



DE LA “A” A LA “AZ”. ESTUDIO DE LAS ANOMALÍAS FORMALES DEL TEXTO EN ESPAÑOL DEL *CAMBRIDGE DICTIONARY ONLINE* (ENGLISH-SPANISH/SPANISH-ENGLISH)

FROM “A” TO “AZ”. A STUDY OF FORMAL ANOMALIES OF THE SPANISH TEXT IN *CAMBRIDGE DICTIONARY ONLINE* (ENGLISH-SPANISH/SPANISH-ENGLISH)

Santiago Rodríguez-Rubio
Universidad Pablo de Olavide
srodmed@upo.es

RESUMEN

En el presente trabajo, se analiza el texto en español de las entradas correspondientes a la letra “A” de las partes English-Spanish y Spanish-English del [Cambridge Dictionary Online](https://www.cambridge.org/dictionary), en función de los siguientes tipos de anomalías formales: mayúscula, puntuación, diseño, acento ortográfico, espacio, *errata de no palabra* (ENP) y *errata de palabra* (EP). Los resultados en términos de incidencia y ubicación de los tipos/subtipos de anomalías demuestran el carácter sistemático de los defectos y reclaman una acción correctora acorde, bajo la premisa de una revisión integral. Más allá del análisis de las anomalías, se abordan aspectos relacionados con la calidad lexicográfica como: a) la posible relación entre los defectos y los métodos (manual, automatizado o semiautomatizado) de introducción de datos; b) el difícil acomodo entre la productividad y la calidad en un contexto de creciente complejidad en que la tecnología o el uso que se hace de ella sigue provocando ajustes; c) la colaboración de los usuarios en la introducción y revisión del contenido lexicográfico; y d) el efecto perjudicial que las imágenes publicitarias pueden tener en los usuarios de los diccionarios en línea gratuitos.

Palabras clave: anomalías formales, calidad lexicográfica, lexicografía en línea gratuita, revisión de diccionarios.

ABSTRACT

The present study analyses the Spanish text of entries corresponding to letter “A” in parts English-Spanish and Spanish-English of the [Cambridge Dictionary Online](https://www.cambridge.org/dictionary), according to the following formal anomaly types: capitalisation, punctuation, design, written accent, spacing, non-word error (*errata de no palabra*/ENP) and real-word error (*errata de palabra*/EP). The results in terms of number of anomalies and location of the anomaly types/subtypes reveal a system of anomalies and ask for appropriate corrective measures, on the premise of a thorough revision. In addition to anomaly analysis, the following lexicographic quality aspects are dealt with: a) the possible relationship between the defects and the data input methods (manual, automated or semi-automated); b) the difficult balance between productivity and quality, in a context of increasing complexity where there are still gaps in relation to technology (or the use humans make of that technology); c) user participation in the input and revision of lexicographic content; and d) the detrimental effect that promotional images in free online dictionaries may have on users.

Keywords: formal anomalies, dictionary quality, free online lexicography, dictionary revision.

Recibido: 20-04-2026
Aceptado: 19-06-2026

DOI: <https://doi.org/10.17561/rilex.9.2.10412>



1. INTRODUCCIÓN

En el presente estudio, se han analizado las anomalías formales del texto en español de las entradas correspondientes a la letra “A” de las partes English-Spanish y Spanish-English del [Cambridge Dictionary Online](#). En una lógica integral que desborda ampliamente la de las colaboraciones puntuales de usuarios, presentamos un necesario análisis de defectos que traslada la importancia de llevar a la práctica una de las ventajas teóricas de los diccionarios en línea: una más “rápida y fácil” revisión lexicográfica.

Las anomalías son fenómenos ampliamente observados que afectan a todo tipo de textos. Los errores tipográficos (una de las formas más comunes de ruido natural en los textos escritos creados por usuarios, normalmente debido a un tecleo erróneo accidental) pueden aparecer en modelos lingüísticos de gran tamaño (Liu, Zhao, Altinger, Schütze & Hedderich, 2025). Debido a la natural habilidad de las erratas para abrirse camino hacia el texto publicado (Wheatley, 1893), no solo los usuarios, sino también desarrolladores o editores pueden generar erratas cuando introducen datos en sus recursos. Además, la errata no solo puede introducirse a través de un error de actuación humano, sino también automáticamente por acción del corrector ortográfico. De hecho, Sharma y Gupta (2015, pp. 99-100) manifiestan que los avances tecnológicos (informáticos) han supuesto una escalada de *spelling errors*, afirmación que, salvando las distancias, concuerda con la lógica de que “as datasets become larger and more complex, it can be challenging to develop anomaly detection methods that can scale effectively to these larger datasets” (Gujral, 2023, p. 7).

Entre los problemas más acuciantes, se encuentra la asunción errónea de que los textos que se introducen (por ejemplo, aquellos que forman parte de modelos lingüísticos) están limpios, dejando el terreno de la robustez desatendido en cuanto al aspecto de las erratas (Liu, Zhao, Altinger, Schütze & Hedderich, 2025). Esa asunción contrasta en cierto modo con la noción de “Dirty Web” como fuente lexicográfica, es decir, el hecho de que el

contenido de Internet puede contener numerosos errores lingüísticos, a pesar de que el propio volumen de información funcione a menudo como filtro de corrección; por ejemplo, la forma *believe* aparece en 3,8 millones de búsquedas, frente a las 41 600 ocurrencias de *beleive* y las 1700 de *beleave*, lo que permite identificar fácilmente la forma correcta (Svensén, 2009, pp. 59-60).

2. CALIDAD DE LOS DICCIONARIOS ELECTRÓNICOS

2.1. INTRODUCCIÓN Y REVISIÓN DE CONTENIDO EN DICCIONARIOS ELECTRÓNICOS

Destacados especialistas han manifestado de manera constante en el tiempo que la tarea de revisión de un diccionario es, en realidad, inacabable (Landau, 2001; Klosa-Kückelhaus & Tiberius, 2024). Herold, Meyer y Wiegand explican que el marco de revisión de los diccionarios en línea suele situarse entre el de los diccionarios en papel (cuya reedición podía tardar años o incluso décadas) y el de la publicación inmediata de actualizaciones individuales (2024, p. 22). No obstante, Landau (2001, pp. 399-400) cuestionó que los diccionarios electrónicos se revisaran de manera continua y advirtió sobre el deterioro de la fase de revisión (*proofreading*) que la espiral y acumulación de tareas propiciada por los propios avances tecnológicos puede ocasionar. Jackson (2018, p. 836) significativamente usa el adverbio *potentially* cuando alude a la superior actualización de los diccionarios online respecto de los realizados en papel, a la par que Lew y De Schryver incluyen (como no podía ser de otra forma) el modal de posibilidad en su afirmación de que “Online dictionaries as well as dictionary apps can be updated as often as needed” (2014, p. 345). Sea como fuere, Herold, Meyer y Wiegand declaran: “In online reference works that are compiled with a limited set of editors, a dictionary entry will only be published after rigorous checking and approval” (2024, p. 22).

Svensén (2009, p. 421) distingue diversos *modus operandi* en cuanto a la introducción y revisión del contenido lexicográfico. En primer lugar, el uso de un procesador (especialmente cuando se acompaña de anotación

tipográfica) limita la búsqueda y la revisión automática de la coherencia y de la conformidad con los principios editoriales establecidos. En segundo lugar, el procedimiento basado en bases de datos léxicas tiene la virtud de que se pueden hacer búsquedas en función del tipo de campo, lo que facilita que el usuario pueda encontrar conexiones basadas en la ausencia de linealidad. Finalmente, se puede usar un lenguaje de marcado (por ejemplo, Extensible Markup Language, XML) que permite un acceso más rápido a los datos almacenados¹. En entradas complejas, la compilación y edición se desgrana en múltiples campos, entre los que se encuentran aquellos relativos al idioma de referencia; en diccionarios basados en lenguaje de marcado, puede usarse una sintaxis determinada para definir el formato en términos de negrita y cursiva, entre otras características (Abel & Meyer, 2024, pp. 233-234). Otros autores incluyen el tipo y tamaño de letra, la negrita, la cursiva, la mayúscula y signos de puntuación como las comillas dentro de la categoría *formal properties* (Koeva & Blagoeva, 2013, p. 59). Por su parte, Bloodgood y Strauss (2016, p. 79, pp. 81-82) expusieron seis métodos para marcar anomalías en diccionarios electrónicos basados en XML, entre los que se encuentran sistemas basados en expectativas sobre contenido de un determinado tipo de campo (*single-field*) y sobre contenido de múltiples campos interrelacionados (*tied-field*). El primer tipo de sistema descubrió la presencia de una barra (/) cuyo valor disyuntivo no encajaba en un campo que no admitía opciones. Se pusieron como ejemplos de campos interrelacionados la ortografía y la transcripción fonética, o la frase ilustrativa y su traducción. De manera análoga,

¹ Basándose en la *Quality Assurance Activity Statement* (World Wide Web Consortium, W3C), Rosenblum y Golfman (2004) manifestaron que el aseguramiento de la calidad es un aspecto fundamental del lenguaje XML y puede usarse para detectar automáticamente si se cumplen determinados requisitos (por ejemplo, que el nombre de una revista aparezca en cursiva o que contenga palabras que no incluyan letras iniciales mayúsculas). Los autores señalan un aspecto fundamental al referir que muchas organizaciones despliegan un aseguramiento de la calidad limitado o no lo despliegan en absoluto, ya que “quality assurance incurs extra cost, and many organizations are unwilling or unable to make that investment” (Rosenblum & Golfman, 2004, p. 5).

Rodrigues, Zajic, Doermann, Bloodgood y Ye (2011) ya habían explorado métodos que permitieran a los lexicógrafos detectar anomalías intraentrada e interentrada bajo el prisma de la regularidad estructural, mediante el uso de medidas de perplejidad. Se mencionó, por ejemplo, el hecho de que en los diccionarios bilingües el texto de los idiomas de referencia estuviera mezclado. Los autores emplearon un *Dictionary Manipulation Language* (DML, en Zajic, Maxwell, Doermann, Rodrigues & Bloodgood, 2011, p. 297), una de cuyas ventajas principales era la posibilidad de marcar los errores en el diccionario original.

En todo caso, según Svensén (2009, pp. 422-423), los programas de edición lexicográfica que se usen deben (entre otras funciones): alertar sobre redundancias de lemas; permitir la función copia-pegar de texto; promover que se respete el orden estructural de manera que los lemas aparezcan antes que las definiciones; posibilitar el uso de plantillas de entrada; y facilitar la consulta de un registro de modificaciones (especialmente útil en la fase editorial del diccionario).

No existe un sistema de edición lexicográfica perfecto. Así, según Simões y Salgado (2022, p. 425), el modelo TEI Lex-0 (versión avanzada de la *Text Encoding Initiative*) refuerza la interoperabilidad y permite un más fácil manejo computacional, pero puede ser más propenso al error cuando se realiza una edición manual de los recursos. Los autores proponen métodos de simplificación de la codificación de fenómenos (por ejemplo, de la sinonimia), al objeto de que el lexicógrafo pueda centrarse en su labor de análisis lingüístico. El modelo TEI Lex-0 facilita la inclusión de ejemplos tanto bibliográficos (entrecomillados) como inventados por el propio lexicógrafo (Simões & Salgado, 2022, p. 426, p. 428). Es de suponer que el lexicógrafo introduce los ejemplos de uso inventados de manera manual, o, al menos de manera híbrida manual-automática.

El lexicógrafo tiene a su disposición programas de gestión lexicográfica (*Dictionary Writing Systems*, DWS) que pueden incrementar la productividad

gracias al uso de plantillas y de la función de copia, además de reducir el riesgo de errores aleatorios (Koeva & Blagoeva, 2013, p. 65). Dichos programas pueden usarse en tareas de revisión o edición (Abel, 2012; Barbierik, Děngeová, Holcová, Jarý, Liška, Lišková & Virius, 2014), pero más recientemente se ha mostrado que, según una encuesta sobre prácticas lexicográficas realizada a nivel europeo, los DWS eran útiles para comprobar la cohesión, pero muchos de ellos no incluían un corrector ortográfico (*spellchecker*) (Kallas, Koeva, Kosem, Langemets & Tiberius, 2019). Algo más recientemente, frente a los problemas de calidad de los grandes conjuntos de datos de corpus, los creadores de la aplicación “Lexical Data API” basada en los *K Dictionaries* (KD) han defendido una combinación de revisión automática y manual de sus recursos, privilegiando la calidad sobre la cantidad (Lonke, Kernerman & Dzhuranyuk, 2022, p. 405).

Por otro lado, las tareas de revisión requieren de una elevada sistematicidad que tenga en cuenta conexiones y la multiplicidad de espacios en los que la misma información pueda encontrarse. Ese condicionante aparece en la literatura de manera constante en el tiempo: desde el estándar denominado *Multi-Dictionary Formatter* o MDF, en cuyo marco los errores deben corregirse no solo en la versión imprimible del diccionario, sino también en la base de datos léxica original (Coward & Grimes, 2000, p. 63), hasta las bases de datos lexicográficas centralizadas, que integran una “constelación” de recursos, con la consecuente complejidad en la gestión de los datos (Klosa-Kückelhaus & Tiberius, 2024, p. 90).

En relación con el uso de corpus, Geyken y Lemnitzer (2024, pp. 219-220) plantean que una de las desventajas de la integración de fuentes primarias en los diccionarios digitales es precisamente la presencia en ellas de errores, equivocaciones (*mistakes*) e incoherencias, a las que se suman “errores” que surgen “during the process of automatically annotating and analysing the data since no language technology tool can operate without making mistakes”. Los autores indican la pertinencia de lo anterior en términos de fiabilidad

(ya que los usuarios esperan encontrar información conforme a las normas lingüísticas) y apuntan que una de las discrepancias puede deberse a la diferencia entre formas extraídas de la fuente primaria y que se desvían de la norma ortográfica (*raw data*) y, por otro lado, las que han de consignarse en el diccionario con arreglo a dicha norma (*processed data*).

2.2. COLABORACIÓN DE LOS USUARIOS DE DICCIONARIOS Y APUNTES EMPRESARIALES

Podemos partir de la idea básica de que la interacción de los usuarios y lexicógrafos es fundamental para crear diccionarios fáciles de usar. Abel y Meyer amplían el foco, expresando que el grado de participación de los usuarios de diccionarios puede abarcar desde compilaciones completas hasta entradas individuales, incluyendo la corrección de errores, en un marco en que “user participation has become an increasingly important factor in the planning, development, and use of a dictionary” y en el que incluso las editoriales lexicográficas pueden ahorrar dinero (Abel & Meyer, 2024, pp. 228-230).

Atendiendo a lo expresado por Klosa-Kückelhaus y Tiberius (2024, pp. 71-73), la interacción de los usuarios con los editores (por ejemplo, reportando errores ortográficos) no solo está asentada en la práctica lexicográfica, sino que puede suponer un impulso para el desarrollo y la mejora de los diccionarios electrónicos. En función de los proyectos, editores jefes, directores de proyecto y lexicógrafos intervienen en varios estadios del proceso de preparación para corregir, entre muchos otros aspectos, errores ortográficos e inconsistencias. En algunos casos, el lenguaje de marcado o *mark-up language* “was subsequently refined in order to identify other units of information in the microstructure of the dictionary text that could be derived from typographical features (examples, for instance, are always coded in italics)” (Klosa-Kückelhaus & Tiberius, 2024, pp. 79-81, p. 83).

En relación con el tipo de diccionario objeto del presente estudio (es decir, un diccionario electrónico institucional respaldado por una editorial asentada), Abel y Meyer (2024) mencionan que no conocen ningún

proyecto activo de participación “directa” de los usuarios y afirman: “this indicates that this type of direct user participation was not successful. This can be due to the economic situation of the institution, the costs of running such a service, the quality of the submissions, or the amount of work that would be needed to make professional use of these contributions” (Abel y Meyer, 2024, p. 237). La participación “indirecta” implica sugerencias, solicitudes y corrección de errores. Las contribuciones conllevan ventajas e inconvenientes. Por un lado, pueden suponer un considerable trabajo para el editor, ya que, por ejemplo, no solo ha de verificar y categorizar la contribución, sino también comprobar si puede localizar los errores a partir de las indicaciones que los usuarios facilitan (Abel y Meyer, 2024, p. 238, p. 241). Por otro lado, desde una perspectiva más empresarial, las aportaciones pueden conllevar una mejora de la calidad de las obras o productos, una oportunidad de publicidad y de atraer a los usuarios hacia su marca y un valor añadido en forma de ahorro de tiempo y dinero (Abel y Meyer, 2024, p. 248, p. 250). No obstante, conviene recordar que el mantenimiento o la actualización a menudo son desatendidos en el campo lexicográfico (Lorentzen & Trap-Jensen, 2016, p. 138).

2.3. TIPOS DE ANOMALÍAS

El concepto de anomalía no ha sido delimitado y suele variar en función del campo de estudio (Foorthuis, 2021, p. 297). En el campo de las ciencias de la informática, Gujral lo define como “rare or unusual events, observations, or patterns that do not conform to the expected behavior or norm within a dataset” (2023, pp. 1-3). El autor señala que las anomalías pueden ser causadas por errores humanos y que su detección puede servir para hacer un seguimiento de los defectos en los procesos de fabricación. Adams y Maxwell establecieron la siguiente distinción: “an anomaly is an unusual structure that has been noticed [...] an error is a structure which a human expert has determined is incorrect. Errors range from typos to information which has been mis-tagged...” (2016, p. 23).

2.3.1. Erratas

Partimos de la distinción general entre *non-word error* (forma errónea que no está en un diccionario) y *real-word error* (forma errónea que adopta el aspecto de una palabra). La dificultad que los correctores ortográficos (*spell-checkers*) y aplicaciones informáticas tienen a la hora de detectar y corregir *real-word errors* (también denominados *grammatical errors*) ha sido referida de manera constante y sostenida en el tiempo (Peterson, 1986, pp. 633-634; Gardner, 1992, p. 51; Fossati & Di Eugenio, 2008, p. 896; Samanta & Chaudhuri, 2013, p. 211; Dashti, Bardsiri & Bardsiri, 2018, p. 488). Dando una vuelta de tuerca, estos últimos autores puntualizan que los propios correctores ortográficos pueden generar *real-word errors* cuando intentan corregir *non-word errors* cometidos por los humanos (p. 488).

Por otro lado, son cuatro las subcategorías de erratas ampliamente adoptadas en el estudio del modelaje del ruido y la evaluación de la robustez de los sistemas: sustitución, adición, supresión y transposición (Liu, Zhao, Altinger, Schütze & Hedderich, 2025). Como se expuso en otro lugar, estas subcategorías fueron sentadas por estudios de los campos de la sicolingüística y del procesamiento del lenguaje natural y, según Blecua (1983), se remontan a la antigüedad griega (Rodríguez-Rubio & Fernández-Quesada, 2021). Dado su carácter universal, son también las cuatro categorías principales que se tienen en cuenta en el campo de las bases de datos en línea (Gardner, 1992, p. 50)².

A pesar de ser universales, ciertos fenómenos motores responsables de equivocaciones del lenguaje escrito han recibido diversas denominaciones en función del período histórico y de la disciplina de estudio. En el campo de la crítica textual, se puso como ejemplo de “atracción” la forma *piereden*

² Desde la perspectiva del usuario, el impacto de las erratas (o, en su caso, de los errores ortográficos) puede variar desde una simple molestia, hasta errores científicos o con implicaciones legales, pasando por una percepción de dejadez que afecte a la credibilidad del comunicador científico/profesional o de la propia base de datos (Gardner, 1992, p. 52).

(por *pierden*) (Blecua, 1983). El fenómeno de la “influencia” entre letras como causa de las erratas se ha mencionado desde el PLN, por ejemplo, *évidemement* (por *évidemment*); *aéoroport* (por *aéroport*) o *bubget* (por *budget*) (Ren & Perrault, 1992, p. 411). En el marco de la sicolingüística, se expuso como ejemplo de “anticipación” (o “migración”) *bfiber* por *fiber* (Grudin, 1983, p. 133) y de “perseveración” *difffer* (por *differ*) (Logan, 1999). Con anterioridad a todos ellos, ya se había descrito desde ese mismo terreno la anticipación (entre palabras, a mayor abundamiento): *wrapid** *writing* (Lashley, 1951)³. Dichas manifestaciones se enmarcan en lo que en psicología se denomina *serial order*, fenómeno que se ha estudiado ampliamente en relación con la producción hablada, pero también con la escrita (Snyder & Logan, 2014, p. 1697).

Por último, poniendo el foco en uno de los elementos fundamentales del terreno metalexicográfico, no debe extrañar que en algunos estudios no se incluyan entre los requisitos para que un ejemplo sea bueno el hecho de que no contenga erratas (Kosem, Koppel, Kuhn, Michelfeit & Tiberius, 2019; Geyken & Lemnitzer, 2024, p. 212-214), pues, entendemos, se trata de una condición tan básica (léase, fundamental) que se da por sentada. Otros autores sí refieren expresamente ese requisito (Leroyer, 2018, pp. 445-446).

2.3.2. Otras anomalías (*mayúscula, acento ortográfico, signos de puntuación y diseño*)

El uso de mayúsculas iniciales en lemas es un aspecto importante que el lexicógrafo ha de tener presente al realizar su obra. Para evitar opciones no acordes, como escribir todos los lemas con letra inicial mayúscula, con letra inicial minúscula añadiendo una nota explicativa (v. gr. *america* por *America*) o con todas las letras en mayúsculas, Svensén formula: “the basic rule is

³En otro lugar, se detectaron en diccionarios especializados bilingües numerosos ejemplos que revelaban problemas motores de los tipos descritos (algunos de ellos relacionados con el acento ortográfico): *calientamiento*, *inmunidad*, *economía agraria**, *también cambién**, *selección jurisdiccional** (Rodríguez-Rubio Mediavilla, 2021, p. 133, p. 141).

that the lemma should be presented in the graphic form normally used within a sentence in continuous text” (2009, pp. 109-110). La relevancia de las letras iniciales mayúsculas en relación con nombres propios fue expuesta anteriormente en el terreno del procesamiento del lenguaje natural (Ren & Perrault, 1992, pp. 409-410). Lorentzen y Trap-Jensen (2016, p. 140) mencionan la problemática de los cambios ortográficos que, en 2012 y en el contexto danés, afectaron al uso de la mayúscula en siglas como bnp/BNP.

En relación con las anomalías en las que interviene el acento ortográfico, Ren y Perrault realizaron la siguiente clasificación de los errores (coincidente con la tipología general que hemos referido arriba en relación con las erratas o errores ortográficos): *insertion* (éssai por essai), *deletion* (agées por âgées), *substitution* (évènement por événement) y *repositioning* (chomâge por chômage) (1992, p. 410). Los mismos autores también incluyeron en su estudio errores de puntuación que tenían que ver con el uso de los guiones en palabras compuestas o con el genitivo sajón (pp. 410-411).

En lo concerniente al diseño (nuevamente con el usuario en mente), cabe citar las palabras de Levy y Steel:

Users appreciate regular patterns of engagement with the text and this is encouraged through precise and, oftentimes, special use of particular font variants, for example bold for the headword, a smaller font for the definition, and italics for the examples etc. Familiarity and routine are important in design, and of course this principle applies for all dictionaries (2015, p. 179).

3. IMÁGENES PUBLICITARIAS Y MULTIMODALIDAD LEXICOGRÁFICA

En general, la idoneidad de las imágenes que aparecen en los diccionarios puede abordarse desde múltiples prismas, con referencia a los beneficios y utilidad de las imágenes, pero también a la sobrecarga cognitiva (*cognitive load*) que pueden acarrear (Dziemianko, 2022, 2024). Retomando conceptos de épocas en las que aún no se había desarrollado la multimodalidad lexicográfica digital, cabe mencionar también el enfoque de la efectividad (*effectivity*) (Swanepoel, 2001).

Los diccionarios electrónicos (aunque habría que matizar a qué tipo de diccionario electrónico se está aludiendo) permiten a los lexicógrafos pasar de una disposición estática a una dinámica, con una estructura de entradas lexicográficas basada en diferentes capas (Gouws, 2014, p. 164). Ahora bien, el dinamismo puede ser un arma de doble filo en lexicografía, ya que, por ejemplo, una animación puede desorientar al consultante (Lew & Doroszewska, 2009, pp. 253-254). Fuertes-Olivera postula que los datos lexicográficos deben poder ser convertidos en información lexicográfica “in a single cognitive process or click”. A continuación, el autor asegura que esa condición no se cumple en la mayoría de los diccionarios españoles, por mor de defectos en su tratamiento o presentación (2023, p. 31).

Dziemianko (2012, p. 197) alude a la merma en cuanto a la utilidad (*usefulness*), la recepción, producción y aprendizaje (Dziemianko, 2019, p. 5) y el tiempo de consulta (Dziemianko, 2020, p. 394) que los anuncios publicitarios generan en las versiones gratuitas de diccionarios en línea como *Longman Dictionary of Contemporary English* (LDOCE) y *Oxford Advanced Learner's Dictionary* (OALD9). No en todos los estudios la autora analizó todos esos diccionarios y, en ocasiones, los resultados interestudio no fueron concordantes (Dziemianko, 2020, p. 394). Por su parte, Ferrett y Dollinger (2021) manifiestan que los anuncios (algunos de los cuales cambian sin cesar e incluyen efectos llamativos como las animaciones) son una de las principales desventajas de los diccionarios electrónicos respecto de los diccionarios en papel: “It is clear that advertisements, ‘noise’, which are generally required for commercial reasons, make demands on presentation space that run counter to lexicographical needs” (Ferrett & Dollinger, 2021, p. 68, p. 74).

Las palabras de Dziemianko ponen de relieve que la cuestión de las imágenes publicitarias (más concretamente, cómo afectan en términos de uso) no se ha resuelto:

This is a significant research gap, as even top-quality free online dictionaries, including the most prestigious English monolingual dictionaries for foreign learners,

allow targeted advertisements. It should be remembered that online dictionaries are not like many other websites, which may be read for pleasure, joy or broad overviews. They are consulted to satisfy punctual linguistic needs (Tarp 2012: 111), which requires focused attention. Yet, at the same time, advertisements vie for dictionary users’ attention and may affect lexicographic information costs, i.e. the difficulty or inconvenience associated with dictionary lookup (Nielsen 1999: 111) (Dziemianko, 2020, p. 378).

4. MATERIALES Y METODOLOGÍA

4.1. MATERIALES

A los diccionarios de Cambridge que se encuentran en el portal <https://dictionary.cambridge.org/> se los ha llamado “*Cambridge Dictionaries Online*” (Jackson, 2018; Engelberg & Storrer, 2024, p. 56), “*Cambridge English Dictionaries, CaED*” (Geyken & Lemnitzer, 2024, p. 225) o “*Cambridge-Online*” (Engelberg & Storrer, 2024, pp. 49-50). En el presente trabajo, se usa la primera de las denominaciones, manteniendo el nombre inglés en singular y, cuando sea oportuno, la sigla “CDO”. En lugar de como diccionarios individuales, las direcciones “English-Spanish” y “Spanish-English” se consideran partes de un mismo diccionario (*Cambridge Dictionary Online*).

Sobre la base de la tipología de Engelberg y Storrer, se puede considerar el *Cambridge Dictionary Online* como un diccionario general (categoría de usuario *learners’ dictionary*) bilingüe, del tipo portal lexicográfico (híbrido entre un motor de búsqueda y una red de diccionarios o *dictionary network*), además de un portal léxico en virtud de la cobertura de aspectos culturales, enseñanza-aprendizaje de vocabulario, listas temáticas de palabras, juegos y concursos de preguntas, secciones de “palabra del año”, etc. (2024, pp. 31-59). A partir de las nociones *persistence* y *identity* (Herold, Meyer & Wiegand, 2024, pp. 24-26), cabe asociar a los diccionarios de Cambridge con una clara vocación de permanencia.

Las categorías de anomalías estudiadas en el presente trabajo son: “mayúscula”, “puntuación”, “diseño”, “acento ortográfico”, “espacio”, “errata de no

palabra” (ENP) y “errata de palabra” (EP)⁴. Para las ENP y EP, se van a usar las subcategorías sustracción, adición-repetición, sustitución o transposición de una o más letras y, por otro lado, sustracción, adición-repetición, sustitución o transposición de una o más palabras, respectivamente.

Las ubicaciones en las que se encuentran las anomalías del presente estudio son: equivalente, explicación, ejemplo (léase, frase ilustrativa), definición y lema.

4.2. METODOLOGÍA

En el presente estudio, no se ha empleado el término *anomalía* en sentido estricto tal y como suele entenderse en el terreno de las ciencias de la informática, sino general: “[defecto de forma o de funcionamiento](#)”.

Se ha realizado una revisión manual de los datos lexicográficos. Se analizaron las anomalías del texto en español de las entradas correspondientes a la letra “A” de las partes English-Spanish (desde *A & E* hasta *azygous*) y Spanish-English (desde *a* hasta *azuzar*) del [Cambridge Dictionary Online](#).

Cuando aparecieron varias anomalías seguidas, computó cada una de ellas y, si procedía, se desglosaron por tipo. Así, en la entrada “[A-bomb](#)” aparece “Bomba Atomica”, que contiene tres anomalías, dos relativas al uso de la mayúscula y una relativa al acento ortográfico (ver Anexo 1, parte English-Spanish).

Se analizaron solo las palabras en español, porque, de haber incluido también aquellas en inglés, la complejidad habría superado los márgenes del trabajo tal y como se concibió, ya que se habría tenido que incluir un aspecto más (en clave bilingüe) a los previstos: revisión de las anomalías en ambas partes del diccionario, análisis de los tipos generales y subtipos de anomalías y análisis de la ubicación de dichas anomalías. Como excepción a la norma de que solo se analizó el texto en español, cabe decir que las anomalías relativas al diseño sí incluyeron texto en inglés, ya que las implicaciones no

⁴ Como alternativas a las categorías del PLN *non-word errors* y *real-word errors*, se han mantenido las denominaciones en español acuñadas en otro lugar: *errata de no palabra* y *errata de palabra* (Rodríguez-Rubio Mediavilla, 2021).

son meramente lingüísticas. Por ejemplo, la Imagen 1 (captura de pantalla del CDO) revela la repetición del lema *a barrage of something*.

Las anomalías de diseño de la parte Spanish-English implican ausencia de espacio, pero, en realidad se trata de un solapamiento de segmentos o campos textuales, de ahí que se incluyeran en la categoría “diseño”. La Imagen 2 muestra que existen dos tipos de texto diferenciados en el contenido de

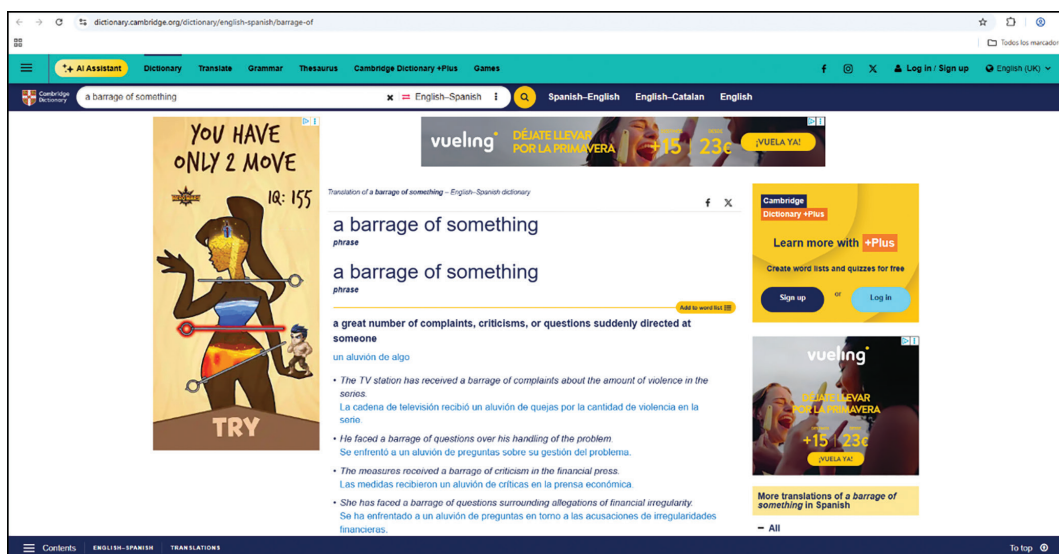


IMAGEN 1: Anomalía de diseño (repetición de lema)

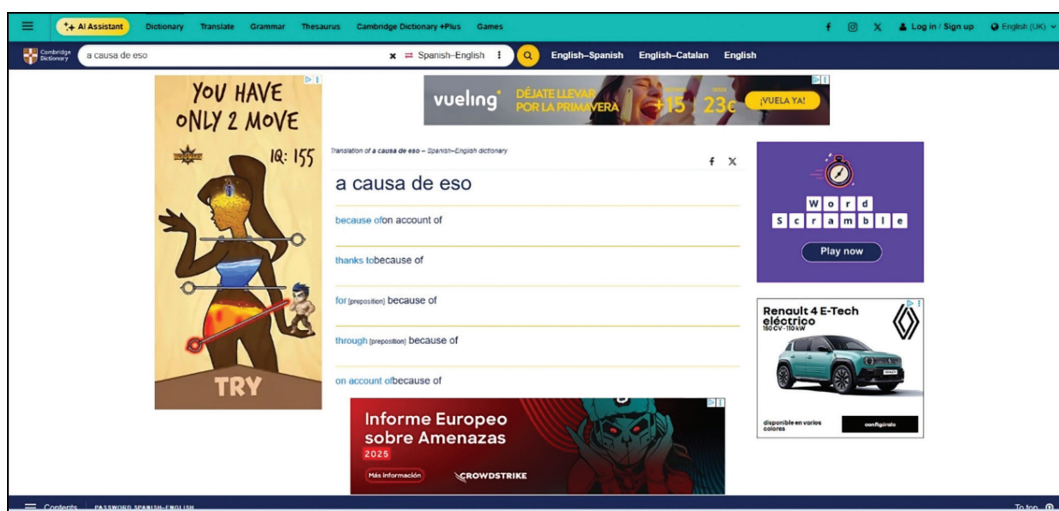


IMAGEN 2: Anomalía de diseño (solapamiento de segmentos textuales)

los equivalentes en inglés de la entrada “a causa de eso”: 1) uno escrito en letra azul, que reza por ejemplo “because of”, clicando en el cual se accede mediante hipervínculo a dicho lema de la parte English-Spanish; y 2) otro, escrito en negro y no vinculado que reza “on account of”. Al seleccionar con el cursor el fragmento “ofon” se observa un solapamiento entre ambos elementos. Los segmentos “thanks tobecause of” y “on account ofbecause of” también presentan el problema, no así “for [preposition] because of” y “through [preposition] because of”. Ver ejemplos similares en el Anexo 1.

En el *Cambridge Dictionary Online*, aparecen imágenes de muy diverso tipo, tanto publicitarias (por ejemplo, anunciando juegos electrónicos) como “lingüísticas” (asociadas a blogs o cuestionarios/juegos). Las publicitarias, ya sean estáticas o animadas, aparecen a menudo de manera errática e intercaladas entre los componentes microestructurales. En la parte Spanish-English en mayor medida, a veces las imágenes no se cargan y no aparecen (por tanto, dejando un vacío que genera una sensación de discontinuidad textual cuando el usuario se mueve hacia arriba y abajo en la página). Otras (si bien pocas), las imágenes centrales (no laterales) no encajan exactamente con el texto escrito, de forma que ambos textos (el escrito y el pictórico) se solapan. En el presente estudio, esos desajustes espaciotemporales no se han considerado como anomalías, pero invitamos a los editores o desarrolladores del diccionario a que revisen este aspecto.

Finalmente, no se ha tratado la problemática del uso de mayúscula o minúscula en función de mudanzas en la norma ortográfica (Lorentzen & Trap-Jensen, 2016, p. 140), sino los casos en los que: 1) un determinado elemento se desvía de la “norma” o costumbre interna del propio *Cambridge Dictionary Online* [en la definición de *aparte*, se indica “Que está separado o es distinto”, vulnerando la norma no escrita pero constatable de forma constante en el diccionario de que las definiciones empiezan con letra inicial minúscula (<https://dictionary.cambridge.org/dictionary/spanish-english/aparte>)]; 2) no se siguen normas asentadas [en la entrada “agent”, figura

“hollywood” con letra inicial minúscula (<https://dictionary.cambridge.org/dictionary/english-spanish/agent>), mientras que en “*Arab*”, aparece “*esta-*dos unidos” en lugar de “Estados Unidos”.

5. RESULTADOS

En la Figura 1, los resultados de incidencia general de anomalías arrojan un número 2,6 veces mayor de anomalías en la parte English-Spanish (771) que en Spanish-English (290). En la parte English-Spanish, sobresalen las categorías de mayúsculas y puntuación (que representan en conjunto un 63,9 % de las anomalías de dicha parte), mientras que en Spanish-English destaca de manera sensiblemente más acusada la categoría diseño (84 %). La proporción de *erratas de palabra* (EP) respecto de las *erratas de no palabra* (ENP) es similar en ambas partes (el número de las EP es 2,1 y 2,5 veces superior al de ENP en la parte English-Spanish y Spanish-English, respectivamente). Las

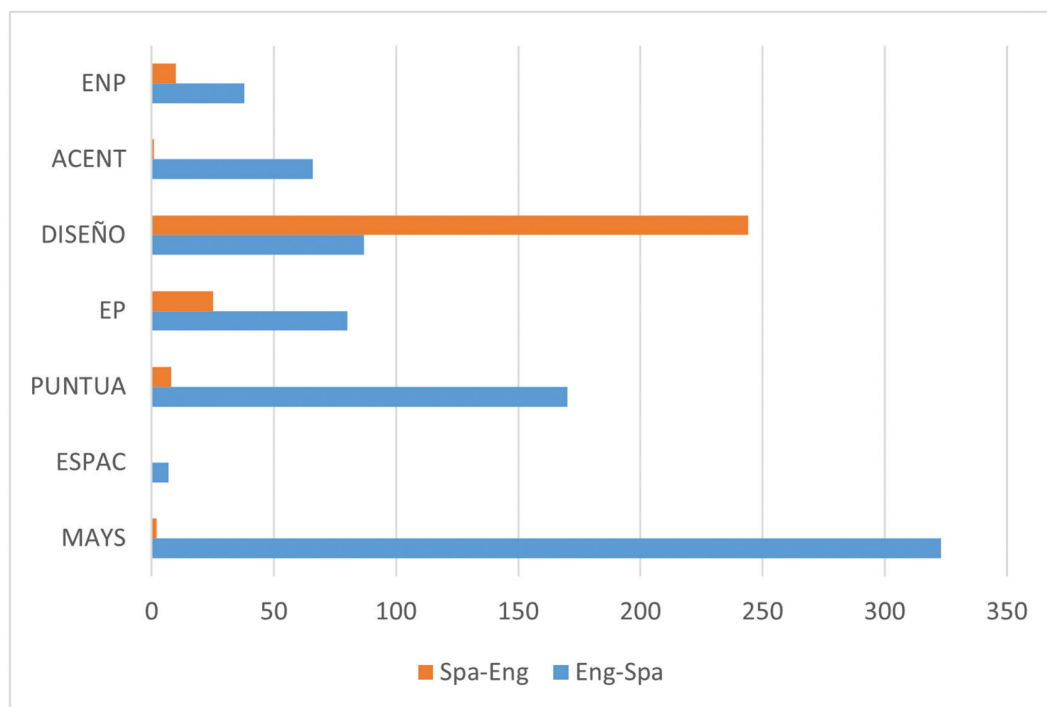


FIGURA 1: Incidencia de anomalías (general)

anomalías relativas al espacio ocupan en ambas partes la última posición (de hecho, en Spanish-English no se han detectado anomalías de este tipo).

Las Figuras 2 y 3 muestran una distribución cualitativamente homogénea de los subtipos de errata, con los siguientes resultados de mayor a menor incidencia en ambas partes: sustitución, sustracción, adición-repetición y transposición en cuanto a las *erratas de palabra* y, por otro lado, adición-repetición, sustracción, sustitución y transposición en lo que se refiere a las *erratas de no palabra*. El desglose en función de las partes revela una cifra parecida a la del dato de incidencia general, con un número agregado de EP y ENP 3,2 veces mayor en la parte English-Spanish (113) en comparación con Spanish-English (35).

En términos de ubicación, las Figuras 4 y 5 expresan claramente que la mayoría de las anomalías se sitúan en el campo del “equivalente”: 438 (56,8 %) en la parte English-Spanish y, de manera más acusada, 244 (84,1 %) en Spanish-English. En el suelo se sitúan las anomalías situadas en el campo “lema”, con 87 (11,3 %) en English-Spanish y un testimonial 1 (0,3 %) en Spanish-English. El hecho de haber

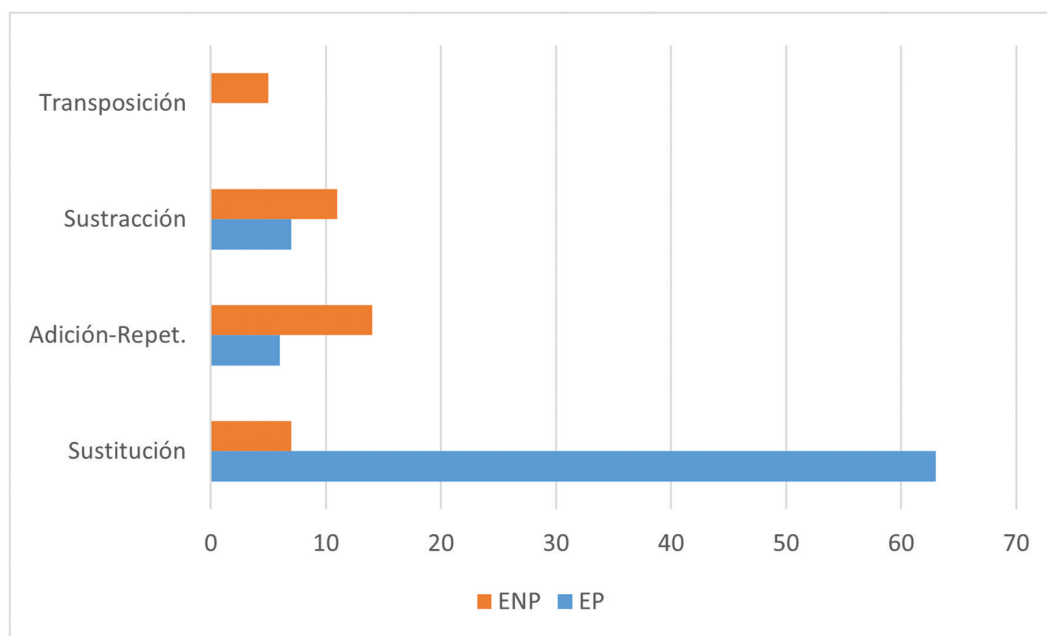


FIGURA 2: Incidencia de subtipos de errata (English-Spanish)

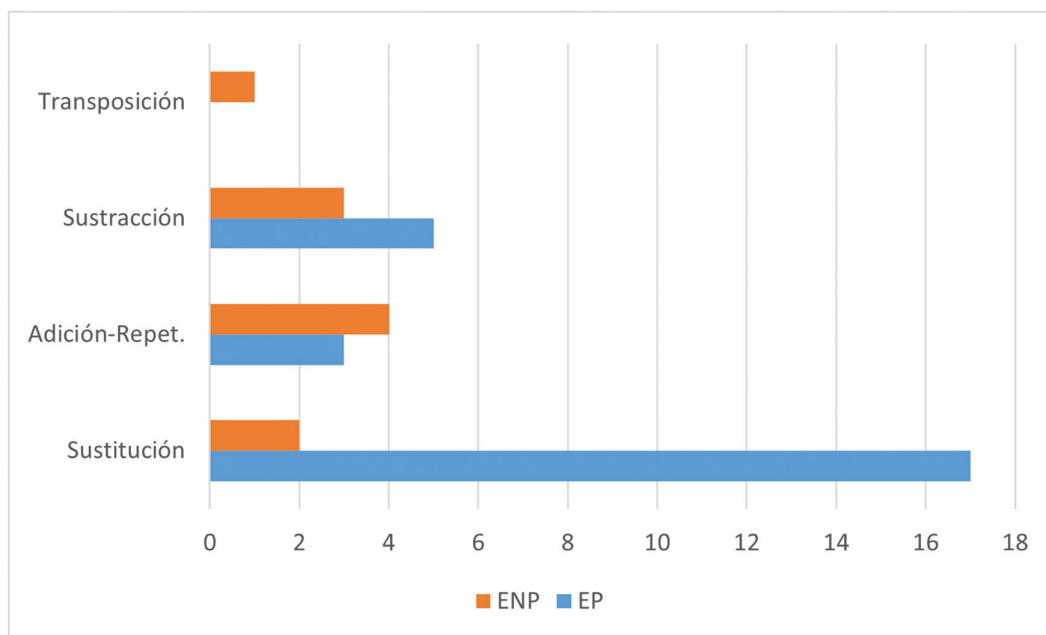


FIGURA 3: *Incidencia de subtipos de errata (Spanish-English)*

analizado el texto en español, pero no en inglés, hace que en English-Spanish no se hayan registrado anomalías en la ubicación “definición”. Por último, en English-Spanish, se observaron anomalías en todas las ubicaciones, pero en Spanish-English la ubicación “lema” es, como acabamos de explicitar, testimonial. El análisis desgranado de los elementos más salientes de las Figuras 4 y 5 muestra que, a menudo, no se trata de una acumulación de defectos inconexos. Así, en la parte English-Spanish:

- a) de las 323 anomalías de mayúsculas, 315 (97,5 %) responden al mismo problema, consistente en haber escrito con letra inicial mayúscula los equivalentes que, según la norma no escrita del propio CDO, deben empezar con minúscula;
- b) de las 129 anomalías en explicaciones, 128 (99,2 %) consisten en la indicación de dos puntos (:) y una coma (,) seguidos. Por ejemplo, en “forma abreviada de ‘also known as’; alias” de la entrada “[aka](#)”;
- c) las 87 (100 %) anomalías de diseño consisten en la repetición del lema de referencia.

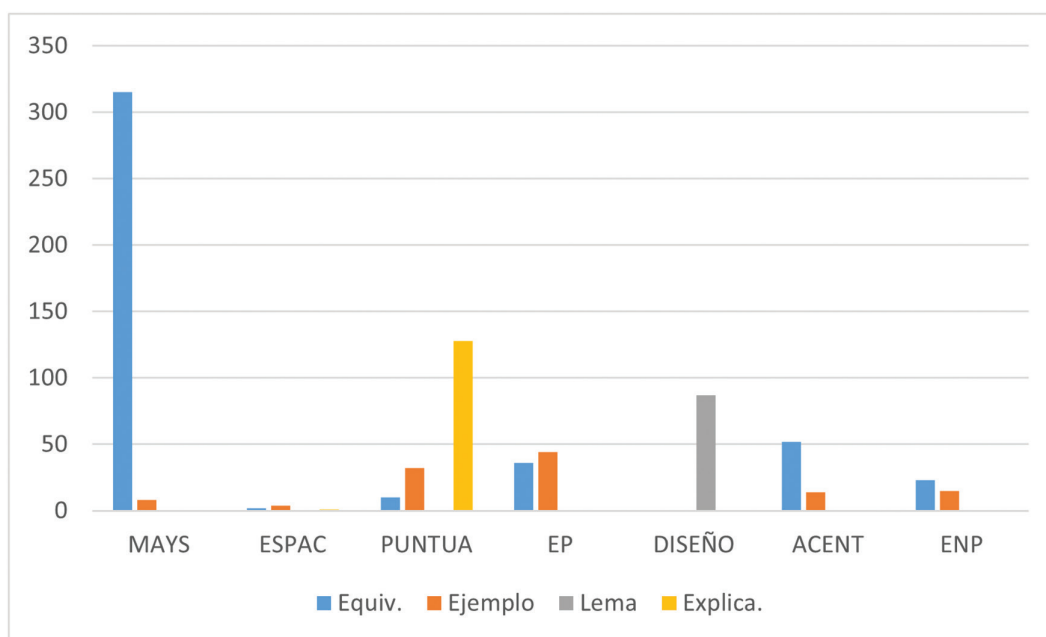


FIGURA 4: Ubicación de las anomalías (English-Spanish)

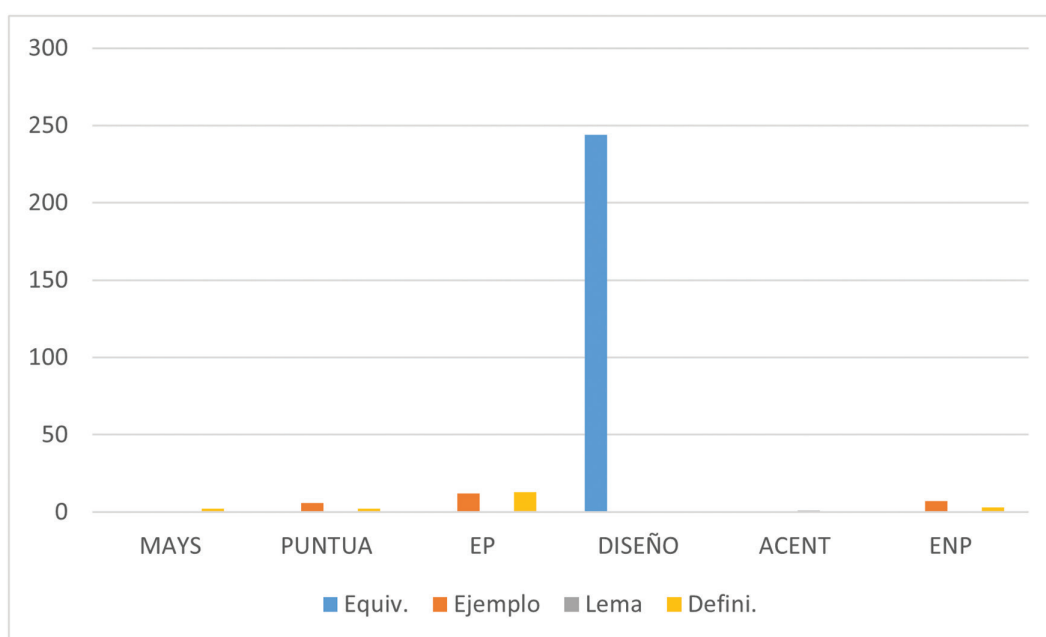


FIGURA 5: Ubicación de las anomalías (Spanish-English)

De manera similar, en la parte Spanish-English, las 244 anomalías de diseño revelan el mismo tipo de solapamiento de segmentos textuales explicado arriba.

En los resultados del presente estudio no podían faltar ejemplos de erratas causadas por fenómenos motores (dada la universalidad de estos últimos). La Tabla 1 contiene un ejemplo de anticipación y otro de perseveración (todos interpalabra) en cada una de las partes del CDO.

	Anticipación	Perseveración
English-Spanish	<i>terreno utilizad<u>a</u>* para la agricul- tura</i> (ejemplo de la entrada “agri- culture”: https://dictionary.cambridge.org/dictionary/english-spanish/agriculture)	<i>le enfad<u>ó</u> much<u>ó</u>*</i> (ejemplo de la entrada “angry”, sección “TRANSLATION of angry ”: https://dictionary.cambridge.org/dictionary/english-spanish/angry)
Spanish-English	<i>negocios<u>os</u>* sucios</i> (ejemplo de la acepción “desper- tar sospecha de ilegalidad” de la entrada “apestar”: https://dictio- nary.cambridge.org/dictionary/ spanish-english/apestar)	<i>una casa de campo de* muy apa- cible</i> (ejemplo de la acepción “agradable, sereno y tranquilo” de la entrada “apacible”: https://dictionary.cambridge.org/dictionary/spani- sh-english/apacible)

TABLA 1: Ejemplos de anticipación y perseveración (Eng-Spa/Spa-Eng)

Ver en el Anexo 1 otros ejemplos de anticipación o perseveración intra o interpalabra, ya sea en English-Spanish o en Spanish-English.

En clave bilingüe, se registraron aparentes interferencias, como:

- *adjectivo* (por *adjetivo*, posible influencia del vocablo inglés *adjective*), en la entrada “adj”.
- *objetivo* (por *objetivo*, posible influencia de *objective*) en la frase ilustrativa *comenzó con el objetivo* de acabar con la crisis* de la entrada “aim”.

Si bien no hemos aplicado un enfoque semántico en este estudio, cabe mencionar un ejemplo de doble negación, un fenómeno universal amplia-
mente observado en contextos de aprendizaje y comunicación tanto cuando
interviene alguna lengua extranjera como cuando nos limitamos a la propia
lengua materna. En las entradas del CDO “antiporn” y “antipornography”
encontramos la formulación “antipornografía, contra la antipornografía*”, que
debería rezar: “antipornografía, contra la pornografía”.

Además, aunque no se realizó un estudio de la posible reproducción de
las anomalías en varias entradas (o en el seno de una determinada entrada)

del diccionario, podemos mencionar que la formulación “con interés en lo artesana*” (por “con interés en lo artesano”) aparece en las entradas vinculadas “[arty-crafty](#)” y “[artsy-craftsy](#)”. Otros casos de aparente reproducción, como la ausencia de espacio en el mismo contenido de las entradas relacionadas “a gatas” y “a cuatro patas” (*on all fourson hands and knees*) o “a voz en cuello” y “a voz en grito” (*at the top of one’s voicevery loudly*) responden en realidad a anomalías de diseño (como se ha explicado en el apartado “Métodos”).

Entre los ejemplos de palabras que deberían tener letra inicial mayúscula en lugar de minúscula, al objeto de diferenciar la palabra del correspondiente nombre propio, se encuentran *Clara* (por *clara*, en el lema [albumen](#)) y *Angustia* (por *angustia*, en el lema [angst](#)).

6. DISCUSIÓN

En el presente estudio, se ha ido más allá de la (beneficiosa y seguramente necesaria) colaboración o interacción de los usuarios de diccionarios explicitada en Klosa-Kückelhaus y Tiberius (2024, pp. 71-73), para presentar un conjunto representativo de anomalías del *Cambridge Dictionary Online* (partes English-Spanish y Spanish-English). Se ha demostrado así la vigencia de la lucha contra las anomalías en los diccionarios electrónicos. Aunque en el presente trabajo solo se han analizado anomalías formales del texto en español, en un futuro podrían abordarse aspectos de trascendencia semántica (por ejemplo, desviaciones o dobles negaciones) o, en clave bilingüe, que tengan que ver con la traducción. A su vez, convendría explorar la posible reproducción de anomalías en diferentes contenidos del diccionario; aspecto, el de la reproducción de anomalías, que ya se abordó en otro lugar en relación con plantillas lexicográficas y erratas de diccionarios especializados bilingües en papel (Rodríguez-Rubio & Fernández-Quesada, 2020a).

Se desconoce el método de introducción y revisión de contenido lexicográfico que se usa en el caso del *Cambridge Dictionary Online*. Se trata

de un aspecto relevante, pues algunos *modus operandi* pueden facilitar la tarea de revisión, mientras que otros pueden dificultarla. Como se ha aclarado arriba, algunos modelos de edición pueden ser más propensos al error cuando se lleva a cabo una edición manual (Simões & Salgado, 2022, p. 425). Cabe suponer que algunos de los errores lingüísticos presentes en los diccionarios electrónicos pueden deberse precisamente a la edición manual.

A su vez, sería interesante explorar la utilidad de las funciones de revisión ortográfica y gramatical de los programas de gestión lexicográfica (DWS) de los diccionarios en línea y si los errores o anomalías están claramente marcados para que el lexicógrafo pueda localizarlos con relativa facilidad a nivel más profundo, pero también mediante una ojeada a la vista WYSISWYG (*what-you-see-is-what-you-get*). Más específicamente, como se sugirió en otro lugar, sería conveniente averiguar cómo consigue el corrector ortográfico diferenciar la granularidad de un texto bilingüe en microestructuras complejas, donde intervienen dos lenguas de revisión (Rodríguez-Rubio & Fernández-Quesada, 2020b, pp. 397-398). Los diccionarios electrónicos pueden incluir funciones de edición que sirven para detectar anomalías, pero sigue quedando una parte de la tarea que ha de realizar el revisor humano. A partir de aquí, una derivada a investigar en futuros estudios sería determinar con qué periodicidad y nivel de detalle se revisan los diccionarios electrónicos, aspecto que planteamos hace unos años en un estado de la cuestión (meta)lexicográfica (Fernández-Quesada & Rodríguez-Rubio, 2022). Esa pretensión se toparía, no obstante, con el impedimento de que, como sugieren de manera constante y extendida en el tiempo los investigadores, los editores lexicográficos tradicionalmente han brindado escasa información sobre sus procesos internos (Swanepoel, 2001, p. 170; Klosa-Kückelhaus & Tiberius, 2024, p. 62). Esa parca información hace, a su vez, que no se sepa en qué medida los ejemplos del *Cambridge Dictionary Online* son inventados o reutilizados.

Existen numerosos manuales de Cambridge, entre otros, un *Handbook 2025-Regulations and guidance for administering Cambridge exams* y una

Editorial Services Style Guide for Academic Books orientado a autores. Pero se echa en falta una guía de uso del *Cambridge Dictionary Online* en consonancia con un recurso de consulta académica de primer orden con vocación de permanencia. Tal vez la megaestructura se asocie en mayor medida a los diccionarios en papel, pero componentes como “advertencias de uso”, “artículos de muestra” o aquellos que expliquen el manejo de la obra también son útiles en el marco de la lexicografía electrónica (Ramírez Sánchez, 2023, p. 850). Esas consideraciones coinciden parcialmente con las de Klosa-Kückelhaus y Tiberius, quienes incluyen el manual de usuario (o *guided tour*) entre los elementos integrantes del proceso de elaboración de diccionarios digitales (*Internet dictionaries*) (2024, p. 67).

Partiendo del desconocimiento del número de editores de los diccionarios de Cambridge, la afirmación de que cuando hay pocos editores las entradas lexicográficas deben someterse a una rigurosa revisión antes de su publicación (Herold, Meyer & Wiegand, 2024) debe aplicarse *a fortiori* a los casos en los que hay más editores.

Como se ha explicado arriba, las imágenes publicitarias pueden provocar ruido. Sin embargo, los anuncios pueden contribuir decididamente a la continuidad de los recursos lexicográficos de acceso gratuito. En el caso del *Cambridge Dictionary Online*, muchas de las imágenes publicitarias móviles o animadas incluyen figuras (por ejemplo, personajes de juegos *online*) y colores en extremo llamativos, lo que puede, en efecto, desorientar o incluso molestar al usuario. Además, esas imágenes van cambiando sin cesar y su tiempo de carga produce a menudo un desacople temporal entre el momento (normalmente breve) de la consulta del usuario y la aparición de la imagen, de manera que quedan espacios aparentemente en blanco intercalados en el contenido lexicográfico de numerosas entradas (especialmente en la parte Spanish-English). A pesar de que se descartó computar como anomalías dichos espacios vacíos, y a pesar de que propiamente hablando esas imágenes no pueden considerarse datos lexicográficos de todas formas, conviene

apuntar que las disfunciones que provoca su comportamiento errático no concuerdan con la lógica de ausencia de defectos en el tratamiento o la presentación que Fuertes-Olivera aplica a los datos lexicográficos (2023, p. 31).

7. CONCLUSIONES

La proporción de 2,1 veces *erratas de palabra* (80) frente a *erratas de no palabra* (38) en la parte English-Spanish y de 2,5 veces *erratas de palabra* (25) frente a *erratas de no palabra* (10) en la parte Spanish-English concuerda con la idea general expresada en la literatura de manera sostenida en el tiempo de que los *real-word errors* son difíciles de detectar o corregir. A efectos prácticos, la distribución cualitativamente homogénea de los subtipos de errata puede servir de guía con vistas a una hipotética revisión del *Cambridge Dictionary Online* en su totalidad. Así, en ambas partes del diccionario, se ha podido diferenciar entre el “comportamiento” de las *erratas de palabra* (con una clara prevalencia de las sustituciones, seguidas de las sustracciones y adiciones-repeticiones) y el de las *erratas de no palabra* (mucho más repartidas, pero siempre siguiendo el mismo orden de adiciones-repeticiones, sustracciones, sustituciones y transposiciones).

Más interesantes para los editores o lexicógrafos del CDO sean tal vez las conclusiones respecto de la ubicación de las anomalías de la letra “A”, pues a partir de ellas se puede prever en cierta medida dónde se situarán las anomalías del resto de letras; máxime cuando se ha observado una coincidencia en las causas de los defectos de mayúscula, puntuación y diseño en la parte English-Spanish y de diseño en Spanish-English. Una sistematicidad de anomalías precisa de un enfoque integral que incluya una revisión y unas medidas correctivas acordes. Nuevamente, los resultados pueden dirigir u ordenar la revisión que se lleve a cabo.

Los avances tecnológicos ofrecen potentes recursos de gestión lexicográfica, pero encierran la necesidad de una creciente atención por parte del lexicógrafo/editor que atienda a la mayor complejidad de los procesos y

a la multiplicación, expansión e interconexión de recursos. En el campo de la revisión y de la calidad de los datos (como en tantos otros) se ha de encontrar un equilibrio entre productividad (léase, cantidad) y calidad, entre la posibilidad de crear productos lexicográficos avanzados y complejos y la pertinencia de realizar un trabajo fino de cohesión y corrección tanto en el fondo como en la forma. La posibilidad de modificar en tiempo real el contenido lexicográfico exige una revisión dinámica y flexible en consonancia, al tiempo que detallista y rigurosa.

Pese a la existencia de un marco tecnológico que ofrece medios avanzados para promover la calidad de los datos, el papel del revisor humano continúa siendo fundamental. En línea con lo manifestado en otro lugar (Fernández-Quesada & Rodríguez-Rubio, 2022, p. 24) la evidencia de que la tarea de revisión de los diccionarios es un proceso inacabable no debe ser óbice sino acicate para que se sigan realizando revisiones periódicas de las grandes obras de referencia. En el plano académico, los resultados del presente estudio en un diccionario en línea, sumados a los de nuestras anteriores investigaciones sobre erratas en diccionarios en papel, muestran claramente la pertinencia de seguir profundizando en la búsqueda de la corrección formal lexicográfica. Tal vez, de hecho, la alianza entre el mundo profesional y el académico pueda alumbrar una oportunidad de conseguir obras o recursos más afinados en una disciplina milenaria como la lexicografía que aún sigue en pie a pesar de las dificultades.

REFERENCIAS BIBLIOGRÁFICAS

- Abel, A. (2012). Dictionary writing systems and beyond. En S. Granger & M. Paquot (eds.), *Electronic Lexicography* (pp. 83-106). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199654864.003.0005>
- Abel, A. & Meyer, M. (2024). User Participation. En A. Klosa-Kückelhaus (ed.), *Internet Lexicography* (pp. 227-260). De Gruyter. <https://doi.org/10.1515/9783111233758-009>
- Adams, N. & Maxwell, M. (2016). Somali Spelling Corrector and Morphological Analyzer. En C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell & E. Wandl-Vogt (eds.), *Proceedings of the LREC 2016 Workshop “CCURL 2016 – Towards an Alliance for Digital Language Diversity”* (pp. 22-29).

- Barbierik, K., Děngeová, Z., Holcová, M., Jarý, V., Liška, T., Lišková, M. & Virius, M. (2014). Simple and Effective User Interface for the Dictionary Writing System. En A. Abel, C. Vettori & N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen, Italy* (pp. 125-136). EURAC research.
- Blecua, A. (1983). *Manual de crítica textual*. Editorial Castalia.
- Bloodgood, M. & Strauss, B. (2016). Data Cleaning for XML Electronic Dictionaries via Statistical Anomaly Detection. En *Proceedings of the IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2016* (pp. 79-86). <https://doi.org/10.1109/ICSC.2016.38>
- Coward, D. F. & Grimes, Ch. E. (2000). *Making Dictionaries: A Guide to Lexicography and the Multi-Dictionary Formatter*. SIL International.
- Dashti, S. M., Bardsiri, A. K. & Bardsiri, V. Kh. (2018). Correcting real-word spelling errors: A new hybrid approach. *Digital Scholarship in the Humanities*, 33(3), 488-499. <https://doi.org/10.1093/llc/fqx054>
- Dziemianko, A. (2012). Why One and Two Do Not Make Three: Dictionary Form Revisited. *Lexikos*, 22, 195-216. <https://doi.org/10.5788/22-1-1003>
- Dziemianko, A. (2019). The role of online dictionary advertisements in language reception, production, and retention. *ReCALL*, 31(1), 5-22. <https://doi.org/10.1017/S0958344018000149>
- Dziemianko, A. (2020). Smart Advertising and Online Dictionary Usefulness. *International Journal of Lexicography*, 33(4), 377-403. <https://doi.org/10.1093/ijl/ecaa017>
- Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL*, 34(2), 218-234. <https://doi.org/10.1017/S0958344021000264>
- Dziemianko, A. (2024). Pictures in Online Dictionaries: Shall We See Them? *GEMA Online - Journal of Language Studies*, 24(2), 31-53. <http://doi.org/10.17576/gema-2024-2402-03>
- Engelberg, S. & Storrer, A. (2024). A Typology of Internet Dictionaries and Portals. En A. Klosa-Kückelhaus (ed.), *Internet Lexicography* (pp. 31-59). De Gruyter. <https://doi.org/10.1515/9783111233758-003>
- Fernández-Quesada, N. & Rodríguez-Rubio, S. (2022). Estudio preliminar. El tratamiento del error textual y de la errata en la era digital: elogio de la corrección. En N. Fernández-Quesada & S. Rodríguez-Rubio (eds.), *Detección y tratamiento de errores y erratas: Un diagnóstico para el siglo XXI* (pp. 13-28). Dykinson. <https://doi.org/10.2307/j.ctv2gz3s5v.3>
- Ferrett, E. & Dollinger, S. (2021). Is Digital Always Better? Comparing Two English Print Dictionaries with Their Digital Counterparts. *International Journal of Lexicography*, 34(1), 66-91. <https://doi.org/10.1093/ijl/ecaa016>
- Foorthuis, R. (2021). On the nature and types of anomalies: a review of deviations in data. *International Journal of Data Science and Analytics*, 12, 297-331. <https://doi.org/10.1007/s41060-021-00265-1>

- Fossati, D. & Di Eugenio, B. (2008). I saw TREE trees in the park: How to Correct Real-Word Spelling Mistakes. En *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco (pp. 896-901). <https://doi.org/10.63317/4qig6sasadgh>
- Fuertes-Olivera, P. A. (2023). Structuring a Collection of Lexicographic Data for Different User and Usage Situations. *Lexikos*, 33(2), 22-42. <https://doi.org/10.5788/33-2-1832>
- Gardner, S. A. (1992). Spelling Errors in Online Databases: What the Technical Communicator Should Know. *Technical Communication*, 39(1), 50-53.
- Geyken, A. & Lemnitzer, L. (2024). The Automatic Extraction of Lexicographic Information. En A. Klosa-Kückelhaus (ed.), *Internet Lexicography* (pp. 191-226). De Gruyter. <https://doi.org/10.1515/9783111233758-008>
- Gouws, R. (2014). Article Structures: Moving from Printed to e-Dictionaries. *Lexikos*, 24, 155-177. <https://doi.org/10.5788/24-1-1256>
- Grudin, J. T. (1983). Error Patterns in Novice and Skilled Transcription Typing. En W. E. Cooper (ed.), *Cognitive Aspects of Skilled Typewriting* (pp. 121-143). Springer. https://doi.org/10.1007/978-1-4612-5470-6_6
- Gujral, E. (2023). *Survey: Anomaly Detection Methods*. Department of Computer Science and Engineering University of California Riverside (pp. 1-16). <https://www.semanticscholar.org/paper/Survey:-Anomaly-Detection-Methods-Gujral/75b78a54132e193b5a-c528645045e9ccbb98395d>
- Herold, A., Meyer, P. & Wiegand, F. (2024). The Technological Context for Internet Lexicography. En A. Klosa-Kückelhaus (ed.), *Internet Lexicography* (pp. 5-29). De Gruyter. <https://doi.org/10.1515/9783111233758-002>
- Jackson, H. (2018). English Lexicography in the Internet Era. En P. A. Fuertes-Olivera (ed.), *The Routledge Handbook of Lexicography* (pp. 835-856). Routledge. <https://doi.org/10.4324/9781315104942-34>
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). Lexicographic practices in Europe: A survey of user needs. *ELEXIS – European Lexicographic Infrastructure*, Deliverable D1.1. <https://bit.ly/2WJwSlC>
- Klosa-Kückelhaus, A. & Tiberius, C. (2024). The Lexicographic Process. En A. Klosa-Kückelhaus (ed.), *Internet Lexicography* (pp. 61-96). De Gruyter. <https://doi.org/10.1515/9783111233758-004>
- Koeva, S. & Blagoeva, D. (2013). The Dictionary Writing System Lexit and its Application in Bilingual Lexicography. *Cognitive Studies*, 13, 57-65. <https://doi.org/10.11649/cs.2013.003>
- Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2019). Identification and Automatic Extraction of Good Dictionary Examples: The Case(s) of GDEX. *International Journal of Lexicography*, 32(2), 119-137. <https://doi.org/10.1093/ijl/ecy014>
- Landau, S. I. (2001). *Dictionaries. The Art and Craft of Lexicography*. Cambridge University Press.
- Lashley, K. S. (1951). The Problem of Serial Order in Behavior. En L. A. Jeffress (ed.), *Cerebral mechanisms in behavior* (pp. 112-146). Wiley.

- Leroyer, P. (2018). The Oenolex Wine Dictionary. En P. A. Fuertes-Olivera (ed.), *The Routledge Handbook of Lexicography* (pp. 438-454). Routledge. <https://doi.org/10.4324/9781315104942-28>
- Levy, M. & Steel, C. (2015). Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL*, 27(2), 177-196. <https://doi.org/10.1017/S095834401400038X>
- Lew, R. & De Schryver, G. M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, 27(4), 341-359. <https://doi.org/10.1093/ijl/ecu011>
- Lew, R. & Doroszewska, J. (2009). Electronic Dictionary Entries with Animated Pictures: Lookup Preferences and Word Retention. *International Journal of Lexicography*, 22, 239-257. <https://doi.org/10.1093/ijl/ecp022>
- Liu, Y., Zhao, R., Altinger, L., Schütze, H. & Hedderich, M. A. (2025). Evaluating Robustness of Large Language Models Against Multilingual Typographical Errors. *arXiv*, 2510.09536. <https://arxiv.org/abs/2510.09536v2>
- Logan, F. A. (1999). Errors in Copy Typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1760-1773. <https://doi.org/10.1037//0096-1523.25.6.1760>
- Lonke, D., Kernerman, I. & Dzhuranyuk, V. (2022). Lexical Data API. En A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs & P. Storzjohann (eds.), *Dictionaries and Society: Proceedings of the XX EURALEX International Congress, Mannheim, Germany* (pp. 401-408). IDS-Verlag.
- Lorentzen, H. & Trap-Jensen, L. (2016). What, When and How? – the Art of Updating an Online Dictionary. En T. Margalitadze & G. Meladze (eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 138-145). Tbilisi University Press.
- Peterson, J. L. (1986). A Note on Undetected Typing Errors. *Communications of the ACM*, 29(7), 633-637. <https://doi.org/10.1145/6138.6146>
- Ramírez Sánchez, I. (2023). Anatomía del diccionario: Análisis descriptivo y propuesta terminológica de las estructuras de las obras lexicográficas. *Boletín de la Real Academia Española (BRAE)* 103(328), 841-861. <https://revistas.rae.es/brae/article/view/622>
- Ren, X. & Perrault, F. (1992). The Typology of Unknown Words: An Experimental Study of Two Corpora. En Ch. Boitet (ed.), *COLING '92 Proceedings of the 14th Conference on Computational Linguistics, Vol. 1* (pp. 408-414). GETA (IMAG) & Association Champollion. <https://doi.org/10.3115/992066.992132>
- Rodrigues, P., Zajic, D., Doermann, D., Bloodgood, M. & Ye, P. (2011). Detecting Structural Irregularity in Electronic Dictionaries Using Language Modeling. En I. Kosem & K. Kosem (coords.), *New Applications for New Users: Proceedings of eLex 2011* (pp. 227-232).
- Rodríguez-Rubio, S. & Fernández-Quesada, N. (2020a). The Dynamics of Typographical Error Reproduction: Optimising Formal Correctness in Three Specialised Bilingual Dictionaries. *ELIA*, 20, 147-190. <https://doi.org/10.12795/elia.2020.i20.06>

- Rodríguez-Rubio, S. & Fernández-Quesada, N. (2020b). Towards Accuracy: A Model for the Analysis of Typographical Errors in Specialised Bilingual Dictionaries. Two Case Studies. *Lexikos*, 30, 386-415. <https://doi.org/10.5788/30-1-1606>
- Rodríguez-Rubio, S. & Fernández-Quesada, N. (2021). El tratamiento de las erratas desde la sicolingüística y desde el procesamiento del lenguaje natural. Un estado de la cuestión. *Tonos Digital*, 40, 1-19.
- Rodríguez-Rubio Mediavilla, S. (2021). *Estudio comparativo de erratas en diccionarios especializados inglés-español en papel de los ámbitos económico-financiero, jurídico e industrial. Una propuesta metodológica* [Tesis Doctoral no publicada]. Universidad Pablo de Olavide.
- Rosenblum, B. & Golfman, I. (2004). Automated Quality Assurance for Heuristic-Based XML Creation Systems. *Proceedings of the Extreme Markup Languages® 2004 Conference, Montréal, Quebec, Canada* (pp. 1-11). <https://www.semanticscholar.org/paper/Automated-Quality-Assurance-for-Heuristic-Based-XML-Rosenblum-Golfman/91948d5b64aee5a91f008249e74066f20ff27b22>
- Samanta, P. & Chaudhuri, B. B. (2013). A simple real-word error detection and correction using local word bigram and trigram. En *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013), Kaohsiung, Taiwan* (pp. 211-220). Association for Computational Linguistics and Chinese Language Processing (ACLCLP). <https://aclanthology.org/O13-1022/>
- Sharma, S. & 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>
- Simões, A. & Salgado, A. (2022). Smart Dictionary Editing with LeXmart. En A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs & P. Storjohann (eds.), *Dictionaries and Society: Proceedings of the XX EURALEX International Congress, Mannheim, Germany* (pp. 423-434). IDS-Verlag.
- Snyder, K. M. & Logan, G. D. (2014). The Problem of Serial Order in Skilled Typing. *Journal of Experimental Psychology: Human Perception and Performance*, 40(4), 1697-1717. <https://doi.org/10.1037/a0037199>
- Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge University Press.
- Swanepoel, P. (2001). Dictionary Quality and Dictionary Design: A Methodology for Improving the Functional Quality of Dictionaries. *Lexikos*, 11, 160-190. <https://doi.org/10.5788/11-0-846>
- Wheatley, H. B. (1893). *Literary Blunders: A Chapter in the "History of Human Error"*. Elliot Stock.
- Zajic, D., Maxwell, M., Doermann, D., Rodrigues, P. & Bloodgood, M. (2011). Correcting Errors in Digital Lexicographic Resources Using a Dictionary Manipulation Language. En I. Kosem & K. Kosem (coords.), *New Applications for New Users: Proceedings of eLex 2011* (pp. 297-301). Institute for Applied Slovene Studies.

ANEXO 1

Muestra de anomalías de las partes English-Spanish y Spanish-English del *Cambridge Dictionary Online*.

ENGLISH-SPANISH

Dice	Debería decir	Ubicación	Observaciones	Enlace
MAYÚSCULA				
Bomba Atómica	bomba atómica	Equivalente		https://dictionary.cambridge.org/dictionary/english-spanish/a-bomb
Lluvia Acida	lluvia ácida	Equivalente		https://dictionary.cambridge.org/dictionary/english-spanish/acid-rain
SIGNOS DE PUNTUACIÓN				
Es una pianista consumada., Es una pintora consumada., Es una experta jinete.	Es una pianista consumada. Es una pintora consumada. Es una experta jinete.	Ejemplo	Sobra la coma	https://dictionary.cambridge.org/dictionary/english-spanish/accomplished
acumular, acumularse acrecentar	acumular, acumularse, acrecentar	Equivalente	Falta la coma	https://dictionary.cambridge.org/dictionary/english-spanish/accrete
DISEÑO				
ablutions <i>phrase</i> ablutions <i>phrase</i>	ablutions <i>phrase</i>	Lema	Repetición de lema	https://dictionary.cambridge.org/dictionary/english-spanish/ablutions
ERRATA DE PALABRA (EP)				
El juez despejó la sala de todos menos ella misma y los testigos	La jueza despejó la sala de todos menos ella misma y el testigo	Ejemplo	Sustitución. En la frase inglesa original, se indica que es una jueza y se usa la palabra <i>witness</i> (no <i>witnesses</i>)	https://dictionary.cambridge.org/dictionary/english-spanish/all
Aunque que	aunque	Equivalente	Adición-repetición. Aparente perseveración del fragmento “que”	https://dictionary.cambridge.org/dictionary/english-spanish/albeit

ERRATA DE NO PALABRA (ENP)				
¡te sigo quie- rendo!	¡te sigo que- riendo!	Ejemplo	Transposición. Se han inter- cambiado los fragmentos “e” e “ie” de la palabra “que- riendo”	https://dictionary.cambridge.org/dictionary/english-spanish/as-bald-as-a-coot
Cabio* Dras- tico	cambio drás- tico	Equivalente	Sustracción.	https://dictionary.cambridge.org/dictionary/english-spanish/about-face
ACENTO ORTOGRÁFICO				
Agnostico	agnóstico	Equivalente	Esta anomalía aparece dos veces seguidas en la sección TRANSLATION of agnostic	https://dictionary.cambridge.org/dictionary/english-spanish/agnostic
Árbitro	árbitro	Equivalente	Esta anomalía aparece dos veces seguidas en la sección TRANSLATION of arbiter	https://dictionary.cambridge.org/dictionary/english-spanish/arbiter
ESPACIO				
por los viajes y las aventura s	por los viajes y las aventu- ras	Ejemplo	Sobra espacio entre la “a” y la “s”	https://dictionary.cambridge.org/dictionary/english-spanish/adventure
el sonido den- tal inicial [t]se pierde	el sonido den- tal inicial [t] se pierde	Ejemplo	Falta espacio entre “[t]” y “se”	https://dictionary.cambridge.org/dictionary/english-spanish/assimilate

SPANISH-ENGLISH

Dice	Debería decir	Ubicación	Observaciones	Enlace
DISEÑO				
on all four- son hands and knees	on all fours on hands and knees	Equiva- lente	Solapamiento de segmentos textuales	https://dictionary.cambridge.org/dictionary/spanish-english/a-cuatro-patas
on all four- son hands and knees	on all fours on hands and knees	Equiva- lente	Solapamiento de segmentos textuales	https://dictionary.cambridge.org/dictionary/spanish-english/a-gatas

ERRATA DE PALABRA (EP)				
<i>una casa de campo de*</i> <i>muy apacible</i>	<i>una casa de campo muy apacible</i>	Ejemplo	Adición-repetición. Aparente perseveración de “de”	https://dictionary.cambridge.org/dictionary/spanish-english/apacible
<i>E* madre asintió orgulloso</i>	<i>La madre asintió orgulloso</i>	Ejemplo	Sustitución. Discordancia de género	https://dictionary.cambridge.org/dictionary/spanish-english/asentir
ERRATA DE NO PALABRA (ENP)				
en una amplia e incontralada* abertura	en una amplia e incontrolada abertura	Ejemplo	Sustitución. Aparente anticipación de la letra “a”	https://dictionary.cambridge.org/dictionary/spanish-english/abertura
Se despide attentamente*	Se despide atentamente	Ejemplo	Adición-repetición. Posible interferencia de la voz inglesa <i>attentively</i>	https://dictionary.cambridge.org/dictionary/spanish-english/atentamente
SIGNOS DE PUNTUACIÓN				
Redactar un acta o documento oficial en por el que se certifica un suceso	Redactar un acta o documento oficial en/ por el que se certifica un suceso	Definición	Falta barra inclinada entre “en” y “por”. También podría tratarse de una adición de “en” o de “por”	https://dictionary.cambridge.org/dictionary/spanish-english/acta
<i>¡Ah, qué bien!.</i>	<i>¡Ah, qué bien!</i>	Ejemplo	Sobra el punto tras el signo de exclamación de cierre	https://dictionary.cambridge.org/dictionary/spanish-english/ah
MAYÚSCULA				
Que está separado o es distinto	que está separado o es distinto	Definición	Las definiciones del CDO empiezan con letra minúscula	https://dictionary.cambridge.org/dictionary/spanish-english/aparte
ACENTO ORTOGRÁFICO				
aliteracion noun	aliteración noun	Lema		https://dictionary.cambridge.org/dictionary/spanish-english/aliteracion