errores específicos.

Finalmente, la naturaleza sensible del corpus no nos ha permitido poder obtener información a nivel sociolingüístico sobre los autores de los informes recogidos, lo que ha impedido poder analizar otros factores que pudiesen influir en la producción de errores en los informes y que pudiese ser relevante tener en cuenta sobre la muestra de lengua.

5.3. Trabajo futuro

Las investigaciones futuras están en estricta relación con la anterior sección, pues se va a trabajar sobre las limitaciones mencionadas anteriormente.

La tipología de errores se ha elaborado de manera empírica y se han extraído errores de contextos lingüísticos reales. En proyectos futuros se van a ampliar los conjuntos de datos y a analizar otras especialidades médicas para compararlas con los resultados obtenidos. Se va trabajar en la actualización y la expansión del módulo, mediante la ampliación del repertorio de reglas lingüísticas para generar errores y el estudio de otros métodos.

Asimismo, se va a continuar investigando sobre la mejora de los parámetros del modelado del lenguaje y la aplicación de distintas técnicas y arquitecturas para la detección de errores *real-word,* como el entrenamiento de los vectores con otras arquitecturas como las basadas en *transformers*.

Se van a construir corpus de datos de entrenamiento con la inclusión de los errores recopilados en este trabajo, contribuyendo a la creación de recursos para el procesamiento de textos médicos en español. Se integrará el módulo en el sistema de detección y corrección, cuya arquitectura está compuesta por otras técnicas estadísticas y puramente

computacionales, y se definirá una fórmula de decisión y ponderación de las alternativas generadas automáticamente para las palabras incorrectas, mediante la combinación de criterios.

Por último, a partir de los conjuntos creados se entrenará un modelo neuronal Seq2seq y se evaluará el rendimiento del sistema.

En síntesis, con esta tesis doctoral hemos querido aportar una mirada lingüística al informe médico y contribuir al estudio de errores lingüísticos que se cometen en este dominio a partir del análisis de casos reales, proporcionando recursos y una nueva tipología de errores para el español.

6. CONCLUSIONS AND FUTURE WORK

6.1. Final considerations

This work has provided an analysis and classification of errors in a corpus made up of clinical reports in Spanish. The results have been collected from four medical specialties; therefore, this comparative approach has made it possible to identify similarities and differences between these specialties. Likewise, it has contributed to the classification of characteristic errors in this domain, through the detection of linguistic patterns, and to the greater coverage of cases for automatic correction, complementing techniques based on statistical analysis and machine learning.

An approach that has included the implementation of a linguistic model based on n-grams, the vector representation of the words of the corpus from *Word2Vec* and the grammatical labeling of the corpus has been carried out. The likelihood of the different bigrams and trigrams has been estimated using a probability distribution, taking into account that the correct sequences have a notably higher probability score than the incorrect ones. The objective of these different techniques has been to work in a complementary way to detect the greatest number of errors. The linguistic model based on n-grams and the vector representation of words have been especially fruitful for error detection.

The results indicate that the specialty with the highest rate of errors in medical reports is emergency medicine. The frequencies and types of errors detected in the corpus have revealed that many of the errors have consistent reproduction patterns that can be systematized. Most of the erroneous words have an editing distance of 1 with respect to the correct word, a large part of the non-word errors detected are concentrated in a reduced number of characters, and the most common type of error with a high incidence is omission. The most frequent error patterns in the corpus are the omission of accent marks, the insertion of accent marks, the substitution of characters with phonetic similarities (such as *s*-*c*, *c*-*z*, *n*-*m*, or *y*-*i*), errors caused by ignorance of the current orthographic norm, and errors derived from the use of the keyboard (such as the substitution and transposition of adjacent characters).

In addition, other errors detected include inconsistency in the formation of words through derivation and composition, errors in the writing of foreign words and proper names, as well as errors in the graphic representation of some phonemes, consonant clusters, and analogy with other forms.

A specific section has also been devoted to real-word errors or context-dependent errors. In this case, the results have shown that the presence of this type of errors is low in the corpus, and considerably lower than non-word errors. Nevertheless, it is essential to develop techniques for the detection of these hidden errors in order to improve the accuracy of correction systems in the medical domain. The analyses carried out have provided a more exhaustive knowledge of the types of real-word errors that usually appear in clinical reports, and patterns that appear frequently in the corpus have been detected. Most of the errors detected have been concentrated in a reduced number of characters, such as vowel characters in which the omission or insertion of an accent mark has occurred. The omission of accent mark is the type of error that occurs most frequently, as in non-word errors. Errors have also been generated due to the omission of characters, especially at the end of the word, or the substitution of vowel characters for others (*o*-*a*, *a*-*o*, *a*-*e* and *e*-*i*), especially giving rise to agreement errors. Other types of errors detected have been the insertion of a character, accent mark or space, while transposition errors have hardly been present as triggers of real-word errors. Likewise, errors motivated by paronymy have been discovered; nominal and verbal discordance errors; errors caused by the incorrect union or separation of words; errors due to ignorance of the current orthographic norm; incorrect verbal forms in the context; and lastly, lexical errors caused by errors of substitution, omission or insertion of a character. This study has made it possible to verify the variability of types of errors that must be taken into account. Therefore, a typology of real-word errors focused solely on gender and number discordances and homophone errors, as has usually been done, is insufficient for the detection process and for training of correction models.

The typology of errors developed will serve as a guide for generating errors in a more exhaustive and systematic way, resulting in better training sets. The definition of these new types of specific errors will contribute to the development of rule-based systems for the generation of artificial errors, which, in turn, are used to produce training material in automatic correction systems. Training data has a determining influence on the performance of models based on deep learning.

Furthermore, this typology provides knowledge that can be used, on the one hand, to improve the detection phase through the use of confusion sets and, on the other, for the weighting of alternatives in the correction phase, taking into account the patterns and data obtained on the frequency of the types of errors.

6.2. Limitations

The approach used has allowed us to detect and analyze a large number of errors and provide an exhaustive classification, but it has some limitations that must be taken into account for future research.

The main obstacle of this work has been found in the detection phase. The core of this has been to classify the existing errors in a large corpus, hence the importance in the execution of the detection phase. If an error has gone unnoticed in our experiment, it has not been detected and, as a consequence, it has not been possible to analyze and quantify it. We are aware that some errors may have remained undetected in the corpus and it is a circumstance that we must bear in mind for the future.

Some real-word errors related to semantic and pragmatic issues may not have been detected, especially in cases where the elements triggering the error were far from each other in the sentence and may have given rise to grammatical and semantic inconsistencies. However, it was inevitable that these restrictions could occur, unless a manual reading and revision of the corpus was carried out, an unfeasible task in large corpora such as the one in this work. This is why we continue working to improve the error detection phase, widening the context window and experimenting with different execution parameters. The future goal is to make it completely exhaustive and to make it possible to detect all types of errors in the different language levels.

The other limitation of this work has to do with evaluation. This doctoral thesis has been situated in a descriptive sphere, and has provided a linguistic analysis and a classification of types of errors to be used in different phases of the automatic correction process. However, it is relevant to be able to go a step further and evaluate the performance that the incorporation of the module based on linguistic knowledge can bring to the systems. An example could be to check if, thanks to the incorporation of the new error patterns in the training corpus, the neural network is able to decide better on a global level or on specific errors.

Finally, the sensitive nature of the corpus did not allow us to obtain sociolinguistic information on the authors of the reports collected. This circumstance has prevented the analysis of other factors that may influence the production of errors in the reports and are relevant for a better understanding of the language sample.

6.3. Future work

Future research is strictly related to the previous section, as it will work on the limitations mentioned above.

The typology of errors has been developed empirically and errors have been extracted from real linguistic contexts. Future projects will attempt to extend the datasets and analyze other medical specialties in order to compare them with the results obtained. Work will be done on updating and extending the module, increasing the repertoire of linguistic rules for generating errors and exploring other methods.

Research will also continue on the improvement of language modelling parameters and the application of different techniques and architectures for real-word error detection, such as vector training with other architectures such as those based on transformers.

Training data corpora will be built with the inclusion of the errors collected in this work, contributing to the creation of resources for medical text processing in Spanish. The module will be integrated into the detection and correction system, whose architecture is composed of other statistical and purely computational techniques, and we will define a decision and weighting formula for the automatically generated alternatives for the erroneous words, by combining criteria.

Finally, a Seq2seq neural model will be trained on the ensembles created and the performance of the system will be evaluated.

7. CONTRIBUCIONES CIENTÍFICAS DERIVADAS DE LA TESIS DOCTORAL

La realización de esta tesis doctoral ha posibilitado la publicación de artículos en revistas de investigación, capítulos de libro y comunicaciones en congresos internacionales, como veremos con más detalle a continuación.

Los artículos científicos derivados de la tesis doctoral han sido sometidos a procesos de revisión por pares y han sido aceptados en revistas científicas de reconocido prestigio, que están indexadas en bases de datos internacionales, como *Journal Citation Report* (JCR), *Scimago Journal & Country Rank* (SJR), *Latindex* y en el listado de revistas científicas españolas FECYT. Cada uno de los trabajos realizados supone la materialización de las distintas fases del desarrollo de la tesis.

Durante la etapa inicial de la tesis doctoral, que estuvo dedicada al estudio teórico y a la investigación sobre el estado del arte en corrección automática y, más concretamente, sobre corrección automática en el dominio médico, se constató la existencia de numerosos estudios que presentaban técnicas, entornos y características muy diferentes, por lo que se consideró apropiado llevar a cabo una revisión sistemática de la literatura para aportar a la comunidad científica una síntesis de todas las investigaciones relevantes hasta la fecha. En este trabajo, que dio lugar a un capítulo de libro (López-Hernández et al., 2019), se plasmó una revisión bibliográfica en torno a la corrección automática y al análisis de errores en el ámbito biosanitario, se investigaron las principales técnicas de detección y recursos utilizados en el área, así como las limitaciones y retos existentes que debían ser superados. El objetivo principal fue conocer los desafíos actuales que la corrección automática presentaba específicamente en el lenguaje médico.

Posteriormente, en López-Hernández et al. (2021), se reflejan los resultados obtenidos tras el primer análisis de errores en informes médicos en español. Es una investigación de carácter exploratorio en la que se desarrolló una aproximación inicial para detectar errores, que fueron analizados posteriormente. La muestra analizada estaba formada por una colección de informes pertenecientes a la especialidad médica de

urgencias. Durante el análisis fueron examinadas variables como la frecuencia de aparición del error y la longitud de la palabra. Se llevó a cabo una clasificación inicial de carácter cualitativo, que permitiese el reconocimiento de patrones y el posterior diseño de una tipología de errores. Este estudio se ocupa exclusivamente de los errores *non-word*, es decir, aquellos que dan lugar a palabras no existentes.

Conforme avanzó la investigación, la variabilidad del corpus fue ampliada y se recopilaron informes clínicos de nuevas especialidades médicas: UCI, cirugía general y psiquiatría. El estudio resultante (López-Hernández y Almela, 2021) supuso el primer análisis cuantitativo de tipos errores en un corpus de informes médicos en español. En él se reflejan cuáles son los tipos de errores más frecuentes, qué parejas de caracteres suelen ser confundidos, si existen diferencias significativas en los resultados de las especialidades médicas analizadas o si hay diferencias entre los errores detectados en el dominio médico y la tipificación existente sobre errores del español general.

En la última etapa de la tesis doctoral, la investigación se centró en la detección y análisis de errores *real-word*. Se investigó sobre estas técnicas, se realizaron diversas pruebas con modelos lingüísticos y *word embeddings*, y también se definieron criterios de análisis de error. En Bravo-Candel et al. (2021) se introdujeron diferentes tipos de errores en un corpus mediante el uso de reglas, para el entrenamiento de un modelo de traducción automática neuronal *Seq2seq* utilizado para corregir errores *real-word* en textos clínicos en español. Finalmente, en López-Hernández et al. (2022), —artículo que a la culminación de esta tesis doctoral se encontraba en proceso de evaluación—, se presentan los resultados tras haber llevado a cabo la identificación, análisis y clasificación sistemática de errores *real-word* contenidos en informes clínicos a partir de las técnicas y herramientas desarrolladas.

7.1. Publicaciones en revistas de investigación

1. López-Hernández, J., Almela, Á. y Valencia-García, R. (2021). Linguistic errors in the biomedical domain: Towards an error typology for Spanish. *Sintagma: revista de lingüística*, 33, 83-100. ISSN: 2013-6455. https://doi.org/10.21001/sintagma.2021.33.05

2. López-Hernández, J., y Almela, Á. (2021). Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas. *Panace@. Revista de medicina, lenguaje y traducción*, 22 (53), 96-108. ISSN: 1537-1964. https://www.tremedica.org/wp-content/uploads/panacea21-53_13_Tribuna_07_LopezHernandez-Almela.pdf

3. Bravo-Candel, D., López-Hernández, J., García-Díaz, J. A., Molina-Molina, F., y García-Sánchez, F. (2021). Automatic correction of real-word errors in Spanish clinical texts. *Sensors*, 21 (9), 2893. https://doi.org/10.3390/s21092893

4. López-Hernández, J., Molina-Molina, F., y Almela, Á. (2022). Detección automática de errores lingüísticos en textos clínicos: análisis y tipificación de errores *real-word* en español. En proceso de evaluación en la revista *Linguamática*. https://linguamatica.com/index.php/linguamatica

7.2. Capítulos de libro

1. López-Hernández, J., Almela, Á. y Valencia-García, R. (2019). Automatic spelling detection and correction in the medical domain: A systematic literature review. En R. Valencia-García, G. Alcaraz-Mármol, J. Del Cioppo-Morstadt, N. Vera-Lucio y M. Bucaram-Leverone (eds.): *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science* (vol. 1124, pp. 104-117). Cham: Springer. ISBN: 978-3-030-34988-2. https://doi.org/10.1007/978-3-030-34989-9_8

7.3. Congresos

1. López-Hernández, J. (2020). Avances en el análisis y tipificación de errores para una propuesta de mejora de informes médicos en español. En *CEUR Workshop Proceedings* (eds.): *Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network* (PLNnet-DS-2020), pp. 58-63,

Universidad de Jaén (España), 16 de diciembre de 2020. http://ceur-ws.org/Vol-2802/paper9.pdf

2. López-Hernández, J., y Almela, Á. (2020). Detección y corrección automática en textos especializados: análisis de patrones de errores en un corpus del dominio médico. En *Actas del III Congreso Internacional de Lingüística Computacional y de Corpus (CILCC 2020) y V Workshop en Procesamiento Automatizado de Textos y Corpus (WoPATeC 2020).* Universidad de Antioquia, Medellín (Colombia), 21-23 de octubre de 2020. https://cilcc20.files.wordpress.com/2020/11/libro-de-resumenes-actas-iii-cilcc-2020-y-v-wopatec-2020-virtual.pdf

3. López-Hernández, J. (2019). Análisis y tipificación de errores para una propuesta de mejora de informes médicos en español. En E. Lloret, E. Saquete, P. Martínez-Barco y R. Sepúlveda-Torres (eds.): *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language* Processing (SEPLN 2019), pp. 38-43, Universidad del País Vasco (España), 25 de septiembre de 2019. http://ceur-ws.org/Vol-2633/paper7.pdf

4. López-Hernández, J. (2019). Análisis y tipificación del contexto relevante para una propuesta de mejora de corpus médicos para la generación de modelos de lenguaje. En *V Jornadas Doctorales. Universidad de Murcia, Campus Mare Nostrum, Escuela Internacional de Doctorado (UM-UPCT)*, Universidad de Murcia (España), 29-31 de mayo de 2019.

5. López-Hernández, J., y Almela, Á. (2019). El papel de la lingüística en la detección automática de errores en el ámbito biosanitario: hacia una propuesta de tipología de errores. En *XI Congreso Internacional de Lingüística de Corpus (CILC2019)*. Universidad de Valencia (España), 15-17 de mayo de 2019. https://adeit-estaticos.econgres.es/19_CILC/book_abstracts.pdf

BIBLIOGRAFÍA

Aduriz, I., Alegría, I., Artola, X., Ezeiza, N., Sarasola, K. y Urkia, M. (1997). A spelling corrector for Basque based on morphology. *Literary and Linguistic Computing*, 12 (1), 31-38.

Aguilar Ruiz, M. J. (2013a). Las normas ortográficas y ortotipográficas de la nueva Ortografía de la lengua española (2010) aplicadas a las publicaciones biomédicas en español: una visión de conjunto, *Panace@*, 14 (37), 101-120. https://www.tremedica.org/wp-content/uploads/n37-tribuna-MJAguilarRuiz.pdf

Aguilar Ruiz, M. J. (2013b). Manual de estilo para la publicación de originales en Revista Hispanoamericana de Hernia. *Revista hispanoamericana de hernia*, 1, 37-43.

Ahmed, F., Luca, E. W. D. y Nurnberger, A. (2009). Revised N-Gram based automatic spelling correction tool to improve retrieval effectiveness. *Polibits*, 40, 39-48.

Al-Jefri, M. M. y Mahmoud, S. A. (2013). Context-sensitive Arabic spell checker using context words and N-Gram language models. En *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*. Al-Madinah: Institute of Electrical and Electronics Engineers (IEEE), pp. 258-263.

Alcaraz, E. y Martínez, M. A. (1997). *Diccionario de lingüística moderna*. Barcelona: Ariel.

Aleixandre-Benavent, R. y Amador-Iscla, A. (2001). Problemas del lenguaje médico actual (II): Abreviaciones y epónimos. *Papeles Médicos*, 10(4), 170-176.

Aleixandre-Benavent, R., Bueno-Cañigral, F. J. y Castelló-Cogollos, L. (2017). Características del lenguaje médico en los artículos científicos. *Educación médica*, 18 (2), 23-29.

Aleixandre-Benavent, R., Valderrama, J. C. y Bueno-Cañigral, F. J. (2015). Utilización adecuada del lenguaje médico: principales problemas y soluciones. *Revista Clínica Española*, 215 (7), 396-400. https://doi.org/10.1016/j.rce.2015.04.001

Alpízar Castillo, R. (2005). *El lenguaje en la medicina. Usos y abusos*. Salamanca: Clavero.

Atserias, J., Fuentes, M., Nazar, R. y Renau, I. (2012). Spell checking in Spanish: The case of diacritic accents. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Estambul: European Language Resources Association (ELRA), pp. 737-742.

Asker, L., Boström, H., Papapetrou, P. y Persson, H. (2016). Identifying factors for the effectiveness of treatment of heart failure: a registry study. En *2016 IEEE 29th international symposium on computer-based medical systems (CBMS)*. Belfast y Dublín: Institute of Electrical and Electronics Engineers (IEEE), pp. 205–206.

Azmi, A. M, Almutery, M. N. y Aboalsamh, H. A. (2019). Real-word errors in arabic texts: a better algorithm for detection and correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27* (8), 1308-1320. https://doi.org/10.1109/TASLP.2019.2918404

Baba, Y. y Suzuki, H. (2012). How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. En H. Li, C. Lin, M. Osborne, G. G. Lee y J. C. Park (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: *Short Papers*. Jeju Island: Association for Computational Linguistics (ACL), pp. 373-377. https://www.aclweb.org/anthology/P12-2073

Balabaeva, K., Funkner, A. y Kovalchuk, S. (2020). Automated spelling correction for clinical text mining in Russian. *Studies in Health Technology and Informatics*, *270*, 43-47. https://doi.org/10.3233/SHTI200119

Bello, P. (2016). Aprendiendo a redactar mejor tus informes. En Asociación Española de Pediatría de Atención Primaria (ed.), *Curso de Actualización de Pediatría.* Madrid: Lúa Ediciones 3.0., pp. 391-400.

Beloki, Z., Saralegi, X., Ceberio, K., y Corral, A. (2020). Grammatical Error Correction for Basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, 65, 13-20.

Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Toronto: Graeme Hirst.

Blázquez, M. (2019). Using bigrams to detect written errors made by learners of Spanish as a foreign language. *CALL-EJ (Computer Assisted Language Learning Electronic Journal)*, 20 (2), 55-69.

Bravo-Candel, D., López-Hernández, J., García-Díaz, J. A., Molina-Molina, F. y García-Sánchez, F. (2021). Automatic correction of real-word errors in Spanish clinical texts. *Sensors*, 21 (9), 2893.

Brill, E. y Moore, R. C. (2000). An improved error model for noisy channel spelling correction. En *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, pp. 286-293.

Brockett, C., Dolan, W. B. y Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. En *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney: Association for Computational Linguistics, pp. 249-256. https://doi.org/10.3115/1220175.1220207

Bustamante-Rodríguez, M. D., Piedrahita-Ospina, A. A., y Ramírez-Velásquez, I. M. (2018). Modelo para detección automática de errores léxico-sintácticos en textos escritos en español. *TecnoLógicas*, 21 (42), 199-209.

Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Antártida.

Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield, UK: Equinox Publishing.

Carlini, R., Codina-Filba, J., y Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. En *Proceedings of the third workshop on NLP for computer-assisted language learning.* Uppsala: LiU Electronic Press, pp. 1-12.

Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua V., Tobin R., Whiteley, W., Wu, H. y Alex, B. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21 (1), 179. https://doi.org/10.1186/s12911-021-01533-7

Casillas, A., Pérez, A., Oronoz, M., Gojenola, K., y Santiso, S. (2016). Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications*, 61, 235-245.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., y Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34 (5), 301-310. https://doi.org/10.1006/jbin.2001.1029

Chen, M., Ge, T., Zhang, X., Wei, F., y Zhou, M. (2020). Improving the efficiency of grammatical error correction with erroneous span detection and correction. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP). Online: Association for Computational Linguistics, pp. 7162-7169. https://doi.org/10.18653/v1/2020.emnlp-main.581

Chipere, N., Malvern, D., y Richards, B. (2004). *Using a corpus of children's writing to test a solution to the sample size problem affecting Type-Token Ratios.* En G. Aston, S. Bernardini y D. Stewart (eds.): *Corpora and language learners*. Amsterdam: John Benjamins, pp. 139-147. https://doi.org/10.1075/scl.17.10chi

Chiu, B. y Baker, S. (2020). Word embeddings for biomedical natural language processing: A survey. *Language and Linguistics Compass*, 14 (12). https://doi.org/10.1111/lnc3.12402

Choe, Y. J., Ham, J., Park, K. y Yoon, Y. (2019). A neural grammatical error correction system built on better pre-training and sequential transfer learning. En *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florencia: Association for Computational Linguistics, pp. 213-227. https://doi.org/10.18653/v1/W19-4423

Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. New York: Praeger Publishers.

Ciosici M. R y Assent I. (2018). Abbreviation expander - a web-based system for easy reading of technical documents. En *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 1-4.

Clark, A., Fox, C. y Lappin, S. (2013). *The handbook of Computational Linguistics and Natural Language Processing*. New York: Wiley.

Corder, S. P. (1967). The Significance of Learners' Errors. *International Review of Applied Linguistics in Language Teaching*, 5, 161-170.

Corpas, G., y Seghiri, M. (2007). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del Lenguaje Natural*, 39, 165-172.

Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., y Xu, F. (2017). Annotation of entities and relations in Spanish radiology reports. En *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*. Varna: INCOMA Ltd., pp. 177-184.

Cruz, N. P., Morante, R., Maña, M. J., Mata, J. y Parra, C. L. (2017). Annotating negation in Spanish clinical texts. En *Proceedings of the workshop computational semantics beyond events and roles (SemBEaR)*. Valencia: Association for Computational Linguistics, pp. 53-58.

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Cham: Springer Nature.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of ACM*, 7 (3), 171-176. https://doi.org/10.1145/363958.363994

Davidson, S., Yamada, A., Fernández, P., Carando, A., Sánchez, C. H., y Sagae, K. (2020). Developing NLP tools with a new corpus of learner Spanish. En *Proceedings of the 12th Language Resources and Evaluation Conference*. Marsella: European Language Resources Association, pp. 7238-7243.

Deléger, L. y Grouin, C. (2012). Detecting negation of medical problems in French clinical notes. En *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. New York: Association for Computing Machinery, pp. 697-702.

Dernoncourt, F., Lee, J. Y., Uzuner, O. y Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24 (3), 596-606.

Díaz Villa, A M. (2005). Tipología de errores gramaticales para un corrector automático. *Procesamiento del Lenguaje Natural*, 35, 409-416. http://hdl.handle.net/10045/1341

Dreisbach C., Koleck T. A, Bourne P. E. y Bakken S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125, 37-46. https://doi.org/ 10.1016/j.ijmedinf.2019.02.008

Dziadek, J., Henriksson, A. y Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction, *Studies in health technology and informatics*, 235, 241-245. https://doi.org/10.3233/978-1-61499-753-5-241

Dzieciątko, M., Spinczyk, D. y Borowik, P. (2019). Correcting Polish bigrams and diacritical marks. En E. Pietka, P. Badura, J. Kawa y W. Wieclawek (eds.), *Information Technology in Biomedicine* (pp. 338-348). Springer: Cham. http://doi.org/10.1007/978-3-030-23762-2_30

D'hondt, E., Grouin, C. y Grau, B. (2016). Low-resource OCR error detection and correction in French clinical texts. En *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis.* Texas: Association for Computational Linguistics, pp. 61-68. https://aclanthology.org/W16-6108.pdf

Domènech-Bagaria, O., Estopà, R. y Vidal-Sabanés, L. (2020). La comprensió dels informes mèdics. *L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió*. Barcelona: Quaderns de la Fundació Dr. Antoni Esteve, 28-45.

Estopà, R. (2020). *L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió*. Barcelona: Fundación Dr. Antoni Esteve.

Estopà, R. y Ruiz, M. A. (2021). Metodología JUNTS de creación de Webapps para el abordaje de barreras en la comunicación médico-paciente: el caso de la aplicación COMJuntos en el ámbito de las enfermedades raras. *Teknokultura*, 18 (2), 157-165.

Faili, H., Ehsan, N., Montazery, M. y Pilehvar, M. T. (2016). Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. *Digital Scholarship in the Humanities*, 31 (1), 95-117. https://doi.org/10.1093/llc/fqu043

Felice, M., Yuan, Z., Andersen, Ø. E., Yannakoudakis, H. y Kochmar, E. (2014). Grammatical error correction using hybrid systems and type filtering. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Maryland: Association for Computational Linguistics, pp. 15-24.

Ferner, R. E. y Aronson J. K. (2016). Nominal ISOMERs (Incorrect Spellings Of Medicines Eluding Researchers)—variants in the spellings of drug names in PubMed: a database review, *BMJ*, 355, i4854. https://goo.gl/z9wTqg

Ferraro, G., Nazar, R., Alonso Ramos, M. y Wanner, L. (2014). Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation*, 48 (1), 45-64. https://doi.org/10.1007/s10579-013-9242-3

Fivez P., Suster, S. y Daelemans, W. (2017). Unsupervised context-sensitive spelling correction of clinical free-text with word and character N-Gram embeddings. En K. B. Cohen, D. Demner-Fushman, S. Ananiadou y J. Tsujii (eds.), *Proceedings of the BioNLP 2017 Workshop*. Vancouver: Association for Computational Linguistics, pp. 143-148. https://doi.org/10.18653/v1/W17-2317

Gamallo, P., García, M., del Río, I. y González, I. (2015). Avalingua: Natural language processing for automatic error detection. En M. Callies y S. Götz (eds.), *Learner Corpora in Language Testing and Assessment* (pp. 35-58). Ámsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.70.02gam

Gamon, M. (2010). Using mostly native data to correct errors in learners' writing. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Los Ángeles: Association for Computational Linguistics, pp. 163-171. https://aclanthology.org/N10-1019

Gimenes, P. A., Roman, N. T. y Carvalho, A. M. (2015). Spelling error patterns in Brazilian Portuguese. *Computational Linguistics*, 41 (1), 175-183. https://doi.org/10.1162/coli_a_00216

Gkoulalas-Divanis, A., Loukides, G. y Sun, J. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, 50, 4-19.

Golding, A. R., y Schabes, Y. (1996). Combining trigram-based and feature-based methods for context-sensitive spelling correction. En *34th Annual Meeting of the Association for Computational Linguistics*. Californa: Association for Computational Linguistics, pp. 71-78. https://doi.org/10.3115/981863.981873

Grundkiewicz, R., Junczys-Dowmunt, M. y Heafield, K. (2019). Neural grammatical error correction systems with unsupervised pre-training on synthetic data. En *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florencia: Association for Computational Linguistics, pp. pp. 252-263. https://doi.org/10.18653/v1/W19-4427

Gutiérrez, B. (2005). *El lenguaje de las ciencias.* Madrid: Gredos.

Gutiérrez, B. (2006). Medicina y diccionarios: ¿para cuándo una buena lexicografía de divulgación? *Panace@: Revista de Medicina, Lenguaje y Traducción*, 7 (24), 279-284.

Gutiérrez, B. y Navarro, F. A. (coords.) (2014). La importancia del lenguaje en el entorno biosanitario. Barcelona: Fundación Dr. Antoni Esteve.

Harremoës, P. y Topsøe, F. (2005). Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics*, 21, 315-318. https://doi.org/10.1109/ISIT.2002.1023479

Henriksson, A., Zhao, J., Boström, H. y Dalianis, H. (2015). Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. En *2015 IEEE international conference on bioinformatics and biomedicine (BIBM)*. Washington, pp. 343-350.

Hernández García, F. (2012). Palabras problemáticas y frases incorrectas: una solución autónoma para detectar lo indetectable. *Revista electrónica de lingüística aplicada*, 11 (1), 41-55.

Hernández Vaquero, D. (1992). *El artículo científico en biomedicina: normas para la publicación de trabajos*. Barcelona: Ciba-Geigy.

Hernández, H. y Bustabad, S. (2009). Características lingüísticas de los trabajos científicos de la medicina de urgencias. *Emergencias*, 21, 133-140.

Huang, C. C., y Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17 (1), 132-144. HTTPS://DOI.ORG/10.1093/bib/bbv024.

Hussain, F., y Qamar, U. (2016). Identification and correction of misspelled drugs' names in electronic medical records (EMR). En *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016)*. Roma: Science and Technology Publications, vol. 2, pp. 333-338. https://doi.org/10.5220/0005911503330338

Instituto Cervantes (2013). *Las 500 dudas más frecuentes del español*. Madrid: Espasa.

Intxaurrondo, A., Pérez-Pérez, M., Pérez-Rodríguez, G., López-Martín, J. A., Santamaría, J., de la Pena, S. y Krallinger, M. (2017). *The Biomedical Abbreviation Recognition*

and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts. En *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval)*. Murcia: Sociedad Española para el Procesamiento del Lenguaje Natural, pp. 230-246.

Iroju, O. y Olaleke, J. O. (2015). A systematic review of natural language processing. *Healthcare. International Journal of Information Technology and Computer Science*, 7, 44-50.

Jacobson, O. y Dalianis, H. (2016). Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. En *ACL Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016)*. Berlín: Association for Computational Linguistics, pp. 191-195.

Jensen, P. B., Jensen, L. J. y Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13 (6), 395-405.

Jensen, K., Soguero-Ruiz, C., Mikalsen, K. O., Lindsetmo, R.-O., Kouskoumvekaki, I., Girolami, M., Skrovseth, S. y Augestad, M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, 7, 46226.

Joopudi, V., Dandala, B. y Devarakonda, M. (2018). A convolutional route to abbreviation disambiguation in clinical text. *Journal of Biomedical Informatics*, 86, 71-78.

Junczys-Dowmunt, M. y Grundkiewicz, R. (2014). The AMU system in the CoNLL-2014 shared task: grammatical error correction by data-intensive and feature-rich statistical machine translation. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Baltimore: Association for Computational Linguistics, pp. 25-33.

Jurafsky, D. y Martin, J. H. (2014). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Education.

Kang, T., Zhang, S., Xu, N., Wen, D., Zhang, X. y Lei, J. (2017). Detecting negation and scope in Chinese clinical notes using character and word embedding. *Computer Methods and Programs in Biomedicine*, 140, 53-59.

Kernighan, M. D., Church, K. W. y Gale, W. A. (1990). A spelling correction program based on a noisy channel model. En *Proceedings of the 13th Conference on Computational Linguistics*, vol. 2. Helsinki: Association for Computational Linguistics, pp. 205-210.

Keselman, A. y Smith, C. A. (2012). A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, 45 (6), 1151-1163. https://doi.org/10.1016/j.jbi2012.07.012

Kilicoglu, H., Fiszman, M., Roberts, K. y Demner-Fushman, D. (2015). An ensemble method for spelling correction in consumer health questions. En American Medical Informatics Association (eds.): *AMIA Annual Symposium Proceedings*. San Francisco: AMIA, pp. 727-736. https://pubmed.ncbi.nlm.nih.gov/26958208/

Kim, M., Jin, J., Kwon, H. y Yoon, A. (2013) Statistical context-sensitive spelling correction using typing error rate. En *IEEE 16th International Conference on Computational Science and Engineering*. Sidney: IEEE Computer Society, pp. 1242-1246. https://doi.org/10.1109/CSE.2013.185

Kukich, K. (1992). Technique for automatically correcting words in text. *ACM Computing Surveys*, 24 (4), 377-439. https://doi.org/10.1145/146370.146380

Lai, K. H., Topaz, M., Goss, F. R. y Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55, 188-195. https://doi.org/10.1109/ICAIBD.2018.8396209

Lawley, J. (2015). New software to help EFL students self-correct their writing. *Language Learning y Technology*, 19 (1), 23-33.

Lehal, G. S. y Bhagat, M. (2007). Spelling error pattern analysis of Punjabi typed text. En *Proceedings of the 2007 International Symposium on Machine Translation, NLP and TSS*. Nueva Delhi: Tata McGraw-Hill, pp. 128-141. http://learnpunjabi.org/pdf/icon2004.pdf

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10 (8), 707-710.

Lima-López, S., Pérez, N., García-Sardiña, L. y Cuadros, M. (2020). Hitzalmed: Anonymisation of clinical text in Spanish. En *Proceedings of the 12th Language Resources and Evaluation Conference*. Marsella: European Language Resources Association, pp. 7038-7043.

Lima-López, S., Pérez, N. y Cuadros, M. (2021). Grammatical error correction for Spanish health records. *Procesamiento del Lenguaje Natural*, 66, 121-132.

Llopart-Saumell, E. y Da Cunha, I. (2020). L'informe mèdic. *L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió*. Barcelona: Quaderns de la Fundació Dr. Antoni Esteve, 14-27.

Liu, J., Liu, C. y Huang, Y. (2017). Multi-granularity sequence labeling model for acronym expansion identification. *Information Sciences*, 378, 462-474.

Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P. J., Huff, S. M. y Chute, C. G. (2015). Towards a semantic lexicon for clinical natural language

processing. En *AMIA Annual Symposium proceedings*. Chicago: AMIA, pp. 568-576. https://pubmed.ncbi.nlm.nih.gov/23304329/

López-Ferrero, C., Renau, I., Nazar, R. y Torner, S. (2014). Computer-assisted revision in Spanish academic texts: Peer-assessment. *Procedia-Social and Behavioral Sciences*, 141, 470-483. https://doi.org/10.1016/j.sbspro.2014.05.083

López-Hernández, J., Almela, Á. y Valencia-García, R. (2019). Automatic spelling detection and correction in the medical domain: A systematic literature review. En R. Valencia-García, G. Alcaraz-Mármol, J. Del Cioppo-Morstadt, N. Vera-Lucio y M. Bucaram-Leverone (eds.): *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science* (vol. 1124, pp. 104-117). Cham: Springer. https://doi.org/10.1007/978-3-030-34989-9_8

López-Hernández, J., Almela, Á. y Valencia-García, R. (2021). Linguistic errors in the biomedical domain: Towards an error typology for Spanish. *Sintagma: revista de lingüística*, 33, 83-100.

López-Hernández, J. y Almela, Á. (2021). Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas. *Panace@. Revista de medicina, lenguaje y traducción*, 22 (53), 96-108.

Lozano, C. y Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. *Automatic treatment and analysis of learner corpus data*, 59, 65-100.

Lu, C. J. y Demner-Fushman, D. (2018). Improving spelling correction with consumer health terminology. En *AMIA Annual Symposium Proceedings*. San Francisco: AMIA, p. 2053.

Lu, C. J., Aronson, A. R., Shooshan, S. E. y Demner-Fushman, D. (2019). Spell checker for consumer language (CSpell). *Journal of the American Medical Informatics Association*, 26 (3), 211-218. https://doi.org/ 10.1093/jamia/ocy171

Mamede, N., Baptista, J. y Dias, F. (2016). Automated anonymization of text documents. En *Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC)*. Vancouver: CEC, pp. 1287-1294.

Manning, C. D. y Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Massachusetts: MIT Press.

Marimon, M., Vivaldi, J. y Bel, N. (2017). Annotation of negation in the IULA Spanish Clinical Record Corpus. En E. Blanco, R. Morante y R. Saurí (eds.): *Computational Semantics Beyond Events and Roles (SemBEaR 2017)*. Stroudsburg: Association for Computational Linguistics, pp. 43-52.

Marimon, M., González-Agirre, A., Intxaurrondo, A., Rodríguez, H., López-Martín, J. A., Villegas, M. y Krallinger, M. (2019). Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*. Bilbao: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), pp. 618-638.

Martín, J. C. (2008) Sinonimia y polisemia en el léxico científico. El caso de las abreviaturas, las siglas y los epónimos. En I. Olza, M. Casado y R. González (eds.): *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)*. Pamplona: Servicio de Publicaciones de la Universidad de Navarra, pp. 509-517.

Martinet, A. (1984). *Elementos de lingüística general*. Madrid: Gredos.

Martínez de Sousa, J. (2008). *Manual de ortografía y ortotipografía del español actual*. Gijón: Ediciones Trea.

Martínez de Sousa, J. M. (2012). Algunas consideraciones sobre la ortografía académica. *Dendra médica. Revista de humanidades*, 11 (1), 9-25.

Mayor Serrano, M. B. (2010). Revisión y corrección de textos médicos destinados a los pacientes... y algo más. *Panace@. Revista de Medicina, Lenguaje y Traducción*, 11 (31), 29-36.

Medlock, B. (2006). An introduction to NLP-based textual anonymisation. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa: European Language Resources Association, pp. 1051-1056.

Merino Torre, R. (2015). *Editor de textos con corrector ortográfico para textos médicos*. Escuela universitaria de ingeniería técnica industrial de Bilbao. Recuperado de https://addi.ehu.es/handle/10810/15733 [16 de septiembre de 2020].

Meystre, S. y Haug, P. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39, 589-599. https://doi.org/10.1016/j.jbi.2005.11.004

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. y Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 35, 128-144.

Miangah, T. M. (2014). Farsispell: A spell-checking system for Persian using a large monolingual corpus. *Literary and Linguistic Computing*, 29 (1), 56-73.

Mitton, R. (1987). Spelling checkers, spelling correctors, and the misspellings of poor spellers. *Information Processing y Management*, 23 (5), 495-505.

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Moalla, Z., Soualmia, L. F., Prieur-Gaston, E., Lecroq, T. y Darmoni, S. J. (2011). Spell-checking queries by combining Levenshtein and Stoilos distances. En *Proceedings of*

Network Tools and Applications in Biology. Palermo: Network Tools and Applications in Biology, pp. 1-6.

Morante, R., Liekens, A. y Daelemans, W. (2008). Learning the scope of negation in biomedical texts. En *Proceedings of the 2008 conference on empirical methods in natural language processing*. Honolulu: Association for Computational Linguistics, pp. 715-724.

Moreira Silva, R. *Errores de prescripción: Ejemplos de errores de prescripción frecuentes y su posible prevención.* CedimCat (Centro de Información de Medicamentos de Cataluña. Recuperado de: https://www.cedimcat.info/index.php?option=com_content&view=article&id=192:errores-de-prescripcion-ejemplos-de-errores-de-prescripcion-frecuentes-y-su-posible-prevencion&lang=es [18 de noviembre de 2021].

Moreno, J. C. (2000). *Curso universitario de lingüística general*. Madrid: Síntesis.

Mykowiecka, A. y Marciniak, M. (2006). Domain-driven automatic spelling correction for mammography reports. En M. A. Klopotek, S. T. Wierzchon y K. Trojanowski (eds.): *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, vol. 35. Berlín: Springer, pp. 521-530. https://link.springer.com/chapter/10.1007/3-540-33521-8_56

Naber, D. (2003). *A rule-based style and grammar checker*. Munich: GRIN Verlag. http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf

Nagata, R., Takamura, H. y Neubig, G. (2017). Adaptive spelling error correction models for learner English. *Procedia Computer Science*, 112, 474-483. https://doi.org/10.1016/j.jbi.2005.11.004

Napoles, C. y Callison-Burch, C. (2017). Systematically adapting machine translation for grammatical error correction. En *Proceedings of the 12th Workshop on Innovative Use*

of NLP for Building Educational Applications. Copenhague: Association for Computational Linguistics, pp. 345-356. https://doi.org/10.18653/v1/W17-5039

Navarro, F. A. (2001). El inglés, idioma internacional de la medicina: causas y consecuencias de un fenómeno actual. *Panace@*, 2 (3), 35-51.

Navarro, F. A. (2015). *Medicina en español. Laboratorio del lenguaje: florilegio de recomendaciones, dudas, comentarios etimológicos, errores, anglicismos y curiosidades varias del lenguaje médico.* Madrid: Fundación Lilly.

Nazar, R. y Renau I. (2012). Google books n-gram corpus used as a grammar checker. En *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*. Aviñón: Association for Computational Linguistics, pp. 27-34.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G. y Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics*, 9 (1): 1-13. https://doi.org/10.1186/s13326-018-0179-8

Nizamuddin, U. y Dalianis, H. (2014). Detection of spelling errors in Swedish clinical text. En *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014)*. Upsala: SLTC, pp. 1-4.

Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D. y Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56, 318-332.

Paggio, P. (2000). Spelling and grammar correction for Danish in SCARRIE. En *Proceedings of the Sixth Conference on Applied Natural Language Processing.* Seattle: Association for Computational Linguistics, pp. 255-261. https://www.aclweb.org/anthology/A00-1035.pdf

Pande, H. (2017). Effective search space reduction for spell correction using character neural embeddings. En *Proceedings 15th Conference of the European Chapter of the Association for Computational Linguistics-EACL 2017*. Valencia: Association for Computational Linguistics, pp. 170-174.

Patrick, J., Sabbagh, M., Jain, S. y Zheng, H. (2010). Spelling correction in clinical notes with emphasis on first suggestion accuracy. En S. Ananiadou, K. Cohen y D. Demner-Fushman (eds.): *Second Workshop on Building and Evaluating Resources for Biomedical Text Mining*. Malta: Association for Natural Language Processing, pp. 2-8.

Pedler, J. (2007). *Computer correction of real-word spelling errors in dyslexic text*. Universidad de Londres. Birkbeck. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.4914&rep=rep1&type=pdf [21 de febrero de 20201]

Pedler, J. y Mitton, R. (2010). A large list of confusion sets for spellchecking assessed against a corpus of real-word errors. En *Language Resources Evaluation Conference*. Malta: European Language Resources Association, pp. 755-762. http://www.lrec-conf.org/proceedings/lrec2010/pdf/122_Paper.pdf

Perera, S., Sheth, A., Thirunarayan, K., Nair, S. y Shah, N. (2013). Challenges in understanding clinical notes: why NLP engines fall short and where background knowledge can help. En *Proceedings of the 2013 international workshop on Data management and analytics for healthcare (DARE)*. San Francisco: Association for Computing Machinery, pp. 21-26.

Pérez Castro, L. C. (1997). Vocabularios científico-técnicos y léxico común en el latín clásico. *Revista española de lingüística*, 27 (1), 107-114.

Pérez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M. y Dalianis, H. (2017). Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical

corpora. *Journal of Biomedical Informatics*, 71, 16-30. https://doi.org/10.1016/j.jbi.2017.05.009

Piñero, L., Aleixandre-Benavent, R. e Ibáñez, A. B. (2006). Uso y abuso de abreviaturas y siglas entre atención primaria, especializada y hospitalaria. *Papeles médicos* 15 (2), 29-37.

Plasencia, S. y Moliner, J. (2012). *Uso y abuso de las abreviaturas en los informes del Hospital Clínico Universitario Lozano Blesa*. Repositorio Institucional de Documentos, Universidad de Zaragoza.

Pollock, J. J. y Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *Journal of American Society of Informatics and Science*, 34 (1), 51-58. https://doi.org/10.1002/asi.4630340108

Pomares-Quimbaya, A., López-Úbeda, P., Oleynik, M. y Schulz, S. (2020). Leveraging PubMed to Create a Specialty-Based Sense Inventory for Spanish Acronym Resolution. *Studies in Health Technology and Informatics*, 270, 292-296. https://doi.org/ 10.3233/SHTI200169

Rambell, O. (1999). Error typology for automatic proof-reading purposes. En A. Sagvall Hein (ed.): *Reports from the SCARRIE project*. Upsala: Universidad de Upsala.

Ramírez, F. y Sánchez, F. (1996). GramCheck: a grammar and style checker. En *Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING '96)*. Copenhague: Association for Computational Linguistics, pp. 175-181. https://doi.org/10.3115/992628.992661

Ramírez, F. y López, E. (2006). Spelling error patterns in Spanish for word processing applications. En *Proceedings of Fifth international conference on Language Resources and Evaluation (LREC 2006)*. Genoa: European Language Resources Association, pp. 93-98. http://www.lrec-conf.org/proceedings/lrec2006/pdf/119_pdf.pdf

Real Academia Española y Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Madrid: Santillana. https://www.rae.es/dpd/

Real Academia Española y Asociación de Academias de la Lengua Española (2010). *Nueva gramática de la lengua española*. Madrid: Espasa.

Real Academia Española y Asociación de Academias de la Lengua Española (2010). *Ortografía de la lengua española*. Madrid: Espasa.

Real Academia Nacional de Medicina. (2012). *Diccionario de Términos Médicos*. Madrid: Panamericana.

Rello, L., Llisterri, J. y Baeza-Yates, R. (2014). DysList: An annotated resource of dyslexic errors. en *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC*). Reikiavik: European Language Resources Association, pp. 1289-1296. https://doi.org/10.13140/2.1.2542.7205

Rello, L., Baeza-Yates, R. y Llisterri, J. (2017). A resource of errors written in Spanish by people with dyslexia and its linguistic, phonetic and visual analysis. *Language Resources and Evaluation*, 51 (2), 379-408.

Ren, X. y Perrault, F. (1992). The typology of unknown words: an experimental study of two corpora. En *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)*. Nantes: ICCL, pp. 408-414.

Ringler, M. D., Goss, B. C. y Bartholmai, B. J. (2017). Syntactic and semantic errors in radiology reports associated with speech recognition software. *Health informatics journal*, 23 (1), 3-13.

Rodríguez-Rubio, S. (2018). Análisis cuantitativo de erratas del *Diccionario Terminológico de las Ciencias Farmacéuticas Inglés-Español/Spanish-English* (Ariel,

2007), *Panace@*, 19 (47), 76-88. https://www.tremedica.org/wp-content/uploads/n47-analisis.pdf

Rodríguez-Rubio, S. y Fernández Quesada, N. (2020). The dynamics of typographical error reproduction: optimising formal correctness in three specialised bilingual dictionaries. *Elia. Estudios de lingüística inglesa aplicada*, 20, 147-190. http://dx.doi.org/10.12795/elia.2020.i20.06

Rojo, G. y Palacios, I. (2016). Learner Spanish on computer: The CAES 'Corpus de aprendices de español' project. Spanish Learner Corpus Research. *Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 55-87.

Romero, V., Serrano, N., Toselli, A. H., Sánchez, J. A. y Vidal, E. (2011). Handwritten text recognition for historical documents. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*. Hissar: Association for Computational Linguistics, pp. 90-96.

Rozovskaya A. y Roth D. (2010). Generating confusion sets for context-sensitive error correction. En *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. Cambridge: Association for Computational Linguistics, pp. 961-970.

Rozovskaya, A. y Roth, D. (2019). Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7, 1-17. https://doi.org/10.1162/tacl a 00251

Ruch, P., Baud R. y Geissbühler. A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine*, 29 (2), 169-184. https://doi.org/10.1016/S0933-3657(03)00052-6

Samanta, P. y Chaudhuri, B. (2013). A simple real-word error detection and correction using local word bigram and trigram. En *Proceedings of the Twenty-Fifth conference*

on computational linguistics and speech processing (ROCLING 2013). Kaohsiung: The Association for Computational Linguistics and Chinese Language Processing, pp. 211-220.

San Mateo-Valdehíta, A. (2016). Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. *Revista Signos,* 49, 94-118. https://doi.org/10.4067/S0718-09342016000100005

Santamaría, J. y Krallinger, M. (2018). Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos. *Procesamiento del Lenguaje Natural*, 61, 49-56.

Santiso, S., Casillas, A., Pérez, A. y Oronoz, M. (2018). Word embeddings for negation detection in health records written in Spanish. *Soft Computing*, 23 (21), 10969-10975. https://doi.org/10.1007/s00500-018-3650-7

Sánchez, S., Pérez, F. D. P., Moreno, J., Gutiérrez, M. C., Martín, J., Rodríguez, G., Pérez, J. A. y Parra, C. L. (2018). Plataforma para la extracción automática y codificación de conceptos dentro del ámbito de la Oncohematología (Proyecto COCO). *Procesamiento del Lenguaje Natural*, 61, 65-71.

Sayle, R. A., Petrov, P., Winter-Holt, J. J. y Muresan, S. (2011). Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction. *Journal of Cheminformatics*, 3 (1), 1. https://doi.org/10.1186/1758-2946-3-S1-O16

Seco, M. (2004). *Diccionario de dudas y dificultades de la lengua española*. Barcelona: Espasa.

Senger, C., Kaltschmidt, J., Schmitt, S. P. W., Pruszydlo, M. G. y Haefeli, W. E. (2010). Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention. *International Journal of Medical Informatics*, 79 (12), 832-839. https://doi.org/ 10.1016/j.ijmedinf.2010.09.005

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*. 27, 379-423.

Sharma, S. y Gupta, S. (2015). A correction model for real-word errors. *Procedia Computer Science*, 70, 99-106. https://doi.org/10.1016/j.procs.2015.10.047

Shickel, B., Tighe, P. J., Bihorac, A. y Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22 (5), 1589-1604. https://doi.org/10.1109/jbhi.2017.2767063

Siklósi, B., Novák, A. y Prószéky, G. (2016). Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech and Language*, 35, 219-233. https://doi.org/10.1016/j.csl.2014.09.001

Skeppstedt, M. (2011). Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2 (3), 1-12.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S. y Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering*, 1-9. https://doi.org/10.1155/2018/4302425

Schwartz, A. S. y Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. En *Proceedings of Pacific Symposium on Biocomputing*. Lihue: Pacific Symposium on Biocomputing, pp. 451-462.

Terroba Reinares, A. R. (2016). *Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico*. [Tesis Doctoral. Universidad de La Rioja]. Dialnet. https://dialnet.unirioja.es/servlet/tesis?codigo=46993

Tetreault, J., Foster, J. y Chodorow, M. (2010). Using parse features for preposition selection and error detection. En *Proceedings of the ACL 2010 - 48th Annual Meeting*

of the Association for Computational Linguistics. Upsala: Association for Computational Linguistics, pp. 353-358.

Thompson, P. M., McNaught, J. y Ananiadou. S. (2015). Customised OCR correction for historical medical text. *Digital Heritage,* 1, 35-42.

Van Peer, W., Hakemulder, J. y Zyngier, S. (2012). *Scientific methods for the humanities*. Amsterdam: John Benjamins Publishing Company.

Veronis, J. (1988). Computerized correction of phonographic errors. *Computers and the Humanities*, 22 (1), 43-56.

Villegas, M., de la Peña, S., Intxaurrondo, A., Santamaría, J. y Krallinger, M. (2017). Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, (59), 141-144.

Villegas M., Intxaurrondo A., González-Agirre A., Marimon M. y Krallinger, M. (2018). The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. En *Proceedings of the LREC 2018 Workshop. MultilingualBIO: Multilingual Biomedical Text Processing*. Miyazaki: European Language Resources Association, pp. 32-39.

Vivaldi, J. (2020). Els errors ortotipogràfics, ortogràfics i de puntuació. *L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió*. Barcelona: Quaderns de la Fundació Dr. Antoni Esteve, 104-111.

Vovk O., Piho G. y Ross P. (2021). Anonymization methods of structured health care data: a literature review. En C. Attiogbé y S. Ben Yahia (eds.): *Model and Data Engineering. MEDI 2021. Lecture Notes in Computer Science*, vol. 12732. Tallin: Springer. https://doi.org/10.1007/978-3-030-78428-7_14

Wen, Z., Lu, X. H. y Reddy, S. (2020). MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. En *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, pp. 130-135.

White, M. y Rozovskaya, A. (2020). A comparative study of synthetic data generation methods for grammatical error correction. En *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle y online: Association for Computational Linguistics, pp. 198-208. https://doi.org/10.18653/v1/2020.bea-1.21

Wilcox-O'Hearn, A., Hirst, G. y Budanitsky, A. (2008). Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. En *Proceedings of 9th International conference on intelligent text processing and computational linguistics (CICLing-2008)*. Haifa: Springer, pp. 605-616.

Wong, W. y Glance, D. (2011). Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. *Artificial Intelligence in Medicine*, 53 (3), 171-180. https://doi.org/10.1016/j.artmed.2011.08.003

Workman, T. E., Shao, Y., Divita, G. y Zeng-Treitler, Q. (2019). An efficient prototype method to identify and correct misspellings in clinical text. *BMC research notes*, 12 (1), 1-5. https://doi.org/10.1186/s13104-019-4073-y

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni S., Wang Q., Wei Q., Xiang Y., Zhao B. y Xu H. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27 (3), 457-470. https://doi.org/ 10.1093/jamia/ocz20

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A. y Xu, H. (2012). A comparative study of current clinical natural language processing systems on handling

abbreviations in discharge summaries. En *AMIA Annual Symposium Proceedings 2012*. Chicago: American Medical Informatics Association, p. 997.

Xu, H., Stetson, P. D. y Friedman, C. (2007). A study of abbreviations in clinical notes. En *AMIA Annual Symposium Proceedings 2007*. Chicago: American Medical Informatics Association, p. 821.

Yannakoudakis, E. J. y Fawthrop, D. (1983). The rules of spelling errors, *Information processing and management*, 19 (12), 101-108. https://doi.org/10.1016/0306-4573(83)90045-6

Yazdani, A., Ghazisaeedi, M., Ahmadinejad, N., Giti, M., Amjadi, H. y Nahvijou, A. (2019). Automated misspelling detection and correction in Persian clinical text. *Journal of digital imaging*, 33, 555-562. https://doi.org/10.1007/s10278-019-00296-y

Yuan, Z. y Briscoe, T. (2016). Grammatical error correction using neural machine translation. En *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego: Association from Computational Linguistics, pp. 380-386. https://doi.org/10.18653/v1/N16-1042

Yuan, Z., Briscoe, T. y Felice, M. (2016). Candidate re-ranking for SMT-based grammatical error correction. En *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego: Association for Computational Linguistics, pp. 256-266.

Zhou, X., Zheng, A., Yin, J., Chen, R., Zhao, X., Xu, W. Chen, W., Xia T. y Lin, S. (2015). Context-sensitive spelling correction of consumer-generated content on health care. *JMIR medical informatics*, 3 (3), e27. https://doi.org/ 10.2196/medinform.4211

Zech, J., Forde J., Titano J., Kaji D., Costa A. y Oermann E. K. (2019). Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-

to-sequence models, *Annals of translational medicine, 7* (11), 233-242. https://doi.org/10.21037/atm.2018.08.11