# Linguistic errors in the biomedical domain: Towards an error typology for Spanish[1]

**Jésica López Hernández** - Universidad de Murcia
jesica.lopez@um.es
**Ángela Almela** - Universidad de Murcia
angelalm@um.es
**Rafael Valencia-García** - Universidad de Murcia
valencia@um.es

Rebut / Received: 30/01/20
Acceptat / Accepted: 27/06/20

**Resum. Errors lingüístics en el domini biomèdic: Cap a una tipologia d'errors per a l'espanyol.** L'objectiu d'aquest treball és l'anàlisi d'errors continguts en un corpus d'informes mèdics en llenguatge natural i el disseny d'una tipologia d'errors, ja que no hi va haver una revisió sistemàtica sobre verificació i correcció d'errors en documentació clínica en castellà. En el desenvolupament de sistemes automàtics de detecció i correcció, és d'interès aprofundir en la naturalesa dels errors lingüístics que es produeixen en els informes clínics per tal de detectar-los i tractar-los adequadament. Els resultats mostren que els errors d'omissió són els més freqüents en la mostra analitzada i que la longitud de la paraula sens dubte influeix en la freqüència d'error. La tipificació dels patrons d'error proporcionats permet el desenvolupament d'un mòdul basat en coneixements lingüístics, actualment en curs, que serà capaç de millorar el rendiment dels sistemes de correcció de detecció i correcció d'errors per al domini biomèdic.

**Paraules clau:** detecció automàtica d'errors, error candidat, patrons d'error, corpus biomèdic, processament del llenguatge natural.

---

1. This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). Furthermore, the main autor is supported by Ministerio de Universidades of Spain through the national program *Ayudas para la formación de profesorado universitario* (FPU), with reference FPU16/03324.

**Abstract. Linguistic errors in the biomedical domain: Towards an error typology for Spanish.** The objective of this work is the analysis of errors contained in a corpus of medical reports in natural language and the design of a typology of errors, as there was no systematic review on verification and correction of errors in clinical documentation in Spanish. In the development of automatic detection and correction systems, it is of great interest to delve into the nature of the linguistic errors that occur in clinical reports, in order to detect and treat them properly. The results show that omission errors are the most frequent ones in the analyzed sample, and that word length certainly influences error frequency. The typification of error patterns provided is enabling the development of a module based on linguistic knowledge, which is currently in progress. This will help to improve the performance of error detection and correction systems for the biomedical domain.

**Keywords**: automatic error detection, candidate error, error patterns, biomedical corpus, natural language processing.

## 1. Introduction

Automatic error detection is a prerequisite for error correction systems, an essential element of state-of-the-art technologies in natural language processing. However, there is no systematic review of error patterns in Spanish biomedical texts, nor do we have quantitative data on them. It is therefore necessary to carry out a study and typification of linguistic errors that allow us to know what types of errors tend to occur in this domain, what their properties are and how we can provide a basis of linguistic knowledge to the existing methods of detection and correction for this purpose. Furthermore, it is not possible to elaborate a universal typology of errors, but rather the types of errors identified may vary depending on the scope and context of use (Díaz, 2005), hence the need to develop a typology for medical language itself, where the use of specific terms and the particularities of the domain are crucial.

There are several existing works that study and analyze medical language from a prescriptive perspective. Among them, we must highlight the *Diccionario de Términos Médicos* (2012) developed by the *Real Academia Nacional de Medicina*, whose purpose is the standardization of medical language, and the work of Aguilar Ruiz (2013), which carries out the systematic collection of the main orthographic novelties of *Ortografía de la lengua española* (2010) which are relevant for biomedical publications in Spanish and should be taken into account to avoid mistakes. These references represent a good starting point and theoretical support for the design of the typology, but we consider it relevant to delve deeper into the types of errors through the analysis of a real corpus of clinical reports.

This paper is divided into five sections: sections 2 and 3 present the state of the art of automatic correction in medical reports and the development of error typologies,

respectively. Section 4 sets out the methodology used for the analysis, which includes the presentation of the corpus and the procedure used. Section 5 studies and discusses the results obtained and presents the typology model. Finally, the last section includes the conclusions drawn, as well as suggestions for further research.

## 2. Automatic correction in the medical domain

In the field of medicine, the digitalization of clinical records in recent years has generated greater availability of various data sets, resulting in an extremely valuable source of information. In this domain, it is particularly important that information is presented as rigorously and accurately as possible to facilitate the process of information extraction, decision making, event prediction or interoperability. However, for the processing of clinical documents there are several drawbacks that must be taken into account: the information is unstructured and often contains spelling errors (Ruch & Geissbühler, 2003), abbreviations (Wong & Glance, 2011), and ambiguities (Meystre & Haug, 2006). The intrinsic lexical characteristics of these texts are also noteworthy: they contain abundant words from ancient Greek and Latin, the creation of neologisms is constant, and in recent years, the incorporation of anglicisms. Also worth mentioning are the special mechanisms of word formation, among which we find eponyms, acronyms and onomatopoeias (Gutiérrez, 2005). This lexical complexity also contributes to the increase in errors.

The literature on automatic correction in clinical reports is limited and heterogeneous (López-Hernández, Almela & Valencia-García, 2019). Although it is true that there is a considerable number of studies in natural language processing (henceforth, NLP) for the biomedical field, most of them are focused on tasks of information extraction, disambiguation and named entity recognition. Regarding automatic error detection and correction, there are works that have applied correction techniques in various types of medical texts: clinical records of emergency department (Patrick, Sabbagh, Jain & Zheng, 2010), progress notes (Wong & Glance 2011), electronic health records (Fivez, Suster & Daelemans, 2016), and queries put by patients or consumers to improve search systems (Senger, Kaltschmidt, Schmitt, Pruszydlo & Haefeli, 2010). All agree in pointing out the large number of errors presented in these texts and the complexity of the treatment of clinical records, both because of the significant amount of abbreviations they contain, the complex terminology, the lack of standardization of forms and the absence of subsequent review (López-Hernández et al., 2019). The correction techniques used include the following: dictionary searches, distance of spelling and phonetic editing (Veronis, 1988), rule-based methods (Naber, 2003), statistical methods (Ahmed, Luca & Nurnberger, 2009), and methods based on machine learning. In recent years, techniques based on deep learning have been incorporated to the list, such as neural embeddings (Shickel, Tighe, Bihorac & Rashidi, 2018). The greatest difficulties are found in the detection and correction of grammatical and semantic errors, that is to say, errors in

which context comes into play (Kilicoglu, Fiszman, Roberts & Demner-Fushman, 2015). The greatest advances have been made on English corpora and there is a larger number of tools available for this language, hence the need to adapt and generate new resources for other languages, especially if they are as widely spoken as Spanish.

## 3. Error typologies

An error typology is a hierarchically organized classification system for all types of errors of a particular language or domain (Rambell, 1999). The development of error typologies has been especially fertile in the field of language learning. There are numerous analysis and error classification works based on corpora of second language learners (Nagata, Takamura & Neubig, 2017). These studies on identification and classification of errors have been carried out for different languages. Among them, the largest number is dedicated to English. The first studies on types of error in relation to automatic correction systems for this language date back to the 1960s. At that time, the concept known as *Levenshtein distance* (1966) was defined, which refers to the minimum number of operations required to transform one string of characters into another. Damerau (1964) established that 80% of misspelled words contained one of the following types of errors: addition or insertion of a character; omission or deletion of a character; substitution, which consists in the elimination of one character and the insertion of a different one in its place; and transposition, which occurs with the exchange of adjacent characters on the keyboard.

Other relevant contributions are found in Yannakoudakis and Fawthrop (1983), Pollock and Zamora (1983), Mitton (1987) or Kukich (1992), among others. Yannakoudakis and Fawthrop (1983) collected 1,377 errors in a corpus of more than 60,000 words and established that these could be explained from seventeen heuristic rules. They also determined that incorrectly written words do not usually contain more than one error, thesis defended by other authors, such as Pollock and Zamora (1983), which also speak of a very low percentage of multiple errors, around 6% in a compilation of 50,000 errors. However, these rules are exclusively applicable to English, so they could not be generalized to other languages. In recent years, studies have also been conducted on error patterns for other languages, namely Portuguese (Gimenes, Roman & Carvalho, 2015), Hungarian (Siklósi, Novák & Prószéky, 2016), Japanese (Baba & Suzuki, 2012), Danish (Paggio, 2000), and Punjabi (Lehal & Bhagat, 2007).

There are different dimensions to classify errors: the nature of the error, the cause of the error, the type of error, the context in which the error appears, and the correction of the error (Rambell, 1999). It can be seen, therefore, that they can present different degrees of granularity according to the purpose for which they will be used. Other typologies distinguish between *non-word*, when the error made gives rise to a word that does not exist in any dictionary, and *real-word*, when the erroneously written word results in another that does exist, which makes its identification as an error more difficult.

It also addresses whether it is a typographical, spelling, style, grammatical or semantic error (Naber, 2003). Some authors also distinguish between cognition and performance errors. These are related to the ignorance of the orthographic norm of the language, that is to say, if the errors have a cognitive motivation or have happened accidentally (Díaz, 2005). This classification is especially interesting in language teaching research; specifically, Corder (1967) first posed the distinction *error* for the former concept, and *mistake* for the latter.

Regarding Spanish, we find two works on typologies focused on automatic correction tasks that deserve to be highlighted. Ramírez and López (2006) discuss previous generalizations of error patterns in studies conducted for other languages and offer a new perspective on error patterns in Spanish. It is a work that is framed in the development of a corrector for Spanish in Microsoft Corporation and is especially relevant because it is the most complete typology about errors in Spanish to date. On the other hand, in Díaz (2005), the treatment of grammatical errors and cognitive motivation is addressed, and the benefit of an error typology to create a grammar and style corrector is defended. The present study is in line with these two works, attempting at a typology or rules that serve to design knowledge-based techniques for automatic detection and correction processes.

## 4. Methodology

### *4.1. Corpus description*

The corpus under analysis is composed of a collection of electronic clinical reports in Spanish belonging to the medical specialty of emergency medicine. It consists of 631,576 tokens and 24,286 types[2]; it is a monolingual corpus and is not POS-tagged. It is a private corpus, belonging to the company Vócali[3], which uses it to generate language models that are applied to the development of voice recognition systems in the medical environment.

The corpus consists of plain unstructured text. The corpus underwent some preprocessing in order to make the format of the reports uniform, and HTML and XML tags were eliminated in those containing them. The clinical reports include the following information: anamnesis, record of tests, physical examination, treatments and procedures adopted. It is worth noting that the reports were directly typed by doctors, not handwritten or transcribed. Figure 1 shows an example of the physical examination section:

2. The term "token" is used to refer to the total number of words in the corpus, regardless of how often they are repeated. The term "type" refers to the number of distinct words in the corpus. In the following sections "word" will be used as a synonym for "token" in the corpus.

3. https://vocali.net/.

```
AC: Ritmico sin soplos ni extratonos destacables.
AP: MVC en ambos campos, no ruidos agregados.
ABD: Blando, depreisble. No masas ni megalias.
Dolor al palpar flancos, fosas iliacas e hipoastrio, más acentuado en lado izquierdo,
no signos de irritación peritoneal, peristaltismo presente.
MMII: Pulsos conservados y simétricos. Buena y simétrica temperatura distal. No edemas.
NRL: Sin focalidad Neurológica.
```

Figure 1. Corpus sample: physical examination section.

Furthermore, in observance of the General Data Protection Regulation (GDPR)[4], the reports have been previously anonymized, and they contain no information about the center, date, place, identity or affiliation number of the patient. Automatic anonymization methods, such as searching for established patterns and training techniques based on machine learning, were used to achieve the total elimination of these terms. This circumstance restricts the analysis of certain linguistic features, such as geographical and dialectal differences, because that information was not accessible.

### *4.2. Data analysis*

For the delimitation and classification of errors, the abovementioned analysis proposal by Ramírez and López (2006) will be taken as a starting point. Certain parameters such as frequency and type of error (omission, substitution, addition or transposition), word length, and context in which it occurs were taken into account. This is an empirical investigation exploratory in nature, as it is the first stage of a broader research project. By means of this analysis, an initial classification was attempted, as well as a tentative process of identification of errors, which may enable the recognition of patterns and the subsequent design of a typology of errors. Specifically, all the cases affecting the quality of the automatic processing of the corpus and those not considered as normative according to the *Real Academia Española* in *Ortografía de la lengua española* (2010) were taken as erroneous words. Thus, the present study deals exclusively with non-word errors, that is to say, those which result in words not included in any dictionary; real-word errors require the analysis of contextual information, which lies beyond the scope of this work.

The extraction process for the potential error list involved a number of tasks:

- First, the corpus, which had previously been preprocessed, was compared to a list of stored words or lexicon, with which the words previously validated as correct were detected. This lexicon was compiled from the Spanish dictionary

4. https://gdpr-info.eu/.

of the Hunspell[5] spellchecker, which also contains inflected and derived forms of the words, as well as specific biomedical terminology in Spanish obtained from different sources, like specialized glossaries, terminological resources, lexical databases and nomenclatures such as Snomed-CT[6] and CIE-10[7], including lists of drugs, acronyms, active ingredients, diseases, symptoms, processes, anatomical structures, and protocols, among others.

- Subsequently, a wordlist was obtained including all the potential errors or error candidates. The vast majority of them were errors, but the list also contained correct words not incorporated to the reference dictionary yet because they were either neologisms or words never seen before; these words can be considered as false positives.
- The manual revision of this wordlist was carried out, from which adapted categories were created that served for the design of the error typology.

As for the quantitative description of the dataset, the raw frequency of error candidates by word length (measured in number of letters) was obtained, as well as the frequency in relation to the total word count, that is to say, their relative frequency. The rationale behind this decision lies in the increasing use of readability indices, which take word complexity as one of the essential variables determining text density, as stated in Cantos and Almela (2019). The automatic calculation of the frequencies by word length was obtained by means of the text classification tool UMU Text Stats (García-Díaz, Cánovas-García & Valencia-García, 2020).

## 5. Results and discussion

### *5.1. Quantitative results*

The results yielded are presented in this section. 8, 882 words out of a total of 631, 576 were identified as error candidates. As shown in Table 1 and Figure 2, the largest number of errors was registered in words consisting of eight, nine and ten letters; in short words (shorter than 5 letters), the most striking case being three-letter words. This may be due to the lack of standardization of acronyms and abbreviations, which has led to false positives. Most of the errors in three-letter words correspond to acronyms created by doctors; they are not incorrect as such, they simply do not have enough consistency to be incorporated into glossaries and reference works, which makes their identification difficult.

5. http://hunspell.github.io/.

6. Systematized Nomenclature of Medicine – Clinical Terms. http://www.ihtsdo.org/snomed-ct/.

7. Clasificación Internacional de Enfermedades, 10th edition, corresponding to the Spanish version of the ICD-10 (International Statistical Classification of Diseases). https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html.

Table 1. Total word count and frequency of error candidates in the corpus.

| Word length | Total word count | Number of error candidates (raw freq.) | Number of error candidates (relative freq.) |
|---|---|---|---|
| 1-L | 34006 | 19 | 0.06 |
| 2-L | 119593 | 230 | 0.19 |
| 3-L | 68140 | 669 | 0.98 |
| 4-L | 46857 | 364 | 0.78 |
| 5-L | 55631 | 339 | 0.61 |
| 6-L | 53799 | 631 | 1.17 |
| 7-L | 60823 | 920 | 1.51 |
| 8-L | 50506 | 1134 | 2.25 |
| 9-L | 48000 | 1063 | 2.21 |
| 10-L | 34057 | 1083 | 3.18 |
| 11-L | 28827 | 817 | 2.83 |
| 12-L | 12194 | 598 | 4.90 |
| 13-L | 10089 | 411 | 4.07 |
| 14-L | 4275 | 274 | 6.41 |
| 15-L | 2558 | 151 | 5.90 |
| 16-L | 1049 | 82 | 7.82 |
| 17-L | 614 | 54 | 8.79 |
| 18-L | 312 | 16 | 5.13 |
| 19-L | 164 | 14 | 8.54 |
| 20-L | 45 | 6 | 13.33 |
| 21-L | 22 | 3 | 13.64 |
| 22-L | 3 | 1 | 33.33 |
| 23-L | 8 | 1 | 12.50 |
| 24-L | 3 | 1 | 33.33 |
| 25-L | 1 | 1 | 100.00 |

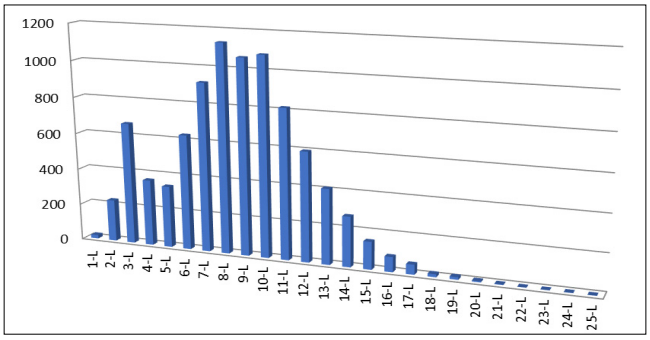


Figure 2. Error candidates by word length (raw frequency).

A one sample Kolmogorov-Smirnov test provided evidence against the null hypothesis of normality, implying that the data sample does not show a normal distribution (Cantos, 2013: 45). Thus, in order to explore any potential correlations between the variables, a Spearman's rank correlation coefficient (non-parametric) was computed with IBM SPSS Statistics 24[8], as proposed by Cantos (2013: 84). As can be seen in Table 2, there is a strong positive correlation between the total count and the potential errors, which means in practice that the larger the number of words of a given length, the higher the frequency of errors registered. In parallel, an even more significant negative correlation is observed between word length and total word count, entailing that, in general terms, the shorter the word, the more frequent it is; this is in line with Harremöes and Topsøe's linguistic findings on Zipf's law (2005).

Table 2. Results from Spearman's rank correlation coefficient.

| | | | **Total Count** | **Potential Errors** | **Word Length** |
|---|---|---|---|---|---|
| Spearman's Rho | Total Count | Correlation coefficient | 1.000 | .832** | -.950** |
| | | Sig. (bilateral) | . | .000 | .000 |
| | | N | 25 | 25 | 25 |
| | Potential Errors | Correlation coefficient | .832** | 1.000 | -.731** |
| | | Sig. (bilateral) | .000 | . | .000 |
| | | N | 25 | 25 | 25 |
| | Word Length | Correlation coefficient | -.950** | -.731** | 1.000 |
| | | Sig. (bilateral) | .000 | .000 | . |
| | | N | 25 | 25 | 25 |
| **. The correlation is significant at level 0.01 (bilateral). | | | | | |

According to the Kruskal-Wallis test (Cantos, 2013), the distributions of grouped word lengths were found to be significant (see Figure 3). Outstandingly, the group of words of length 6 to 10 is especially interesting, since despite not being the largest group, it is the one in which the highest number of potential errors has been registered (see Figure 4).

8. https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-24.

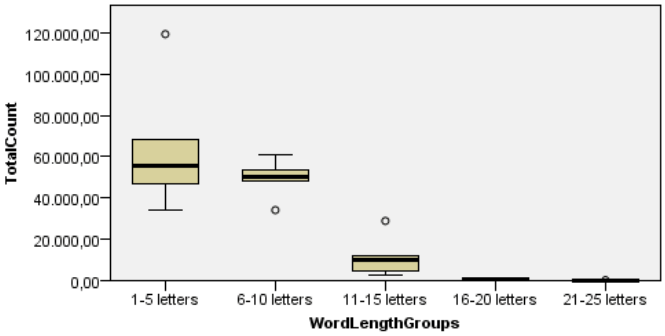


Figure 3. Results from Kruskal-Wallis test for independent samples (total count and word length groups).

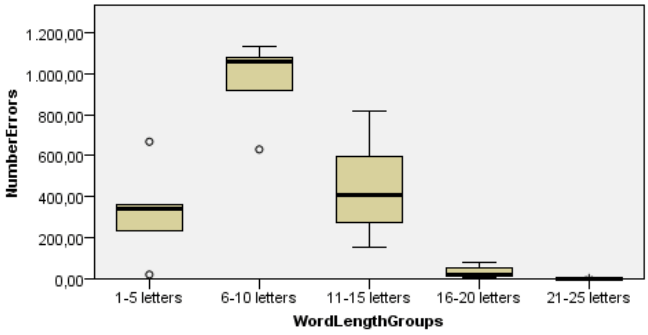


Figure 4. Results from Kruskal-Wallis test for independent samples (number of potential errors and word length groups).

### *5.2. Qualitative results and error typology proposal*

Regarding the qualitative description of the errors detected, they can be classified into the following categories:

- Errors due to intentional use of non-standard spelling:

Fast writing often results in spelling mistakes, especially the use of accent marks, with their omission being the most common error. The ten errors with the highest frequency of occurrence, without taking into account the case of acronyms, are the

following: *dias* (días)[9] [1577],[10] *dia* (día) [948], *analitica* (analítica) [679], *bioquimica* (bioquímica) [604], *neutrofilos* (neutrófilos) [546], *infeccion* (infección) [437], *sindrome* (síndrome) [415], *hematies* (hematíes) [406], *colico* (cólico) [365], y *toracico* (torácico) [359].

- Typing errors derived from the use of a keyboard:

They occur due to slips of the finger and failures in motor coordination, which implies that adjacent letters are used on the keyboard instead of the pertinent ones, or that nearby letter keys are accidentally pressed, which is known as keyboard adjacency effects (Ramírez & López, 2006). Some examples are: *palapción* (palpación) [23], *bilñateral* (bilateral) [2], *hroas* (horas) [18], *comrpimido* (comprimido) [17], or *evolcuion* (evolución) [9].

- Errors in the use of lowercase and uppercase letters:

Initial capitalization is used in common words that should be written in lowercase, in many cases with an emphatic character: *Hipertrofia ventricular* (hipertrofia ventricular) [1]; *Antecedente de Tos* (antecedente de tos) [1]. As with the name of some diseases, we find a tendency to write with initial uppercase letters the name of months and days of the week, possibly due to the influence of English and to Spanish obsolete spellings.

- Errors in the use of accent marks:

As mentioned above, the most common error in the corpus is the omission of accent marks. Furthermore, there are also errors in accents that have been added to syllables where they are not appropriate. Several demonstrative pronouns and monosyllables are marked, probably due to ignorance of spelling reform: *dió* (dio) [7], or *vió* (vio) [3].

- Errors in the use of abbreviations:

A lack of uniformity is observed in the treatment of the abbreviations in our corpus, since some are marked with a period and others are not, when its use is prescriptive. The abbreviations must maintain the accent mark if they include the vowel that carries it in the word, but it is not very frequent in our corpus. A good example is "célula": *cel.* [119] and *cél.* [7]. There is high variability of abbreviations for the same word (e. G., *actv* [2] or *act* [44] for "actividad", *cefale* [3] or *cef* [1] for "cefalea"), and procedures of truncation (*comp* [2602] for "comprimido", *vasc* [3] for "vascular") and contraction (*tto* [539] for "tratamiento", *hrs* [395] for "horas") are commonly used. A noteworthy

9. The right version of the word is offered in brackets.

10. The figures in square brackets indicate the raw frequency of misspelled words in the corpus.

example is "neo", as in the corpus we can find examples like *neo mama* [4], *neo gastric* [3] and *neo pharyngo-laryngeal* [1]. In these cases, this abbreviation stands for "neoplasia", but the absence of full stop and the fact that it is not a standardized or widely used abbreviation can lead to confusion. There are cases of extreme abbreviation, such as *sd* [384], standing for "síndrome", or *qx* [300], standing for "quirúrgico".

- Errors whose solution requires context:

We find instances which need context in order to be properly corrected, as the word itself does not provide enough information to choose the adequate correction. This happens in cases like *altenar* [3], in which we cannot know with certainty whether it is "alternan" or "alteran"; *alérgias* [3], which may be the result of an accent mark insertion (intended word: "alergias") or of a consonant omission (intended word: "alérgicas"); *otrso* [2], where the intended word may be "otros" or "torso"; or *perdida*, in words which can be marked for accent or not ("perdida" or "pérdida"). Likewise, incomplete word endings often hinder the recognition of the morpheme needed: *crónic* [2], *cabez* [2], *bacterian* [2], *neurol* [4], etc.

- Errors in the treatment of acronyms:

Acronyms must be written in capital letters and they have neither accent marks nor stops, and they do not vary for plural. In the corpus, we find acronyms written in lowercase letters: *itu* [12] (ITU) or *got* [11] (GOT).

- Errors in the treatment of symbols:

Symbols are scientific abbreviations set by institutions with international validity. They must be written without full stops and accent marks, and their forms do not vary for plural. Nevertheless, a frequent mistake is turning to plural measurement units: (*grs*, *kgs*, *cms* or *mls*). Furthermore, there must be a space between them and the corresponding numerical value, which is not respected in many instances of the corpus.

- Prefixes, suffixes and other compositional elements:

Prefixes must not be separated by a space or by a hyphen from the stem they precede. A large number of prefixes can be found in the corpus linked by a hyphen, e. G., *ex-fumador* [22], or separated from the stem through a space, e. G., *post traumática* [27]. Moreover, the use of "pseudo" and "seudo" is heterogeneous, like the use of "post" and "pos".

In accordance with the results obtained in the analysis, an adapted typology was designed and Figures 5 and 6 provide samples for the categories defined. The typology presents a set of classification categories according to the type of error and the operation required for its correction, and it is made up of five general

categories. Four of them include errors whose solution needs a single step, that is to say, where there is an edit distance of 1 from the ungrammatical word (source word) to the grammatical one (target word). The last category gathers the so-called multi-errors, those requiring more than one step or operation in order to reach the target word. There are four types of operations: omission, insertion, substitution and transposition. Furthermore, a section devoted to abbreviations has been added, due to their strong presence in biomedical corpora, a difference from that of Ramírez and López (2006), and the subcategories have been distributed differently, based on operations with characters, orthographic symbols or spaces. Types of errors involving operations with grave accent, diaeresis, parenthesis or omission of the full stop between sentences have been included too. This typology includes all types of errors without overlapping and covering most of the patterns presented in the corpus at hand, assigning each word to a specific position in the hierarchy. All in all, this is a preliminary typology which is open to modifications.

| **Omission** | |
|---|---|
| **Omission of a character** | |
| | *normoreactivas* (normorreactivas) |
| **Omission of orthographic sign** | |
| Acute accent | *paralisis* (parálisis) |
| Grave accent | *Lasegue* (Lasègue) |
| Diaeresis | *linguísticos* (lingüísticos) |
| Inverted question mark | *AIT? Hematemesis de pequeña cantidad?* |
| **Omission of space** | |
| Space between independent words | *esque* (es que) |
| Space between number and unit of measure | *25mg* (25 mg) |
| Space after orthographic sign | *-melanoma* (- melanoma) |

Figure 5. Error typology: omission.

| **Insertion** | |
|---|---|
| **Insertion of a character** | |
| Same letter | *een* (en) |
| Different letter | *laringuectomía* (laringectomía) |
| **Insertion of orthographic sign** | |
| Accent mark | *fué* (fue) |
| Hyphen between prefix and stem | *ex-fumador* (exfumador) |
| **Insertion of space** | |
| Space between letters of the same word | *bue na* (buena) |
| Space between prefix and stem | *post profilaxis* (postprofilaxis) |
| **Substitution** | |
| **Substitution of a letter** | |
| Different letter | *artritid* (artritis) |
| Lowercase for uppercase letter | *sjögren* (Sjögren) |
| Uppercase for lowercase letter | *PEndiente cita* (Pendiente cita) |
| **Substitution of orthographic sign** | |
| Grave accent for acute accent | *informaciòn* (información) |
| **Transposition** | |
| Transposition of a letter | *tratamietno* (tratamiento) |
| Transposition of a space | *hayq ue* (hay que) |
| **Multi-Error** | |
| Abbreviations | *tto* (tratamiento) |
| Lowercase acronyms | *itu* (ITU) |
| Omission + insertion | *àciente* (paciente) |
| Omission + omission | *tnel* (túnel) |
| Omission + substitution | *balido* (válido) |
| Omission + transposition | *anlagesico* (analgésico) |
| Substitution + insertion | *hepaticoyiyunostommía* (hepaticoyeyunostomía) |

| | |
|---|---|
| Substitution + substitution | *sinbastatina* (simvastatina) |
| Substitution + transposition | *antihieprtebsivo* (antihipertensivo) |
| Insertion + insertion | *hepatoniesplenomegalia* (hepatoesplenomegalia) |
| Insertion + transposition | *sobregaregadoos* (sobreagregados) |
| Transposition + transposition | *cerivcodosrolumbalgia* (cervicodorsolumbalgia) |

Figure 6. Error typology: insertion, substitution, transposition and multi-error.

## 6. Conclusions and future work

In this work, an error typology proposal has been provided, including the description of its components and posing real examples from a biomedical corpus in Spanish. It certainly reveals the beneficial effects of working with ground-truth data as a basis for the elaboration of the typology. The preliminary results are promising, as it has been shown that errors arising from the use of a keyboard are more frequent than other types of errors, as well as omission of accent marks; in addition, it can be stated that word length certainly influences error frequency. Furthermore, there is a tendency to separate prefixes and lack of consistency in the creation of abbreviations, in the use of upper and lowercase letters, and in the writing of acronyms and symbols. This classification of errors, specific to the biomedical field, contributes to the description and organization of errors in this area and to the greater coverage of cases for spelling correction, complementing the techniques based on statistical analysis and machine learning.

As for future research, the present authors will attempt to validate the representativeness of the typology both by means of further case reports in emergency medicine and of other corpora containing medical specialties in Spanish. Furthermore, our research agenda includes the application of the results to the implementation of a spell checker able to detect misspellings in biomedical texts, providing isolated-word error correction by offering a set of context-dependent candidate corrections.

## 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. *Panace@*, *14*(37), 101-120.

Ahmed, F., Luca, E. W. D., & Nurnberger, A. (2009). Revised N-Gram based automatic spelling correction tool to improve retrieval effectiveness. *Polibits*, 39-48.

Baba, Y., & Suzuki, H. (2012). How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. In H. Li, C. Lin, M. Osborne, G. G. Lee & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: *Short Papers* (pp. 373-377). Jeju Island: Association for Computational Linguistics (ACL).

Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield, UK: Equinox Publishing.

Cantos, P., & Almela, A (2019). Readability indices for the assessment of textbooks: a feasibility study in the context of EFL. *Vigo International Journal of Applied Linguistics,* 16, 31-52.

Corder, S. P. (1967). The Significance of Learners' Errors. *International Review of Applied Linguistics in Language Teaching, 5*, 161-170.

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of ACM*, *7*(3), 171-177.

Díaz Villa, A. (2005). Tipología de errores gramaticales para un corrector automático. *Procesamiento del Lenguaje Natural, 35*, 409-416.

Fivez P., Suster, S., & Daelemans, W. (2016). Unsupervised context-sensitive spelling correction of clinical free-text with word and character N-Gram embeddings. In K. B. Cohen, D. Demner-Fushman, S. Ananiadou & J. Tsujii (Eds.), *Proceedings of the BioNLP 2017 Workshop*, (pp. 143-148). Vancouver: Association for Computational Linguistics (ACL).

García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*, *112*, 641-657. doi:10.1016/j.future.2020.06.019.

Gimenes, P. A., Roman, N. T., & Carvalho, A. M. (2015). Spelling Error Patterns in Brazilian Portuguese. *Computational Linguistics*, *41*(1), 175-183. doi:10.1162/coli_a_00216.

Gutiérrez Rodilla, B. (2005). *El lenguaje de las ciencias.* Madrid: Gredos.

Harremoës, P., & Topsøe, F. (2005). Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics*, *21*, 315-318. doi:10.1109/ISIT.2002.1023479.

Kilicoglu, H., Fiszman, M., Roberts, K., & Demner-Fushman, D. (2015). An ensemble method for spelling correction in consumer health questions. In American Medical Informatics Association (Eds.), *AMIA Annual Symposium Proceedings* (pp. 727-736). San Francisco: AMIA.

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

Lehal, G. S., & Bhagat, M. (2007). Spelling error pattern analysis of Punjabi typed text. In *Proceedings of the 2007 International Symposium on Machine Translation, NLP and TSS* (pp. 128-141). New Delhi: Tata McGraw-Hill.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 707-710.

López-Hernández J., Almela Á., & Valencia-García R. (2019). Automatic Spelling Detection and Correction in the Medical Domain: A Systematic Literature Review. In R. Valencia-García, G. Alcaraz-Mármol, J. Del Cioppo-Morstadt, N. Vera-Lucio & M. Bucaram-Leverone (Eds.) *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science* (vol. 1124, pp. 104-117). Cham: Springer.

Meystre, S., & Haug, P. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39, 589-599.

Mitton, R. (1987). Spelling checkers, spelling correctors, and the misspellings of poor spellers. *Information Processing & Management*, *23*(5): 495-505.

Naber, D. (2003). *A rule-based style and grammar checker.* Munich: GRIN Verlag.

Nagata, R., Takamura, H., & Neubig, G. (2017). Adaptive spelling error correction models for learner English. *Procedia Computer Science*, *112*, 474-483. doi:10.1016/j.procs.2017.08.065

Paggio, P. (2000). Spelling and grammar correction for Danish in SCARRIE. In Association for Computational Linguistics (Eds.), *Proceedings of the Sixth Conference on Applied Natural Language Processing,* (pp. 255-261). Seattle.

Patrick, J., Sabbagh, M., Jain, S., & Zheng, H. (2010). Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *Second Workshop on Building and Evaluating Resources for Biomedical Text Mining* (pp. 2-8). Malta: Association for Natural Language Processing.

Pollock, J. J., & Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *Journal of American Society of Informatics and Science*, *34*(1), 51-58.

Rambell, O. (1999). Error typology for automatic proof-reading purposes. In A. Sagvall Hein (Ed.), *Reports from the SCARRIE project* (pp. 1-29). Uppsala: Uppsala University.

Ramírez, F., & López, E. (2006). Spelling Error Patterns in Spanish for Word Processing Applications, In *Proceedings of Fifth international conference on Language Resources and Evaluation LREC'06* (pp. 93-98). Genoa: European Language Resources Association.

Real Academia Española y Asociación de Academias de la Lengua Española. (2010). *Ortografía de la lengua española*. Madrid: Espasa.

Real Academia Nacional de Medicina. (2012). *Diccionario de Términos Médicos*. Madrid: Panamericana.

Ruch, B., & Geissbühler, A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine,* 29(2), 169-84. doi:10.1016/s0933-3657(03)00052-6.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589-1604. doi:10.1109/jbhi.2017.2767063.

Senger, C., Kaltschmidt, J., Schmitt, S. P. W., Pruszydlo, M. G., & Haefeli, W. E. (2010). Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention. *International Journal of Medical Informatics*, *79*(12), 832–839. doi: 10.1016/j.ijmedinf.2010.09.005.

Siklósi, B., Novák, A., & Prószéky, G. (2016). Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech & Language*, *35*, 219-233.

Veronis, J. (1988). Computerized correction of phonographic errors. *Computers and the Humanities*, *22*(1), 43-56.

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.

Yannakoudakis, E. J., & Fawthrop, D. (1983). The rules of spelling errors. *Information processing and management*, *19*(12), 101-108.