Original Research

SAGE Open
January-March 2023: 1–11

DOI: 10.1177/21582440221148454
journals.sagepub.com/home/sgo
SAGE

Analysis of Context-Dependent Errors in the Medical Domain in Spanish: A Corpus-Based Study

Jésica López Hernández[1], Fernando Molina Molina[2], and Ángela Almela[1]

Abstract

This corpus-based study aimed to investigate the presence of context-dependent linguistic errors in a corpus of clinical reports. The data were taken from a corpus comprising more than 2 million words and made up of clinical reports from emergency medicine, intensive care unit, general surgery, and psychiatry. Quantitative and qualitative analyses were carried out. A language model based on n-grams was developed for the detection of errors, parameters for the selection of cases were defined, and a classification tool was implemented. The findings indicated that emergency medicine was the medical specialty with the highest number of context-dependent errors and that the most frequent type of error was omission of written accent. Furthermore, the analysis revealed the presence of errors of competence due to the incorrect application of the linguistic norm of Spanish, phenomena of phonetic similarity, and composition of words; it is also worth noting that performance errors occurred due to rapid typing on the keyboard. This study constituted the first analysis and creation of a typology of context-dependent errors for the medical domain in Spanish. It contributed to the design of a module based on linguistic knowledge that can be used for the development and improvement of automatic correction systems that, in turn, are used for data processing in medicine.

Keywords

automatic spell checking, classification of linguistic errors, context-dependent errors, error analysis, medical corpus

Introduction

Medical language is language for specific purposes, therefore, it must convey concepts and ideas with accuracy, precision, and clarity to facilitate the exchange of information (Bello Gutiérrez, 2016). In the medical domain, it is especially important to use technologies based on automatic data processing to facilitate the extraction and classification of clinical information. However, medical language presents linguistic features that make its automatic processing difficult, such as its richness and lexical complexity, since it is an extremely productive specialized language (Navarro, 2015).

The difficulties are increased by the presence of numerous linguistic errors in the case of clinical reports. Clinical reports show frequent errors caused by fast typing on the keyboard, due to the short time available for their composition by specialists, among other reasons.

Errors are usually classified into two large groups in automatic correction: non-word errors and real-word errors (Jurafsky & Martin, 2018). The first are those that result in a word that does not exist in that language (e.g., *coeficient* for *coefficient*), that is to say, spelling errors. Contrastingly, real-word errors give rise to an existing word that is orthographically correct, but incorrect in that particular context (*dessert* for *desert*, or *their* for *there*); for that reason, they are also called context-dependent errors. These errors affect the syntactic and semantic level, and their causes can be cognitive, typographic or phonetic, among others (Kim et al., 2013). The concepts of context-dependent errors and real-word errors will be used interchangeably throughout this work. Numerous tools have been developed to address spelling error correction, but context-dependent errors

[1]Universidad de Murcia, Spain
[2]VÓCALI Sistemas Inteligentes, S.L., Murcia, Spain

Corresponding Author:
Jésica López Hernández, Facultad de Informática, Universidad de Murcia, Campus de Espinardo, Murcia 30100, Spain.
Email: jesica.lopez@um.es

still remain a challenge for researchers. There is no previous research analyzing the presence of this type of errors in clinical texts in Spanish. Currently, only Bravo-Candel et al. (2021) implements a model based on deep learning to correct context-dependent errors in clinical texts in Spanish. This study confirms the need to better know the variability of the errors that appear in this domain for improvement.

Therefore, the principal objective of this research is the detection of context-dependent linguistic errors present in clinical reports in Spanish for their subsequent analysis and classification. The research questions of this study are as follows:

- How many context-dependent errors occur in clinical reports in Spanish?
- What type of errors occur most frequently?
- In which specialty are there more mistakes?
- Are the results consistent with previous studies?
- What are the main difficulties in detecting context-dependent errors?

The paper is divided into the following parts: Section 2 presents the latest research on automatic error correction and analysis of errors in clinical documentation. Section 3 is devoted to the methodological framework, which includes a description of the corpus used, the research design and the procedures for error detection, analysis, and classification. Section 4 presents the results obtained, which are discussed in Section 5. Last, Section 6 provides the reader with the most relevant conclusions and suggestions for future research.

Error Analysis and Automatic Correction in the Medical Domain

In recent years, various research projects have arisen in the field of natural language processing that aim to develop resources for healthcare. They intend to process large amounts of data by developing systems for the classification of information or the recognition of named entities. However, the effectiveness of these systems is conditioned by the presence of errors. Thus, the correction process becomes an essential phase and is usually part of larger applications, such as voice recognition or information synthesis.

Various investigations have been carried out on automatic detection and correction in clinical documentation and, specifically, in clinical reports. Most of these studies have been developed with corpora in English (e.g., Fivez et al., 2017; Workman et al., 2019), but these issues have been also explored for French (D'Hondt et al., 2016), Russian (Balabaeva et al., 2020), Swedish (Dziadek et al., 2017), Dutch (Fivez et al., 2017), Hungarian (Siklósi et al., 2016), and Persian (Yazdani et al., 2020). These works highlight the substantial number of linguistic errors that the corpora comprising clinical reports usually contain, with studies whose error rate is around 5% (Lai et al., 2015) or even 10% (Ruch et al., 2003). Most of these works explore spelling errors, which tend to occur more frequently than syntactic or semantic errors. Furthermore, studies on errors that depend on the context are limited due to the difficulties of detection that they pose.

Among the most widely used detection and correction techniques, we must highlight dictionary search and editing distance, which appear in all the studies revised by the present authors. Likewise, the use of statistical methods and the analysis of n-grams stand out (Lu & Demner-Fushman, 2018; Siklósi et al., 2016; Wong & Glance, 2011). Some studies also make use of methods based on rules and regular expressions (Lai et al., 2015; Thompson et al., 2015); others incorporate POS-tagging (Hussain & Qamar, 2016; Zhou et al., 2015); or the use of ontologies (Dziadek et al., 2017). Finally, some of the most recent works add techniques based on neural networks (Fivez et al., 2017; Workman et al., 2019). Hybrid approaches and a combination of different techniques are used by most of them to improve system performance. A common feature of most systems that correct context-dependent errors is the use of confusion sets and language models based on n-grams (Azmi et al., 2019; Sharma & Gupta, 2015). According to Pedler (2007), a confusion set is a collection of linguistic units that can be confused (actual and predictions) and are represented in a contingency table. Words can be confused with each other due to phonological, graphic, or grammatical similarity. Jurafsky and Martin (2018) provide some examples of prototypical confusions in English such as *peace*/*piece*, *among*/*between*, and *affect*/*effect*.

In the case of Spanish, Bravo-Candel et al. (2021) proposes the use of a neural machine translation seq2seq model to correct these errors. The corpus used to train and evaluate the model was compiled from Wikicorpus and a medical corpus made up of clinical cases from CodiEsp, MEDDOCAN, and SPACC. However, these corpora were reviewed and did not contain real errors, so synthetic errors were created and added by applying linguistic rules. Among the types of errors introduced are gender and number mismatch errors, and confusion between homophone words.

Studies on types of errors in Spanish clinical reports have been recently carried out; these works have focused on spelling errors and non-word errors. For instance, in López-Hernández et al. (2021), a corpus of emergency medicine was analyzed and the authors concluded that the length of the words is a variable that influences the frequency of errors. This paper also highlights the

Table 1. Tokens, Types, and STTR in the Corpus.

Medical specialty	Tokens	Types	Standardized type/token ratio (STTR)
Emergency medicine	730,468	25,870	40.91
ICU	725,690	22,897	43.63
Psychiatry	424,775	19,489	45.25
General surgery	440,893	14,697	42.04

Source. Taken from López-Hernández and Almela (2021).

significant frequency with which the prefixes are separated from the lexical base which they complement, and the lack of consistency in the use of abbreviations. Likewise, López-Hernández and Almela (2021) provides the first quantitative analysis of non-word errors and compile the most frequent erroneous substitution patterns. On the other hand, errors occurring in biomedical publications in Spanish (Aguilar Ruiz, 2013) or in terminological dictionaries (Rodríguez-Rubio, 2018) have also been analyzed. However, there are no works investigating the presence of context-dependent errors or real-word errors in clinical reports.

Furthermore, spelling, grammar, and style checkers have been developed for Spanish, such as Stilus or LanguageTool; however, they are systems developed to detect errors in the general domain, not for texts from specialized domains such as health, which may affect performance. Most of them are commercial systems whose use is not publicly available. On the other hand, CorrectM and Spellex Medicine have been specifically designed for the medical domain. The first one is a plain text editor with a spell checker that contains medical terminology, and the second is a spell checker that is integrated into Microsoft Office tools. Therefore, they recognize medical and pharmacological terms and provide suggestions to replace incorrect words; both work with spelling errors, checking for spelling and recognizing specialized terms that common checkers do not contain, but detecting no grammatical or semantic errors.

Methodology

Corpus Information

The corpus analyzed is made up of a collection of electronic clinical reports. These reports contain information on the medical specialties of emergency medicine, intensive care unit (henceforth, ICU), psychiatry, and general surgery. The clinical report is a document issued by the medical practitioner with information on a healthcare process provided to the patient. In terms of structure and content, this type of reports present specific details according to the specialty, but all include sections such as the following: background, reason for consultation, examinations, diagnosis, treatment, and guidelines for action. The corpus used in the present study is monolingual and private, owned by Vócali (https://vocali.net/), a company which uses it for the development of speech recognition software solutions in the biomedical field. It is made up of a set of unstructured plain text files and is not tagged. It was subjected to preprocessing to standardize its format and eliminate HTML and XML tags. In addition, it is worth noting that the clinical reports were completely anonymized and did not contain information about patients, doctors, dates, or places, as established by the General Data Protection Regulation (https://gdpr-info.eu). The reports were typewritten on a computer directly by the doctors, they were not transcribed or subsequently revised; thus, it is a real reference of errors in a professional context.

The corpus contains a total of 2,321,826 words; Table 1 (taken from López-Hernández & Almela, 2021) presents the total amount of tokens and types that the corpus contains according to medical specialty. The concept "token" refers to the total number of words in a corpus, while "type" designates the number of distinct words. From those counts, the type/token ratio (TTR) was calculated in order to measure the lexical richness of the sample (Chipere et al., 2004). Due to the difference in size between the emergency and ICU subcorpus with respect to psychiatry and general surgery, the corresponding standardized proportion was calculated with WordSmith Tools in order to appropriately compare corpora with different lengths (López-Hernández & Almela, 2021). As far as this indicator is concerned, the specialty with the greatest lexical richness is psychiatry, due to the fact that the reports have a narrative structure; on the contrary, emergency and general surgery specialties have a much more synoptic structure, which influences the results.

Data Analysis

Spelling Error Detection and Correction. In order to detect real-word or context-dependent errors, it was first necessary to detect and correct all the spelling errors in the corpus. To detect spelling errors, dictionary searches were used, which are based on the automatic comparison of the corpus with a lexicon made up of previously

validated words. The lexicon was compiled from Hunspell's Spanish dictionary, an open source spell checker, containing inflected and derived word forms. A dictionary of general Spanish was insufficient because it did not include specialized lexicon, so biomedical terminology compiled from various sources was added. Among them, the systematized nomenclatures of Snomed-CT and CIE-10, specialized glossaries, and lexical resources obtained from sources such as the Spanish Agency for Medicines and Health Products (AEMPS) and the Spanish Society for Medical Documentation (SEDOM) can be highlighted. The words from the corpus included in the lexicon were considered as correct, while those that did not appear on the list were considered as error candidates.

Subsequently, the errors detected were corrected. For the generation of correction suggestions, the minimum editing distance technique was used, known as Damerau-Levenshtein distance. This distance corresponds to the number of operations required to transform one group of characters into another, and the basic types are insertion, omission, substitution, and transposition (Damerau, 1964; Levenshtein, 1966). Insertion occurs when an extra character is added to the word; omission takes place when the word is missing a character; substitution occurs when the word contains a wrong character instead of the corresponding one, and transposition occurs when the position of two adjacent characters is swapped. The frequency of appearance of each suggestion in the corpus and the analysis of n-grams were also taken into account for the ordering of the list of suggestions. Finally, the corrections obtained were validated by assisted manual review and, when required, the context of the error was also checked.

Context-Dependent or Real-Word Error Detection. After correcting the spelling errors in the corpus, we went through the experiment phase to detect real-word errors. These errors had gone unnoticed in the first phase of automatic correction. Accordingly, a proposal based on statistical language modeling using n-grams were developed for the detection of errors in context. The probability distribution of corpus elements was obtained by means of a language model and, therefore, it was based on a probabilistic description of the language phenomena.

SRI Language Modeling (SRILM) was used, with which it is possible to create statistical language models for various natural language processing applications, such as voice recognition, summarization, named-entity recognition, and machine translation, among others. The execution of the algorithm resulted in a linguistic model of order 3 that was trained from the elements of the corpus (2,321,826 words). The output file in ARPA format was divided into n-gram sections according to length. Each n-gram appeared on a line accompanied by a logarithm (in base 10) of the conditional probability of that n-gram in the corpus. The linguistic model resulted in probabilities that accurately reflected the frequency of combination of the words that make up the corpus. This made it possible to detect anomalous combinations that must be analyzed to find possible errors. A low frequency of appearance of the bigram or trigram in the corpus or a frequency lower than expected may be an indication that it is an abnormal combination of words and that there is an error.

Table 2. Examples of Cases Detected in the Linguistic Model.

abundantes líquido (−3.270262)
abundantes liquido (−3.480262)
abundante líquido (−0.203338)
abundantes líquidos (−0.293051)
gasometría vernosa (−3.966578)
gasometría venosa (−0.759682)
antecedente médicos (−2.818827)
antecedentes médicos (−0.800204)
localidad neurológica (−3.380874)
focalidad neurológica (−0.135904)
anorexia nervios (−2.87922)
anorexia nerviosa (−0.328551)

The next step in the process was to automatically generate alternatives using minimum editing distance, and compare the values that these alternatives had in the linguistic model. The minimum editing distances 1 and 2 were applied to each n-gram (with insertion, substitution, omission, and transposition operations) to obtain alternative bigrams and trigrams. The new bigrams and trigrams had to be made up from existing words, so the system was asked to only return candidates made up of correct words that were in the dictionary. A different number of alternatives were obtained from each original bigram and trigram, depending on the number of existing words close to their spelling. The objective of this phase was to check the scores of each original bigram and trigram and its nearby alternatives to analyze which ones were more likely and if there was a significant difference between them. The value that appeared for each of them was compared and, if the difference was above the set threshold, it was detected as a bigram or trigram that could contain an error (see Table 2).

Classification and Analysis of Real-Word Errors. In the final phase, a calculation and error classification tool were developed. The calculation tool was based on searching for text strings and regular expressions in the corpus to quantify how many times previously detected errors occurred in the corpus. The frequency of appearance of errors in each of the specialties was obtained and a list of results was compiled. This list of results was classified

Table 3. Non-Word and Real-Word Errors According to Medical Specialty.

Medical specialty	Tokens	Non-word errors		Real-word errors	
		Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.
Emergency medicine	730,468	35,819	4.90	1,222	0.16
ICU	725,690	19,890	2.74	511	0.07
Psychiatry	424,775	7,271	1.71	477	0.11
General surgery	440,893	11,175	2.53	264	0.05
Total	2,321,826	74,155	3.19	2,474	0.11

Table 4. Real-Word Errors According to the Type of Editing Operation.

Type of error	Emergency medicine		ICU		Psychiatry		General surgery	
	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.
Omission	936	76.60	404	79.06	388	81.34	214	81.06
Insertion	188	15.38	67	13.11	58	12.16	17	6.44
Substitution	95	7.77	39	7.63	31	6.5	32	12.12
Transposition	3	0.25	1	0.20	0	0	1	0.38

with a tool generated for this task. The operation was based on the comparison between the wrong character string and the respective assigned correction to obtain information about the type of editing operation and the editing distance.

Results

Quantitative Analysis

The quantitative data obtained after the detection and classification of errors are presented below. The presence of context-dependent errors or real-word errors is notably lower than spelling errors or non-word errors; a total of 76,629 errors have been detected, of which 2,474 (3.23%) are real-word errors. Table 3, which includes absolute and relative frequency (in percentage form), shows the results obtained for each specialty. The specialty with the highest number of non-word and real-word errors is emergency medicine.

Context-dependent error types are quantified according to the editing operation in Table 4. The type of error that is made most frequently in all specialties is omission, with a significant difference over the rest. The next most common type of error is insertion in emergency medicine, ICU, and psychiatry, and substitution in third place, while in general surgery the second most frequent type of error is substitution, followed by insertion. Finally, the transposition error is hardly present in the corpus.

As can be seen in Figure 1, at a higher specification level, it is detected that the subtype of error with the strongest presence is omission of the written accent (*diagnostico principal* [diagnóstico principal]), followed by omission of letter (*ambos brazo* [ambos brazos]). The correct version of the examples is offered in brackets. In emergency medicine, ICU and psychiatry, the third subtype of error is letter insertion (*cardiofrénicos liberes* [cardiofrénicos libres]), while in general surgery it is letter substitution (*color local* [calor local]).

Likewise, Table 5 includes some in-context examples of all the types of errors detected according to the editing operation.

Qualitative Analysis

Next, a qualitative classification and description of the errors detected is provided (see Table 6), as well as an explanation of the possible causes. Many of the real-word errors seem to be caused by fast typing on the computer keyboard. As a result, two keys can be pressed instead of one, an adjacent key can be accidentally pressed on the keyboard instead of the corresponding one, a letter can be omitted, or they can be typed in the wrong order. The practitioner must type fast and usually does not have time for subsequent review. Therefore, there could be performance errors that occur accidentally and are motivated by non-linguistic factors, such as distractions or mechanical problems, not by ignorance of the linguistic norm or cognitive motivation. However, some errors that could be considered as cognitively motivated, or competence errors, are also detected, since they imply the ignorance of orthographic rules that regulate the functioning of the language (Díaz Villa, 2005). The corpus contains words with a written accent that should

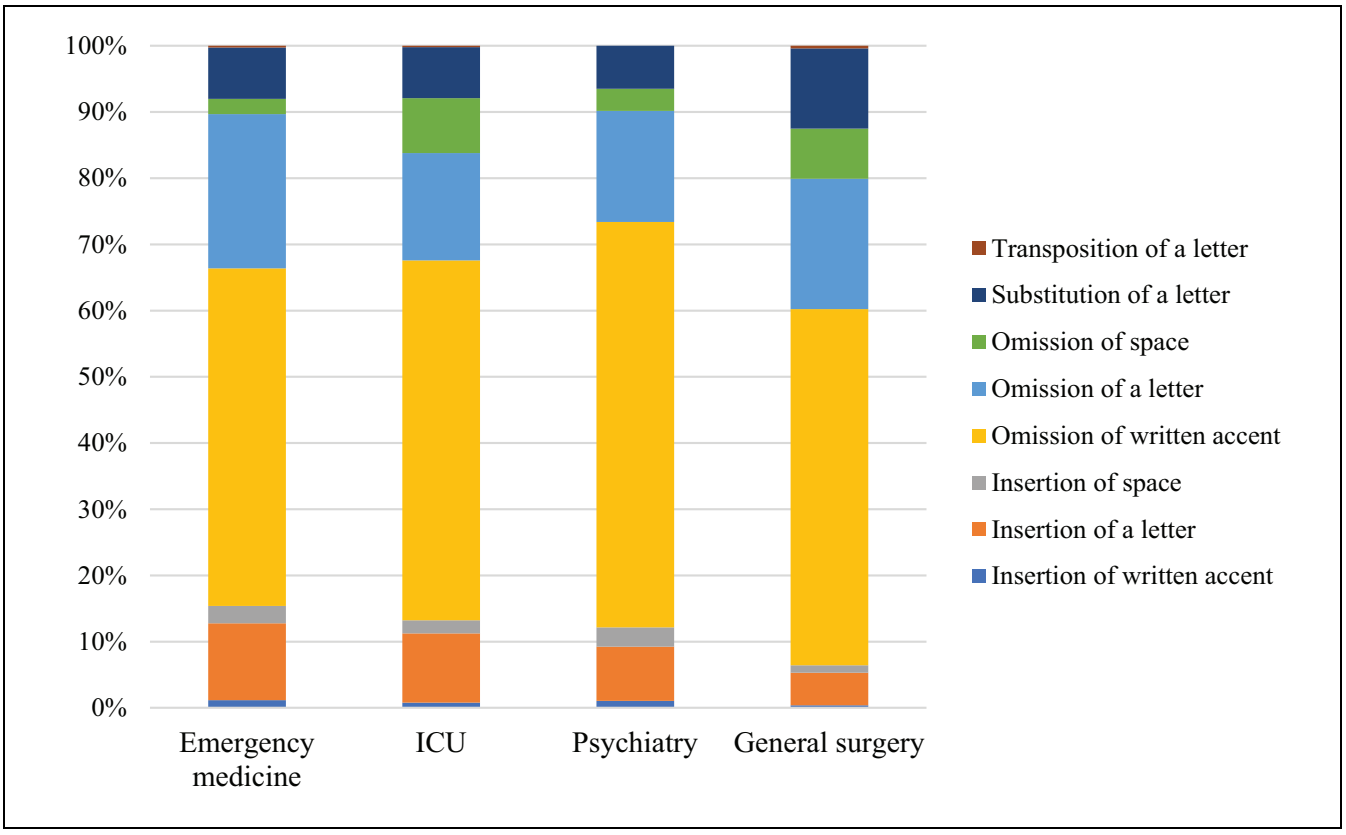

Figure 1. Real-word errors according to the editing operation subtype.

not include it, such as monosyllables and demonstratives (*dió* [dio], *fué* [fue], *éstos* [estos]), or that have it written on the inappropriate syllable for that context of use (*segmentarías* [segmentarias], *mamá* [mama], *colón* [colon], *broté* [brote]). However, there are multiple cases in which it is difficult to define with complete certainty whether it is a cognitive or a performance error, such as in the use of B/V, because they are very close on the QWERTY keyboard, or on the omission of written accents.

Furthermore, an intentional use of non-standard writing is noteworthy. An illustrative example of non-standard spelling is the omission of the written accent because its use involves pressing two keys on the keyboard. This omission creates ambiguity to natural language processing systems. Another common example is capitalizing common nouns to emphasize them.

Paronymy. Paronyms are words that are similar in sound or form (graphic or phonetic proximity), or even by their etymology, but have different meanings. Most phonetic errors occur due to similarities in pronunciation. This is the case of homophones, which are pronounced identically but are written differently and their meanings also differ. Confusions also arise from verbs (*halla*/*haya, haber*/*a ver, echo*/*hecho*) and phonemes with more than one graphic possibility to be represented (*bulbar*/*vulvar*). In the group of errors motivated by paronymy, words whose difference is the insertion of a written accent, or words whose form is similar but differ in the stressed syllable, can also be included. In the case of the latter, when the stressed syllable varies, it entails a change in meaning, giving rise to accentual paronyms, or verbs conjugated in different person, modes, or tenses (*diagnostico*/*diagnosticó*/*diagnóstico*).

Grammatical Agreement. The presence of syntactic errors in the corpus is particularly notable, specifically errors related to grammatical agreement. Nominal agreement refers to the coincidence of gender and number between constituents, while verb agreement refers to number and person, and is the one that occurs between verb and subject. The cases detected occur frequently between nouns and adjectives that accompany them. In order to analyze the possible errors of verbal discordance between verb and subject, it would be necessary to expand the context window.

Word Formation. In addition to errors related to issues of accentuation and tonicity, phonetic proximity or

Table 5. Real-Word Errors According to the Type of Editing Operation.

Type of editing operation	Original example	Translated example
Omission of a letter	*Poliartrosis con lumbalgia cónica** [crónica].	- Polyarthrosis with conical* low back pain. - Polyarthrosis with chronic low back pain.
	*Con diuresis mantenida, con buena situación cínica** [clínica].	- With sustained diuresis, with good cynical* situation. - With sustained diuresis, with good clinical situation
Omission of written accent	*Fractura de vertebra** [vértebra] *lumbar.*	- Lumbar fracture support*. - Lumbar vertebra fracture.
	*Habito** [hábito] *enólico ocasional.*	- Occasional enolic inhabit*. - Occasional enolic habit.
Omission of space	*Tensiones arteriales entorno** [en torno] *a 100/40 mmHg.*	- Surrounding* arterial tensions 100/40 mmHg. - Arterial tensions around 100/40 mmHg.
	*Doloroso en hemiabdomen inferior sobretodo** [sobre todo] *fosa ilíaca izquierda.*	- Painful in the lower hemiabdomen raincoat* left iliac fossa. - Painful in the lower hemiabdomen especially left iliac fossa.
Insertion of a letter	*Paciente muestras** [muestra] *rasgos destacados en esta escala.*	- Patient samples* prominent features on this scale. - Patient shows prominent features on this scale.
Insertion of written accent	*Última cita está** [esta] *mañana.*	- Last date is* tomorrow. - Last date this morning.
	*Monitorización pulsioximétrica continúa** [continua].	- Pulse oximetry monitoring continues*. - Continuous pulse oximetry monitoring.
Insertion of space	*No ingresos hospitalarios a parte** [aparte] *de las cirugías.*	- No hospital admissions of a part* of the surgeries. - No hospital admissions apart from surgeries.
	*No enfermedades médico quirúrgicas** [medicoquirúrgicas] *de interés.*	- No doctor surgical* diseases of interest. - No medico-surgical diseases of interest.
Substitution of a letter	*STOP-COLD dada** [cada] *ocho horas.*	- STOP-COLD given* eight hours. - STOP-COLD every eight hours.
	*Presenta crisis mioclónicas sobre todo en hombre** [hombro] *derecho.*	- He presents myoclonic crisis especially in a right man*. - He presents myoclonic seizures especially in the right shoulder.
Transposition of a letter	*Se decide lata** [alta] *a planta.*	- It is decided can* to the plant. - Discharge to the plant is decided.
	*Fue dado de lata** [alta] *con hematoma hepático.*	- He was canned* with a liver hematoma. - He was discharged with a liver hematoma.

grammatical mismatches, errors related to the union and separation of words are found. Some pairs of words used wrongly in the corpus are: *por que*/*porque*/*por qué, si no*/*sino, acerca*/*a cerca, sobre todo*/*sobretodo, entorno*/*en torno, aparte*/*a parte*. Furthermore, the presence of specialized terminology formed by composition and prefixation mechanisms is very common in the medical domain. In these cases, the error occurs when the two elements that make up the lexical unit are separated by space. Both words are independently correct, which means that they go unnoticed in the usual detection process (*ansioso depresivo* [ansiosodepresivo], *médico quirúrgicas* [medicoquirúrgicas]). On the other hand, in Spanish a prefixed word must be written as one whole word without a hyphen, but cases are also detected in which there is a space between the prefix and the word.

Abnormal Verb Forms in the Domain. In addition to the errors already classified, some verb forms that appear in the corpus should be marked as error candidates in a medical corpus because they tend to mask errors. First, clinical reports are usually written in an impersonal style and third person is mostly used. Consequently, it is not usual to find forms in the first person, beyond the use of so-called communication verbs. Therefore, first-person forms of the present indicative hide errors. Second, there are forms in the imperative in the corpus, a verb form that is used to give orders and advice. Specifically, the appearance of this form occurs due to the omission of the final character of the word, turning the adjective or participle into an imperative form in the second person plural. It also occurs by substituting the consonant "s" for the consonant "d" when writing words in the plural, due to the adjacency of both letters on the keyboard. Third, subjunctive forms are detected, which are used to express suppositions, wishes or possibilities. Therefore, its use in clinical reports will not be common. The error is usually the omission of the character "s" or the insertion of the vowel "e." Finally, second person verbs, in which a recipient is appealed directly, are error candidates too, as their presence in the corpus is improper and usually masks an error.

Table 6. Qualitative Classification of Errors.

Paronymy	
Homophones	*cayo/callo* *maya/malla*
Diacritical written accent	*dé/de* *sí/si*
Accentual paronyms	*liquido/líquido/liquidó* *vomito/vómito/vomitó*
Phonetic proximity	*cociente/consciente* *consiente/consciente*
Grammatical agreement	
Nominal gender mismatch	*niveles hidroaéreas* [niveles hidroaéreos] *varias días* [varios días]
Nominal number mismatch	*cólico nefríticos* [cólicos nefríticos] *tres paquete* [tres paquetes]
Nominal gender and number mismatch	*hepatopatías crónico* [hepatopatías crónicas] *antecedentes traumática* [antecedentes traumáticos]
Verb discordance	*se diagnostica patologías* [se diagnostican patologías] *se visualiza lesiones* [se visualizan lesiones]
Word formation	
Word formation by composition	*uretero vesical* [ureterovesical] *dorso lumbar* [dorsolumbar]
Word formation by prefixation	*post traumático* [postraumático] *pre menstrual* [premenstrual]
Confusion between pairs of words	*entorno/en torno* *a ver/haber*
Abnormal verb forms in the domain	
First person present	*transito* [tránsito] *trafico* [tráfico] *homologo* [homólogo]
Second person	*orbitas* [órbitas] *ultimas* [últimas]
Plural imperative	*colapsad* [colapsado] *etiquetad* [etiquetas]
Subjunctive	*comprobare* [comprobar] *señale* [señales]

Discussion

In this section, the results presented above are discussed further as they relate to each research question posed:

How Many Context-Dependent Errors Occur in Clinical Reports in Spanish?

First of all, the number of errors detected in clinical reports is significant. A total of 76,711 errors have been detected in a corpus of 2,321,826 tokens, which represents an error rate of 3.3%. It is verified that the presence of context-dependent errors (real-word errors) is notably lower than that of non-word errors in the corpus. 2,474 real-word errors were detected, a percentage of 3.23% of the total errors. It is an expected result, since they are a type of error that requires specific conditions, namely, that the error caused gives rise to an existing word. However, it is a relevant type of error due to the problems of detection and, as a consequence, of automatic processing that it poses.

What Type of Errors Occur Most Frequently?

Most of the errors detected have been concentrated in a small number of characters, such as vowel characters in which the omission or insertion of a written accent has occurred. The omission of a written accent is the type of error that has occurred most frequently. Errors have also been generated by the omission of characters, especially at the end of the word, or the substitution of vowel characters for others (o-a, a-o, a-e, and e-i), giving rise to agreement errors in particular. Other types of errors detected have been the insertion of a letter, written accent or space, while transposition errors have hardly been present as triggers of real-word errors. Likewise, errors motivated by paronymy phenomena have been detected; nominal and verbal mismatch errors; errors caused by the incorrect union or separation of words; errors due to ignorance of the current orthographic norm; incorrect verb forms in the context; and, lastly, lexical errors caused by errors of substitution, omission, or insertion of a character.

Are the Results Consistent With Previous Studies?

The results obtained are consistent with the previous studies revised, which coincide in highlighting the high presence of writing errors in clinical documentation. This phenomenon represents an obstacle to patient understanding and hinders automated computer processing. In the case of Ruch et al. (2003), the percentage of error reaches 10% in the French clinical reports analyzed; Nizamuddin and Dalianis (2014) report an error rate of 7.6% in a corpus of medical reports in Swedish; Lai et al. (2015) detect 4% in reports in English; Ringler et al. (2017) mention 3.2% in radiology reports in English; and Liu et al. (2012) estimate that there is around 0.4% of spelling errors in medical documents in English. Therefore, fluctuations in the results are observed depending on the language and environment. In addition, these authors point out that the most common error patterns involve the addition, omission and transposition of characters, the inappropriate use of punctuation marks, grammatical errors, and overuse of abbreviated forms (Siklósi et al., 2016). According to the studies for non-word or spelling errors by Damerau (1964), Kukich (1992), Ramírez (2006), Gimenes (2015), and López-Hernández and Almela (2021), most of the errors present in the corpus tend to be unique cases of error in the word. These investigations provide quantitative data on errors in clinical documentation in languages such as English, French, Swedish, or Hungarian, but the results obtained in the present study cannot be directly contrasted with previous data on Spanish. No research has been found that has particularly focused on the automatic detection and quantitative analysis of context-dependent errors in medical reports in Spanish.

In Which Specialty Are There More Mistakes?

The specialty with the highest number of errors of both types is emergency medicine. This result may be due to the circumstances of this specialty, such as the limitation of time to write the reports and the need for an immediate response to the patient. In addition, it is a specialty that covers a wide spectrum of medical language, with great variability of etiologies and diseases, which can cause more difficulties at the terminological level.

What Are the Main Difficulties in Detecting Context-Dependent Errors?

Although the approach used allows detecting a large number of context-dependent linguistic errors and providing a first classification, it has some limitations that must be considered in future research. Among them, the detection of false positives is worth noting, that is to say, bigrams and trigrams whose score in the linguistic model was likely to be an anomalous combination, but in reality they were correct cases with little presence in the corpus. The opposite case can also occur, namely errors that have remained undetected in our investigation, such as errors related to punctuation and especially those cases whose elements are far from each other in the sentence and give rise to semantic inconsistencies.

Conclusions

All in all, this work provides the first analysis and compilation of real-word errors or context-dependent errors in electronic clinical reports in Spanish. To this end, a method based on an n-gram language model has been implemented. The results indicate that the presence of this type of error is considerably weaker than that of spelling errors. However, to improve the accuracy of correction systems in the medical domain, it is essential to address and develop techniques for the detection of these hidden errors. Likewise, this study has made it possible to verify the variability of types of errors that must be taken into account. Patterns that appear frequently in the corpus have been detected, most of the errors detected are concentrated in a reduced number of characters, such as vowel characters. The omission of written accent is the type of error that occurs most frequently. Errors also arise from the omission of characters, especially at the end of the word, or the substitution of vowel characters for others (o-a, a-o, a-e, and e-i), giving rise especially to concordance errors. Other types of errors detected are insertion of letter, written accent, or space, while transposition errors have little presence as triggers of context-dependent errors.

Likewise, errors caused by paronymy, verbal and nominal mismatch errors, errors caused by the incorrect joining or separation of words, errors due to ignorance of the current spelling norm, and incorrect verb forms in context have been detected. Therefore, a real-word errors typology focused solely on gender, number, and homophone mismatch errors, as has been done previously, is insufficient for the detection process and for training correction models.

The type of errors developed will serve as a guide for the detection and generation of errors in a more complete and systematic way, giving rise to better training sets. The definition of these new types of specific errors will contribute to the development of rule-based systems for the generation of artificial errors that, in turn, are used to produce training material in automatic correction systems. Training data has a decisive influence on the efficiency of deep learning-based models. In addition, this typology provides knowledge that can be used to improve the detection phase through the use of confusion sets and the correction phase using the patterns and data

obtained for the weighting of alternatives of the decision architecture (López-Hernández, 2022).

Last, future research will continue to investigate the improvement of language modeling parameters, as well as vector training with other architectures such as FastText. Other specialties will also be analyzed to compare them with the results obtained. Training data sets will be built with the inclusion of the errors collected in this work, contributing to the creation of resources for the processing of medical texts in Spanish.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Spanish National Research Agency (MCIN/ AEI /10.13039/501100011033) through project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033). Furthermore, Jésica López-Hernández has been supported by the Ministry of Education of Spain through the national program "Ayudas para la Formación de Profesorado Universitario" (FPU), with reference FPU16/ 03324, of State Programme for the Promotion of Talent and its Employability.

Ethics Statement

This is not applicable. It does not apply because we have worked with a totally anonymized corpus belonging to a private company, VÓCALI Sistemas Inteligentes, which allowed us to use it thanks to a collaborative project and only for academic and scientific purposes.

ORCID iD

Jésica López Hernández https://orcid.org/0000-0003-4115-9248

References

Aguilar Ruiz, M. J. (2013). Las normas ortográficas y ortotipográficas de la nueva Ortografía de la lengua española (2010) aplicadas a las publicaciones biomédicas en español: una visión de conjunto [The orthographic and orthotypographic norms of the new "Ortografía de la lengua española" (2010) applied to biomedical publications in Spanish: An overview]. *Panace@*, *14*(37), 101–120. https://www.tremedica.org/wp-content/uploads/n37-tribuna-MJAguilarRuiz.pdf

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

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

Bello Gutiérrez, P. (2016). Aprendiendo a redactar mejor tus informes [Learning to write better reports]. In AEPap (Ed.), *Curso de Actualización de Pediatría* (pp. 391–400). Lúa Ediciones 3.0.

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

Chipere, N., Malvern, D., & Richards, B. (2004). Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. In G. Aston, S. Bernadini & D. Stewart (Eds.), *Corpora and language learners* (pp. 139–147). John Benjamins Publishing Company. https://doi.org/10.1075/scl.17.10chi

D'Hondt, E., Grouin, C., & Grau, B. (2016, November 5). *Low-resource OCR error detection and correction in French clinical texts* [Paper presentation]. *Proceedings of the seventh international workshop on health text mining and information analysis*, Austin, TX, United States (pp. 61–68). Association for Computational Linguistics. https://aclanthology.org/W16-6108.pdf

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of ACM*, *7*(3), 171–177. https://doi.org/10.1145/363958.363994

Dziadek, J., Henriksson, A., & Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, *235*, 241–245. https://doi.org/10.3233/978-1-61499-753-5-241

Fivez, P., Suster, S., & Daelemans, W. (2017, August). *Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings* [Paper presentation]. Biomedical natural language processing workshop, Vancouver, Canada (pp. 143–148). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2317

Gimenes, P. A., Roman, N. T., & Carvalho, A. (2015). Spelling error patterns in Brazilian Portuguese. *Computational Linguistics*, *41*(1), 175–183. https://doi.org/10.1162/COLI_a_00216

Gutiérrez Rodilla, B. (2005). *El lenguaje de las ciencias* [The language of science]. Gredos.

Hussain, F., & Qamar, U. (2016, April 25). *Identification and correction of misspelled drugs' names in electronic medical records (EMR)* [Conference session]. *Proceedings of the 18th international conference on enterprise information systems (ICEIS 2016)*, Rome, Italy (*Vol. 2*, pp. 333–338). Science and Technology Publications. https://doi.org/10.5220/00059115033 30338

Jurafsky, D., & Martin, J. H. (2018). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Kim, M., Jin, J., Kwon, H., & Yoon, A. (2013, December 3–5). *Statistical context-sensitive spelling correction using typing*

error rate [Conference session]. IEEE 16th international conference on computational science and engineering, Sydney, NSW, Australia (pp. 1242–1246). IEEE Computer Society. https://doi.org/10.1109/CSE.2013.185

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, *24*(4), 377–439. https://doi.org/10.1145/146370.146380

Lai, K. H., Topaz, M., Goss, F., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, *55*, 188–195. https://doi.org/10.1016/j.jbi.2015.04.008

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P. J., Huff, S. M., & Chute, C. G. (2012). Towards a semantic lexicon for clinical natural language processing. AMIA Annual Symposium Proceedings, *2012*, 568–576.

López-Hernández, J. (2022). *Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español* [Analysis and classification of linguistic errors for a proposal to improve medical reports in Spanish] [Doctoral thesis]. Universidad de Murcia. https://digitum.um.es/digitum/handle/10201/121607

López-Hernández, J., & Almela, A. (2021). Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas [Automatic detection of linguistic errors in clinical texts: Analysis of error patterns in various medical specialties]. *Panace@*, *22*(53), 96–108. https://www.tremedica.org/wp-content/uploads/panacea21-53_13_Tribuna_07_LopezHernandez-Almela.pdf

López-Hernández, J., Almela, A., & Valencia-García, R. (2021). Linguistic errors in the biomedical domain: Towards an error typology for Spanish. *Sintagma*, *33*, 83–100. https://doi.org/10.21001/sintagma.2021.33.05

Lu, C. J., & Demner-Fushman, D. (2018, November 3–7). *Improving spelling correction with consumer health terminology* [Conference session]. AMIA 2018 annual symposium proceedings, San Francisco, CA, United States (p. 2053).

Navarro, F. (2015). *Medicina en español. Laboratorio del lenguaje: florilegio de recomendaciones, dudas, comentarios etimológicos, errores, anglicismos y curiosidades varias del lenguaje médico* [Medicine in Spanish. Language laboratory: Florilegium of recommendations, doubts, etymological comments, errors, anglicisms and various curiosities of medical language]. Fundación Lilly.

Pedler, J. (2007). *Computer correction of real-word spelling errors in dyslexic text* [PhD thesis]. Birkbeck College, London University.

Ringler, M. D., Goss, B. C., & Bartholmai, B. J. (2017). Syntactic and semantic errors in radiology reports associated with speech recognition software. *Health Informatics Journal*, *23*(1), 3–13.

Rodríguez-Rubio, S. (2018). Análisis cuantitativo de erratas del diccionario terminológico de las ciencias farmacéuticas inglés-español/Spanish-English (Ariel, 2007) [Quantitative analysis of errors in the English-Spanish/Spanish-English terminology dictionary of pharmaceutical sciences (Ariel, 2007)]. *Panace@*, *19*(47), 76–88. https://www.tremedica.org/wp-content/uploads/n47-analisis.pdf

Ruch, P., Baud, R., & Geissbühler, A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, *29*, 169–184. https://doi.org/10.1016/S0933-3657(03)00052-6

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

Siklósi, B., Novák, A., & Prószéky, G. (2016). Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech & Language*, *35*, 219–233. https://doi.org/10.1016/j.csl.2014.09.001

Thompson, P. M., McNaught, J., & Ananiadou, S. (2015). Customised OCR correction for historical medical text. *Digital Heritage*, *1*, 35–42. https://doi.org/10.1109/DigitalHeritage.2015.7413829

Wong, W., & Glance, D. (2011). Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. *Artificial Intelligence in Medicine*, *53*(3), 171–180. https://doi.org/10.1016/j.artmed.2011.08.003

Workman, T. E., Shao, Y., Divita, G., & Zeng-Treitler, Q. (2019). An efficient prototype method to identify and correct misspellings in clinical text. *BMC Research Notes*, *12*(1), 1–5. https://doi.org/10.1186/s13104-019-4073-y

Yazdani, A., Ghazisaeedi, M., Ahmadinejad, N., Giti, M., Amjadi, H., & Nahvijou, A. (2020). Automated misspelling detection and correction in persian clinical text. *Journal of Digital Imaging*, *33*, 555–562. https://doi.org/10.1007/s10278-019-00296-y

Zhou, X., Zheng, A., Yin, J., Chen, R., Zhao, X., Xu, W., Cheng, W., Xia, T., & Lin, S. (2015). Context-sensitive spelling correction of consumer-generated content on health care. *JMIR Medical Informatics*, *3*(3), e27. https://doi.org/10.2196/medinform.4211