

# Avances en el Análisis y Tipificación de Errores para una Propuesta de Mejora de Informes Médicos en Español

## *Progress in the Analysis and Classification of Errors for a Proposal to Improve Medical Reports in Spanish*

**Jésica López-Hernández**  
Departamento de Informática y Sistemas  
Universidad de Murcia  
jesica.lopez@um.es

**Resumen:** Las tareas de detección y corrección automática de errores constituyen un elemento esencial en las tecnologías del procesamiento del lenguaje natural. Sin embargo, los lenguajes de especialidad presentan particularidades lingüísticas que dificultan en gran medida este cometido, como el uso de terminología especializada. En el caso del lenguaje médico, y más concretamente de los informes clínicos, estas particularidades son cruciales, debido a que los textos también suelen contener abreviaturas, siglas y errores lingüísticos. Sin embargo, no existen datos estadísticos previos sobre patrones de error en textos en español que procedan del ámbito biosanitario. Resulta de gran interés investigar sobre la naturaleza de los errores que ocurren en los informes clínicos para mejorar el proceso de detección y la generación de listas de sugerencias y decisión en la fase de corrección. Por tanto, este proyecto surge con la intención de aportar un módulo basado en conocimiento lingüístico que pueda añadir otra capa de información a los métodos actuales de corrección en informes médicos y, en consecuencia, contribuir a la mejora de corpus pertenecientes al ámbito de la medicina. Con el fin de investigar sobre el objetivo mencionado, se utilizan herramientas de detección, extracción y análisis de errores en un corpus de informes clínicos electrónicos de diversas especialidades médicas.

**Palabras clave:** Corrección automática de errores, detección automática de errores, tipología de error, lenguaje médico.

**Abstract:** Automatic error detection and correction tasks are an essential element in natural language processing technologies. However, specialized languages have linguistic peculiarities that make these tasks more difficult, such as the use of specialized terminology. These particularities are crucial in the case of medical language, and more specifically of clinical reports, because the texts usually have abbreviations, acronyms and linguistic errors. However, there are no previous statistical data on error patterns in biomedical texts in Spanish. It is of great interest to delve into the nature of errors that occur in clinical reports to improve the detection process and the generation of suggestion lists in the correction phase. Therefore, this project arises with the intention of providing a module based on linguistic knowledge that can add another layer of information to current methods of correction in medical reports and, consequently, contribute to the improvement of corpora belonging to the field of medicine. In order to investigate the aforementioned objective, error detection, extraction and analysis tools are used in a corpus of electronic clinical reports from various medical specialties.

**Keywords:** Automatic error correction, automatic error detection, error typology, medical language.

## ***1 Justificación de la investigación propuesta***

La detección y corrección automática de errores continúa siendo un problema de actualidad en el campo del procesamiento del lenguaje natural, a pesar de ser uno de los primeros en comenzar a abordarse (Jurafsky y Martin, 2014). La efectividad de los correctores ortográficos, como la mayoría de aplicaciones que se construyen para procesamiento de textos, se ve condicionada en gran medida por las características del dominio donde se van a aplicar. La dificultad de esta tarea aumenta cuando se trabaja con corpus pertenecientes al dominio médico, debido a las características intrínsecas que estos textos poseen. Entre estas peculiaridades destacan el uso de terminología específica, abreviaturas, siglas y la presencia de errores de escritura (Meystre y Haug, 2006). Son diversos los estudios (Wong y Gance, 2011; Ruch, Baud y Geissbühler, 2003; Siklósi, Novák y Prószycki, 2016; entre otros) que señalan el importante número de errores lingüísticos que presentan los informes clínicos, debido al limitado tiempo que los profesionales de la salud disponen para la redacción de los mismos y a la ausencia de revisión posterior. Sin embargo, no encontramos investigaciones que se centren en el estudio de los errores lingüísticos en documentación clínica para el español.

Nuestra experiencia previa trabajando con informes médicos procedentes de diversas especialidades nos ha permitido comprobar que estos poseen un elevado número de errores ortográficos, tipográficos, gramaticales y semánticos. Los profesionales de la salud suelen sufrir sobrecarga de trabajo y disponen de poco tiempo para redactar estos documentos, por lo que no pueden atender a la forma, sino únicamente al contenido en muchas ocasiones. En un área como medicina es de especial importancia poder hacer uso de las tecnologías basadas en procesamiento automático de datos, para facilitar la extracción de información, la interoperabilidad semántica, la toma de decisiones o la predicción de sucesos, entre otros. Es en esta realidad donde surge la motivación que determina en gran medida este trabajo: ¿qué puede aportar la lingüística de corpus en la detección y corrección automática de errores en el campo de la medicina? En la actualidad no existen datos cuantitativos sobre

patrones de error en textos que procedan del ámbito biosanitario, y tampoco hay una revisión sistemática sobre la naturaleza de los mismos. Por consiguiente, es necesario llevar a cabo un estudio y tipificación de errores que nos permita saber qué tipos de errores tienden a cometerse en este dominio, cuáles son sus propiedades y cómo podemos aportar una base de conocimiento lingüístico a los métodos de detección y corrección existentes para tal fin. De esta manera, vamos a poder definir rasgos de manera explícita, lo que puede ayudar a la toma de decisiones en aquellos casos que plantean dificultades o conflictos en la elección de alternativas a la palabra errónea. A partir de este análisis de errores puede añadirse un nuevo criterio al motor de sugerencias del corrector automático, y así contribuir a tener una mayor precisión y cobertura en este dominio especializado.

## ***2 Antecedentes y trabajo relacionado***

Los primeros trabajos sobre detección y corrección automática de errores se remontan a la década de los sesenta, cuando se define el concepto de *distancia de Levenshtein*, que alude al número mínimo de operaciones requeridas para transformar una secuencia de caracteres en otra. Se establecen cuatro operaciones básicas de edición (Damerau, 1964; Levenshtein, 1966): inserción, omisión, sustitución y trasposición. Los métodos convencionales de detección y corrección automática se han basado principalmente en el uso de diccionarios y en la distancia de edición mínima entre un error ortográfico y sus candidatos de corrección. Con el paso de los años se han ido sumando a estos métodos nuevas técnicas, como las basadas en similitud fonética (Veronis, 1988); técnicas probabilísticas, como el análisis de n-gramas (Ahmed, Luca y Nürnberger, 2009); técnicas basadas en reglas y heurísticas (Naber, 2003); técnicas basadas en modelos de canales ruidosos o noisy channel model (Brill y Moore, 2000); o las más actuales basadas en aprendizaje automático y redes neuronales (Pande, 2017).

La literatura sobre corrección automática en informes clínicos es más limitada y heterogénea, aunque el procesamiento automático de textos médicos es un tema emergente y de actualidad en procesamiento del lenguaje natural. Si bien es cierto que existen multitud de trabajos realizados en el área, la

mayoría están enfocados actualmente en tareas de extracción de información médica, desambiguación o reconocimiento de entidades nombradas (Patrick, Sabbagh, Jain, & Zheng, 2010). En muchos casos, el proceso de corrección es parte de investigaciones mayores que incluyen otras técnicas. Todas las investigaciones previas coinciden en señalar el importante número de errores que presentan los informes clínicos y la complejidad de su tratamiento, tanto por el gran número de abreviaturas que contienen, como por la compleja terminología, la falta de estandarización de las formas y la ausencia de revisión posterior (Patrick et al, 2010; Lai et al, 2015; Siklósi, Novák, y Prószéky, 2016; Fivez, Suster y Daelemans, 2017; entre otros).

A su vez, como hemos mencionado anteriormente, son inexistentes los estudios realizados sobre patrones de error en documentación clínica en español. No obstante, encontramos dos trabajos enfocados en tareas de corrección automática para el español común o general: *Spelling Error Patterns in Spanish for Word Processing Applications* (Ramírez y López, 2006) y *Tipología de errores gramaticales para un corrector automático* (Díaz, 2005). El primero discute sobre generalizaciones previas de patrones de error en estudios realizados para otros idiomas y ofrece una nueva perspectiva sobre patrones de error en español, mientras que el segundo se centra en el estudio de errores gramaticales y de motivación cognitiva.

Por último, se han llevado a cabo estudios sobre identificación y clasificación de errores en otros idiomas y dominios. Entre ellos, el mayor número está dedicado al inglés (Kukich, 1992; Yannakoudakis y Fawthrop, 1983; Pollock y Zamora, 1983; Mitton, 1985; Verberne, 2002; entre otros). En los últimos años, también han sido publicados estudios sobre patrones de error en portugués (Gimenes, Roman y Carvalho, 2015), en húngaro (Siklósi, Novák, y Prószéky, 2016), en japonés (Baba y Suzuki, 2012), en danés (Paggio, 2000) o en punjabi (Lehal y Bhagat, 2007). Es destacable el número de tipologías y estudios sobre patrones de error desarrollados en el ámbito de aprendizaje de lenguas (Nagata, Takamura y Neubig, 2017).

### 3 Descripción de la investigación propuesta

La tesis doctoral tiene como objetivo principal la tipificación y el análisis de errores lingüísticos en corpus del dominio médico, para contribuir así al desarrollo y perfeccionamiento de sistemas de corrección automática mediante el diseño de un módulo basado en conocimiento lingüístico.

En la primera etapa de la tesis, centrada en el estudio teórico, se ha investigado sobre el estado del arte en corrección automática y, más concretamente, sobre corrección automática en el dominio médico. Asimismo, se ha investigado sobre análisis de errores, criterios de clasificación y diseño de tipologías de error. Durante esta fase inicial comprobamos que existen numerosos trabajos sobre detección y corrección ortográfica automática, pero son extremadamente divergentes y con características muy dispares, por lo que se consideró útil realizar una revisión sistemática de la literatura para conocer los desafíos actuales que la corrección automática presenta específicamente en el lenguaje médico (López-Hernández, Almela & Valencia-García, 2019). Los objetivos principales fueron: identificar las técnicas y métodos utilizados en detección y corrección automática en el ámbito médico; recopilar recursos, corpus y bases de datos existentes; y conocer las principales limitaciones y problemas a los que se deben enfrentar los investigadores en esta área de investigación. Los resultados reflejaron que la combinación de métodos es esencial para lograr efectividad en el proceso de corrección. También es fundamental tener un diccionario lo más exhaustivo y completo posible, una tarea que no siempre es fácil debido a las características del dominio y la creación constante de neologismos por parte de los profesionales de la salud. Las técnicas utilizadas son diversas, el punto de partida es la búsqueda en diccionarios, seguido de la aplicación de algoritmos de distancia de edición ortográfica y de distancia de edición fonética, y la aplicación de métodos estadísticos y aprendizaje automático. La mejora del rendimiento en los últimos años es especialmente relevante debido al uso de sistemas que tienen en cuenta el contexto, como el modelado del lenguaje o el uso de redes neuronales. Los resultados de la investigación

indicaron que todavía queda trabajo por hacer para mejorar las medidas de precisión y exhaustividad (López-Hernández, Almela & Valencia-García, 2019).

En segundo lugar, se ha llevado a cabo el primer análisis de errores en un corpus formado por una recopilación de informes clínicos pertenecientes a la especialidad de urgencias (publicación pendiente de aparecer en la revista *Sintagma*<sup>1</sup>). En este trabajo se ha realizado un análisis cualitativo y una descripción de los tipos de errores detectados, y un análisis cuantitativo teniendo en cuenta variables como la frecuencia de aparición del error y la longitud de la palabra. Los distintos tipos de errores han sido identificados mediante el desarrollo de una herramienta de extracción y clasificación de errores. Teniendo en cuenta estos datos, se ha diseñado una primera tipología de error centrada en errores “non-word”, es decir, errores que dan lugar a palabras no existentes (por ejemplo, “antihipertensivo” en lugar de “antihipertensivo”). Hemos creado categorías adaptadas a las características del dominio y hemos establecido una serie de convenciones para que la tipología sea consistente, especialmente en el tratamiento de abreviaturas, siglas, acrónimos, anglicismos, neologismos, errores de puntuación, etc.

Posteriormente, la variabilidad de la muestra ha sido ampliada con la incorporación de informes clínicos de otras especialidades, como UCI, cirugía general y digestiva o psiquiatría. Por tanto, contamos con los primeros datos cuantitativos sobre patrones de error en corpus médicos en español y una tipología provisional de errores. Además, se han creado las primeras matrices de confusión para el tipo de error de sustitución.

Actualmente, estamos trabajando en la mejora de la fase de detección de errores, para que los resultados sean lo más precisos y representativos posibles. Para conseguirlo, entre otras tareas, vamos a incluir nuevas categorías, como la detección de errores en el uso de mayúsculas y minúsculas en el corpus.

Uno de los mayores retos tiene que ver con la detección de errores “real-word”, es decir, errores que dan lugar a palabras existentes en diccionarios (por ejemplo, “respeto” en lugar de “respecto” o “hemitórax” en lugar de “hemotórax”). Su identificación como error es sumamente compleja y ha sido muy poco

abordada, especialmente para el español, pues no existe tipificación previa. El análisis de este tipo de errores requiere de técnicas de procesamiento a nivel sintáctico y semántico que tengan en cuenta el contexto. De forma paralela, estamos investigando sobre errores motivados por similitud fonética y errores de tipo cognitivo.

La última parte de la investigación se centrará en la generación de candidatos, en ella se incorporará el módulo basado en conocimiento lingüístico (previamente definido a partir de los datos obtenidos en la fase descriptiva) al sistema de corrección, contribuyendo a la ponderación de alternativas y elección de sugerencias.

#### **4 Metodología y experimentos propuestos**

Los pasos principales y experimentos que componen la metodología de la tesis son los siguientes:

1. Investigación teórica sobre el estado del arte, delimitación del proyecto y estudio de diversas metodologías.
2. Constitución del corpus de estudio. En esta fase se lleva a cabo la recopilación de informes clínicos digitalizados, la compilación del corpus objeto de estudio y el preprocesamiento del mismo.
3. Análisis. Se realiza la identificación de patrones de errores, mediante el desarrollo de una herramienta de extracción, cómputo y clasificación. Los errores van a ser analizados sistemáticamente teniendo en cuenta criterios como frecuencia de aparición, distancia de edición, tipo de error (omisión, sustitución, inserción o trasposición), subtipo de error, existencia de multierror en la palabra, posición del error en la palabra, longitud de la palabra o contexto en el que se produce. También se llevará a cabo el diseño y estabilización de una tipología de error específica para el dominio médico, tras analizar si se presentan diferencias significativas entre las especialidades analizadas. En cuanto a los errores “real-word”, se detectarán mediante técnicas como el análisis de n-gramas y el uso de modelos lingüísticos. Además, se hará una recopilación de casos para

---

<sup>1</sup> <http://www.sintagma.udl.cat/es/>

el diseño de un set de confusión, formado por parejas de palabras que tienden a ser confundidas de forma significativa en el corpus. Tras la obtención de los resultados en esta fase, crearemos matrices de confusión que serán utilizadas para el desarrollo del módulo basado en conocimiento lingüístico.

4. Diseño e implementación de propuesta de mejora basada en conocimiento lingüístico. Formalización del módulo basado en conocimiento a partir de los datos recopilados en la fase de identificación, análisis y clasificación. Integraremos el módulo en el sistema de detección y corrección, cuya arquitectura está compuesta por otras técnicas estadísticas y puramente computacionales, y definiremos una fórmula de decisión y ponderación de las alternativas generadas automáticamente para las palabras erróneas. Finalmente, llevaremos a cabo la prueba y validación del prototipo, con distintos experimentos y métricas que reflejen la cobertura real y el grado de precisión que podemos alcanzar mediante la combinación de técnicas y la integración del módulo diseñado.

## 5 Cuestiones de investigación propuestas para discusión

En correspondencia con el propósito de trabajo establecido, consideramos de interés plantear las siguientes preguntas para discusión:

- ¿La incorporación de un módulo basado en conocimiento junto con la combinación de técnicas y criterios de elección va a aumentar la precisión de los métodos de corrección?
- ¿Es pertinente hacer uso de sets de confusión para la detección de errores real-word? ¿Cómo se podría establecer el límite idóneo en la incorporación de pares de palabras que generan confusión o son susceptibles de ser confundidas?
- ¿Cómo podría aplicarse el análisis sintáctico para la detección de errores real-word? ¿O para la detección de coocurrencias de errores en la oración?
- ¿Qué factores del análisis de errores pueden ser más útiles para un método de corrección automática? ¿Posición del error, longitud de la palabra, distancia de edición, tipo de error, etc.?

- ¿De qué forma puede complementar la tipificación y análisis de errores a las técnicas basadas en redes neuronales?

## Agradecimientos

Esta investigación está financiada por el Ministerio de Universidades de España a través del Programa Nacional de Ayudas para la Formación de Profesorado Universitario (FPU).

## Bibliografía

- Ahmed, F., E. W. Luca, y A. Nürnberger. 2009. Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. *Polibits*, 40:39–48.
- Baba, Y. y H. Suzuki. 2012. How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, páginas 373–377, Jeju Island (Corea).
- Brill, E. y R. C. Moore. 2000. An improved error model for noisy channel spelling correction. En *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics - ACL*, páginas 286–293, Hong Kong (China).
- Damerau, F.J. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of ACM*, 7(3):171–176.
- Díaz Villa, A. 2005. Tipología de errores gramaticales para un corrector automático, *Procesamiento del Lenguaje Natural*, 35:409–416.
- Fivez P., S. Suster y W. Daelemans. 2017. Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings. En *Proceedings of the BioNLP workshop – Association for Computational Linguistics*, páginas 143–148, Vancouver (Canada).
- Gimenes, P. A., N. T. Roman y A. M. Carvalho. 2015. Spelling Error Patterns in Brazilian Portuguese. *Computational Linguistics*, 41(1):175–183.
- Jurafsky, D. y J. Martin. 2014. *Speech and Language Processing*, Pearson Education, 480-493.

- Kukich, K. 1992. Technique for automatically correcting words in text. *ACM Computing Survey*, 24(4):377–439.
- Lai, K. H., M. Topaz, F. R. Goss, y L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55:188–195.
- Lehal G. S. y M. Bhagat. 2007. Spelling Error Pattern Analysis of Punjabi Typed Text. En *Proceedings of the 2007 International Symposium on Machine Translation, NLP and TSS*, páginas 128–141.
- Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.
- López-Hernández J., Á. Almela y R. Valencia-García (2019). Automatic Spelling Detection and Correction in the Medical Domain: A Systematic Literature Review. En R. Valencia-García, G. Alcaraz-Mármol, J. Del Cioppo-Morstadt, N. Vera-Lucio y M. Bucaram-Leverone (Eds.) *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science*, 1124, páginas 104–117. Cham: Springer.
- Mitton, R. 1987. Spelling Checkers, Spelling Correctors, and the Misspellings of Poor Spellers, *Information Processing & Management*, 23(5):495–505.
- Nagata, R., H. Takamura, y G. Neubig. 2017. Adaptive Spelling Error Correction Models for Learner English. *Procedia Computer Science*, 112:474–483.
- Naber D. 2003. A Rule-Based Style and Grammar Checker. Diploma thesis, University of Bielefeld.
- Paggio, P. 2000. Spelling and grammar correction for Danish in SCARRIE. En *Proceedings of the Sixth Conference on Applied Natural Language Processing*, páginas 255–261, Seattle (Washington).
- Pande, H. 2017. Effective search space reduction for spell correction using character neural embeddings. En *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 170–174, Valencia (España).
- Patrick, J., M. Sabbagh, S. Jain y H. Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. En *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, páginas 1–8, Valletta (Malta).
- Pollock, J. J. y A. Zamora. 1983. Collection and characterization of spelling errors in scientific and scholarly text, *Journal of American Society of Informatics and Science*, 34(1):51–58.
- Ramírez, F. y E. López. 2006. Spelling Error Patterns in Spanish for Word Processing Applications. En *Proceedings of Fifth international conference on Language Resources and Evaluation, LREC*, páginas 93–98, Genoa (Italy).
- Ruch, P., R. Baud y A. Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, *Artif. Intell. Med.* 29 (1):169–184.
- Siklósi, B., A. Novák, y G. Prószéky. 2016. Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, 35:219–233.
- Verberne, S. 2002. Context-sensitive spell checking based on trigram probabilities. Master's thesis, University of Nijmegen.
- Veronis, J. 1988. Computerized correction of phonographic errors. *Computers and the Humanities*, 22(1):43–56.
- Wedbjer Rambell, O. 1999. Error typology for automatic proof-reading purposes. En A. Sagvall Hein, editor, *Reports from the SCARRIE project*, Uppsala.
- Wong, W. y D. Glance. 2011. Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. *Artificial Intelligence in Medicine*, 53(3): 171–180.
- Yannakoudakis, E. J. y D. Fawthrop. 1983. The rules of spelling errors, *Information processing and management*, 19(12):101–108.