

DetECCIÓN AUTOMÁTICA DE ERRORES LINGÜÍSTICOS EN TEXTOS CLÍNICOS: ANÁLISIS DE PATRONES DE ERROR EN VARIAS ESPECIALIDADES MÉDICAS

Jésica López Hernández* y Ángela Almela**

Resumen: El objetivo de este trabajo es aportar el primer análisis cuantitativo de tipos de errores contenidos en un corpus formado por informes clínicos en español. Se han analizado informes clínicos pertenecientes a las especialidades de urgencias, UCI, psiquiatría y cirugía general. Los errores fueron estudiados teniendo en cuenta criterios como distancia de edición, tipo de error o existencia de multierror en la palabra. Para tal cometido, se desarrolló una herramienta de identificación y clasificación de errores, se utilizaron técnicas estadísticas y se compararon los resultados con trabajos previos sobre patrones de errores. Los resultados indican que el tipo de error más frecuente es el de omisión de tilde y la mayoría de los errores ocurren a distancia de edición 1, entre parejas de caracteres con similitudes fonéticas y parejas de caracteres adyacentes en el teclado.

Palabras clave: análisis de errores, corrección ortográfica automática, informes clínicos, procesamiento del lenguaje natural, técnicas basadas en conocimiento lingüístico.

Automatic detection of linguistic errors in clinical texts: analysing error patterns in several medical specialties

Abstract: The aim of this study is to conduct the first quantitative analysis of errors in a corpus of clinical reports in Spanish. The clinical reports analysed belong to four medical specialties: emergency medicine, ICU, psychiatry, and general surgery. Errors will be studied according to criteria such as edit distance, error type, and presence of multiple-error words. To this end, we developed a tool for identifying and classifying errors and used statistical techniques, comparing the results with previous studies on error patterns. The results show that the most frequent error type is the omission of accent marks and that most errors occur at edit distance 1 between pairs of characters with phonetic similarities and pairs of characters that are adjacent on the keyboard.

Key words: automatic spellchecking, clinical reports, error analysis, linguistic-knowledge based techniques, natural language processing.

Panace@ 2021; XXII (53): 96-108

Recibido: 11.1.2021. Aceptado: 08.11.2021.

1. Introducción

El procesamiento del lenguaje natural, entre otras tareas, busca 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 corpus. Los sistemas de detección y corrección automática de errores constituyen un elemento esencial en las tecnologías del procesamiento del lenguaje natural (Jurafsky y Martin, 2014). El rendimiento de los sistemas utilizados habitualmente se ve condicionado cuando se trabaja con corpus pertenecientes a lenguajes de especialidad, como ocurre en el campo de la medicina, debido a las particularidades intrínsecas que estos textos poseen.

En el ámbito biosanitario es de especial importancia poder hacer uso de las tecnologías basadas en procesamiento automático de datos para facilitar la extracción y clasificación de información clínica, la interoperabilidad semántica, la toma de decisiones o la predicción de sucesos; tareas fundamentales para mejorar la investigación en la práctica clínica. Entre las principales dificultades para el tratamiento automatizado del lenguaje médico destacan el empleo de terminología específica, los neologismos, los extranjerismos y la utilización de abreviaturas y siglas no estandarizadas (Aleixandre *et al.*, 2015). Además, en el caso de los informes clínicos, se suma la presencia de abundantes errores de escritura (Meystre y Haug, 2006). Concretamente, son múltiples los estudios (Wong y Glance, 2011; Ruch *et al.*, 2003; Lai *et al.*, 2015; Zech *et al.*, 2019; entre otros) que señalan el importante número de errores ortográficos que presentan los informes clínicos debido al limitado tiempo que los profesionales de la salud disponen para la redacción de estos y a la ausencia de revisión posterior. Estas investigaciones se han realizado mayoritariamente para el inglés, aunque también

* Universidad de Murcia (España). Dirección para correspondencia: jesica.lopez@um.es.

** Universidad de Murcia (España). Dirección para correspondencia: angelalm@um.es.

encontramos estudios sobre errores en el dominio médico para el sueco (Dziadek *et al.*, 2017) o el húngaro (Siklósi *et al.*, 2016), entre otras lenguas. Sin embargo, no existen datos cuantitativos previos sobre patrones de error en documentación clínica en español. Por tanto, resulta de gran interés investigar sobre la naturaleza de los errores que ocurren en los informes clínicos para contribuir a la mejora en el proceso de detección y para la generación de listas de sugerencias en la fase de corrección automática. El análisis de patrones de errores lingüísticos nos va a permitir identificar rasgos y descubrir tendencias probabilísticas características en este dominio, como qué tipo o subtipo de error es más común o qué pares de caracteres suelen ser confundidos entre sí con más frecuencia. Asimismo, la identificación de rasgos va a ayudar al desarrollo de un módulo basado en conocimiento lingüístico que contribuirá a la mejora del rendimiento de los sistemas de detección y corrección de errores.

Tras la introducción del estudio, la segunda sección presenta el estado del arte sobre corrección automática en textos clínicos y análisis de errores, las líneas de investigación que fundamentan esta investigación. Las secciones tercera y cuarta están dedicadas a los objetivos y al marco metodológico, que incluye la descripción detallada del corpus usado y los procedimientos de extracción y análisis llevados a cabo. Los apartados quinto y sexto están dedicados a la presentación y posterior discusión de los resultados obtenidos. Se finaliza con las conclusiones más relevantes y futuras investigaciones.

2. Fundamentación teórica

2.1. Corrección automática y análisis de errores

El proceso de detección y corrección automática se divide en tres fases fundamentales: la detección del error, la generación de posibles candidatos para la corrección, así como la clasificación de esos candidatos en un *ranking*, de forma que el candidato apropiado para el error aparezca al principio del listado de sugerencias (Kukich, 1992). A los candidatos generados se les asigna una puntuación, que se calcula de acuerdo con una serie de criterios. Para la generación de sugerencias, se utiliza la distancia de edición, conocida como la distancia de Damerau-Levenshtein, que alude al número mínimo de ediciones necesarias para transformar una cadena de caracteres en otra. Son cuatro las operaciones básicas de edición: inserción, omisión, sustitución y transposición (Damerau, 1964; Levenshtein, 1966). El error de inserción ocurre cuando se añade un carácter extra a la palabra; el error de omisión, cuando a la palabra le falta un carácter; el error de sustitución, cuando la palabra contiene un carácter erróneo en lugar del correspondiente, y el error de transposición, cuando se intercambia la posición de dos caracteres adyacentes. Damerau (1964) estableció que el 80% de las palabras escritas incorrectamente contenían un único error, es decir, necesitaban una sola operación de edición.

Son múltiples las técnicas desarrolladas para detectar y corregir los errores, entre ellas destacan la búsqueda en diccionario, la distancia de edición mínima, el uso de *n*-gramas, la generación de modelos lingüísticos, el uso de algoritmos de

aprendizaje automático (*machine learning*) y, más recientemente, el entrenamiento de redes neuronales y arquitecturas basadas en aprendizaje profundo (*deep learning*). En las últimas investigaciones, se aboga por sistemas híbridos que combinen varias de estas técnicas junto con las más novedosas basadas en aprendizaje profundo, al ser las que están arrojando mejores resultados.

El análisis de errores y el diseño de tipologías de error también han sido utilizados para aportar información lingüística y estadística que ayude al diseño de técnicas basadas en reglas y conocimiento lingüístico. En el caso del español, pueden mencionarse dos trabajos sobre análisis de errores desarrollados en el marco de las tareas de corrección. Ramírez y López (2006) realizan una tipificación de errores para contribuir al desarrollo de un corrector para el español en Microsoft Corporation, y ofrecen la primera cuantificación de errores en un corpus en español. Díaz (2005) presenta una propuesta de tipología de errores gramaticales y de motivación cognitiva. Díaz (2005: 409) pone de manifiesto que «en diferentes lenguas y en diversos estudios pueden variar considerablemente los tipos de error identificados, así como los ejes en torno a los cuales se defina la clasificación propuesta», lo que dificulta la creación de una tipología de errores universal y conlleva la necesidad de estudiar los errores en contexto. Es también destacable el estudio de Gimenes *et al.* (2015), en el que se analiza un corpus en portugués de Brasil y se busca averiguar si las estadísticas establecidas por Damerau (1964) sobre la distribución de errores para el inglés son válidas en otras lenguas. Además, los autores comparan los resultados obtenidos con los de Ramírez y López (2006).

El análisis de errores ha sido especialmente relevante para el estudio y desarrollo de herramientas en el campo de aprendizaje de idiomas (Nagata *et al.*, 2017). También se ha trabajado con análisis de errores para el desarrollo de correctores automáticos que puedan ayudar a personas con dificultades o trastornos específicos de aprendizaje. Un ejemplo es DysList (Rello *et al.*, 2014), un recurso lingüístico formado por una lista de errores extraídos de textos escritos por personas con dislexia.

Los trabajos sobre patrones de error en inglés han sido numerosos (Yannakoudakis y Fawthrop, 1983; Pollock y Zamora, 1983; Naber, 2003; Pedler y Mitton, 2010; entre otros). No obstante, en los últimos años, también han sido publicados estudios sobre patrones de error en húngaro (Siklósi *et al.*, 2016), en japonés (Baba y Suzuki, 2012), en danés (Paggio, 2000) o en punjabi (Lehal y Bhagat, 2007).

Los errores pueden ser clasificados atendiendo a diversas dimensiones de análisis según el propósito de este: tipo de error, causa del error, frecuencia de aparición, contexto en el que aparece, posición del error o el tipo de corrección que se le aplica. Una de las divisiones fundamentales distingue entre los llamados *errores non-word* y *errores real-word*. Los errores *non-word* son aquellos que dan como resultado una palabra o conjunto de caracteres no existente en esa lengua —como *hieprtensio* en lugar de *hipertenso*—, mientras que los *real-word* son aquellos que producen una palabra existente, pero no la palabra adecuada en ese contexto —*costa* en lugar de *costal*—. Estos últimos son errores semánticos y sintácticos, por lo que necesitan de

técnicas basadas en contexto que permitan detectarlos. Asimismo, en los estudios relacionados con aprendizaje de lenguas resulta de gran interés distinguir entre errores de competencia y errores de actuación (Chomsky, 1986). Los errores de competencia tienen una motivación cognitiva e implican el desconocimiento de la norma ortográfica y gramatical que regula una lengua (Díaz, 2005), mientras que los de actuación ocurren accidentalmente y están motivados por factores no lingüísticos, como distracciones o problemas mecánicos.

2.2. Detección y corrección automática de errores lingüísticos en informes médicos

El tratamiento automático de textos médicos es un tema emergente y de suma actualidad en el procesamiento del lenguaje natural. Son diversos los proyectos de investigación centrados en desarrollar recursos para procesar datos en el dominio biosanitario mediante tareas como la recuperación y extracción de información, la desambiguación o el reconocimiento de entidades nombradas. Sin embargo, los errores afectan a la efectividad de los sistemas automatizados, por lo que son necesarias la detección y corrección previas. En la mayoría de los casos, el proceso de corrección forma parte de investigaciones mayores que engloban otras tareas, como la extracción de información o el reconocimiento de voz (Sayle *et al.*, 2012). Son diversos los trabajos que señalan el alto porcentaje de errores que contienen los corpus formados por informes médicos, con tasas de error que van desde el 0,4% (Liu *et al.*, 2012), el 4% y el 7% (Lai *et al.*, 2015), o hasta el 10% (Ruch *et al.*, 2003). Además, las investigaciones anteriores coinciden en destacar la dificultad en el tratamiento de los informes clínicos debido a la importante presencia de abreviaturas, la falta de estandarización de formas, la complejidad léxica y la ausencia de revisión posterior (López-Hernández *et al.*, 2019). Estos trabajos aportan datos sobre la presencia de errores en informes médicos en lenguas como el inglés, el francés, el sueco o el húngaro. Además de informes médicos pertenecientes a especialidades como urgencias o cirugía, también encontramos corpus formados por notas de progreso (Wong y Glance, 2011), o las consultas realizadas por pacientes y consumidores en sistemas de búsqueda (Kilicoglu *et al.*, 2015).

Según Siklósi *et al.* (2016) los tipos de errores más frecuentes tienen que ver con la inserción, omisión y transposición de caracteres. Otros que destaca son el uso inadecuado de los signos de puntuación, los errores gramaticales, las abreviaturas y la terminología con errores ortográficos. La mayoría de los trabajos se centran en errores *non-word*, que suelen ocurrir con un porcentaje muy superior a los *real-word*; los estudios sobre errores de tipo gramatical o semántico son limitados.

Como hemos mencionado anteriormente, en la actualidad no existen investigaciones que aborden el análisis de errores en informes médicos en español, debido, en gran parte, a las dificultades para recopilar un corpus anonimizado con estas características, aunque sí podemos mencionar trabajos que analizan errores aparecidos en publicaciones biomédicas en español (Aguilar Ruiz, 2013) o en diccionarios terminológicos (Rodríguez-Rubio Mediavilla, 2018).

3. Objetivo del estudio

Este trabajo tiene como objetivo principal investigar sobre los tipos de errores presentes en un corpus de informes clínicos en español de varias especialidades médicas. Este objetivo principal puede ser desglosado en una serie de objetivos específicos entre los que se encuentran: aportar el primer análisis cuantitativo de tipos de errores en un corpus del dominio médico, descubrir cuáles son los tipos de errores más frecuentes y qué combinaciones de caracteres suelen generar mayor confusión, averiguar si hay diferencias significativas entre los errores de las especialidades médicas analizadas y, por último, comprobar si hay diferencias entre los errores en el dominio médico y la tipificación existente sobre errores del español general.

4. Metodología

El diseño del presente estudio fue motivado por Ramírez y López (2006), trabajo en el que se realizó una primera tipificación de errores en un corpus en español para la mejora de un sistema de corrección. Ha sido tomado como referencia por ser el único trabajo previo que contiene datos cuantitativos sobre tipos de errores para el español común o general. Además de fundamentar nuestra investigación con trabajos esenciales de detección y corrección automática, como Damerau (1964) y Kukich (1992), hemos partido de investigaciones recientes sobre análisis de errores para sistemas de procesamiento automatizado del lenguaje realizadas en otros dominios e idiomas, como Baba y Suzuki (2012), Gimenes *et al.* (2015) o Nagata *et al.* (2017). Siguiendo el planteamiento establecido en estos trabajos, este estudio se ha centrado en los errores *non-word* para que la comparación de resultados fuese coherente.

Registramos como palabra errónea cualquier caso que dificulta el correcto procesamiento automático de un corpus y no se rige por las normas vigentes definidas en fuentes de autoridad, como las recogidas en la *Ortografía de la lengua española* (2010), la *Nueva gramática de la lengua española* (2010) y el *Diccionario panhispánico de dudas* (2005), así como otras obras especializadas que contribuyen a la estandarización del lenguaje médico, como el *Diccionario de términos médicos* (2011), desarrollado por la Real Academia Nacional de Medicina, o revistas y recursos profesionales para la traducción y la redacción médicas (Navarro, 2015). Como veremos con más detalle en la sección de «Procedimiento», la identificación y cuantificación de los errores se llevó a cabo mediante una serie de tareas de detección y corrección. La clasificación de los errores se realizó mediante el desarrollo de una herramienta de *software* propia y se tuvieron en cuenta varios criterios de análisis:

— Tipo de error según la operación de edición que lo ha provocado: omisión, sustitución, inserción o transposición.

— Subtipo de error: es un nivel de concreción mayor en el que se indica a qué tipo de carácter ha afectado el error de omisión, sustitución, inserción o transposición. Los niveles definidos son letra, tilde y espacio.

— Existencia de multierror en la palabra: se contabiliza el número de errores que contiene la palabra y el número de operaciones de edición necesarias para convertirla en la palabra de destino.

4.1. Descripción del corpus

El corpus utilizado en la investigación está compuesto por una colección de informes médicos electrónicos pertenecientes a cuatro especialidades médicas distintas: urgencias, unidad de cuidados intensivos (UCI), psiquiatría y cirugía general. El informe clínico es un documento emitido por el médico responsable con la información sobre la atención prestada al paciente. A grandes rasgos, en él se suele encontrar una anamnesis, un diagnóstico y unas pautas de actuación; no obstante, cada especialidad presenta concreciones específicas en su estructura.

El corpus total, formado por los cuatro conjuntos, contiene un total de 2 321 826 palabras. En la tabla 1 puede verse el desglose de información por cada especialidad:

TABLA 1. Datos estadísticos del corpus

Subcorpus	Palabras/ 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 gral.	440 893	14 697	42,04

Las muestras están equilibradas y son representativas de cada especialidad a nivel cualitativo debido a los criterios de compilación previamente establecidos y a las características especí-

ficas del corpus. Asimismo, tras llevar a cabo un análisis con el programa ReCor¹ (gráfico 1), creado por Gloria Corpas Pastor, Miriam Seghiri Domínguez y Romano Maggi, podríamos considerar que el corpus es representativo cuantitativamente para el propósito de esta investigación. Observamos que la tendencia es descendente y que la inclusión de nuevos ficheros y palabras no incorpora novedades significativas a partir de 1,5 millones de tokens. El número de documentos es limitado porque cada uno de ellos contiene un gran número de informes clínicos.

En la tabla 1 se muestra el número total de palabras o formas léxicas (tokens) y el número total de palabras distintas o tipos (types) que contiene el corpus. Asimismo, hemos obtenido por medio de WordSmith Tools² la proporción estandarizada correspondiente para poder comparar corpus con distinta longitud debido a la diferencia de tamaño entre los subcorpus de urgencias y UCI con respecto a psiquiatría y cirugía general. La relación type/token estandarizada (STTR) calcula cada *n* palabras del corpus (por defecto, *n* = 1000) y computa la media de esos fragmentos para así obtener una proporción promedio de type/token basada en fragmentos consecutivos de 1000 palabras (Chipere *et al.*, 2004). La especialidad que posee una mayor riqueza léxica es psiquiatría debido a que los informes tienen una estructura más narrativa; por el contrario, especialidades como urgencias o cirugía general presentan una estructura mucho más sinóptica, lo que ha podido influir en los resultados.

Se trata de un corpus monolingüe en español, formado por ficheros de texto plano no estructurado y no etiquetado. Los informes electrónicos fueron sometidos a un preprocesamiento para uniformar su formato y eliminar etiquetas HTML y XML en aquellos que las contenían. El corpus pertenece a la empresa Vócali³, que lo utiliza para generar modelos de lenguaje que se emplean para el desarrollo de sistemas de reconocimiento de voz en el ámbito médico. Los informes fueron escritos a ordenador directamente por los médicos, no transcritos. En diver-

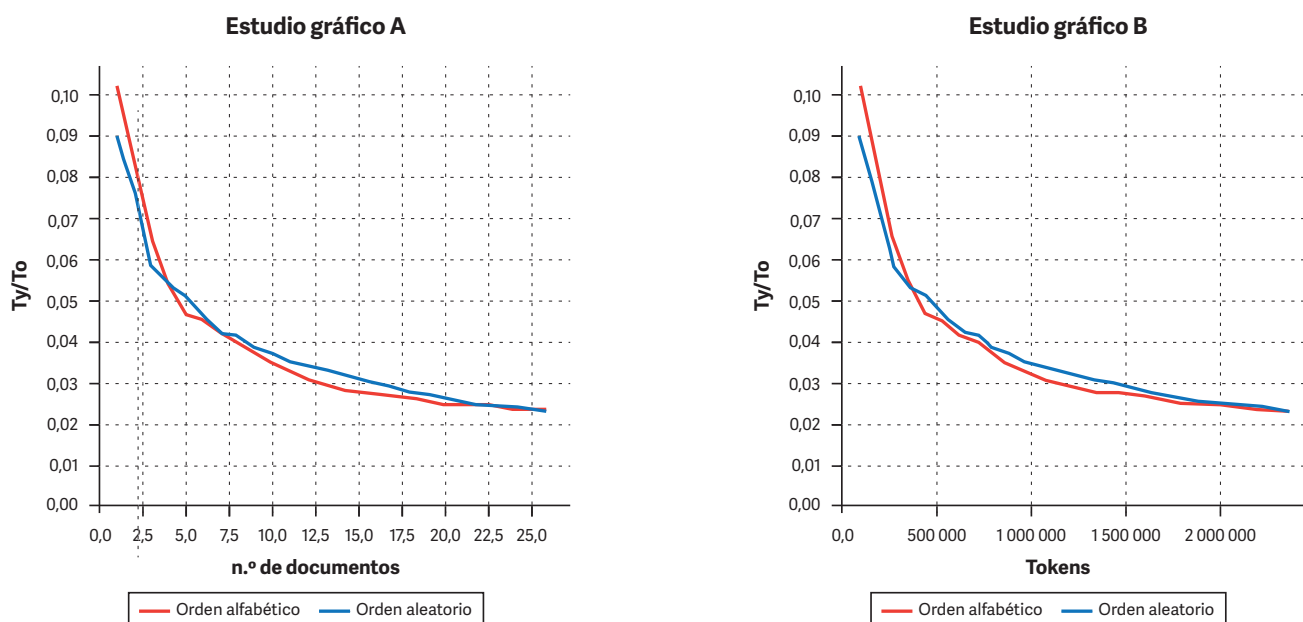


GRÁFICO 1. Análisis cuantitativo de la representatividad del corpus mediante ReCor.

sas investigaciones previas sobre sistemas de detección y corrección en el dominio médico, los errores lingüísticos fueron generados de forma artificial (Dziadek *et al.*, 2017), debido a la dificultad para acceder a este tipo de textos especializados. Sin embargo, una de las fortalezas de este corpus es que está formado por informes clínicos que contienen errores reales. Por último, es necesario destacar que, en cumplimiento del *Reglamento General de Protección de Datos* (RGPD)⁴, los informes están totalmente anonimizados y no contienen datos sobre el centro, fechas, lugares, identidad de pacientes o médicos.

4.2. Procedimiento

La identificación y análisis de errores del corpus conllevó el desarrollo de tres fases, integradas a su vez por una serie de tareas. El proceso comenzó con la fase de detección de errores. En primer lugar, se utilizó la técnica de búsqueda en el diccionario, mediante la cual se comparó automáticamente el corpus con un lexicón constituido por palabras previamente validadas. Todas aquellas palabras que coincidían en el corpus y en la lista fueron consideradas correctas y descartadas del proceso; por el contrario, todas aquellas palabras que no aparecían en el listado fueron consideradas candidatas a error. Entre las candidatas a error se encontraron falsos positivos, es decir, palabras que eran correctas pero no estaban recogidas en el diccionario —porque eran neologismos o palabras que no habían aparecido anteriormente—, y que hubo que excluir de la lista de errores. El lexicón fue compilado a partir del diccionario para el español de HunsPELL⁵, un corrector ortográfico multiplataforma de código abierto que contiene formas flexionadas y derivadas de las palabras. Al tratarse de un corpus del dominio médico, un diccionario del español general era insuficiente, por lo que hubo que añadir terminología biomédica en español, que fue recopilada a partir de diversas fuentes. Entre ellas podemos destacar las nomenclaturas sistematizadas de Snomed-CT⁶ y CIE-10⁷, glosarios especializados y recursos léxicos obtenidos de fuentes como la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)⁸ y la Sociedad Española de Documentación Médica (SEDOM)⁹. Gracias a estos recursos se añadieron listados de medicamentos, siglas, principios activos, enfermedades, síntomas, procesos, estructuras anatómicas y protocolos. Además, se incorporó un módulo específico con expresiones regulares para detectar errores en el uso de símbolos y palabras formadas con guiones.

El segundo paso consistió en la fase de corrección. En esta fase se asignó la corrección correspondiente a cada palabra errónea. Para ello, se generaron automáticamente sugerencias de corrección a través de la técnica de distancia de edición mínima, es decir, a cada palabra errónea se le aplicaron las cuatro operaciones básicas de edición mencionadas anteriormente con el objetivo de hallar una palabra correcta que estuviese recogida en nuestro lexicón. Para la ordenación de la lista de sugerencias, también se tuvo en cuenta la frecuencia de aparición de cada sugerencia en el corpus y el análisis de bigramas. Finalmente, se validaron las correcciones obtenidas mediante la revisión manual asistida y la comprobación del contexto del error en aquellos casos que lo requerían.

En el siguiente paso, tuvo lugar la fase de clasificación de

errores. Fue necesario el desarrollo de una herramienta propia que obtuviese esa información. El funcionamiento de la herramienta se basa en la comparación entre la palabra errónea y la corrección asignada para obtener información sobre el tipo de operación de edición que ha convertido una cadena de caracteres en otra, el subtipo y la distancia de edición. Mediante esta herramienta de clasificación de errores también fue posible contabilizar las palabras erróneas y, a su vez, cuántas veces se repetía cada palabra errónea. De esta forma, obtuvimos la frecuencia de aparición de errores en cada una de las especialidades. Además, se implementó la creación de una matriz de confusión a partir de los resultados obtenidos en cada especialidad. Finalmente, se llevó a cabo la prueba de ji al cuadrado a través del programa IBM SPSS Statistics¹⁰, un *software* que permitió el análisis estadístico de los datos obtenidos y de las mediciones de las distintas variables descritas anteriormente.

5. Resultados

A continuación, se muestran los resultados obtenidos tras la clasificación de errores y posterior análisis. En la tabla 2, donde aparece el número total de palabras y de errores en cada corpus, se observa que la especialidad con un mayor porcentaje de errores es urgencias. Hay factores que pueden explicar el mayor número de errores en esta especialidad, como la sobrecarga de trabajo, disponer de un tiempo muy limitado para redactar los informes o la necesidad de dar respuesta inmediata al paciente. También puede influir que se trata de una especialidad que cubre un espectro del lenguaje médico más amplio, con una mayor variabilidad de etiologías y enfermedades, lo que puede plantear más dificultades terminológicas.

TABLA 2. Errores en el corpus según la especialidad médica

Especialidad	Palabras	Errores	Porcentaje de error
Urgencias	730 468	35 819	4,90 %
UCI	725 690	19 890	2,74
Psiquiatría	424 775	7 271	1,71
Cirugía gral.	440 893	11 175	2,53

Los resultados mostrados en la tabla 3, que incluye la frecuencia relativa y absoluta, confirman que la mayor parte de las palabras erróneas encontradas en el corpus contienen un error único, es decir, la distancia de edición entre la palabra errónea y la palabra meta implica una sola operación. En todas las especialidades, el porcentaje de errores con distancia 1 es superior al 95 %, por tanto, el porcentaje de errores múltiples en las palabras es muy reducido, inferior al 4 % en el cómputo total de errores e inferior al 0,2 % en el total de palabras del corpus.

En la tabla 4 aparecen los resultados obtenidos en cada especialidad según el tipo de error en distancia 1. Observamos una

TABLA 3. Errores según la distancia de edición

	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	6 981	96,01	10 823	96,85
Multierror	1 275	3,56	503	2,53	290	3,99	352	3,15

TABLA 4. Errores según el tipo de operación de edición

Tipo de error	Urgencias		UCI		Psiquiatría		Cirugía General	
	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	5 762	82,54	9 816	90,70
Inserción	2 471	7,15	733	3,78	553	7,92	620	5,73
Sustitución	1 741	5,04	336	1,73	543	7,78	235	2,17
Transposición	1 049	3,04	361	1,86	123	1,76	152	1,40

TABLA 5. Errores según el subtipo de operación de edición

Subtipo de error	Urgencias		UCI		Psiquiatría		Cirugía General	
	Frec. abs.	Frec. rel.	Frec. abs.	Frec. rel.	Frec. abs.	Frec. rel.	Frec. abs.	Frec. rel.
Inserción de tilde	602	1,74	230	1,19	204	2,92	109	1,00
Inserción de letra	1 869	5,41	503	2,59	349	4,99	511	4,72
Omisión de tilde	26 494	76,67	15 668	80,82	5 049	72,33	6 846	63,25
Omisión de letra	2 631	7,62	1 122	5,79	585	8,38	2 847	26,30
Omisión de espacio	158	0,46	1 167	6,02	128	1,83	123	1,14
Sustitución de letra	1 726	5,00	333	1,71	529	7,59	231	2,13
Sustitución de tilde	15	0,04	3	0,02	14	0,20	4	0,04
Transposición de letra	1 049	3,04	361	1,86	123	1,76	152	1,40

clara predominancia de los errores de omisión sobre el resto. En las cuatro especialidades del dominio médico, es el tipo de error más habitual. Urgencias, psiquiatría y cirugía general coinciden en la jerarquía (omisión > inserción > sustitución > transposición), mientras que en UCI el error de transposición ocupa el tercer lugar, y el de sustitución, el cuarto lugar, aunque con una diferencia no significativa.

Si abordamos un nivel de concreción mayor en la jerarquización de errores (tabla 5), observamos que el subtipo de error que posee un mayor porcentaje en las cuatro especialidades es la omisión de tilde. El segundo lugar lo ocupa la omisión de letra, excepto en UCI, cuyo segundo puesto lo ocupa la omisión de espacio. En cuanto al 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.

Como apoyo visual a la relación entre los datos, el gráfico 2 muestra un gráfico de barras apiladas que permite tener una mejor visión de los resultados, puesto que la escala se ha ajustado al 100 %.

Para poner a prueba la independencia de los datos, se ha empleado el estadístico de ji al cuadrado, generando para ello una tabla cruzada en la que las dos variables son de tipo cualitativo (tabla 6), clasificándose en especialidad médica y subtipo de error, con la intención de determinar si ambas variables se encuentran relacionadas. En otras palabras, con ello se pretende testar la hipótesis alternativa que implicaría que existe una

TABLA 6. Tabla cruzada Subtipo_Error*Especialidad

		Especialidad				Total	
		Urgencias	UCI	Psiquiatría	Cirugía		
Subtipo_Error	Inserción de tilde	Recuento	602	230	204	109	1 145
		Recuento esperado	551,4	309,4	111,4	172,8	1 145,0
	Inserción de letra	Recuento	1 869	503	349	511	3 232
		Recuento esperado	1 556,4	873,5	314,5	487,6	3 232,0
	Omisión de tilde	Recuento	26 494	15 668	5 049	6 846	54 057
		Recuento esperado	26 031,2	14 609,4	5 260,6	8 155,8	54 057,0
	Omisión de letra	Recuento	2 631	1 122	585	2 847	7 185
		Recuento esperado	3 459,9	1 941,8	699,2	1 084,0	7 185,0
	Omisión de espacio	Recuento	158	1167	128	123	1 576
		Recuento esperado	758,9	425,9	153,4	237,8	1 576,0
	Sustitución de letra	Recuento	1 726	333	529	231	2 819
		Recuento esperado	1 357,5	761,9	274,3	425,3	2 819,0
	Sustitución de tilde	Recuento	15	3	14	4	36
		Recuento esperado	17,3	9,7	3,5	5,4	36,0
	Transposición de letra	Recuento	1 049	361	123	152	1 685
		Recuento esperado	811,4	455,4	164,0	254,2	1 685,0
	Total	Recuento	34 544	19 387	6 981	10 823	71 735
		Recuento esperado	34 544,0	19 387,0	6 981,0	10 823,0	71 735,0

TABLA 7. Pruebas de ji al cuadrado

	Valor	gl	Significación asintótica (bilateral)
Ji al cuadrado de Pearson	6 753,317 ^a	21	,000
Razón de verosimilitud	5 804,407	21	,000
Asociación lineal por lineal	21,696	1	,000
N de casos válidos	71 735		

a. Se esperaba que una casilla (3,1%) tuviera un recuento menor que 5. El recuento mínimo esperado es 3,50.

relación entre la especialidad médica contenida en cada corpus y el subtipo de error cometido. A la hora de codificar los datos en el paquete de *software* estadístico IBM SPSS, se generó la tabla cruzada que incluía una tercera variable denominada *frecuencia*, la cual se corresponde con los valores observados. Puesto que la prueba de ji al cuadrado principalmente se basa en detectar posibles diferencias entre dichos valores observados y los valores esperados, la tabla 6 incluye ambos tipos de valor.

La tabla 7 muestra los resultados de la prueba de ji al cuadrado, en la que puede verse 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. De este modo, podemos rechazar la hipótesis nula, concluyendo que,

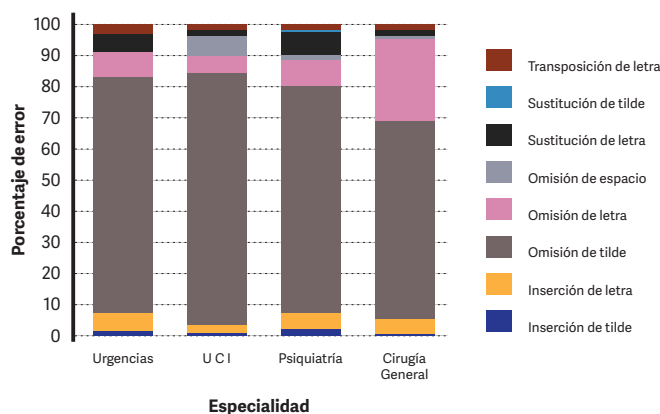


GRÁFICO 2. Errores según el subtipo de operación de edición

TABLA 9. *Patrones de sustituciones erróneas más frecuentes*

Patrón	Ejemplos
E → A	alteraciones (alteraciones), creneoencefálico (craneoencefálico)
O → A	Citolopram (Citalopram), Alprozolam (Alprazolam)
A → E	anaxectomía (anexectomía), atelactasia (atelectasia)
I → E	Seritide (Seretide), craniotomía (craneotomía)
E → O	estonosis (estenosis), goserolina (goserelina)
O → I	tóxicos (tóxicos), cofosis (cifosis)
I → O	autilimitada (autolimitada), neutrifilia (neutrofilia)
S → C	leucositosis (leucocitosis), residivante (recidivante)
C → S	doxazocina (doxazosina), distención (distensión)
S → X	estrasistolia (extrasistolia), hallus (hallux)
X → S	herpex (herpes), extrabismo (estrabismo)
C → Z	Cyprexa (Zyprexa), Becocyme (Becozyme)
Z → C	cabezera (cabecera), enzías (encías)
I → Y	Enantium (Enantyum), Misoline (Mysoline)
Y → I	Betadyne (Betadine), Secalyp (Secalip)
V → B	fivrinógeno (fibrinógeno), varicoflevitis (varicoflebitis)
B → V	bancomicina (vancomicina), cabidad (cavidad)
M → N	parémquima (parénquima), preferemte (preferente)
N → M	abdoninal (abdominal), descompensación (descompensación)
L → R	dolsalgia (dorsalgia), Omeplazol (Omeprazol)
R → RR	bilirubina (bilirrubina), hipereactividad (hiperreactividad)
RR → R	aéreo (aéreo), contracturra (contractura)
C → V	cognitico (cognitivo), cercical (cervical)
G → F	ecogragía (ecografía), diclogenaco (diclofenaco)
S → A	hallszgos (hallazgos), orsles (orales)
S → D	calensario (calendario), absomen (abdomen)
T → R	dolot (dolor), femotal (femoral)
R → T	durandre (durante), gastroenteriris (gastroenteritis)
R → S	verpertino (vespertino), cuerpor (cuerpos)

académica que regula su uso. Finalmente, aquellos casos cuyo error es motivado por la adyacencia en el teclado de esas letras y que evidencian que son errores de actuación o de tipo mecánico. En la tabla 11 se señalan los patrones de error más frecuentes y algunos ejemplos:

6. Discusión

Los análisis nos han permitido detectar una serie de rasgos destacables. En primer lugar, urgencias es la especialidad con mayor número de errores. Como hemos mencionado anteriormente, una posible explicación es el tiempo disponible por consulta y las condiciones más estresantes que existen en esta especialidad para el facultativo. En psiquiatría, la especialidad

con un porcentaje de error más reducido, suelen disponer de mayor tiempo para cada consulta. Además, el subcorpus de psiquiatría contiene más palabras comunes debido a un tipo de escritura más narrativa y descriptiva, lo que ha podido influir en la menor aparición de errores. Las pruebas realizadas han confirmado que existe relación entre las variables de especialidad médica y subtipo de error, lo que implica que hay diferencias significativas entre las cuatro especialidades analizadas.

Como ya hemos señalado anteriormente, tomamos el trabajo de Ramírez y López (2006) como punto de referencia para comparar el español general y el dominio médico, al ser el único estudio que contiene una tipificación cuantitativa de errores para el español general. Dicho corpus está formado por tres conjuntos de datos: el primero contiene textos editados y no editados, el segundo contiene textos altamente editados y el tercero contiene solo textos no editados. Este último conjunto, con alrededor de dos millones de palabras, es el que hemos utilizado para la comparación del total del número de errores debido a que es el que tiene una naturaleza similar a los informes clínicos al no estar editado ni presentar revisión posterior.

Al tratarse de una misma lengua o código, se dan similitudes entre el corpus de español general y el corpus del dominio médico. No obstante, son varias las diferencias que pueden ser mencionadas si entramos en un nivel de precisión mayor. En primer lugar, el número de errores es superior en nuestro corpus con respecto al conjunto del español general no editado, cuyo porcentaje es de 2,86 %, como puede observarse en la tabla 10. Ambos corpus están formados por textos no revisados, pero el corpus de informes clínicos posee un porcentaje de error superior al corpus de español general debido a sus particularidades lingüísticas y contextuales.

TABLA 10. *Comparación de frecuencia de errores con corpus de español común no editado*

Especialidad	Palabras	Errores	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

La tabla 11 muestra que los errores de omisión son el tipo de error más frecuente en ambos trabajos, aunque con un porcentaje superior en el corpus de informes médicos. La mayoría de los errores ortográficos están relacionados con la omisión de tildes en ambos estudios. Ramírez y López (2006) afirman que la omisión es el error más común, seguido de sustitución, inserción y transposición; sin embargo, en nuestro corpus, el orden es omisión, inserción, sustitución y transposición. En el dominio médico, los errores de inserción son superiores a los de sustitución. No obstante, es necesario precisar que, para el cómputo de los tipos de errores, Ramírez y López (2006) obtienen los resultados a partir de la suma de los tres conjuntos de

datos, incluyendo también el corpus de textos editados, por lo que una parte de los resultados no es directamente comparable con nuestro estudio debido a la naturaleza de la lengua recogida, pero lo tomamos como referencia al ser el único trabajo previo sobre cuantificación de errores en español.

TABLA 11. Comparación de tipos de errores con 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

Asimismo, y de acuerdo con los resultados de Damerau (1964), Kukich (1992), Ramírez y López (2006) o Gimenes *et al.* (2015), la mayoría de los errores presentes en el corpus tienden a ser casos únicos de inserción, omisión, sustitución o transposición en la palabra. A diferencia de otros estudios y dominios, la variabilidad en las posiciones de los errores es mayor en nuestro corpus. Esto es debido a que el léxico del lenguaje especializado de la medicina contiene vocablos de mayor longitud que las palabras usadas en la lengua común.

En nuestro corpus, hemos observado que tienen una fuerte presencia los errores de acentuación. Gran parte de los errores se deben al uso intencional de una ortografía no estándar, motivado por la necesidad de llevar a cabo una escritura rápida. El error de omisión de tilde tiene una alta incidencia y puede ser debido a que implica tener que pulsar dos teclas, por lo que se omite para agilizar el proceso. De igual manera, la rapidez en la escritura implica deslizamientos en la coordinación motora, provocando que se presione accidentalmente una tecla adyacente en el teclado en lugar de la pertinente (*hallazgos*, ‘hallazgos¹¹’), que se presionen dos teclas a la vez en lugar de una (*resposo*, ‘reposito’), o que se escriban en el orden equivocado (*laterla*, ‘lateral’). Ramírez y López (2006) consideran que los efectos de adyacencia del teclado son relevantes en su corpus, pero que las frecuencias obtenidas en su estudio sugieren que otros factores cognitivos deben ser tenidos en cuenta. No tenemos un experimento controlado que nos permita afirmar con rotundidad si se trata de errores cognitivos o de acentuación, aunque sí podemos hablar de casos susceptibles de ser errores cognitivos. Entre ellos, podemos destacar aquellos casos que reflejan un desconocimiento de la norma actual, como la formación de palabras con prefijo mediante guion o espacio (*ex-consumidor*, ‘exconsumidor’), o la acentuación de demostrativos y monosílabos (*vió*, ‘vio’ o *fué*, ‘fue’). Otros casos pueden ser tildar dos veces una palabra compuesta (*médicoquirúrgicas*, ‘medicoquirúrgicas’) o el uso incorrecto de la erre si la

posición del sonido vibrante múltiple es intervocálica (*prerenal*, ‘prerrenal’, *normoreactivas*, ‘normorreactivas’ o *microrotura*, ‘microrrotura’). También son destacables los errores motivados por similitudes fonéticas entre consonantes (*sujestivas*, ‘sugestivas’, *clabulánico*, ‘clavulánico’, *tarzalgia*, ‘tarsalgia’ o *polialtralgias*, ‘poliartralgias’). De igual forma, se sustituyen vocales como la a por la e (*amigdelectomizada*, ‘amigdalectomizada’) o la i por la e (*alimenticeo*, ‘alimenticio’), por procesos de asimilación fonética que se trasladan a la escritura. En estos casos, no son letras adyacentes en el teclado, pero se sustituyen de forma significativa, como puede observarse en la matriz de confusión.

Es importante detenernos en los errores relacionados con la inserción del guion en la construcción de palabras mediante prefijación y composición. Hemos detectado que es un uso erróneo extendido en el corpus analizado. Algunos ejemplos son: *pre-anestesia* (‘preanestesia’), *pre-síncope* (‘presíncope’) o *sub-condral* (‘subcondral’). Encontramos también casos en los que se deja un espacio entre el prefijo y la palabra (*pre menstrual*, ‘premenstrual’) o se combina el uso del guion con la inserción de espacio (*post-quirúrgica*, ‘postquirúrgica’). De igual forma ocurre con las palabras compuestas, con casos como *dorso-lumbar* (‘dorsolumbar’) o *uretero-vesical* (‘ureterovesical’).

Es también relevante mencionar la presencia de errores relacionados con el uso de abreviaturas, símbolos y siglas en el corpus. Un error frecuente es escribir los símbolos acompañados de un punto (*gr.*, ‘gr’) o añadir -s para crear el plural (*kgs.*, ‘kg’). Las letras que componen las siglas deben ser escritas en mayúsculas, pero muchas de las presentes en los corpus fueron escritas en minúsculas (*itu.*, ‘ITU’), lo que dificulta su identificación. El uso de abreviaturas (*cél.*, ‘célula’, *postqx.*, ‘postquirúrgico’, o *sd.*, ‘síndrome’) es continuo en los informes clínicos para acelerar la escritura; sin embargo, se observa falta de estandarización y uniformidad en la creación y uso de estas. El tratamiento de las abreviaciones gráficas requiere una investigación propia debido a la amplitud y complejidad que plantea su presencia en el corpus, por lo que no han sido incluidas en los cómputos de este estudio.

Los hallazgos respecto a la distancia de edición, el tipo de errores y la matriz de confusión permiten dar indicaciones a los sistemas de corrección para optimizar el tiempo de búsqueda durante el proceso y mejorar la precisión en la generación de candidatos para la corrección. Por ejemplo, si sabemos que más del 90 % de los errores ocurren a distancia 1 de la palabra meta y que la mayoría son errores de omisión de tilde, puede indicarse al sistema que comience la búsqueda de candidatos por ese tipo de error y a un solo movimiento de edición de la palabra errónea. También puede otorgarse una puntuación mayor en la lista de sugerencias de corrección a aquellos pares de caracteres que planteen confusión entre sí frecuentemente. De igual forma, puede aportarse un módulo con reglas para detectar y corregir los errores más comunes provocados por desconocimiento de la norma, según criterios posicionales, léxicos o morfológicos, como los que tienen que ver con la prefijación y composición de palabras.

7. Conclusiones

Este trabajo aporta el primer análisis cuantitativo de errores lingüísticos en informes clínicos en español. Además, los datos han sido obtenidos a partir de cuatro especialidades médicas distintas, contribuyendo a la tipificación de los errores característicos en este dominio y a la mayor cobertura de casos. Como resultado de este enfoque comparativo, se han identificado similitudes y diferencias entre las especialidades seleccionadas, lo que nos ha permitido detectar patrones lingüísticos y los pares de error o confusión más comunes. La especialidad con un mayor número de errores detectados es urgencias. La mayoría de las palabras erróneas tienen distancia de edición 1 con respecto a la palabra correcta y el tipo de error más común con una alta incidencia es el de omisión. Entre los patrones de error más frecuentes destacan la omisión e inserción de tilde, la sustitución de caracteres con similitudes fonéticas (como *s/c*, *c/z*, *n/m* o *y/i*), los motivados por desconocimiento de la norma ortográfica actual y la sustitución y transposición de caracteres adyacentes en el teclado.

En cuanto a futuras investigaciones, se intentarán ampliar los conjuntos de datos y analizar otras especialidades para compararlas con los resultados obtenidos. El proyecto pretende incorporar los llamados errores *real-word* al análisis mediante el uso de técnicas basadas en contexto y profundizar en el tra-

tamiento de siglas y abreviaturas. Por último, los resultados obtenidos van a ser utilizados para la implementación de un módulo basado en conocimiento lingüístico, que será añadido a la arquitectura del corrector para aportar más información al listado de sugerencias usado en la generación de candidatos, complementando a las técnicas basadas en aprendizaje automático.

Agradecimientos

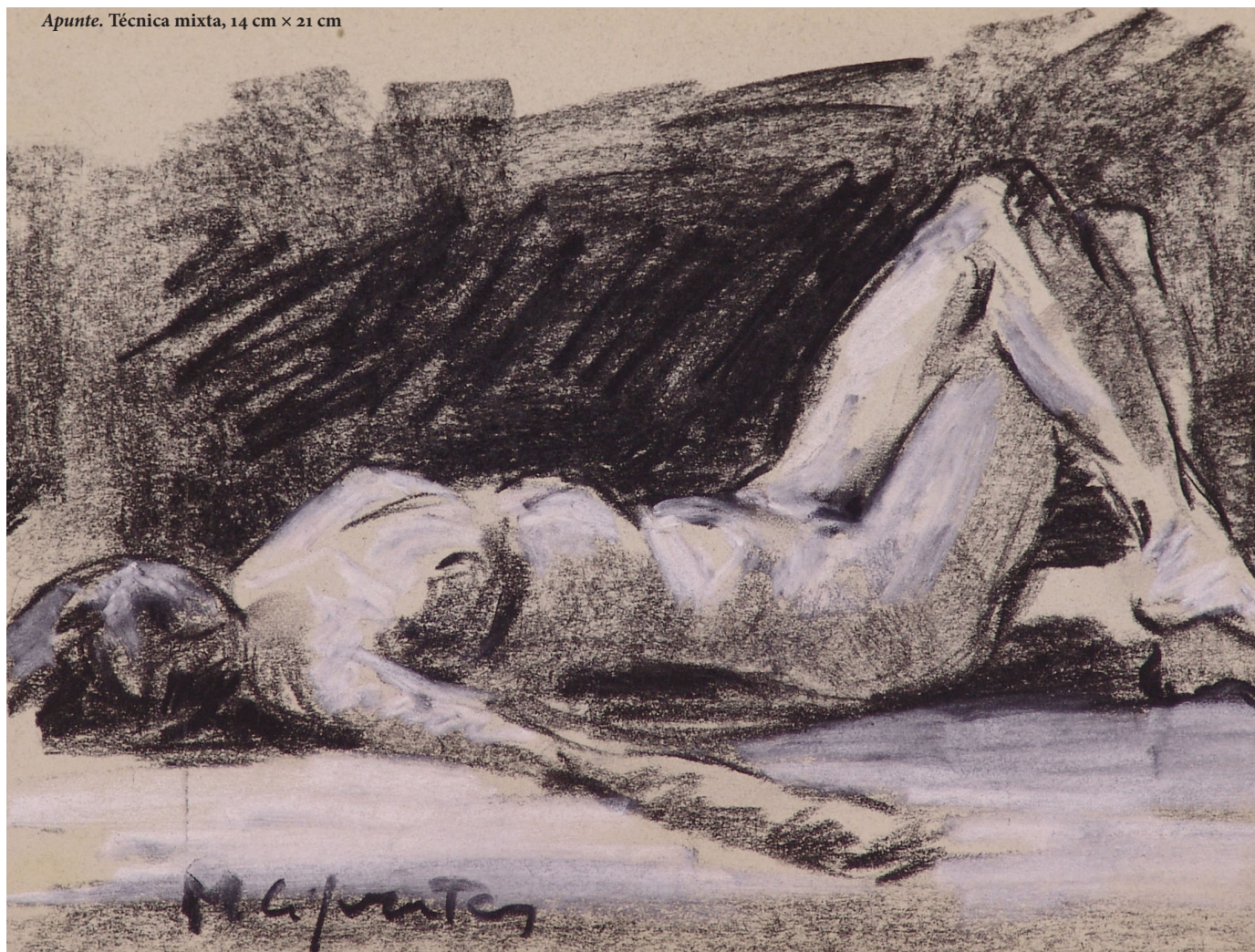
Este trabajo ha sido financiado por la Agencia Estatal de Investigación (AEI) a través del proyecto LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). Además, la investigadora predoctoral Jéssica López Hernández ha sido financiada por el Ministerio de Universidades de España a través del programa nacional de «Ayudas para la formación de profesorado universitario» (FPU), con referencia FPU16/03324.

Extendemos nuestro agradecimiento a la empresa Vócali (www.vocali.net) y al Dr. Fernando Molina por su inestimable ayuda y contribución en el desarrollo técnico de este trabajo.

Notas

1. ReCor es un programa que mide la representatividad de los corpus, creado por Gloria Corpas Pastor, Miriam Se-

Apunte. Técnica mixta, 14 cm × 21 cm



ghiri Domínguez y Romano Maggi, Universidad de Málaga. Disponible en <<http://www.lexytrad.es/es/recursos/recor-2/>> [consulta: 24.II.2021].

- WordSmith Tools: <<https://www.lexically.net/wordsmith/>> [consulta: 6.I.2021].
- Vócali: <<https://vocali.net/>> [consulta: 6.I.2021].
- Reglamento General de Protección de Datos (RGPD): <<https://gdpr-info.eu/>> [consulta: 6.I.2021].
- Sitio web de Hunspell con los recursos necesarios para implementar el corrector ortográfico y sus distintas herramientas: <<http://hunspell.github.io/>> [consulta: 6.I.2021].
- 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/>> [consulta: 6.I.2021].
- 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> [consulta: 6.I.2021].
- Agencia Española de Medicamentos y Productos Sanitarios (AEMPS): <<https://www.aemps.gob.es/>> [consulta: 6.I.2021].
- Sociedad Española de Documentación Médica (SEDOM): <<http://www.sedom.es/>> [consulta: 6.I.2021].
- Software estadístico IBM SPSS Statistics 24: <<https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-24>> [consulta: 6.I.2021].
- Entre comillas simples se ofrece la versión correcta de los ejemplos.

Referencias bibliográficas

Aguilar Ruiz, Manuel José (2013): «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>> [consulta: 5.X.2020].

Aleixandre, Rafael; Juan Carlos Valderrama y Francisco Jesús Bueno-Cañigral (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>> [consulta: 8.XII.2020].

Baba, Yukino y Hisami Suzuki (2012): «How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs», en Haizhou Li *et al.* (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>> [consulta: 12.X.2020].

Chipere, Ngoni; David D. Malvern y Brian J. Richards (2004): «Using a corpus of children's writing to test a solution to the

sample size problem affecting Type-Token Ratios», en Guy Aston, Silvia Bernardini y Dominic Stewart (eds.): *Corpora and language learners*. Amsterdam: John Benjamins, pp. 139-147. <<https://doi.org/10.1075/scl.17.10chi>> [consulta: 6.IX.2020].

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

Damerau, Frederick Jacob (1964): «A technique for computer detection and correction of spelling errors», *Communications of ACM*, 7 (3): 171-177. <<https://doi.org/10.1145/363958.363994>> [consulta: 26.IX.2020].

Díaz Villa, Ana María (2005): «Tipología de errores gramaticales para un corrector automático», *Procesamiento del Lenguaje Natural*, 35: 409-416. <<http://hdl.handle.net/10045/1341>> [consulta: 23.IX.2020].

Dziadek, Juliusz; Aron Henriksson y Martin Duneld (2017): «Improving terminology mapping in clinical text with context-sensitive spelling correction», *Studies in health technology and informatics*, 235: 241-245. <<https://pubmed.ncbi.nlm.nih.gov/28423790/>> [consulta: 10.XI.2020].

Gimenes, Priscila Azar; Norton Trevisan Roman y Ariadne M. Carvalho (2015): «Spelling error patterns in Brazilian Portuguese», *Computational Linguistics*, 41 (1): 175-183. <<https://www.aclweb.org/anthology/J15-1011.pdf>> [consulta: 10.XI.2020].

Jurafsky, Daniel y James Martin (2014): *Speech and Language Processing*. Upper Saddle River, NJ: Pearson Education.

Kilicoglu, Halil; Marcelo Fiszman, Kirk Roberts y Dina Dement-Fushman (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/>> [consulta: 12.XII.2020].

Kukich, Karen (1992): «Technique for automatically correcting words in text», *ACM Computing Surveys*, 24 (4): 377-439. <<https://doi.org/10.1145/146370.146380>> [consulta: 12.XII.2020].

Lai, Kenneth H.; Maxim Topaz, Foster R. Goss y Li Zhou (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>> [consulta: 3.XII.2020].

Lehal, Gurpreet y Meenu Bhagat (2007): «Spelling error pattern analysis of Punjabi typed text», en *Proceedings of the 2007 International Symposium on Machine Translation, NLP and TSS*. New Delhi: Tata McGraw-Hill, pp. 128-141. <<http://learnpunjabi.org/pdf/icon2004.pdf>> [consulta: 20.XII.2020].

Levenshtein, Vladimir Iosifovich (1966): «Binary codes capable of correcting deletions, insertions and reversals», *Soviet Physics Doklady*, 10 (8): 707-710.

Liu, Hongfang; Stephen T. Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Waghlikar, Peter J. Haug, Stanley Mark Huff y Christopher G. Chute (2012): «Towards a semantic lexicon for clinical natu-

- ral language processing», en *AMIA Annual Symposium proceedings*. Chicago: American Medical Informatics Association, pp. 568-576. <<https://pubmed.ncbi.nlm.nih.gov/23304329/>> [consulta: 20.XII.2020].
- López-Hernández, Jéscica; Ángela Almela y Rafael Valencia-García (2019): «Automatic spelling detection and correction in the medical domain: A systematic literature review», en Rafael Valencia-García *et al.* (eds.): *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science*, vol. 1124. Cham: Springer, pp. 104-117. <https://doi.org/10.1007/978-3-030-34989-9_8> [consulta: 29.X.2020].
- Meystre, Stephane y Peter Haug (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>> [consulta: 15.XII.2020].
- Naber, Daniel (2003): *A rule-based style and grammar checker*. Munich: GRIN Verlag. <http://www.danielnaber.de/lanaguagetool/download/style_and_grammar_checker.pdf> [consulta: 9.XI.2020].
- Nagata, Ryo; Hiroya Takamura y Graham Neubig (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>> [consulta: 20.XII.2020].
- Navarro, Fernando 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.
- Paggio, Patrizia (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 (ACL), pp. 255-261. <<https://www.aclweb.org/anthology/A00-1035.pdf>> [consulta: 22.XII.2020].
- Pedler, Jennifer y Roger Mitton (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 (ELRA), pp. 755-762. <http://www.lrec-conf.org/proceedings/lrec2010/pdf/122_Paper.pdf> [consulta: 20.XII.2020].
- Pollock, Joseph J. y Antonio Zamora (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>> [consulta: 20.XII.2020].
- Ramírez, Flora y Enrique López (2006): «Spelling error patterns in Spanish for word processing applications», en *Proceedings of Fifth international conference on Language Resources and Evaluation (LREC)*. Genoa: European Language Resources Association, pp. 93-98. <http://www.lrec-conf.org/proceedings/lrec2006/pdf/119_pdf.pdf> [consulta: 3.I.2021].
- Rodríguez-Rubio Mediavilla, Santiago (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>> [consulta: 14.IX.2020].
- Real Academia Española y Asociación de Academias de la Lengua Española (2005): *Diccionario panhispánico de dudas*. Madrid: Santillana. <<https://www.rae.es/dpd/>> [consulta: 3.XII.2020].
- 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 (2011): *Diccionario de Términos Médicos*. Madrid: Panamericana.
- Rello, Luz; Ricardo Baeza-Yates y Joaquim Llisterra (2014): «DysList: An annotated resource of dyslexic errors», en *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. Reykjavik: European Language Resources Association, pp. 1289-1296. <<https://doi.org/10.13140/2.1.2542.7205>> [consulta: 8.XI.2020].
- Ruch, Patrick; Robert Baud y Antoine Geissbühler (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](https://doi.org/10.1016/S0933-3657(03)00052-6)> [consulta: 25.X.2020].
- Sayle, Roger Anthony; Plamen Petrov, Jon James Winter-Holt y Sorel Muresan (2012): «Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction», *Journal of Chemical Information and Modeling*, 3: 51-62. <<https://doi.org/10.1186/1758-2946-3-S1-O16>> [consulta: 13.XII.2020].
- Siklósi, Borbála; Attila Novák y Gábor Prószéky (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>> [consulta: 2.IX.2020].
- Wong, Wilson y David Glance (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>> [consulta: 27.IX.2020].
- Yannakoudakis, Emmanuel J. y David Fawthrop (1983): «The rules of spelling errors», *Information processing and management*, 19 (12): 101-108. <[https://doi.org/10.1016/0306-4573\(83\)90045-6](https://doi.org/10.1016/0306-4573(83)90045-6)> [consulta: 4.X.2020].
- Zech, John; Jessica Forde; Joseph Titano; Deepak Kaji; Anthony Costa y Enric Karl Oermann (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>> [consulta: 22.X.2020].