EL TRATAMIENTO DE LAS ERRATAS DESDE LA SICOLINGÜÍSTICA Y DESDE EL PROCESAMIENTO DEL LENGUAJE NATURAL. UN ESTADO DE LA CUESTIÓN

Santiago Rodríguez-Rubio

santirrm@hotmail.com

Universidad Pablo de Olavide

Nuria Fernández-Quesada

nferque@upo.es

Universidad Pablo de Olavide

THE TREATMENT OF TYPOGRAPHICAL ERRORS FROM PSYCHOLINGUISTICS AND FROM THE NATURAL LANGUAGE PROCESSING PERSPECTIVE. A STATE OF AFFAIRS

Fecha de recepción: 17.06.2020 / Fecha de aceptación:18.02.2021

Tonos Digital, 40, 2021 (I)

RESUMEN:

A lo largo del tiempo, el estudio de las erratas se ha abordado desde una variedad de disciplinas tales como la bibliografía, la crítica textual, la edición, la metalexicografía, la sicolingüística y el procesamiento del lenguaje natural (PLN). Sin embargo, las erratas siguen siendo consustanciales a los textos escritos, y no existe todavía un método multidisciplinar e integrador para su tratamiento. Por este motivo, presentamos una revisión teórica sobre las erratas y su categorización, especialmente desde la sicolingüística y el PLN. Se trata, por una parte, de dos campos que han protagonizado algunos de los estudios de mayor calado a partir del siglo XX. Por otra parte, presentan coincidencias reseñables, tanto en sus conclusiones (por ejemplo, la menor frecuencia de erratas en la primera letra de las palabras) como en sus limitaciones (la ambigüedad en la categorización de las erratas). Así, partiremos de la

distinción teórica entre los errores motores o de ejecución (*mistypings*) y las faltas de ortografía o errores cognitivos (*misspellings*). A su vez, analizaremos las concomitancias entre las referidas disciplinas a la hora de identificar las operaciones erróneas (sustracción, adición, sustitución y transposición) que dan lugar a estos fenómenos. Por último, veremos cómo, en la práctica de la escritura, a veces resulta imposible distinguir entre errores de ejecución y errores de competencia.

Palabras clave: ambigüedad; categorización de erratas; *non-word error*; procesamiento del lenguaje natural (PLN); *real-word error*; sicolingüística

ABSTRACT:

Throughout time, the study of typographical errors has been approached from a variety of fields, such as bibliography, textual criticism, publishing, metalexicography, psycholinguistics and natural language processing (NLP). However, typographical errors are inherent in written texts, and there is no multidisciplinary and all-encompassing method for their treatment. For that reason, we present a theoretical review of typographical errors and their categorisation, especially from the perspective of psycholinguistics and NLP. On the one hand, some of the most thorough studies from the XX century onwards have been conducted in those two fields. On the other hand, there are remarkable similarities, regarding both the conclusions reached (e.g. the low frequency of errors in the first letter of words) and the limitations encountered (e.g. ambiguity in error categorisation). Thus, our starting point will be the theoretical distinction between motoric or execution errors (mistypings) and orthographic or cognitive errors (misspellings). Moreover, we will analyse the commonalities between the abovementioned disciplines when it comes to identifying the wrong operations (omission, insertion, substitution, and transposition) behind those phenomena. Finally, we will confirm that sometimes execution errors and competence errors cannot be distinguished in the practice of writing.

Keywords: ambiguity; natural language processing (NLP); *non-word error*; psycholinguistics; *real-word error*; typographical error categorisation

1. INTRODUCCIÓN

Afirma Moure (2006, p. 11) que "El error ha estado ligado a la historia de la escritura y de la transmisión textual desde sus comienzos" y asocia la voz "errata" a la

aparición de la imprenta (p. 18). Sin ir más lejos, Martínez de Sousa (2000) apunta que en la primera obra impresa (*Psalmorumcodex*, 1457) figuraba la errata *Spalmorum*.

Desde un punto de vista terminológico, en cambio, el consenso es algo menor. En español encontramos: "errata", "error de copia", "error de imprenta", "error mecanográfico", "error tipográfico", "lapsus cálami". En inglés, el abanico es aún más amplio: "*clerical error*", "*copying error*", "*erratum*" (pl. "*errata*"), "*lapsus calami*"[1], "*misprint*", "*mistype*", "*mistyping*", "*scribal error*" (Wheatley, 1893), "*slip of the pen*", "*typing error*", "*typing mistake*", "*typographical error*" (o simplemente "*typo*")[2]. Además, la literatura en lengua inglesa distingue entre *misspellings* (*spelling errors*) y *mistypings* (*typographical errors*), para aludir a errores ortográficos y a erratas, respectivamente (Véronis, 1988, p. 708). Esa distinción también puede expresarse en los siguientes términos: un error ortográfico (*orthographical error*) es un error cognitivo, mientras que uno tipográfico (*typographical error*) es un error motor (Van Berkel y De Smedt, 1988, p. 77). Sin embargo, como veremos en la parte dedicada a la ambigüedad en la clasificación de las erratas, esas dos categorías de errores suelen confundirse. De hecho, no siempre es posible distinguir un tipo de error del otro (Mitton, 1987).

Sea como fuere, el estudio de las erratas se ha abordado desde diversas esferas a lo largo del tiempo: la bibliografía (Beall, 2005; Ballard, 2008), la crítica textual (Blecua, 1983; Fernández-Ordóñez, 2002; Kordić y Avilés, 2016), la edición (MacKellar, 1893; Wheatley, 1893; Martínez de Sousa, 2005), la metalexicografía (Read, 1978; Mugglestone, 2005; Iamartino, 2017), la sicolingüística (Wells, 1916; MacNeilage, 1964; Grudin, 1983) y el procesamiento del lenguaje natural (Damerau, 1964; Mitton, 1987; Kukich, 1992). En este trabajo, nos centraremos en los dos últimos campos, la sicolingüística y el procesamiento del lenguaje natural. El primero se centra en los mecanismos causantes del error, mientras que el PLN tiende hacia el desarrollo de herramientas y algoritmos para la detección y corrección automática de los errores. Sin embargo, estas fronteras son permeables, puesto que existen, por ejemplo, estudios del campo de la lingüística computacional que sí analizan con cierto detalle los factores generadores de errores, como Ren y Perrault (1992) o Kano, Read, Dix y MacKenzie (2007).

[1] MacKellar (1893) ofrece la siguiente definición de *lapsus calami*: "*A slip of the pen; an error in writing*" (p. 365).
[2] Etymonline.com sitúa la voz *typo* como forma abreviada de *typographer* en 1816, y como forma abreviada de *typographical error* en 1892 (https://bit.ly/2wUQZmk). Merriam-Webster.com señala 1878 como fecha del primer uso de la voz con el sentido de *typographical error* (https://bit.ly/3eHg0SY). El uso de *typo* en el sentido de *typographer* aparece, por ejemplo, en Oldfield (1890, pp. 24, 85, 161).

Asimismo, comprobaremos que los autores que hemos tomado como primeros referentes en cada uno de esos dos campos, Wells (1916) y Damerau (1964), respectivamente, pese a la distancia en el tiempo, coincidieron en su descripción de las cuatro categorías fundamentales de "errores" (ya se tratara de erratas o de errores ortográficos): sustracción, adición, sustitución y transposición de letra. Quizá no sea casualidad, puesto que, desde el punto de vista de la crítica textual, Blecua también alude a cuatro tipos básicos de errores, que, según "las categorías modificativas aristotélicas", serían "*a*) por adición *(adiectio); b*) omisión *(detractatio); c*) alteración del orden *(transmutatio),* y *d*) por sustitución *(immutatio)*" (1983, pp. 19-20).

2. LITERATURA DEL CAMPO DE LA SICOLINGÜÍSTICA

La literatura del ámbito de la sicolingüística ha sido fundamental en el tratamiento y en la categorización de las erratas. El análisis de estas últimas ha servido para conocer los mecanismos sicomotrices que operan en la actividad de mecanografía.

Wells (1916) estableció la clasificación básica de errores mecanográficos: "*The errors fall naturally into four sorts, —omissions, substitutions, transpositions (metatheses) and additions*" (p. 59). El autor apuntó dos pautas generales en cuanto a la comisión de errores de mecanografía: en primer lugar, un mecanógrafo novel escribirá más lento y cometerá más errores que uno experto; en segundo lugar, un mecanógrafo (novel o experto) cometerá más errores tanto si escribe por encima de su velocidad óptima de mecanografiado como si lo hace por debajo de ella (p. 55). Wells explicó en términos generales por qué se producen errores cuando mecanografiamos:

> *The process of typewriting from copy involves a great number of fairly complicated psychomotor adjustments following upon each other in rapid succession. These adjustments do not always run smoothly, but on various occasions incorrect and false adjustments occur. These false adjustments result in " errors ;"…* (p. 56)

La errata *tyrranized* (por *tyrannized*) sirvió al autor para describir el fenómeno conocido como *doubling* (pp. 68-69). Este mecanismo se puede definir como el hecho de realizar en la mente la función de repetición prevista, pero ejecutarla erróneamente, repitiendo la letra equivocada. A continuación, indicamos algunas de las descripciones que Wells hizo de las posibles causas de los errores por sustitución y transposición de letra de su estudio:

- Los errores por sustitución pudieron deberse a: 1) que se pulsó la tecla equivocada (normalmente vecina) con el dedo correcto de la mano correcta; 2) que se usó el dedo equivocado de la mano correcta; 3) que se pulsó la tecla correcta con el dedo equivalente de la mano equivocada ("e" por "i"); 4) que se anticipó una tecla que aparece en seguida en la palabra (*mumor* por *humor*) (1916, pp. 60-64). El fenómeno de anticipación (junto con el opuesto de perseveración) sería estudiado y desarrollado ampliamente por posteriores autores.
- Los errores por transposición pudieron implicar el uso de dedos diferentes de diferentes manos, de dedos diferentes de la misma mano o incluso del mismo dedo de la misma mano (p. 67). Según el autor, algunos errores pudieron deberse a hábitos motores, como en *viture* por *virtue*. Wells razonó que tanto *viture* como *vitures* (por *virtue* y *virtues*) probablemente revelaban una tendencia a escribir "*-ture*", al ser este último un sufijo habitual en la lengua inglesa (p. 68).

Medio siglo más tarde, MacNeilage (1964) distinguiría tres categorías generales de errores de mecanografiado (*typing errors*): "errores espaciales" (tecla adyacente pulsada en lugar de la correcta), "errores temporales" en el orden de las letras (entre los que se incluyeron los errores por anticipación) y "errores misceláneos" (p. 146). El autor dividió los errores espaciales en:

- "Errores horizontales", resultado de pulsar una tecla inmediatamente a la derecha o izquierda de la tecla correcta, en la misma fila del teclado ("e" por "r").
- "Errores verticales", resultado de pulsar una tecla inmediatamente encima o debajo de la letra correcta, en la misma columna del teclado ("f" por "r").
- "Errores oblicuos", resultado de pulsar una tecla de una fila y columna adyacente a la de la letra correcta ("d" por "r").

Dentro de la categoría miscelánea, MacNeilage incluyó lo que denominó *type errors* (*that* por *than*, es decir, un tipo de lo que en el campo del PLN se conoce como *real-word error*, como veremos más adelante). El autor también incluyó en la categoría miscelánea lo que denominó *dynamics errors* (*eroors* por *errors*, es decir, lo que Wells había descrito como *doubling errors*), y lo que denominó *contralateral errors*: "*A stroke is typed using the same row and the corresponding finger to the correct one, but with the other hand*…" (p. 146). Recordemos que Wells también había descrito este fenómeno en su obra pionera, aunque no lo había denominado.

MacNeilage concluyó que la frecuencia de aparición de errores temporales en la primera letra de una palabra era menor que la de errores en la segunda, tercera y cuarta letra (1964, p. 156).

Rumelhart y Norman (1982) describieron cuatro categorías de errores mecanográficos: *transposition errors* (*becuase* por *because*), *doubling errors* (*scholl* por *school*), *alternation reversal errors* como variación del tipo anterior (*thses* por *these*) y *other errors* (pp. 4-5). Los autores indicaron que un 76 % de los errores por transposición implicaban el uso de la mano equivocada (*cross-hand*, en contraposición a *within-hand*). Entre los ejemplos de transposición *cross-hand*, Rumelhart y Norman indicaron *speical* por *special*, mientras que en los de transposición *within-hand* se incluiría *masetr* por *master* (1982, pp. 29-30). En la categoría "otros errores", los autores incluyeron, entre otras, la subcategoría *homologous errors*, que describieron así: "*In this class of errors, the wrong hand is selected, but then within the hand, the correct finger and the correct finger movement is performed. Thus, the erroneous stroke is anatomically homologous to the correct one*" (p. 31). Se trata de la misma categoría que MacNeilage denominaba *contralateral errors* (1964, p. 146) y que Wells (1916) había descrito por primera vez sin denominarla.

Grudin (1983) comparó el desempeño de mecanógrafos profesionales con el de mecanógrafos neófitos. En cuanto a la incidencia de la adyacencia de teclas en los errores por sustitución de letra, sus resultados en cuanto al grupo de mecanógrafos expertos arrojaron un 31 % de incidencia en errores horizontales y un 16 % en errores verticales (p. 127)[3]. El autor comparó sus resultados con los del experimento de Lessenberry (1928), que arrojó un 43 % de errores de fila (*row errors*) y un 15 % de errores de columna (*column errors*) (1983, p. 122). Según Grudin, el segundo factor más frecuente del experimento de Lessenberry después del de la adyacencia de las teclas era el del mecanismo sicomotor que provocaba los *homologous errors*. Este mecanismo ya había sido objeto de estudio en anteriores trabajos: Wells (1916), Rumelhart y Norman (1982), MacNeilage (1964). Grudin (1983, p. 122) estableció tres momentos en los que podrían darse esos "errores homólogos": 1) en la selección del programa motor (*motor program*), es decir, el conjunto de instrucciones dadas a los grupos de músculos; 2) en la especificación de la posición del dedo (*finger position specifications*) que determina la tecla que se va a pulsar; y 3) en la representación abstracta del teclado (por ejemplo, su representación espacial). El autor confirmó la tendencia expresada por MacNeilage (1964) de que se daban pocos errores por

[3] A diferencia de otros autores (que solo hablan de errores verticales), Grudin (1983) distingue entre errores verticales (*same column*) y diagonales (p. 130).

sustracción en la primera letra de las palabras. A este respecto, Grudin apuntó: "*The low incidence of omissions in the first-letter position in a word suggests that that letter is particularly strongly activated, and for that reason possibly subject to less noise*" (1983, p. 133).

Salthouse (1986) señala que el estudio de la mecanografía tiene un gran potencial para el conocimiento de complejos procesos perceptuales, cognitivos y motores (p. 317). El autor analizó las posibles causas de los errores, combinando las cuatro fases del proceso de mecanografiado (*input*, *parsing*, *translation*, *execution*) con las cuatro categorías principales descritas por Wells (*substitutions*, *intrusions*, *omissions*, *transpositions*). Así, un error de sustitución puede deberse a una confusión perceptual (*perceptual confusion*) en la primera fase del proceso, a una asignación errónea de la especificación del movimiento (*faulty assignment of movement specification*) en la tercera fase o a una posición errónea de los dedos (*misplaced finger position*) en la cuarta fase. Un error de adición puede deberse a una confusión perceptual (*perceptual confusion*) en la primera fase del proceso, a una falta de desactivación del carácter anterior (*failure to deactivate prior character*) en la tercera fase o a un pulsado simultáneo de dos teclas adyacentes (*simultaneous depression of two adjacent keys*) en la cuarta fase, y así sucesivamente. En relación con el factor de la adyacencia de las teclas en los errores por sustitución, Salthouse manifestó "*Many substitution errors involve adjacent keys […] Values from the typists in the Salthouse studies were that 35% of all substitution errors involved a horizontally adjacent key, and 17% involved a vertically adjacent key*" (1986, p. 312).

A finales del siglo XX, Logan (1999) argumentaba que existían en la literatura excelentes estudios sobre errores de mecanografía, como los de Gentner, Grudin, Rumelhart, MacNeilage y Wells, entre otros, pero que en ellos se combinaban errores de diferentes mecanógrafos. El autor introdujo la novedad de analizar errores cometidos por un único mecanógrafo durante varios años, a fin de contribuir a un mejor entendimiento de los procesos y mecanismos que subyacen a las operaciones de mecanografía (p. 1760). Logan estableció una tipología de errores de gran granularidad, con tres niveles de subdivisión (p. 1762):

- "Errores de respuesta", subdivididos en *omission*, *substitution*, *insertion*. Dentro de la subcategoría *substitution*, el autor incluyó entre otros los *homologous errors* (*suffecient* por *sufficient*), adoptando la terminología de Rumelhart y Norman (1982).

- "Errores temporales", subdivididos en *transposition*, *antedating response*, *interchange* (*endur* por *under*), *migration* (*runde* por *under*), *alternation* (*suffcicent* por *sufficient*), *doubling* (*sufiicient* por *sufficient*).
- "Errores lingüísticos", subdivididos en *antedating response* (*fsufficient* por *sufficient*), *perseveration* (*sufficicient*), *another word* (*pressing* por *pressure*), *spelling* (*suficient* por *sufficient*).
- "Miscelánea" (*suffcnint* por *sufficient*).

Logan (1999) vinculó los errores del habla con los errores de mecanografía, a través del concepto de "habla privada" o "habla interna" (*inner speech*)[4]:

> *More generally, Luria (1961), for example, has theorized that all voluntary acts are mediated by inner speech. This in turn permits the postulation of an articulatory-acoustic generalization error factor in typewriting, which implies that some errors that might appear to be slips of the fingers are the result, at least in part, of slips of the tongue*… (1999, pp. 1762-1763)

De hecho, Logan señaló que el habla y la escritura se retroalimentan, y que al mismo tiempo que la lengua le indica a los dedos lo que tienen que hacer, estos le indican a la primera lo que tiene que decir (p. 1769-1770).

Entre los fenómenos sicomotores hasta ahora citados, las repeticiones suelen entenderse en términos de "anticipación" o de "perseveración", según los casos. Ambos fenómenos a veces transcienden los límites de una determinada palabra. Así, *sufficicient* y *difffer* podrían considerarse ejemplos de perseveración dentro de la palabra (Logan, 1999, pp. 1762, 1767) y *under rsufficient* un ejemplo de perseveración entre palabras (p. 1762). Por otro lado, *wrapid writing* sería un claro ejemplo de anticipación (Lashley, 1951, p. 119)[5].

Los fenómenos de anticipación y perseveración también han sido observados desde perspectivas diferentes a la sicolingüística. Desde la crítica textual, Blecua (1983) señala que uno de los mecanismos de generación de errores es: "La adición de un fonema por atracción de otro anterior o posterior de la misma palabra o de la palabra contigua", como se observa en el término erróneo *piereden*. Asimismo, explica que estos errores suelen producirse en la fase del proceso de copia denominada "dictado interior" (p. 21).

[4] Oppenheim (2009) describe el concepto *inner speech*: "*Most people hear a little voice inside their head when thinking, reading, writing, and remembering. This voice is inner or internal speech, mental imagery that is generated by the speech production system (Sokolov, 1972)*" (2009, p. 1).

[5] Desde el punto de vista de la lingüística, Luelsdorff (1986) dividió las anticipaciones en *lexical* ("*the anticipation of a word or group of words*") y *literal* ("*the anticipation of a letter or group of letters*") (p. 202).

3. LITERATURA DEL CAMPO DEL PROCESAMIENTO DEL LENGUAJE NATURAL (PLN)

La detección y corrección automática de errores ortográficos y tipográficos es uno de los campos de trabajo habituales del PLN. En este terreno, se distingue entre *non-word errors* (palabras idiomáticamente incorrectas) y *real-word errors* (palabras correctas, pero inválidas desde el punto de vista del contexto).

El establecimiento de algoritmos para la corrección de errores ortográficos se remonta, al menos, a Miller y Friedman (1957) y a Blair (1960). En esta última obra, el autor desarrolló un programa con procedimientos heurísticos en el que se asignaba un peso a cada letra y a cada posición de la letra en la palabra, a partir de lo que se obtenían las probabilidades de que se dieran errores (1960, p. 62). Damerau (1964) es una de las obras seminales del PLN en lo que se refiere a la categorización de los errores ortográficos y tipográficos. El autor describió un método para detectar y corregir automáticamente lo que él denominaba *spelling errors* y estableció las cuatro categorías básicas de operaciones incorrectas que Wells ya había descrito en 1916. Para Damerau, más del ochenta por ciento de los *spelling errors* eran ejemplos singulares (no múltiples) de sustitución de una única letra (*wrong letter*), sustracción de una única letra (*missing letter*), adición de una única letra (*extra letter*) o transposición de dos letras adyacentes (*single transposition*) (1964, p. 171).

Mitton (1987) abordó de manera más sistematizada la cuestión de los *real-word errors*, también señalada por autores de la sicolingüística (Wells, 1916; MacNeilage, 1964; Logan, 1999), o incluso desde el PLN cuando Peterson (1986) se refirió a estos errores aunque sin usar el término *real-word error* (pp. 633-634). Mitton explica: *"A checker that detects errors simply by looking up words in a dictionary will obviously fail to spot errors that happen to match dictionary words, such as 'wether' for 'whether.' I call these 'real-word errors.'"* (1987, p. 496). El autor identificó tres clases de *real-word error*: *wrong-word errors* (*know* por *now*), *wrong-form-of-word errors* (*thing* por *things*, *use* por *used*) y *word-division errors* (*miss dress* por *mistress*) (pp. 497-498). Los primeros implican la sustitución de la palabra correcta y válida por una palabra no relacionada. Los segundos implican que la palabra correcta y válida se sustituye por una forma correcta pero inválida de dicha palabra. En los terceros, la palabra correcta y válida se divide erróneamente en dos palabras correctas pero inválidas. Entre las categorías *real-word error* y *non-word error*, Mitton estableció una categoría intermedia, a saber, *half real-word errors*. Se trata de *word-division errors* que forman una palabra correcta y una incorrecta (siendo ambas inválidas), como en

to gether por *together* o en *evry body* por *everybody* (p. 499). El autor estableció la relación porcentual entre los *real-word errors* (39,1 %), los *non-word errors* (59,7 %) y los *half real-word errors* (1,2 %) de su estudio (p. 497).

En 1992, Kukich realizó un análisis pormenorizado de las técnicas de corrección automática de errores (tanto *non-word errors* como *real-word errors*). La autora desarrolló de manera notable la casuística de *real-word errors* descrita en Mitton (1987), definiendo los siguientes mecanismos generadores de error: "simples erratas" (*from* por *form*), "lapsus cognitivos o fonéticos" (*there* por *their*) y "equivocaciones sintácticas o gramaticales". Estas últimas incluirían formas flexivas incorrectas (*arrives* por *arrive*), palabras sincategoremáticas erróneas (*his* por *her*), anomalías semánticas (*minuets* por *minutes*), adición o sustracción de palabras (*the system has been operating system for*…) y uso incorrecto del espaciado (*ad here* por *adhere*) (1992, p. 412). La autora argumentaba que para la detección y corrección de las erratas era necesario disponer de información contextual. Entre los hallazgos de su investigación se encuentran los siguientes: 1) en la primera letra de las palabras se dan pocas erratas (una consideración recurrente tanto en el campo del PLN como en el de la sicolingüística); y 2) existe un fuerte vínculo entre las erratas por sustitución de letras y la adyacencia de las teclas (p. 392). En relación con este último aspecto, Kukich apuntaba: "*For example, transcription typing errors, which are for the most part due to motor coordination slips, tend to reflect typewriter keyboard adjacencies, e.g., the substitution of* b *for* n" (p. 387).

De manera contemporánea al trabajo de Kukich, Ren y Perrault (1992, pp. 410-411) realizaron una clasificación de errores a partir de textos en inglés y en francés, en la que incluían:

a) errores con el acento ortográfico, subdivididos en adición del acento (*éssai* por *essai*), sustracción del acento (*agées* por *âgées*), sustitución de un acento por otro (*àgées* por *âgées*) y desplazamiento del acento (*chomâge* por *chômage*);

b) errores de puntuación, relativos a guiones y apóstrofes;

c) adición de letras, subdividida en duplicación de letras (*paartnership* por *partnership*), adición de una letra contigua (*profcessional* por *professional*) e interferencia de otra letra de la misma palabra (*aéoroport* por *aéroport*)[6];

[6] La interferencia de otra letra de la misma palabra que describen Ren y Perrault (1992) responde a lo que en sicolingüística se denominan anticipaciones (*aéoroport*) y perseveraciones (*bubget*). Nótese que los errores en *paartnership* y *gouvervement* (que los autores incluyeron en los subapartados "duplicación de letras" y "uso de la mano equivocada", respectivamente) también pueden considerarse perseveraciones. De hecho, los mecanismos sicomotores a los que nos hemos referido en el subapartado de literatura de la sicolingüística (v. gr. anticipación, perseveración,

d) sustracción de una letra;

e) sustitución de una letra, subdividida en letra sustituida por otra adyacente (*esperience* por *experience*), uso de la mano equivocada (*gouvervement* por *gouvernement*), interferencia de otra letra de la misma palabra (*bubget* por *budget*), error ortográfico (*maintenence* por *maintenance*) y otras sustituciones;

f) transposición de letra o de letras, subdividida en inversión de letras adyacentes (*commerical* por *commercial*), inversión de letras no adyacentes (*condiser* por *consider*) y desplazamiento de una única letra (*avalaible* por *available*);

g) errores gramaticales;

h) otros errores.

Desde el ámbito hispano, y ya en el siglo XXI, Ramírez y López (2006) elaboraron una tipología de cerca de 76 000 *spelling errors* en español. Tras comparar los patrones de errores de su investigación con los de estudios realizados por otros investigadores, los autores llegaron a la conclusión de que algunos de esos patrones no podían ser extrapolados a idiomas distintos del inglés (p. 97). Ramírez y López confirmaron algunas de las tendencias apuntadas por Kukich (1992), por ejemplo, que la mayoría de los errores presentaban una distancia de edición de una letra y que en las primeras letras de las palabras se daban pocos *misspellings* (aunque Ramírez y López indicaron que, en su investigación, dicho porcentaje era superior al de otros estudios). Por otro lado, los autores sostenían que en español los errores más frecuentes son (por este orden): sustracción, sustitución, adición y transposición (2006, p. 95). En relación con el factor de la adyacencia de las teclas en los errores en general (no solo en los producidos por sustitución de letra), Ramírez y López reconocieron un fuerte efecto de la colindancia, pero lo matizaron: "*In summary, although keyboard adjacency effects are indeed relevant for a taxonomy on the nature of human misspellings, the frequencies in Table 6 suggest that other factors can be much more relevant*" (p. 97).

Kano y Read (2009) realizaron una categorización de errores con una elevada granularidad. Además de establecer dos normas para la resolución de ambigüedades (que veremos en el siguiente apartado), los autores repasaron en su trabajo métodos de categorización de errores aplicados por diversos autores (incluidos ellos mismos) desde 1945 hasta 2007 (2009, p. 294). Además, analizaron el factor de la adyacencia

doubling) son muy habituales en cualquier texto escrito, cuestión distinta es poder determinar en todos los casos que fueron esos mecanismos (y no otros) los que causaron los correspondientes yerros.

de las teclas en la comisión de errores de mecanografía. Las sustituciones en las que intervienen teclas horizontales las denominaron *Next To error* (*NT*), mientras que aquellas en las que operan teclas verticales u oblicuas recibieron el nombre de *Close To error* (*CT*) (p. 298).

4. LA AMBIGÜEDAD COMO LIMITACIÓN EN LA CLASIFICACIÓN DE LAS ERRATAS

Las erratas pueden generar ruido cuando se usan herramientas de procesamiento automático del lenguaje (Grouin, 2008, p. 1083). Luelsdorff (1986) señaló que, a menudo, los datos compuestos de errores contienen ruido y son irresolubles desde el punto de vista teórico, ya que pueden dar lugar a clasificaciones contradictorias (p. 53). Ello puede conllevar que en una clasificación de errores algunas categorías se solapen.

En ocasiones, la imposibilidad de establecer la causa o causas últimas de las erratas supone un escollo considerable a la hora de clasificarlas. En estudios de los campos del procesamiento del lenguaje natural (PLN) y de la sicolingüística se ha abordado este punto de manera extensa.

Según MacNeilage (1964):

> *Of the two language producing processes, typing and speech, typing has the methodological advantage of occurring in discrete response units that are automatically recorded. But in an ordinary typing situation, many kinds of error are possible, and a number of variables could be influential in any type of error. This possibility of multivariate determination of errors could serve to make analysis of typing behaviour in a free situation a difficult task*. (p. 144)

Desde una perspectiva más general del ámbito de la sicología, Norman (1981) incide en que la mayoría de las equivocaciones (*action slips*) tienen múltiples causas, pues se supone que una acción concreta debe ser el resultado de muchas fuentes de información que interactúan: "*When the act is an error, it is apt to be the result of numerous underlying forces, so that the resulting slip is multiply determined and consistent with a number of constraints and explanations*" (p. 2).

Por otra parte, en un estudio que pretendía identificar los factores que inciden en errores de mecanografía, Logan (1999) señalaba ambigüedad en un porcentaje amplio de los errores analizados, ya que podían achacarse a diferentes causas. En la línea de Salthouse (1986), Logan indicaba que un error mecanográfico podía producirse en cualquiera de las cuatro fases del proceso: *input stage*, *parsing stage*, *translation stage*, *execution stage* (1999, p. 1769).

Desde la perspectiva del PLN, Pollock y Zamora (1984) reconocían la ambigüedad como el principal problema del algoritmo de corrección que habían empleado. Según los autores, entre un 10 %-15 % de los *misspellings* se veía afectado por la ambigüedad, ya que existían varias correcciones válidas (p. 362).

En la misma línea, Deorowicz y Ciura (2005) afirman que, ante determinados errores, no se puede estar seguro de qué palabra correcta se pretende teclear, si no se dispone de información sobre el contexto. Lo ejemplifican a través del término erróneo *stat* que, en un texto sobre astronomía, podría ser con bastante probabilidad una forma errónea de *star*, pero que, en textos de otros campos, podría tener como término subyacente válido *stay* o *state*. Por tanto, incluso disponiendo de un contexto, hay situaciones en las que no se puede estar seguro de corregir una errata de una determinada manera, puesto que se ignora qué operación errónea la produjo (p. 278)[7].

A propósito de los problemas de contexto, Hartley (2009) aborda la cuestión desde la perspectiva de los denominados "lenguajes naturales controlados" (*controlled natural languages*, CNLs). Estos lenguajes son versiones simplificadas de lenguajes naturales, con unas condiciones gramáticas restrictivas que pueden reducir la ambigüedad. Según Hartley, existe un cierto consenso respecto a que los lenguajes controlados pueden mejorar la calidad de los textos traducidos por humanos y la de los textos traducidos automáticamente. Sin embargo, el autor señala que, incluso en entornos controlados, la corrección automática puede ser una tarea difícil, porque no se trata solo de detectar un error, sino de ofrecer la corrección válida. Hartley ilustró dicha dificultad refiriéndose a que la frase antigramatical *The train depart* tiene dos correcciones posibles (*The train departs* y *The trains depart*), pero solo una es válida (p. 116).

Kano et al. (2007) sostienen que los algoritmos que usen información probabilística y contextual pueden contribuir a limitar la ambigüedad de las causas de los errores ortográficos y tipográficos. Además, Kano y Read (2009) establecen dos reglas (una en el plano temporal y otra en el espacial), para eliminar la ambigüedad de los errores por adición de una letra. La primera regla (*Zero Time Rule*) sirve para determinar si dos teclas se pulsaron al mismo tiempo, aunque no para discernir si se pulsaron con un dedo o con dos. Por este motivo, establecen una segunda regla

[7] Baba y Suzuki (2012) distinguen entre las erratas que el propio mecanógrafo percibe y corrige sobre la marcha (*corrected errors*) y aquellas que pasan desapercibidas (*uncorrected errors*). Los autores manifiestan que una de las razones por las que es importante analizar las erratas corregidas por el mecanógrafo es que se puede conocer todo el proceso y saber en qué consistió el yerro, a diferencia de lo que sucede con las erratas que corrigen terceras personas o medios automáticos (p. 373).

(*Impossible NT/CT Rule*), basada en la asunción de que si las dos teclas son colindantes (horizontal, vertical u oblicuamente) se pulsaron con el mismo dedo (pp. 298-299).

La ambigüedad también puede influir a la hora de determinar si un error es ortográfico (*misspelling*) o tipográfico (*mistyping*)[8]. Sobre el papel, se trata de dos tipos de yerro diferentes, ya que el primero es fruto del desconocimiento del que escribe, mientras que el segundo es un error fortuito o equivocación material (dicho de otro modo, una errata)[9]. Pero en la práctica resulta imposible distinguir uno de otro en determinados casos, puesto que la persona que detecta o corrige los errores no puede conocer sus causas. A la confusión entre *misspellings* y *mistypings* puede contribuir también la circunstancia de que, en la literatura, se hayan podido mezclar ambas denominaciones, con independencia de si el investigador era consciente o no de ello. De la estrecha relación existente entre ambos tipos de "error" da cuenta el hecho antes referido de que, en los estudios pioneros sobre *mistypings* (Wells, 1916) y en los relativos a *misspellings* (Damerau, 1964), se determinaron las mismas cuatro categorías fundamentales. Por ejemplo, en la lista de *spelling errors* del estudio de Damerau, muchos términos son a todas luces erratas (v. gr. *Britian* por *Britain*); en palabras del propio autor: "*These are the errors one would expect as a result of misreading, hitting a key twice, or letting the eye move faster than the hand*" (1964, p. 171). Pero también figuran hipotéticos errores ortográficos (v. gr. *Antartic* por *Antarctic*).

Para concluir, un breve repaso de la literatura en torno a las denominaciones *spelling error* y *typographical error* nos lleva a la diferencia establecida por Véronis (1988):

> *We introduce a distinction between competence and performance errors. Performance errors are simply due to mechanical or neuro-motor problems (typographical errors, 'slips of the pen'), whereas competence errors reflect ignorance about language rules or misconceptions about the domain*. (p. 708)

Sin embargo, Mitton (1987) puntualiza: "*Although it is possible in many cases to decide whether an error was a slip or a wrong spelling, there is no general way of distinguishing between them*" (p. 496). En un estudio posterior, el autor reitera: "*Studies of uncorrected typos face the same data-collection problem as studies of slips of the pen; it is easy enough to collect errors from keyboarded text, but it is*

[8] En un estudio del ámbito de la enseñanza de lenguas, Mendikoetxea, Murcia y Rollinson (2010) indican la imposibilidad de acceder al autor de los errores como la razón por la que resulta imposible distinguir las erratas de los errores ortográficos (p. 181).

[9] Arroyo (2017) apunta que a los errores ortográficos se les puede llamar "errores de competencia", mientras que las erratas pueden recibir la denominación de "errores de actuación" (p. 48).

impossible to separate the typos from the misspellings" (1996, p. 88)[10]. En relación con la investigación *SPEEDCOP* realizada por Pollock y Zamora en 1983, Mitton señala cómo los autores habían considerado la mayor parte de los errores como erratas, pero que en una "minoría significativa" de casos podría tratarse de errores ortográficos (1996, p. 89). Min, Wilson y Moon (2000) sostienen que *typographical errors*, *orthographical errors* y *scanning errors* son subcategorías de la categoría genérica *spelling errors* (p. 1). En la misma línea, Peterson (1980) apunta que los *spelling errors* pueden producirse de tres formas: por ignorancia del que escribe, por errores tipográficos al mecanografiar y por problemas de transmisión y almacenaje de datos (v. gr. OCR) (p. 677).

5. CONCLUSIONES

Es imposible desligar las erratas de los textos, ya sean impresos o manuscritos. Las erratas se han estudiado desde muy diversos ángulos a lo largo del tiempo. En este trabajo se ha revisado la producción científica sobre su tratamiento y categorización desde dos perspectivas diferentes, pero con aspectos en común: la sicolingüística y el procesamiento del lenguaje natural (PLN). El enfoque multidisciplinar enriquece el conocimiento de las erratas, pues revela rasgos de universalidad, como las operaciones erróneas básicas, la baja frecuencia de errores mecanográficos en la primera letra de las palabras o el impacto de la adyacencia de las teclas en los errores (especialmente, en aquellos por sustitución de letra).

Podemos concluir, sin embargo, que la categorización de las erratas siempre estará sujeta a limitaciones. Hemos señalado la de la ambigüedad o dependencia del contexto porque puede afectar a la identificación de las causas (y, por tanto, a la corrección) de las erratas, incluyendo la determinación del error como motor o como cognitivo. Es decir, si se trata de un error de actuación o de uno de competencia. El entendimiento y la aceptación de esas limitaciones es un paso decisivo para una mejor comprensión de las erratas y de su categorización.

BIBLIOGRAFÍA

Arroyo, L. (2017). *Lingüística de errores con fines computacionales. PatErr, un recurso para la revisión textual del español basado en patrones de error codificados*.

[10] De la misma manera que, a menudo, es imposible diferenciar entre errores ortográficos y erratas, Blecua (1983) apunta desde la perspectiva de la crítica textual que muchas veces no se puede diferenciar entre una errata (léase "un error accidental") y una intervención voluntaria del copista (p. 20). Kordić y Avilés (2016) sostienen, de hecho, que muchos supuestos "errores de copista" responden a "fenómenos lingüísticos vulgares, absolutamente legítimos e históricamente reconocidos, si bien de baja ocurrencia o escaso testimonio algunos de ellos" (pp. 212-213).

Tesis de doctorado, Facultad de Traducción e Interpretación, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, España.

Baba, Y. & Suzuki, H. (2012). How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. En H. Li, C-Y. Lin, M. Osborne, G. G. Lee & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Vol.2* (pp. 373-377). Jeju Island, Korea: Association for Computational Linguistics (ACL).

Ballard, T. (2008). Systematic Identification of Typographical Errors in Library Catalogs. *Cataloging and Classification Quarterly, 46(1)*, 27-33.

Beall, J. (2005). Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information, 6(3)*, 1-20.

Blair, C. R. (1960). A Program for Correcting Spelling Errors. *Information and Control, 3(1)*, 60-67.

Blecua, A. (1983). *Manual de crítica textual*. Madrid: Editorial Castalia.

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM, 7(3)*, 171-176.

Deorowicz, S. & Ciura, M. G. (2005). Correcting Spelling Errors by Modelling their Causes. *International Journal of applied mathematics and computer science (AMCS), 15(2)*, 275-285.

Fernández-Ordóñez, I. (2002). Tras la *collatio* o cómo establecer correctamente el error textual. *La corónica: A Journal of Medieval Spanish Language and Literature, 30(2)*, 105-180.

Grouin, C. (2008). Certification and cleaning up of a text corpus: towards an evaluation of the "grammatical" quality of a corpus. En N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)* (pp. 1083-1090). Marrakech, Morocco: European Language Resources Association (ELRA).

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). New York: Springer.

Hartley, T. (2009). Technology and Translation. En J. Munday (Ed.), *The Routledge Companion to Translation Studies* (pp. 106-127). Abingdon/New York: Routledge.

Iamartino, G. (2017). Lexicography, or the Gentle Art of Making Mistakes. *Altre Modernità (Numero speciale — Errors: Communication and its Discontents)*, 48-78.

Kano, A., Read, J. C., Dix, A. & MacKenzie, S. (2007). ExpECT: An Expanded Error Categorisation Method for Text Input. En L. J. Ball, M. A. Sasse, C. Sas, T. C. Ormerod, A. Dix, P. Bagnall & T. McEwan (Eds.), *People and Computers XXI – HCI… but not as we know it*. *Proceedings of HCI 2007* (pp. 147-156). Swindon: British Computer Society.

Kano, A. & Read, J. C. (2009). Text Input Error Categorisation: Solving Character Level Insertion Ambiguities Using Zero Time Analysis. En A. F. Blackwell (Ed.), *People and Computers XXIII – Celebrating People and Technology: HCI 2009* (pp. 293-302). Cambridge: British Computer Society.

Kordić, R. & Avilés, T. (2016). ¿Variante lingüística o error de copista? *Hipogrifo, 4(1)*, 199-215.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys, 24(4)*, 377-439.

Lashley, K. S. (1951). The Problem of Serial Order in Behavior. En L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112-146). New York: Wiley.

Lessenberry, D. D. (1928). *Analysis of Errors*. Syracuse/New York: L. C. Smith and Corona Typewriters, School Department.

Logan, F. A. (1999). Errors in Copy Typewriting. *Journal of Experimental Psychology: Human Perception and Performance, 25(6)*, 1760-1773.

Luelsdorff, P. (1986). *Constraints on Error Variables in Grammar: Bilingual Misspelling Orthographies*. Philadelphia: John Benjamins.

MacKellar, T. (1893). *The American Printer: A Manual of Typography*. Philadelphia: MacKellar, Smiths & Jordan Foundry.

MacNeilage, P. F. (1964). Typing Errors as Clues to Serial Ordering Mechanisms in Language Behaviour. *Language and Speech, 7(3)*, 144-159.

Martínez de Sousa, J. (2000). Las erratas. *Centro Virtual Cervantes* (Rinconete > Lengua). Recuperado el 14 de Junio, 2020 de http://bit.ly/2v82eqn

Martínez de Sousa, J. (2005). *Manual de edición y autoedición* (2ª ed. quinta impresión 2016). Madrid: Ediciones Pirámide.

Mendikoetxea, A., Murcia, S. & Rollinson, P. (2010). Focus on Errors: Learner Corpora as Pedagogical Tools. En M. C. Campoy-Cubillo, B. Bellés-Fortuño & M. Ll. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching* (pp. 180-194). London/New York: Continuum.

Miller, G. A. & Friedman, E. A. (1957). The Reconstruction of Mutilated English Texts. *Information and Control, 1*, 38-55.

Min, K., Wilson, W. & Moon, Y-J. (2000). Typographical and Orthographical Spelling Error Correction. *2nd International Conference on Language Resources and Evaluation* 221. Athens, Greece.

Mitton, R. (1987). Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management, 23(5)*, 495-505.

Mitton, R. (1996). *English spelling and the computer*. London: Birkbeck ePrints.

Moure, J. L. (2006). Errores deseables y erratas cohonestadas. *Páginas de guarda: revista de lenguaje, edición y cultura escrita, 1*, 11-25.

Mugglestone, L. (2005). *Lost for Words: The Hidden History of the Oxford English Dictionary*. New Haven/London: Yale University Press.

Norman, D. A. (1981). Categorization of Action Slips. *Psychological Review, American Psychological Association, 88(1)*, 1-15.

Oldfield, A. (1890). *A Practical Manual of Typography and Reference Book for Printers*. London: E. Menken, Technical Publisher.

Oppenheim, G. M. (2009). *The Little Voice in your Head: Error-based Investigations of Abstracted and Articulated Inner Speech*. B. A. Thesis, Grinnell College, University of Illinois, EUA.

Peterson, J. L. (1980). Computer Programs for Detecting and Correcting Spelling Errors. *Communications of the ACM, 23(12)*, 676-687.

Peterson, J. L. (1986). A Note on Undetected Typing Errors. *Communications of the ACM, 29(7)*, 633-637.

Pollock, J. J. & Zamora, A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text. *Communications of the ACM, 27(4)*, 358-368.

Ramírez, F. & López, E. (2006). Spelling Error Patterns in Spanish for Word Processing Applications. En *5th International Conference on Language Resources and Evaluation* (pp. 93-98). Genoa, Italy: European Language Resources Association (ELRA).

Read, A. W. (1978). The Sources of Ghost Words in English. *Word, 29(2)*, 95-104.

Ren, X. & Perrault, F. (1992). The Typology of Unknown Words: An Experimental Study of Two Corpora. En *COLING '92 Proceedings of the 14th Conference on Computational Linguistics, Vol. 1* (pp. 408-414). Nantes, France: Association for Computational Linguistics (ACL).

Rumelhart, D. E. & Norman, D. A. (1982). Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance. *Cognitive Science, 6*, 1-36.

Salthouse, T. A. (1986). Perceptual, Cognitive, and Motoric Aspects of Transcription Typing. *Psychological Bulletin, 99(3)*, 303-319.

Van Berkel, B. & De Smedt, K. (1988). Triphone Analysis: A Combined Method for the Correction of Orthographical and Typographical Errors. En *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP)* (pp. 77-83). Austin, Texas, EUA: Association for Computational Linguistics (ACL).

Véronis, J. (1988). Morphosyntactic correction in natural language interfaces. En J. Vargha (Ed.), *COLING '88 Proceedings of the 12th Conference on Computational Linguistics, Vol. 2* (pp. 708-713). Budapest: Hungary. Association for Computational Linguistics (ACL).

Wells, F. L. (1916). On the Psychomotor Mechanisms of Typewriting. *The American Journal of Psychology, 27(1)*, 47-70.

Wheatley H. B. (1893). *Literary Blunders: A Chapter in the* "History of Human Error". London: Elliot Stock.