Towards Orthographic and Grammatical Clinical Text Correction: a First Approach

Author: Salvador Lima López

Advisors: Montse Cuadros, Olatz Pérez de Viñaspre

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2020

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Akats Gramatikalen Zuzenketa (GEC, ingelesetik, *Grammatical Error Analysis*) Hizkuntza Naturalaren Prozesamenduaren azpieremu bat da, ortografia, puntuazio edo gramatika akatsak dituzten testuak automatikoki zuzentzea helburu duena. Orain arte, bigarren hizkuntzako ikasleek ekoitzitako testuetara bideratu da gehien bat, ingelesez idatzitako testuetara batez ere. Master-Tesi honetan gaztelaniaz idatzitako mediku-txostenetarako Akats Gramatikalen Zuzenketa lantzen da. Arlo espezifiko hau ez da asko esploratu orain arte, ez gaztelaniarako zentzu orokorrean, ezta domeinu klinikorako konkretuki ere. Hasteko, IMEC (gaztelaniatik, *Informes Médicos en Español Corregidos*) corpusa aurkezten da, eskuz zuzendutako mediku-txosten elektronikoen bilduma paralelo berria. Corpusa automatikoki etiketatu da zeregin honetarako egokitutako ERRANT tresna erabiliz. Horrez gain, hainbat esperimentu deskribatzen dira, zeintzuetan sare neuronaletan oinarritutako sistemak ataza honetarako diseinatutako baseline sistema batekin alderatzen diren.

Abstract

Grammatical Error Correction (GEC) is a subfield of Natural Language Processing that aims to automatically correct texts that include errors related to spelling, punctuation or grammar. So far, it has mainly focused on texts produced by second language learners, mostly in English. This Master's Thesis describes a first approach to Grammatical Error Correction for Spanish health records. This specific field has not been explored much until now, nor in Spanish in a general sense nor for the clinical domain specifically. For this purpose, the corpus IMEC (Informes Médicos en Español Corregidos)—a manually-corrected parallel collection of Electronic Health Records—is introduced. This corpus has been automatically annotated using the toolkit ERRANT, specialized in the automatic annotation of GEC parallel corpora, which was adapted to Spanish for this task. Furthermore, some experiments using neural networks and data augmentation are shown and compared with a baseline system also created specifically for this task.

Acknowledgements

This project is the result of a whole year of hard work, a lot of love and even some tears, all sandwiched between a global pandemic. All in all, I'm really happy about the final result, which would not have been the same without the help and support of some people.

First of all, I would like to thank my almost-supervisor and one of the most fundamental supports I've had, Naiara Pérez. Thank you for being there to answer so many of my small, stupid questions and to provide some of the greatest advice. Even though your name couldn't be in the cover of this thesis (as it should be), it will be in many of them pretty soon as I'm sure you have a really bright future ahead of you. I'm really happy I got to meet you and that you accompanied me through this whole journey, as I learned so much from you.

Another pillar of this work was my supervisor at Vicomtech, Montse Cuadros. Even though you are probably one of the busiest people I know, you always found the time to be there and your suggestions really helped shape up this thesis. Olatz Pérez de Viñaspre, my supervisor at the University of the Basque Country, was another vital part of this work. Thank you for listening to my crazy ideas and for believing me in me so much. I don't know what would have been of this work without the help of these three people, so thanks again to the three of you.

Of course, I couldn't not thank Vicomtech for giving me the chance to work with such great people and for allowing me to have access to many more resources than I would have had on my own.

I am truly thankful as well to my partner in crime during the entire Master's Degree, Eneritz García Montero, for always being there for me. We've had countless fun on the good days and lifted each other up on the bad days, and I'm sure we will always be a part of each other's lives even if we are apart. I must also thank my friend Ander González Docasal for his infinite patience and wisdom. I can't wait to see you two succeed even further.

I would not have reached this far without the continued support of my best friend, Álvaro Fernández López. Thank you for always being there and listening to me even when you had no idea what I was talking about. You have been part of my life for so long already and I hope you will continue to be there even longer.

Due to certain global circumstances, I spent around three months in lockdown with three very special people: Angy, Marc and Frank. You really were able to make such a stressful situation fun and I consider myself lucky to have found a home in you. Thank you too for working alongside me each morning in the living room, it really made the process much easier.

I could not forget some of my friends who have intermittently stood my endless rants: Ada, Dani, Josu, Jon Mikel, Mikel, Pepe, Allende, ... You played a bigger role than you may think in this project. Special thanks to Pepe for helping me with the intricacies of medical language. I would also like to thank the professor who introduced me to the field of Natural Language Processing some years ago, Manuel Alcántara Plá. Even though you

had no direct involvement in this project, I would not be here were it not for you.

Last, but absolutely not least, I would like to thank my parents. They are the most loving and supportive parents I could wish for, and I know that they have made many sacrifices for my education. Without them raising and rising me up, I would not have reached these new heights. I hope to someday be able to give back to them all they have given to me.

Contents

List of Figures

List of Tables

1 Introduction

Communication is a key aspect of the relationship between doctors and patients, up to the point that there are compulsory subjects on the topic in medical schools. Despite its importance, at times it is unsuccessful and can become the source of some misunderstandings (Terroba Reinares, 2015). Facts such as the time restrictions medical professionals have to face or the lack of medical knowledge on the patients' side only aggravate this situation.

When it comes to the written medium, health records are the main method of communication. They are documents where doctors write their impressions and diagnosis of a patient during or after their visit to their office. Health records are really valuable as they serve as a bridge between doctors and patients, as well as a reference for other health professionals.

These records are not perfect either. On the one hand, its content can be lacking from the patients' point-of-view. Some usual complaints include the abundance of technical information and lack of advice regarding more pressing matters for them such as how to take the prescribed drugs or dietary restrictions (Silver (1999), as cited in Terroba Reinares (2015)). On the other hand, since doctors usually work under heavy time restrictions, there is a certain carelessness about correctness when writing. This devolves into a series of errors in aspects such as orthography, punctuation or grammar.

Consider, for instance, these sentences taken from health records:

- 'A su llega a urgencia el paciente no refire sintomatología alguna, no recuerda lo acontecido.'
  'At their arrival at the emergency services, the patient does not refer any symptoms nor remembers what happened.'
- 'A.F: Negativos para tu digestivos, o EII'
  'Family history: negative for the digestive tract or inflammatory bowel disease.'
- 'Sigue trat. con omeprazol 20 mgrs 0-0-1, mastical D comp.'
  '(The patient) is undergoing a treatment with omeprazole 20 mg (0-0-1) and Mastical D tablets.'

The writing style of these sentences can make them hard to understand, especially for non-experts. Features such as the heavy use of abbreviations (often with multiple forms for the same concept) or the introduction of foreign elements (such as the *0 - 0 - 1*) do not make them very friendly. On top of that, one of the main characteristics of health records as a genre is a desire to encapsulate as much information as possible in as little space as possible. This often comes at the expense of dismissing the syntactic structure of language as well as orthotypographic conventions.

While the problem with health records' content should be discussed and solved by health professionals themselves due to its complexity, there might be simpler solutions for the problems with their form highlighted above. One of the potential solutions for this communicative issue would necessarily have to include the correction and standardization of written records. Consider now, for a change, these possible corrections to the examples presented earlier:

- 'A su llegada a Urgencias el paciente no refiere sintomatología alguna, no recuerda lo acontecido.'
- 'AF: negativos para tus. digestivos o las EEII.'
- 'Sigue un trat. con omeprazol 20 mg (0-0-1), Mastical D comp.'

By making simple corrections to health records at different levels (syntactic, orthographic, ...), they become clearer and more accessible. Still, manually correcting them is hardly a solution, as it would require a group of people especially dedicated to this task. According to a document released by the Spanish Ministry of Health, in 2018 there were over 350 million Primary Health Care and nursing consultations (Ministerio de Sanidad, 2018, p. 11). Considering that according to a Royal Decree released in 2015 writing a health record is compulsory for each visit (Boletín Oficial del Estado, 2015), it is safe to assume that at least 350 million records were produced in the same year. Because of this, we may want to automate the process.

In Natural Language Processing, the automatic correction of orthographic, lexical and grammatical errors in text has been undertaken by a sub-field called Grammatical Error Correction (GEC). It has traditionally focused on educational purposes, such as correcting second language learners' texts, especially in English, which means there is currently no literature on the topic in Spanish nor any works on texts of specific domains such as the clinical. For this same reason, there are hardly any resources that could be used for exploring this task.

This work addresses and explores how to improve health records' form by presenting an initial approach to the topic from the perspective of GEC, in an attempt to automatize the correction process and alleviate the situation. Due to the scarceness of previous research on this topic, multiple resources had to be developed. All in all, this thesis has the following contributions: (i) the presentation of the IMEC corpus ('Informes Médicos en Español Corregidos'), a compilation of parallel corrected health records, and its annotation guidelines; (ii) an adaptation to Spanish of the automatic annotation toolkit ERRANT; (iii) the introduction of a simple software that automatically induces errors into clean free text in order to create artificial parallel corpora; (iv) an exploitation of the presented resources using Deep Learning and data augmentation techniques with competitive results.

Its structure is the following: Section 2 presents some necessary background and previous research from both a linguistic and computational perspective; Section 3 gives an overview of the steps and stages of this work. Section 4 presents the IMEC corpus, its correction and annotation process and its content. Next, Section 5 provides an explanation of the different experiments that took place as part of this work and their results. Finally, Section 6 wraps up the thesis and presents some possible lines of future work.

2 Related Work

Within Natural Language Processing (NLP), the treatment of medical texts has long been considered a task deserving of a special consideration. A good example of this fact is the existence of multiple NLP congresses specifically focused on medical application. This is not only due to a special interest in the information that can be retrieved from them, but also because clinical language is significantly harder to process. It has some special characteristics, such as specialized vocabulary or the use of recurrent syntactic structures not found elsewhere, that make them closer to scientific language than everyday language. For this reason, many of the tools used for general domain text processing need to be fine-tuned in order to achieve a comparable performance in texts of this genre.

This section is divided in two parts: one explains the characteristics of medical texts and their context; the other reviews the NLP sub-field of Grammatical Error Correction and related topics.

2.1 Linguistic background

This section provides an overview of some of the linguistic features of Electronic Health Records, as well as their strengths and weaknesses and some of the proposals for their improvement.

First, it should be explained that not all medical texts are the same. Terroba Reinares (2015) divides clinical texts into two big groups: specialized texts (scientific papers, manuals, leaflets, ...) and general texts (clinical histories, Electronic Health Records (EHR), ...). Inside the EHR genre, further divisions could be made depending on the medical specialty (cardiology, neurology, ...) and the purpose of the text (report a test's results, make recommendations to the patient, ...).

Even then, these sub-genres share some linguistic characteristics, such as the use of certain patterns that appear so often they could even be considered fixed formulas (e.g. starting a report with a sentence such as '60-year-old patient with no known allergies...' (Terroba Reinares, 2015)). Some other linguistic common points are (*ibid.*)[1]:

- Usage of many abbreviated forms, usually with different forms for one concept.
  'Solicito rx lumbar, RNM y EMG.'
  'I request a lumbar X-ray, NMR and EMG.'

- Usage of terms that come from Latin or are influenced by English, not always properly spelled.
  'El 21 de Agosto 2006 By-pass gástrico.'
  'Gastric bypass on August 21st 2006'.

- Irregular eponyms.
  'Indice de Barthel con dependencia moderada.'

[1]These examples are taken from the uncorrected section of the IMEC, which is presented later on in Section 4.

'Barthel index with moderate dependency'.
'Al alta barthel en torno a 45.'
'Following discharge, Barthel index around 45.'

- Irregular use of upper and lowercase.
  'El dia 17 de Julio pasa al S. de Unidad Hospitalaria de media-larga Estancia.'
  'On July 17th, (the patient) is moved to the medium/long stay unit.'

- Alternation of numeric forms.
  'Sobre las 19:31h del día 15/07/2010...'
  'Around 19:31 on the 15/07/2010...'
  'El dia 19 de junio a las 23'26 horas la paciente...'
  'On June 19th at 23'26, the patient...'

- Non-orthodox usage of gerunds.
  'La paciente ingresa por el motivo reseñado presentando la exploración descrita...'
  'The patient is admitted to the hospital due to the aforementioned motive, presenting the described examination...'
  'La paciente mejora en las proximas 24 horas estabilizandose la TA y presentando buena diuresis y cediendo la fiebre.'
  'The patient improves in the next 24 hours, with the stabilization of their blood pressure, good urine output and their fever going down.'

- Irregular spelling of prefixes and compound words.
  'El postoperatorio discurre...'
  'The post-op passes...'
  'Posoperatorio sin complicaciones.'
  'Post-op with no problems.'
  'Post-operatorio inmediato sin incidencias.'
  'Immediate post-op with no events.'

- Omission of verbs, articles and prepositions.
  'No dolor cervical, no perdida de fuerza ni de sensibilidad en extremidades.'
  'No cervical pain, no loss of limb strength or sensitivity.'
  'Al alta sat 02 basal 95 %.'
  'At discharge basal oxygen saturation of 95 %.
  'Ayer febrícula.'
  'Yesterday low-grade fever.'

- Not many verbs and the ones that are used get repeated a lot.
  'Se inicia tratamiento con Ceftriaxona (ev)'
  'A treatment with endovenous ceftriaxone is initiated.'
  'Se inicia tratamiento inmunosupresor con Simulect.'
  'An immunosuppressive treatment with Simulect is initiated.'

These recurrent phenomena make health records very heterogeneous. One of the reasons for this is that health specialists, the main source of the genre, do not only communicate between them, but also with patients and other non-specialists. Their language is, up to some point, flexible: sometimes it is full of technical terms that require medical knowledge to understand, and others it is brimming with informal usages of words and structures. This means that some of the constructions they use can be unorthodox at best, and, sometimes, plainly wrong.

This heterogeneity also arises partly from the fact that, despite their importance, there is no single, official document that dictates both the form and content of Electronic Health Records. There have been some attempts to standardise them. For example, in 2015 a Royal Decree was released in the Official State Gazette (B.O.E.) by the Spanish government where EHR were deemed compulsory. It included a list of the necessary information they should include (e.g. name, age, type of visit, ...), but it does not address the form of the document from a linguistic point of view (Boletín Oficial del Estado, 2015). Gutiérrez et al. (2010), a group of doctors who studied EHR from different hospitals, developed a set of recommendations for the writing of hospital discharge reports —which could be considered a specific type of EHR.

They start their document by describing discharge reports as a precise and concise summary written in medical terms whose main recipient is the patient. Most of their focus is placed on communication, both with patients (e.g. treatment plans should be clear and in a different page where the objective of the treatment and drugs used is explained; aspects such as treatment duration, dosage and drug names should be clearly stated) and with other doctors (e.g. key aspects—such as the cause of the admission, background information, diagnosis, comorbidity or social valuation—should be part of the document; abbreviations that are not widely known should be avoided). Finally, they also stress the importance of using computer tools that make the writing process simple and logical.

However, as an analysis by Terroba Reinares (2015) of records from public hospitals in La Rioja (Spain) shows, not all of the aforementioned suggestions are carried out.

Another attempt to standardize medical language comes from style guides. They provide insight into what a health record should be and how it should be written. For instance, Bello Gutiérrez (2016) describes from a theoretical point of view three desirable simple principles that any medical writing should show:

- **Veracity:** what is mentioned in the text should correspond both to reality and to what the author meant.
- **Precision:** ambiguous terms should be avoided so there is only one possible interpretation of a given message.
- **Clarity:** Texts should be easy to understand for someone with some knowledge of the field. This implies not only using precise terms, but also avoiding rare grammatical structures that difficult the reader's understanding of the text.

From a more practical perspective, these style guides provide specific suggestions on how to approach certain linguistic phenomena. They are usually based on the suggestions

provided by the Royal Spanish Academy, an official institution dedicated to the Spanish language, and so they are a useful reference for writing clinical texts.

2.2 Computational background

As explained in the introduction, Grammatical Error Correction (GEC) is the task of automatically correcting orthographic, lexical and grammatical errors in text. Throughout this section, different aspects of this field will be reviewed: the usual architectures used and the state of the art, the different corpora available, as well as the annotation and evaluation toolkit ERRANT. Finally, data augmentation is presented at the end as a solution to data sparsity.

2.2.1 Architectures

Grammatical Error Correction is by no means a new discipline. Back in the 1980's there were already some early systems that were able to detect erroneous grammar constructions and return suggestions using rule-based pattern recognisers and dictionary-based systems (Macdonald, 1983; Richardson and Braden-Harder, 1988). These systems use string-matching and linguistic information obtained from the text to detect errors. Even though rule-based systems can be precise, they are not always easy to implement as one must be careful of how rules interact with each other and their priority. Thus, they can be difficult to design and maintain.

A different approach is to use language models to detect errors by calculating the likelihood of a sequence of words. A language model is a probability distribution that is learnt from a corpus. Their application to GEC is based on the idea that correct sequences are bound to have a higher probability score than incorrect ones. They are very dependent on the data that is used to build them, and so large corpora like Wikipedia or Common Crawl are often used. Language models are frequently used in combination with other systems due to their versatility.

Early machine learning approaches attempted to solve this task using statistical classifiers (Bryant, 2019). These systems try to classify a given input into a category by learning patterns from features extracted from the text (e.g. linguistic information such as part-of-speech tags) or engineered by humans. However, as the number of possible corrections in error correction is so large, many of these classifiers focused on specific error types, such as articles (Gamon et al., 2008), prepositions (Tetreault et al., 2010) or verbs (Lee and Seneff, 2008). Eventually, classifiers became impractical, as having to combine multiple of them proved to be hard to use as a general solution, and were replaced with more advanced machine learning architectures.

Probably, over the last few years the most common approach to GEC has been to treat it as a machine translation (MT) task. Both tasks are indeed comparable, as they involve transforming input in one language to another. While in MT this is done using any language pair (e.g. Spanish and French), in GEC correct and incorrect texts are treated

as different languages. The two main MT techniques used in GEC are statistical machine translation and neural machine translation.

Statistical Machine Translation (SMT) uses parallel annotated data to train a translation model that outputs the probability that a sequence of words maps to another. The idea of using this type of architecture for an error correction problem was first proposed by Brockett et al. (2006), who used a noisy channel model to detect and correct mass noun errors made by English as a Second Language (ESL) students. SMT opened the possibility to correct more complex errors, as well as whole sentences with multiple error types at once (e.g. Madnani et al. (2012)), and quickly became a staple of state-of-the-art systems at the time. Furthermore, SMT systems were the first to offer the possibility of generating an N-best list of alternative corrections (see, for instance, Shen et al. (2004)). Re-ranking the possible corrections using text features, classifiers or language models usually leads to an improvement in performance and thus it has become a research topic on its own (Yuan et al., 2016).

Neural Machine Translation (NMT) uses neural networks to learn vector representations of data. The most popular architecture both in NMT and GEC is the Encoder-decoder (Cho et al., 2014; Chen et al., 2017). This framework consists on two parts: an encoder, whose job is to map raw inputs to a mathematical representation of language, and a decoder, that converts the operations performed by the neural network into a sequence of words.

Other types of models are also used. For instance, Yannakoudakis et al. (2017) use a neural sequence-labelling model to experiment with N-best list re-ranking. Lately, NMT has started to use Transformers (Vaswani et al., 2017), a bigger, more powerful Deep Learning architecture that is being used by the state-of-the-art systems of many different NLP tasks. Accordingly, this architecture is also being used in some state-of-the-art GEC systems. One of the advantages of Transformers is that they can be pre-trained with huge, general corpora and be fine-tuned afterwards using a smaller, more specific corpus. This means that one pre-trained model can be fine-tuned for multiple different tasks. This is known as transfer learning. However, their main disadvantage is that their bigger size also makes them more computationally expensive to train. As a consequence, they also have slower inference speed and are harder to interpret.

Currently, the state of the art for the main GEC benchmarks (described in Section 2.2.2) is a model called GECToR (Omelianchuk et al., 2020), a sequence tagger that uses a Transformer encoder which was pre-trained on parallel sentences with artificial errors and fine-tuned on parallel GEC corpora.

Even though the systems described earlier have been used in languages other than English, there is almost no literature on GEC for Spanish. One early system is GramCheck (Ramírez Bustamante and Sánchez León, 1996), a "grammar and style checker" that is mainly rule-based for Spanish and Greek. More recently, Davidson et al. (2020) present a recurrent network with which they test the validity of their corpus, the COWS-L2H.

In the end, both statistical and neural techniques (based or not on machine translation) are the most popular approaches, mainly due to their potential. Nevertheless, they require a large amount of corrected text in order to be able to learn, something that is not always

easy to find. The following part provides a short overview of the available datasets in English, while Section 2.2.4 explores some artificial solutions to this data sparsity problem.

2.2.2 Corpora

GEC as a task has traditionally been focused on educational applications. As a result, most of the existing corpora consists on non-native texts written by language learners.

For instance, the **First Certificate in English (FCE)** corpus (Yannakoudakis et al., 2011), a freely available subset of the private Cambridge Learner Corpus, is a collection of written answers to questions of the Cambridge exam of the same name. The **Lang-8** corpus (Tajiri et al., 2012; Mizumoto et al., 2012) is a multilingual collection of learner texts from the language exchange website[2] with the same name, where users can ask for corrections for their writings. The **National University of Singapore Corpus of Learner English (NUCLE)** (Dahlmeier et al., 2013) is a collection of manually corrected student essays written by students of the National University of Singapore. The **WikEd** corpus (Grundkiewicz and Junczys-Dowmunt, 2014) is somewhat different, as it consists on data-mined Wikipedia sentences and their revisions. This is a process known as corpora generation, which is explained in more detail in Section 2.2.4.

Additionally, there are three corpora that are used for benchmarking. The **CoNLL-2014 shared task test set** (Ng et al., 2014), released as part of said conference, is probably the most widely used. It includes 50 manually corrected essays written by students of the National University of Singapore specifically for this task. Models tested against this dataset use the $F_{0.5}$ metric, where precision weights twice as much as recall. Next, the **JFLEG (JHU FLuency-Extended GUG)** corpus (Napoles et al., 2017) aims to extend GEC corrections to include not only "minimal edits" for incorrect grammar, but also fluency edits. This means that annotators were allowed to rewrite sentences more freely if needed in order to make them sound natural. For this reason, models that use this dataset are evaluated using GLEU (Napoles et al., 2016), a fluency metric derived from BLEU.

Finally, the **W&I+LOCNESS** (Bryant et al., 2019), which was released in 2019 as part of the BEA (Building Educational Applications) shared task, joins text from two different sources: the Cambridge English Write and Improve website (Yannakoudakis et al., 2018)—where English students can get feedback for their texts—and the LOCNESS corpus (Granger, 1998), a collection of essays written by native speakers. By joining both sources, this corpus covers a wider range of topics and language proficiency levels than the CoNLL-2014 test set. Evaluation is performed using span-based correction $F_{0.5}$.

The W&I+LOCNESS corpus was automatically annotated using the ERRor ANnotation Toolkit (ERRANT) (Bryant, 2019), a toolkit developed specifically to annotate parallel GEC datasets. As part of the shared task, some of the corpora explained earlier (namely the FCE, Lang-8 and NUCLE corpora) were also re-annotated using ERRANT and re-released in an attempt to standardize their annotation format.

This section has described the most prominent corpora available in English, but cor-

[2] https://lang-8.com/

pora in languages such as German (Boyd, 2018), Russian (Rozovskaya and Roth, 2019) or Czech (Náplava and Straka, 2019) also exists. In Spanish, the learner corpus COWS-L2H (Davidson et al., 2020) was recently released.

The following section will explain more in detail what ERRANT is, how it works and its usages.

2.2.3 ERRANT

The **ERRor ANnotation Toolkit (ERRANT)** (Felice et al., 2016; Bryant et al., 2017) is a tool that automatically annotates original and corrected parallel sentence pairs. It was designed with the aim of making the annotating process easier and more homogeneous, as well as facilitating error type evaluation.

One of the main characteristics of this framework is that it is "dataset-agnostic" (Bryant et al., 2017): it does not depend on any type of labelled data like a machine learning model would do. Instead, it relies on linguistic information that is extracted from the text itself, such as part-of-speech, dependency tags or lemmas and stems. To extract this information, ERRANT uses freely available tools like spaCy[3] and NLTK[4]. It also uses Damerau-Levenshtein distance calculations and a vocabulary list for spelling-related errors.

Thanks to the way it automatizes the whole process, ERRANT is a valuable tool to create a standard annotation scheme for GEC. For this reason, since its release it has been used to both annotate new corpora (e.g. Hagiwara and Mita (2020)'s GitHub Typo Corpus) and re-annotate existing ones (e.g. the re-release of the FCE, Lang-8 and NUCLE corpora for the BEA 2019 shared task (Bryant et al. (2019) mentioned in Section 2.2.2)). Its rule-based system ensures that annotations are consistent and it makes it easy to trace back the reason why an edit has been classified in a certain way.

ERRANT works in the following way: to start with, original and corrected sentences are compared in order to automatically extract edits. This is an alignment task that is performed using a version of the Damerau-Levenshtein algorithm that incorporates linguistic features into its cost function and a set of merging rules, as proposed by Felice et al. (2016).

Once the edits in a parallel sentence have been aligned, each of them is assigned an error type by means of a rule-based classifier. The classifier attaches different levels of granularity depending on the type of error it finds. First, edits are assigned an edit operation: *missing* (M), *unnecessary* (U) or *replacement* (R). These are analogous to the classic edit operations: insertion, deletion and replacement.

Next, a general category is assigned using part-of-speech tagging, syntactic dependencies and token information. There are categories for verb-related errors (tagged as VERB), determiners (DET), punctuation (PUNCT), spelling (SPELL), ... The error types are based on language-agnostic Universal Dependency POS tags. Edits that do not fit into any of the categories are classified as OTHER. Finally, some POS-tagged errors can re-

[3]https://spacy.io
[4]http://www.nltk.org/

		Operation Tier		
	Type	**Missing**	**Unnecessary**	**Replacement**
Part Of Speech Tier	Adjective	M:ADJ	U:ADJ	R:ADJ
	Adverb	M:ADV	U:ADV	R:ADV
	Conjunction	M:CONJ	U:CONJ	R:CONJ
	Determiner	M:DET	U:DET	R:DET
	Noun	M:NOUN	U:NOUN	R:NOUN
	Particle	M:PART	U:PART	R:PART
	Preposition	M:PREP	U:PREP	R:PREP
	Pronoun	M:PRON	U:PRON	R:PRON
	Punctuation	M:PUNCT	U:PUNCT	R:PUNCT
	Verb	M:VERB	U:VERB	R:VERB
Token Tier	Contraction	M:CONTR	U:CONTR	R:CONTR
	Morphology	-	-	R:MORPH
	Orthography	-	-	R:ORTH
	Other	M:OTHER	U:OTHER	R:OTHER
	Spelling	-	-	R:SPELL
	Word Order	-	-	R:WO
Morphology Tier	Adjective Form	-	-	R:ADJ:FORM
	Noun Inflection	-	-	R:NOUN:INFL
	Noun Number	-	-	R:NOUN:NUM
	Noun Possessive	M:NOUN:POSS	U:NOUN:POSS	R:NOUN:POSS
	Verb Form	M:VERB:FORM	U:VERB:FORM	R:VERB:FORM
	Verb Inflection	-	-	R:VERB:INFL
	Verb Agreement	-	-	R:VERB:SVA
	Verb Tense	M:VERB:TENSE	U:VERB:TENSE	R:VERB:TENSE

Table 1: Table with all 55 error types in the English ERRANT, taken from Bryant (2019).

ceive a more specific classification in order to highlight a specific phenomenon. For example, subject-verb agreement errors receive the VERB:SVA tag.

All in all, around 50 rules are used, resulting in 55 error type combinations. These types are presented in Table 1. ERRANT does not define all possible error types as it would always be possible to increase the level of granularity. Instead, it "aims to be a compromise between informativeness and practicality" (Bryant, 2019).

At the end of the process, ERRANT outputs a new file with the sentences annotated in M2 format, the current standard annotation format for GEC. Figure 1 shows an example sentence in M2. In this format, each sentence is presented as a block along with the extracted corrections. The original sentence is introduced by an S, while each edit is in its own line, preceded by an A. Edit lines have the following fields, separated by three vertical bars: start and end token offset, error type, corrected string, whether the edit is optional or required, a comment and an annotator ID. The optional/required and comment fields

```
S This are a sentence .
A 1 2|||R:VERB:SVA|||is|||-REQUIRED-|||NONE|||0
A 3 3|||M:ADJ|||good|||-REQUIRED-|||NONE|||0
A 1 2|||R:VERB:SVA|||is|||-REQUIRED-|||NONE|||1
A -1 -1|||noop|||-NONE-|||REQUIRED|||-NONE-|||2
```

Figure 1: Example of a sentence annotated in M2 format, taken from Bryant et al. (2019). This example shows three different corrections for the sentence 'This are a sentence': 'This is a good sentence' (annotator 0), 'This is a sentence' (annotator 1) and 'This are a sentence' (annotator 2 made no changes; the term 'noop' indicates that there were no corrections made).

are no longer used but kept for historical reasons.

ERRANT also incorporates its own scorer that uses the $F_{0.5}$ metric. It is able to evaluate a system's overall performance as well as provide a more detailed evaluation in terms of error types. According to Bryant (2019), this is something that had not been done until the release of this framework because manually annotating a system's hypothesis is expensive and impractical. The automatic annotation process makes it possible to effortlessly perform both edit operation and error type analysis, allowing us to discover the strengths and weaknesses of a model.

ERRANT's performance was evaluated by 5 researchers who rated 200 random edits as 'Good', 'Acceptable' or 'Bad'. Their test showed that at least 95% of the predictions were rated as 'Good' or 'Acceptable', and that many of the 'Bad' edits were the results of part-of-speech tagging or parsing errors. It might be concluded that the toolkit's annotations are comparable to those performed by humans, specially given that GEC is often a highly subjective task (Bryant and Ng (2015), as cited in Bryant (2019)).

Even though it was originally developed for English, ERRANT is a flexible toolkit and can easily be adapted to other languages by performing some adjustments. Linguistic information can be obtained using the same tools just by specifying a different language. The most challenging change is adapting the rules it uses for categorization to the kind of errors one may expect in the new language. Still, not all of approximately 50 rules the framework uses need to be changed as some may be general enough to be universal. A new vocabulary list should also be provided, which gives us the chance to include any domain specific words that our task may use. More concrete details on the changes made for the adaptation of ERRANT to Spanish that is part of this thesis are given in Section 4.2.

2.2.4 Data augmentation

Data augmentation is a common technique used in fields that make use of neural networks, such as computer vision, that consists in artificially creating more data that stems from the data we have available. It has been demonstrated that, even if the resulting data

Origin	the **price** of alcohol is **ramped** up at every budget .
Generated	the **puice** of alchool is **ramping** up at every budget .

Table 2: Example sentence generated with a general AEG system that makes use of linguistic features, taken from Xu et al. (2019).

is not true to real life, these techniques can help create more robust models and reduce overfitting even when using smaller datasets (Anaby-Tavor et al., 2019; Xie et al., 2019). Data augmentation has proven itself to be a powerful technique to boost performance in machine learning tasks in an economic and efficient way.

In computer vision, the simplest versions of augmentation may amount to flipping, cropping or rotating images (Shorten and Khoshgoftaar, 2019). However, when it comes to natural language processing, aspects like syntax or semantics make it significantly harder to come up with universal ways to alter text without it losing its meaning. Over the last few years many proposals have been made, amongst the most popular being lexical substitution, back-translation or noise injection (see Zhang et al. (2015); Wang and Yang (2015); Kobayashi (2018)).

In Grammatical Error Correction, data augmentation has become a popular solution to the lack of parallel data. It has been shown that artificial errors and corpora can be of great help to improve performance (Junczys-Dowmunt et al., 2018; Kiyono et al., 2019). The most popular techniques are mostly specific to GEC: artificial error generation and corpora generation. As GEC often uses architectures from machine translation, back-translation is also used at times.

Artificial error generation (AEG) is the act of corrupting error-free sentences in order to create parallel correct/incorrect sentence pairs (Felice, 2016). Essentially, the premise that augmentation for NLP is hard due to any errors we may introduce becomes a non-issue as errors are precisely what we want to learn. The biggest challenge here is finding a corpus that is actually well-written. AEG can be performed on any kind of text, which gives us some control over the characteristics of the text we use. Both rule-based systems (Felice and Yuan, 2014) and machine learning models (Rei et al., 2017) can be used for augmentation. An example sentence generated using AEG is shown in Table 2.

There are two different variables that are crucial when it comes to artificial error generation: what kind of errors we want to introduce and how to introduce them. On the one hand, the type of errors can be either general or specific. General errors are based on some general operations that do not follow the same error typology of a reference corpus. The most basic implementation of this approach would simply introduce changes in our sentences based on edit operations (insertion, deletion, replacement and swap). For replacement, confusion sets can be generated for each word in the vocabulary (Grundkiewicz et al., 2019). Other features, such as part of speech or word length, are also used.

Introducing specific errors requires us either to study the error typology in our reference corpus and create a rule-based system in order to replicate them or to learn them automatically using machine learning in an unsupervised manner.

On the other hand, the way in which these errors are distilled into a text can also be

decided. Each error can be given a probability which can be assigned at random, follow a existing distribution (uniform, normal, ...) or use the distribution of the reference corpus we want to work with (which could be smoothed to leverage poorly populated classes). A study by Felice (2016) shows that random generation increases recall while decreasing precision, and probabilistic generation increases precision while decreasing recall. They argue that this is to be expected as "generating errors at random is likely to produce errors in new contexts and achieve more coverage, while following the same distributions as in the reference corpus will make a system more confident in flagging known errors".

Corpora generation consists on retrieving text from the web which has revisions that can also be retrieved, allowing us to have a parallel corpus almost from the get-go (Cahill et al., 2013). The main source used for this technique is Wikipedia, as it is easy to access articles' revision histories. The problem with this approach is that the resulting parallel text has not been curated for GEC. Revision histories may include changes made because of vandalism or simply to rephrase a sentence or add a citation. For this reason, the authors who use Wikipedia for corpora generation usually apply some kind of filtering, such as the reason of the revision, edit distance, sentence length or more complicated heuristics (Cahill et al., 2013). Some authors even use ERRANT annotations as a filter for the edits (Boyd, 2018). Sometimes, synthetic errors are introduced in the resulting corpora (Lichtarge et al., 2018) as a final step. Another possible approach is to use machine translation to translate an existing corpus (Katsumata and Komachi, 2019).

The main advantage of corpora generation is that it allows us to have big amounts of text for a small cost. Some of the disadvantages are that the parallel sentences may not include the kind of errors that our original corpus does and that it can be hard to find an in-domain source for more specific tasks. The WikEd corpus (Grundkiewicz and Junczys-Dowmunt, 2014) is an example of a freely available corpus generated using this technique.

Back-translation is also used in GEC as a way to introduce noise in clean corpora (Kasewa et al., 2018). However, this technique requires large amounts of data in order to train a model which can create the noise in an unsupervised manner, which is something that is not always available.

3 Methodology

This section lays out the objectives of this work in more detail and the steps taken to meet them. The overall workflow of the thesis is described visually in Figure 2.

Before starting this project, I worked on the annotation of the negation and uncertainty corpus NUBes (Lima López et al., 2020a). During that time, we realized that the source texts were often poorly written, including many orthographic and grammatical errors. It was then decided to explore this problem further in order to try to improve the overall quality of these texts.

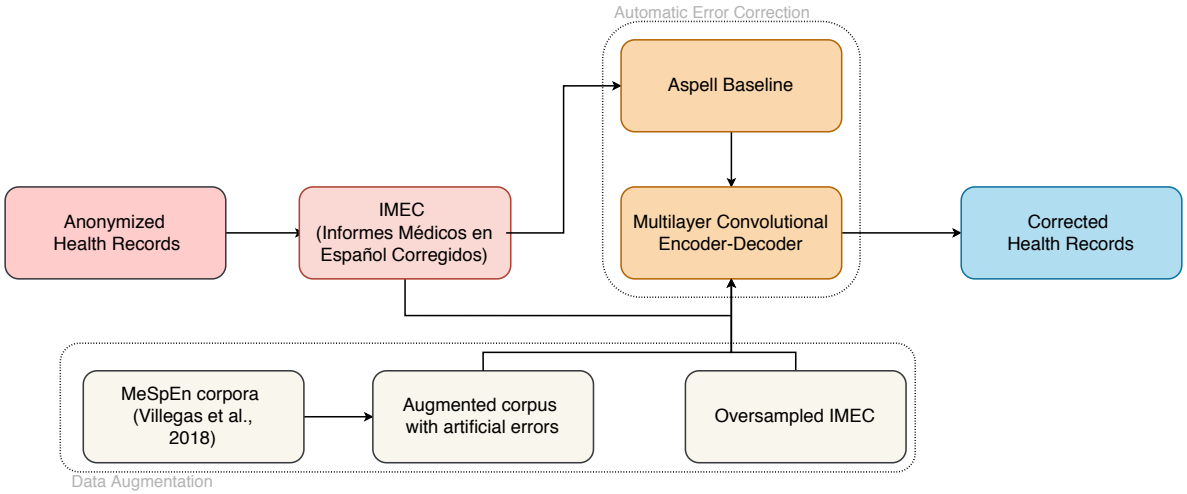

Figure 2: Overview of the different stages of this work.

The first step was to consider how to approach the task. Different disciplines, such as automatic post-editing or text normalization, were initially considered. However, due to the diversity of errors we had observed in the corpus, GEC was thought to be the most appropriate approach.

After doing some research on GEC in Spanish, as well as in specific domains, it became clear that there was a lack of resources that could be used for our purpose. Since we had the source texts of the NUBes corpus (Lima López et al., 2020a), this was no problem as we had enough material to prepare our own corpus. Following the convention of the latest GEC research papers, the texts were manually corrected and annotated automatically using the toolkit ERRANT, which had to be adapted to Spanish specifically for this task. The result was named IMEC: 'Informes Clínicos en Español Corregidos' (Corrected Health Records in Spanish). The correction and annotation process, as well as its content, are described in Section 4.

The corpus was divided into test, train and dev sections and used for experimenting. This whole process, as well as the results obtained, are explained in Section 5. First, a baseline system was created using a spellchecker (described in Section 5.1. Next, the Multilayer Convolutional Encoder-decoder system presented by Chollampatt and Ng (2018) was used to train a convolutional neural network (Section 5.2).

Still, we were presented with the problem that, when compared with other GEC corpora, the IMEC corpus is not too large. For that reason, data augmentation techniques were also explored. Two techniques were used: automatic error generation on some of the MeSpEn corpora collection (Villegas et al., 2018) and oversampling of the IMEC corpus. These augmentations were used in different ways, along with the original IMEC train set, to train new models using the same Multilayer Convolutional Encoder-decoder system. Section 5.3 explains this part of the experimentation.

At the end, the results of all of the experiments and their output are explored in more detail in Section 5.4.

4 Corpus Presentation

IMEC ('Informes Médicos en Español Corregidos' or *Corrected Health Records in Spanish*) is a collection of corrected anonymized Electronic Health Records, presented in a parallel fashion. The records were originally provided by a Spanish private hospital, which were also used for the negation and uncertainty corpus NUBes (Lima López et al., 2020a). They were anonymized in two steps: first, all Personal Health Information (PHI; identifiers such as names, dates, locations, ...) was manually annotated using the annotation tool BRAT; then, these items were replaced with similar items using a system based on rules and dictionaries specifically designed for this task (Lima López et al., 2020b).

The corpus is made up of 10,007 sentences, out of which 7,801 have at least one correction. The original, or source file, has over 160,000 tokens, while the corrections, or target file, has almost 180,000 tokens. The total number of corrections is somewhat over 27,000.

This section describes how the corpus was annotated, focusing on the logic behind the corrections and the automatic annotation using ERRANT, and the different kind of errors found in it.

4.1 Correction process

The annotation process was carried out by a single annotator, the author of this work, who corrected all sentences manually. An annotation guide, available in Annex A, was written as part of the annotation process. These sentences were later automatically annotated using an adaptation to Spanish of the annotation toolkit ERRANT developed specifically for this work.

Even though having a corpus corrected by only one annotator is not ideal, reannotating the corpus in the future is a possibility. For instance, the CoNLL-2014 test set was reannotated multiple times, up to a total of 18 overlapping annotations Bryant (2019).

Before starting to correct the sentences, a thorough analysis of the special characteristics of clinical texts—introduced in Section 2—was performed. This was a necessary step in order to become familiar with the genre and become aware of the do's and don'ts.

For instance, the three principles laid out by Bello Gutiérrez (2016) were specially useful in order to set boundaries of what should or should not be part of our corrections. Each principle contributed in its own way: (i) the veracity principle led me to forfeit annotating any possible lexical or semantic errors, as I considered that my expertise in medicine is not sufficient to correct them without unintentionally changing the meaning of the original sentence; (ii) in order to respect the precision principle, it was decided not to disambiguate abbreviations. Even though they are the largest source of ambiguity in the corpus, many of them have multiple possible disambiguations and treating them manually requires specific knowledge. On top of that, disambiguation per se is a whole different task on itself which would make the scope of this research much bigger. Instead, it was decided that they should simply be normalized by adapting their spelling following the Real Academia Española y Asociación de Academias de la Lengua Española (2014) guidelines. This should make treating abbreviations simpler for future works or disambiguators; (iii)

the last principle, clarity, is the most crucial. We consider that clarity emerges not only from the text's content, but also from its structure. This led us to include some aspects, such as enumerations, as part of the task's scope.

Regarding specific corrections and changes to the text, two style guides, Bello Gutiérrez (2016) and Aguilar Ruíz (2013), as well as the Real Academia Española's dictionary (Real Academia Española y Asociación de Academias de la Lengua Española, 2014) were the main references. Many of the decisions taken during the annotation process are rooted in these documents. However, as the main focus of IMEC is the correction of orthographical and grammatical errors, their suggestions weren't blindly followed. Some of the cases they describe are stylistic choices rather than errors, which sometimes means that multiple forms of the same phenomena are accepted. One example of this is the spelling of the prefixes *pos-*/*post-*, where both forms are accepted, and thus were not changed whenever they appeared.

4.2 ERRANT Adaptation

After correcting the sentences, the annotation toolkit ERRANT (Felice et al., 2016; Bryant et al., 2017) was adapted to Spanish in order to annotate the corpus. As explained in Section 2, ERRANT allows us to automate the annotation process by extracting differences from parallel texts and categorizing them using a rule-based system. Some of the changes that had to be made in order to adapt it to Spanish include finding adequate resources such as dictionaries, selecting language-specific tools for tasks like tokenization (namely, spaCy[5] and NLTK[6]) and the development of new rules that describe some of the language's idiosyncrasy, as well as the removal of some English-specific rules.

Since many of the recent corpora released for GEC has been annotated using ERRANT, it was decided that we should follow on their footsteps. Ultimately, one of the main reasons we decided to use it to annotate our corpus was its implementation in the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19). As explained in Section 2.2.2, as part of the Workshop multiple datasets were re-released after being re-annotated with ERRANT in order to standardize them. We think that the adoption of the framework might push its annotation format as a standard for the GEC task. Having a standard is important as it allows for easier and fairer comparisons, and the framework's apparent quality makes it a suitable candidate for it. We agree with this idea and support it by adapting ERRANT for Spanish GEC.

Only a few rules had to be manually adapted to account for language-specific structures. For example, in the English ERRANT the ADJ:FORM category is used for comparative and superlative adjective errors. This phenomenon works differently in Spanish, where these type of adjectives are formed by simply adding 'más' or 'el más' to any adjective instead of using the suffixes '-er' and '-est'. Thus, it does not make sense to apply this spe-

[5]https://spacy.io
[6]https://nltk.org

cific category to Spanish texts. Another instance is the addition of a rule for determiners' agreement in genre, something that does not exist in English.

Some rules specific to errors found in the corpus were also added, as some cases kept getting misclassified. These are rules mainly related to spelling mistakes, such as the correction of wrongly spelt abbreviations (e.g. 'kilograms' abbreviated as 'kgr.' instead of the correct spelling 'kg'). Nevertheless, this was avoided as much as possible as the addition of more rules also increases the chance of conflict between them.

One of the major changes, however, is the distinction between the SPELL and ORTH categories. In English, SPELL is only used for spelling mistakes, while ORTH includes case and whitespace errors. For the Spanish adaptation, these are all categorized as SPELL. ORTH has been re-purposed for errors regarding orthotypographic errors such as enumerations or the usage of symbols (e.g. '<', '>' or '=' used as abbreviations instead of their written counterparts).

The complete list of changes is the following:

- ADJ:FORM category renamed to ADJ:INFL, includes gender and number agreement errors.
- New category called DET:INFL added for agreement errors.
- NOUN:INFL now encompasses all agreement errors; NOUN:NUM is deprecated.
- NOUN:POSS category was eliminated as it does not apply to Spanish.
- Addition of new rules for specific spelling (SPELL) mistakes.
- Changed scope of SPELL and ORTH categories, as described above.
- Subordinating conjunctions have their own category (SCONJ). This was added on purpose, but rather automatically as some of the rules depend on part-of-speech tags.

It must be kept in mind that, as helpful as ERRANT may be, its output is not perfect and one may expect certain errors. For example, consider the following sentences:

1. 'Se pone colocó sng con salida de abundante contenido intestinal , pero sin mucha mejoria , por lo que el 3/7 se realiza TAC con sospecha de isquemia de ciego'
   'Se pone SNG con salida de abundante contenido intestinal , pero sin mucha mejoría , por lo que el 3/7 se realiza un TAC con sospecha de isquemia de ciego.'
   'A nasogastric tube is inserted with an outflow of abundant intestinal contents, but not many improvements, for this reason on 3/7 a CT scanning is performed with the suspicion of a cecum ischemia.'

2. 'Intercurre enn su evoluci` ´ on con una celulitis del MIIzq por lo que recibe 14 dias de trat ATB con buena respuesta clinica y curacion de la infeccion'
   'Intercurre en su evolución con una celulitis del MII , por lo que recibe 14 días de trat. ATB con buena respuesta clínica y curación de la infección .'

'Intercurrent left lower limb cellulitis during their evolution, because of which they receive 14 days of antibiotic treatment with good clinical response and healing of their infection.'

On the one hand, Example 1 showcases a correction that happens in a specific sentence or context. There is a repeated verb that was deleted, 'pone colocó', which is the only instance of U:VERB in the entire corpus. On the other hand, in the latter example (2) an example of a misannotation on both spaCy's and ERRANT's side. The word 'evoluci` ´ on', with a span of 3 tokens, was corrected to 'evolución', with a span of only one token. Let's have a look at the annotations created by ERRANT for this specific correction.

```
A 3 4|||R:NOUN|||evolución|||REQUIRED|||-NONE-|||0
A 4 5|||U:NOUN||||||REQUIRED|||-NONE-|||0
A 5 6|||U:ADJ||||||REQUIRED|||-NONE-|||0
```

The first correction's span corresponds to 'evoluci`' and was annotated as R:NOUN following ERRANT's rules. The second one corresponds only to '´' and the third one to 'on'. These last two, as expected, were incorrectly parsed by spaCy, resulting in them being categorized as noun and adjective respectively.

Even if the examples above are specially complicated and not the general rule, it would be interesting to evaluate this adaptation in the same way Bryant (2019) evaluated the original English version (mentioned in Section 2.2.3) as part of future work to test its performance.

The next section presents IMEC's error distribution and analyze the phenomena found in the corpus.

4.3 Error analysis

In contrast with the usual corpora used for GEC (see Section 2.2.2), our corpus has a lot fewer grammatical mistakes. This is to be expected, as those corpora originate in second language learners' speech, while it can be assumed that (most) of the writers of the text in our corpus are native Spanish speakers. Due to the sometimes rushed nature of their work, however, it is natural that mistakes may arise in their writing. This could shed light on the reason why the corpus is so focused on orthographic and orthotypographic mistakes, and, specially, why many words are omitted. The error type distributions are shown in Table 3.

From a linguistic point of view, we may divide the errors found in the corpus into two big groups: orthotypographic and syntactic errors.

4.3.1 Orthotypographic errors

These are the most common errors in the corpus, including categories such as SPELL (44.38 %), PUNCT (14.23 %) or ORTH (1.43 %), which accumulate over half of the corrections in the corpus.

Type	Number	%
M	12,505	46.13
R	14,414	53.20
U	184	0.67
ADJ	42	0.15
ADJ:INFL	66	0.24
ADV	15	0.06
AUX	76	0.28
CONJ	57	0.21
DET	6,959	25.68
DET:INFL	56	0.21
MORPH	81	0.30
NOUN	367	1.35
NOUN:INFL	65	0.24
ORTH	388	1.43
OTHER	787	2.90
PREP	1,254	4.63
PRON	37	0.14
PUNCT	3,829	14.13
SCONJ	7	0.03
SPELL	12,024	44.36
VERB	921	3.40
VERB:FORM	19	0.07
VERB:SVA	25	0.09
VERB:TENSE	24	0.09
WO	4	0.01

Table 3: IMEC's edit and error type distribution.

Spelling errors are very diverse. The most common cases include missing accents ('codeina' instead of 'codeína') and typos due to missing ('refire' instead of 'refiere'), excessive ('elitilca' instead of 'etílica') or transposed characters ('esca3a' instead of 'escala').

Incorrect spelling of proper names belonging to diseases, disorders or drugs were also corrected after checking whether they were correct. According to (Terroba Reinares, 2015, p. 145), these errors are common because of phonetic or orthographic similarities.

SPELL also encompasses casing mistakes and inconsistencies. For instance, some sentences start in lowercase and end all in uppercase (see Example 3), while others use uppercase all the time (5):

3. 'Desde las 18 horas deL DIA 22/11/2018...'
   'Desde las 18 horas del día 22/11/2018...'
   'Since 22/11/2018 at 18 hours...'

4. 'PAciente que refiere desde hace 2 dias odinofagia'
   'Paciente que refiere desde hace 2 días odinofagia.'
   'Patient that refers odynophagia for the past 2 days.'

5. 'Intervenciones Secundarias - 53.00, REPARACION UNILATERAL DE HERNIA INGUINAL'
   'Intervenciones secundarias: 53.00, reparación unilateral de una hernia inguinal'
   'Secondary surgery: 53,00, unilateral reparation of an inguinal hernia'

There are also errors that are caused by other languages' influence (e.g. months are spelled using uppercase in English, but not in Spanish). Example 6 shows two different mistakes in this regard: the spelling of the month *abril* using uppercase letters and the incorrect spelling of an English loanword (*screening*).

6. 'En Abril de 2003 en escreenning de cancer colorrectal...'
   'En abril de 2003 en un screening de cáncer colorrectal...'
   'On April 2003, in a colorectal cancer screening...'

Casing errors are also found in drugs' names, as well as in scientific names that often have both a Latin and Spanish names, as there are specific rules as to how these should be written. Drug names should be in uppercase if they refer to a brand-name drug and in lowercase if they refer to a generic drug (as in Example 7). Scientific names that use the Latin equivalent are written in uppercase; however, if they are composed of more than one word, only the first word is in uppercase (see Example 9). Spanish terms, in contrast, are always written in lowercase. Oftentimes, we also find orthographic errors in the cases described above.

7. 'Ha tomado paracetamol esta mañana y hace 30 minutos enantyum.'
   'Ha tomado paracetamol esta mañana y hace 30 minutos Enantyum.'
   'The patient took acetaminophen this morning and Enantyum 30 minutes ago.'

8. 'Hace 2 días es valorado por MAP e inicia tto con Amoxicilina 500/125 + AInes i.m. [...]'
   'Hace 2 días es valorado por MAP e inicia un tto. con amoxicilina 500/125 más AINE IM [...]'
   '(The patient) was examined by their PCP two days ago and started a treatment with amoxicillin 500/125 and NSAIDs IM [...]'

9. 'Varón de 75 años, remitido desde la CCEE de urología (Dr. Sanchez) por nueva ITU por Morganella Morganii.'
   'Varón de 75 años, remitido desde la CCEE de Urología (Dr. Sánchez) por una nueva ITU por Morganella morganii.'
   '75-year-old male referred by Urology's external consultations (Dr. Sánchez) due to a new UTI caused by Morganella morganii.'

As Example 9 shows, symbols are also corrected whenever they are used as abbreviations. Some sentences are abbreviated by means of replacing actual words with symbols that represent the same meaning, an usage that is not recommended by the Real Academia Española y Asociación de Academias de la Lengua Española (2014). The following are some more examples of this phenomenon:

10. 'La cefalea se ha controlado intermitentemetne con Maxalt + Nolotil 27 8h.'
    'La cefalea se ha controlado intermitentemente con Maxalt más Nolotil 27 8 h.'
    'The headache was controlled sporadically using Maxalt and Nolotil 27 8 h.'

11. 'En la analítica previa al alta la urea era de 32 mg %, la creatinina de 0,81 mg % , sodio de 134 meq/ly potasio de 2,54 meq/l.'
    'En la analítica previa al alta la urea era de 32 mg por ciento, la creatinina de 0,81 mg por ciento, el sodio de 134 mEq/l y el potasio de 2,54 mEq/l.'
    'In the lab tests performed before discharge, urea was 32 mg per cent, creatinine was 0.81 mg per cent, sodium was 134 mEq/l and potassium was 2.54 mEq/l.'

Abbreviations are actually one of the most common elements in clinical text in general (Terroba Reinares, 2015). They are so frequent that sometimes their use is simply excessive. For instance, in Example 12 'izquierda' and 'evolución' are common words that have probably been shortened due to their frequency, but that could have perfectly been written in their full form. The problem is that, at times, this kind of abbreviations are unnecessary and may hinder the understanding of the sentence (see Example 13, 14).

12. 'Temporal izda [izquierda] con basocelular nodular de más de 1 año de evol [evolución].'
    'Left temporal with a nodular basal cell with more than one year of evolution.'

13. 'Peor en espacios cerrados (st [sobre todo] tumbado en la cama) y con cambios de T^{a} [temperatura] y corrientes de aire.'
    'It gets worse in enclosed spaces (specially when [the patient is] lying down in bed) and with temperature changes and air currents.'

14. 'se informa a la Flia [familia] de la gravedad del cuadro clinico.'
    'The patient's family is informed of the clinical picture's seriousness.'

If we include spelling errors in the formula, interpreting the meaning of some sentences can be incredibly difficult for someone with little medical background:

15. 'El resto de l expl NLG no es valorable por la afasia.'
    'El resto de la expl. NLG no es valorable por la afasia.'
    'The rest of the neurologic exploration cannot be evaluated due to the aphasia.'

On top of that, multiple forms are often used to refer to the same concept. The opposite is also true, multiple concepts are referred to using the same form:

16. ‘Paciente de 44 años sin A.p. [antecedentes personales] de interés’
    ‘Paciente de 44 años sin AP de interés.’
    ‘44-year-old patient with no personal medical history of interest.’

17. ‘Mejoria de la ap [arteria pulmonar/atresia pulmonar/auscultación pulmonar/...], con menos crepitantes, SatO2 100 [...]’
    ‘Mejoría de la AP, con menos crepitantes. Sat. de 02 del 100 % [...]’
    ‘Improvement of the *AP*, with fewer crepitations., 100 % oxygen saturation [...]

18. ‘Derivada de AP [atención primaria] por absceso periamigdalino.’
    ‘Derivada de AP por un absceso periamigdalino.’
    ‘Referred by Primary Care due to a peritonsillar abscess,’

Normalizing the different forms of an abbreviation to a single, unified spelling is a task on its own in clinical NLP. Because of this, even if developing them might make texts easier to understand, whenever abbreviations appear only their spelling is corrected:

19. ‘A.C.P: Pulso ritmico.’
    ‘ACP: pulso rítmico.’
    ‘Cardiopulmonary auscultation: rhythmic pulse.’

Measurement units’ abbreviations were the only ones that were normalized. These are often written using multiple spellings: ‘kilograms’ may be spelled ‘kg’, ‘kgr’, ‘Kg’, ‘k.g.’, ‘kgs’, ‘kgrs’, ... There are, however, specific rules laid down for this exception. The Orthography released in 2014 (Real Academia Española y Asociación de Academias de la Lengua Española, 2014) states that they should always be written in lowercase and that, even if they are in plural, their form does not change. They should not generally ever end with a dot nor have one inside its components. The following are some examples of these corrections:

20. ‘- Paracetamol 1 gr., 1 c. cada 8 h.’
    ‘- Paracetamol 1 g, 1 c. cada 8 h.’
    ‘- Acetaminophen 1 g, 1 pill every 8 hours.’

21. ‘Pasó de 127 a 93 Kgr.’
    ‘Pasó de 127 a 93 kg.’
    ‘(The patient) went from 127 to 93 kg.’

The last major SPELL error type has to do with whitespaces: sometimes they are missing and sometimes they are used when they should not. For instance, measurement units should always be written separately from the quantity they measure, as in Example 22. A major example of whitespaces being used incorrectly regards the spelling of prefixes and suffixes, as well as of compound words.

22. 'Comienzan tratamiento con Levofloxacino 500/24h hace 4 días con cierta mejoría.'
'Comienzan un tratamiento con levofloxacino 500/24 h hace 4 días con cierta mejoría.'
'A treatment with levofloxacin 500/24 h was started 4 days ago with some improvements.'

Another important category is **punctuation**. Punctuation is a double-edged sword: if used correctly, it can be of great help, specially for long-winded sentences; at the same time, it can quickly become a hurdle if it's used in an incorrect situation.

An example of punctuation being helpful is given in 23, where multiple sentences were separated with dots, making them easier to read. Example 24 shows how punctuation may be used incorrectly.

23. 'TA: 128/65; Tª 37,1 sat 98 % Auscultación cardio pulmonar normal Heridas torácicas de drenajes previos'
'TA: 128/65; Tª de 37,1 ºC, sat. del 98 %. Auscultación cardiopulmonar normal. Heridas torácicas de drenajes previos.'
'Blood pressure: 128/65; temperature of 37.1 ºC, oxygen saturation of 98 %. Normal cardiopulmonary auscultation. Thoracic wounds from a previous drainage.'

24. 'Estando previamente bien, hace unas horas, después de comer,. comienza con un dolor abdominal en hipocondrio derecho, contínuo, acompañado de malestar gral. y naúseas.'
'Estando previamente bien, hace unas horas, después de comer, comienza con un dolor abdominal en el hipocondrio derecho, continuo, acompañado de malestar gral. y náuseas.'
'(The patient) was fine earlier, but a few hours ago, after lunch, they started showing abdominal pain in the right upper quadrant, continuous, together with physical discomfort and stomach sickness.'

Weird usage of punctuation was also corrected, as some sentences included unusual symbols used in places where others may be more fitting (see Example 5 where a dash is replaced by a colon).

There's a special case of punctuation being missing that is specific to the medical domain: drugs' dosage schedules. These are numeric patterns that are usually introduced mid sentence, sometimes within parenthesis and sometimes without any presentation, almost as a foreign element. In an effort to standardize them, parenthesis were added when they were missing (see Example 25).

25. 'Actualmente se encuentra en tratamiento con tramadol 50mg 1-1-1 + tetrazepam y paracetamol 650mg'
'Actualmente se encuentra en tratamiento con tramadol 50 mg (1-1-1) más tetrazepam y paracetamol 650 mg.'
'(The patient is) currently in treatment with tramadol 50 mg (1-1-1) and tetrazepam and acetaminophen 650 mg.'

4.3.2 Syntactic errors

The second major group of errors in the corpus includes syntactic errors. These can be the most disruptive errors, as they may hinder communication by making the meaning of a sentence harder to understand. Even if the content of our corpus has been produced by native speakers, there are some recurrent syntactic errors. They can sometimes be attributed to language economy reasons and hastiness on the doctor's behalf. Within ERRANT's taxonomy, all error types that are referred to using a part-of-speech tag (e.g. ADJ, AUX, CONJ, ...) are considered syntactic errors.

Categories named after part-of-speech tags are assigned either when both sides of an edit have the same part of speech or one of them is missing. If each side has a different part of speech, it is considered to be related to morphology (MORPH) and is annotated as such. The most common syntactic errors found are related to three categories: DET, PREP and VERB. The reason behind it is that these words are often omitted in order to make sentences shorter. This creates sentences that may be understandable but ungrammatical. Thus, many corrections simply add the missing word in the correct place. This is easy for prepositions and determiners, as they are closed class words. They are subject to restrictions such as collocations or grammatical genre, which can be used as clues to guess the missing word.

26. 'Extirpación lesión pabellón auricular derecho'
    'Extirpación de una lesión en el pabellón auricular derecho.'
    'Removal of an injury in the right pinna.'

27. '[...] ésta viene presentando desde las pasadas Navidades, cuadro consistente en episodios de alucinaciones auditivas [...]'
    '[...] esta viene presentando desde las pasadas Navidades un cuadro consistente en episodios de alucinaciones auditivas [...]'
    '(The patient) has shown since last Christmas a history of auditory hallucionations.'

However, open class words such as verbs are harder to reconstruct. Many different words can be used in the same context, at times with almost no change in meaning. Still, even if a word is perfectly valid, it cannot be claimed that said word reflects the speaker's original intention. For this reason, and in order to make the corpus more homogeneous, it was then decided that only a small subset of verbs should be used. The list includes some verbs that could be used in broad contexts, such as 'haber' (*there is/are*) or 'estar' (*to be*). In some cases, sentences structures found in the corpus were mimicked, which led to the inclusion of verbs like 'mostrar' (*to show*) or 'referir' (*to recount*). More verbs were added to this list if none of the already used ones fit into the sentence (see Example 29).

28. 'No evidencia de sangrado activo'
    'No muestra evidencia de sangrado activo.'
    'There is no evidence of active bleeding.'

29. 'El 21 de Agosto 2006 By-pass gástrico.'
    'El 21 de agosto de 2006 se coloca un bypass gástrico.'
    'On August 21st 2006 a gastric bypass is placed.'

Further corrections of the IMEC corpus performed by new annotators may correct these same sentences using different verbs in order to have more varied examples.

Moving on, the more specific error types (such as ADJ:INFL or DET:INFL) usually involve genre or number agreement issues (see Example 30 and 31).

30. 'IC descompensada probablemente por un infeccion'
    'IC descompensada probablemente por una infección.'
    'Decompensated heart failure probably due to an infection.'

31. 'No clinica de infeccion respiratorio'
    'No muestra una clínica de infección respiratoria.'
    '(The patient) does not show symptoms of a respiratory infection.'

Specific verb error types, even if not too common, are a little more detailed. Other than subject-verb agreement (VERB:SVA), there are also categories for incorrect tense choices (VERB:TENSE) and finiteness changes (VERB:FORM). Example 32 shows a classic agreement error, where the verb should be in plural but is not.

32. 'se aprecia los datos ya conocidos en la ecografía abdominal de una hepatopatía crónica ya conocida'
    'Se aprecian los datos ya conocidos en la ecografía abdominal de una hepatopatía crónica ya conocida'
    'The already-known data of an already-known chronic liver disease is appreciated in the abdominal ultrasound.'

It is interesting to note that this sentence could be further corrected, as at first sight it seems to include redundant information ('datos ya conocidos' *already reported data* and 'hepatopatía crónica ya conocida' *already reported chronic liver disease*). However, as explained in Section 4.1, we forfeit from correcting this type of errors out of fear of accidentally removing relevant medical information. It might be the case, for example, that each of these phrases refer to different events.

Finally, the WO (word order) tag is assigned whenever an edit consists of the same two words in different order. This is not common, as there are only four cases in the whole corpus. These are two of them:

To sum up, this section has described the IMEC corpus, its annotation process and its content. Next, the different experiments performed with it will be explained.

5 Experimentation

After correcting and annotating the corpus, some experiments were undertaken in order to ascertain its validity. This section explains the different experiments that were performed: setting up a baseline system, training a convolutional neural network and using data augmentation to improve performance.

The results obtained by each model are also presented. For this part, the IMEC corpus was divided into three parts: train, dev and test. The size of each partition is shown in Table 4.

Partition	Sentences	%
Train	7506	75 %
Dev	1500	15 %
Test	1000	10 %

Table 4: Partitions' size of the IMEC corpus.

For each experiment, a corrected version of the same test set was generated using the corresponding system. Then, this output was annotated using ERRANT and compared to an annotated version of the Gold Standard test set. As explained in Section 2, the models were evaluated using the $F_{0.5}$ measure, which values precision twice as recall.

5.1 Baseline

In English GEC, earlier research and corpora (mentioned in Section 2.2.2) can serve as a comparison point for new experiments in the field. However, up to this day no research has been written proposing GEC systems for Spanish nor for clinical texts. Therefore, there is no benchmark or baseline against which the performance of the experiments using IMEC can be compared. Having a baseline is crucial when experimenting with machine learning. A baseline is a simple model that sets a minimum score for the rest of our trials. The aim of our experiments is to outperform this basic model to prove the effectiveness of our model.

It is hard, however, to create a model that is both simple and able to correct all error types. Thus, given the number of spelling mistakes in the corpus, it was decided that a spell checker would be enough for the task. Since IMEC contains such a high number of orthographic errors, it was deemed that a spellchecker could be a good starting point.

In Spanish, there are some spellcheckers focused on the clinical domain that could be used to correct IMEC. These are, namely, CorrectM (Flores, 2020)—a free plain text editor with a specialized dictionary—, Spellex Medicina (Corp, 2020)—a paid extension for Microsoft Word that incorporates medical terms into its spellchecker—and COM (Corrector Ortográfico Médico) (Merino Torre, 2015)—a spellchecker developed by a student of the University of the Basque Country (UPV/EHU) as their degree's final project.

However, these spellcheckers have certain downsides: first of all, not all of them are freely available. Next, only COM provides any data on how they were evaluated and how

efficient they are. Finally, none of them can be used for an entire text at once. Instead, a replacement must be manually chosen from a list of options for each word that is deemed incorrect.

For these reasons, it was decided that using our own spellchecker would be more useful. Ultimately, the free software Aspell[7] was chosen, as it is a renowned spellchecker that allows for some customization.

Aspell works in the following way: each word in a sentence is first checked against a dictionary; a word is considered to be misspelled if it is not in the dictionary Aspell uses. Whenever an incorrect word is found, it is converted to its soundslike equivalent, an approximate representation of its pronunciation, using the Metaphone algorithm developed by Philips (1990). Then, Aspell tries to find all words with a soundslike that is within one or two edit distances from the original. The results are scored using "the weighed average of the weighed edit distance of the word to the misspelled word and the soundslike equivalent of the two words" (Atkinson, 2020). A list of ranked suggestions is returned, being the result with the lowest score the most likely correction.

Our baseline experiment consists on correcting sentences using Aspell's best suggestion. In addition to using Aspell on its own, three further tweaks were made: (i) the dictionary was expanded using a vocabulary list extracted from IMEC's train set; (ii) a Levenshtein distance threshold for suggestions was set; (iii) the suggestions provided were re-ranked once more using a language model. This language model was trained on the MedLine Plus corpus, which is part of the MeSpEn collection (Villegas et al., 2018), using the KenLM toolkit (Heafield, 2011) with a window size of 5.

System	Precision	Recall	$F_{0.5}$
Aspell	26.44	**16.44**	23.57
Aspell +TRAIN VOCAB	52.62	14.99	**35.03**
Aspell +LM	17.27	10.69	15.38
Aspell +LM +TRAIN VOCAB	30.01	08.59	20.06
Aspell +LEV=1 +LM	37.95	14.65	28.79
Aspell +LEV=1 +LM +TRAIN VOCAB	**54.80**	13.23	33.65

Table 5: Results of the Aspell baseline.

The baseline's results obtained using Aspell are presented in Table 5. On its own, the spellchecker achieved acceptable results. As the table shows, using a specialized, in-domain vocabulary greatly boosted performance. However, any attempts at re-ranking the results, either using 3-gram model trained on the MedLine Plus corpus or cap the suggestions at a given Levenshtein distance, seem to only interfere with Aspell's own re-ranking and lower performance. In general, using Aspell returns good precision scores but really low recall.

[7] https://aspell.net

5.2 Multilayer Convolutional Encoder-decoder

Other than this baseline, one of the main goals of this work is to learn to correct texts automatically using deep learning. As explained in Section 2.2.1, in the past years most of the state-of-the-art systems have used Transformers as their deep learning architecture of choice. However, given their computational expensiveness and the range and time restrictions of this thesis, this kind of model was not seen fit for experimenting. A somewhat smaller architecture can still provide good results while allowing the experimentation process to be swifter and more flexible.

Due to code availability and its former status as state-of-the-art at the time on its release, Chollampatt and Ng (2018)'s multilayer convolutional encoder-decoder neural network was chosen for the experiments. This is an architecture that is widely used in machine translation, and that has been proven to work well for GEC tasks. The authors argue that their approach is effective for this task for two reasons: convolutional neural networks (CNN) are better at capturing local context than the frequently-used recurrent neural networks; at the same time, using multiple layers allows the network to capture wider contexts and interactions as well. Figure 3 illustrates the general idea of this network.

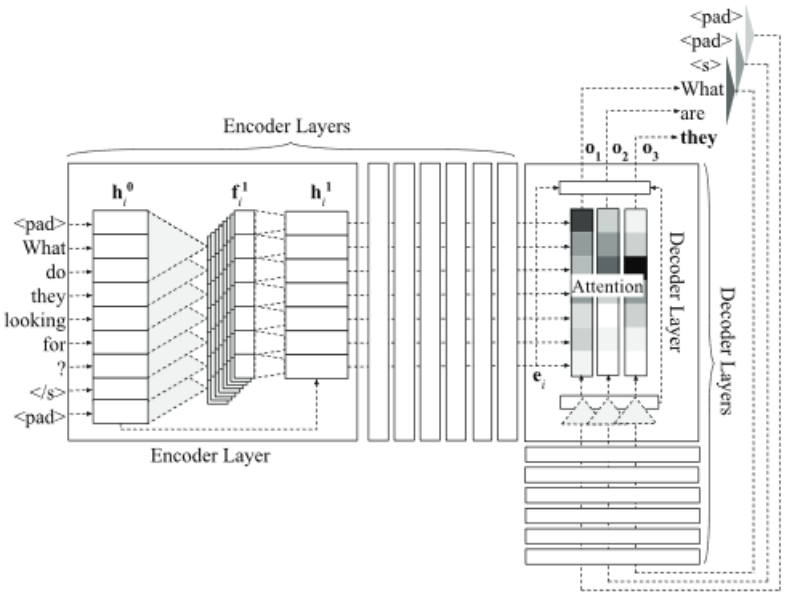

Figure 3: Architecture of the multilayer convolutional model with seven encoder and seven decoder layers, taken from Chollampatt and Ng (2018).

In GEC, the encoder network encodes "the potentially erroneous source sentence in vector space", while "the decoder network generates the corrected output sentence by using the source encoding" Chollampatt and Ng (2018). To begin with, each source token is embedded into a continuous space that is calculated using pre-trained word embeddings, as well as position embeddings. The resulting embeddings, which are also trained together

along with other parameters of the network, are linearly mapped to obtain what is called an input vector. Next, sequences of three consecutive input vectors are mapped to a feature vector using convolutional filters. The result is then followed by a non-linear gated linear unit (GLU) and added to the input vectors to an encoder layer. Finally, "each output vector of the final encoder layer is linearly mapped to get the encoder output vector" Chollampatt and Ng (2018).

As for the decoder, new embeddings are generated, starting with two paddings and beginning-of-sentence marker. Next, "each embedding is linearly mapped to and passed as input to the first decoder layer'. For each layer, 'convolution operations followed by non-linearities are performed on the previous decoder layer's output vectors". Decoder layers have their own attention module, which uses a combination of its own weights and biases and are used to obtain a source context vector. Finally, "the decoder output vector is mapped to the target vocabulary size and softmax is computed to obtain target word probabilities".

Overall, the model is trained using the negative log-likelihood loss function and its parameters are optimized using Nesterov's Accelerated Gradient Descent (Bengio et al. (2012), as cited in Chollampatt and Ng (2018)). At the end of the decoding process, left-to-right beam search is used to obtain the most adequate sequence of target words. As the authors explain, "the top-scoring candidate in the beam at the end of the search will be the correction hypothesis" Chollampatt and Ng (2018).

Additionally, a few tricks are used to improve performance. On the one hand, the authors pre-process their data using the byte-pair encoding (BPE) algorithm (Sennrich et al., 2016), which splits rare words into sub-words. This helps minimize the number of out-of-vocabulary words, which results extremely helpful in a specialized domain such as medicine. Word embeddings are trained on a large corpora split using the same BPE model obtained from the task dataset. On the other hand, after training is done, an n-best list of corrections is generated and the multiple candidates are scored using a large language model and edit operation features.

For this project's experiments, the word embeddings were trained with fastText (Bojanowski et al., 2017) using a dump of the Spanish Wikipedia. For the re-ranker, in an effort to test the difference between using in- and out-of-domain corpora, three different language models were used. For in-domain, the same medical corpus used for the Aspell baseline, MedLine Plus (Villegas et al., 2018), was used. For out-of-domain, a joint version of multiple NewsCrawl dumps in Spanish released as part of the Conference on Machine Translation (WMT) 2019 (Barrault et al., 2019) was chosen. In comparison with MedLine Plus—which has a little over 400,000 lines and 6 million tokens, NewsCrawl is pretty big. For this reason, a smaller subset of the same size as the former was extracted in order to make the comparison between both fairer. The entire NewsCrawl, with a size of over one and a half billion tokens, was also used to train an additional language model to test whether and how corpus size affects performance.

The results of using the multilayer convolutional encoder-decoder on the IMEC corpus are presented in Table 6. This table (as well as the subsequent ones) also present the outcome of performing n-best re-ranking on the model with different language models. As

System	Precision	Recall	$F_{0.5}$
IMEC	42.36	**41.26**	42.14
IMEC +MedLine	45.23	38.79	43.78
IMEC +News220k	45.21	38.27	43.63
IMEC +NewsAll	**46.41**	41.03	**45.22**

Table 6: Results of the multilayer convolutional encoder-decoder.

a reminder, three different language models were used: MedLine Plus, a small subset of the Spanish NewsCrawl (signalled by the term *News220k*) and the entire NewsCrawl (called *NewsAll* in the tables below).

Overall, the results obtained from training a CNN using the IMEC corpus are better than those obtained by the baseline, especially when it comes to recall. Re-ranking also seems to have an effect, giving a performance boost of up to three points.

5.3 Data augmentation

5.3.1 Automatic Error Generation

Additionally, as the original corpus is somewhat small for the standards that neural networks follow these days, we opted for artificially creating more data that could be used to reinforce our network. The idea was to corrupt well-written medical documents in a way that replicated the error types and distribution found in the corpus. These expanded datasets were used to train new models using the same architecture described in the section above.

A short Python program was developed for inducing errors into clean text. It uses a set of handcrafted rules that introduce different edits into the text, such as adding or removing words based on their POS tag, introducing typos or changing the inflection of a word. Each rule has its own probability that is manually assigned. The number of changes in a sentence is randomly chosen withing a range of (0, 4].

Not all error types in IMEC were included, however. Some categories with few members were hard to recreate because many items were the result of an accumulation of errors in the pipeline, either due to spaCy's preprocessing (mistakes in POS-tagging) or to ERRANT's segmentation. Others were not necessarily hard to recreate, but they occurred only a few times and in a specific context (e.g. changing the conjunction 'y' to 'e' and vice-versa).

A total of 24 different rules were developed. They were used to induce errors in four different clinical corpora: IBECS, SciELO, Pubmed (all three are part of the MeSpEn collection by Villegas et al. (2018)) and SPACCC (Intxaurrondo, 2018). These were chosen because they are from the biomedical domain, which should be close to clinical notes. Before augmenting them, they were heavily preprocessed for two reasons: firstly, to eliminate any unwanted text (some of the text was in English, and it also included a lot of bibliographical references); secondly, to normalize some aspects in order to make it as similar as possible to our corpus (for instance, punctuation use or measurement units' abbreviations'

spelling).

Type	Number	%
M	3,021,303	38.76
R	4,728,619	60.66
U	44,099	0.56
ADJ	9,521	0.12
ADJ:INFL	10,845	0.14
ADV	7,174	0.09
AUX	1,420	0.02
CONJ	23,973	0.31
DET	2,017,786	25.89
DET:INFL	95,764	1.23
MORPH	39,468	0.51
NOUN	50,418	0.65
NOUN:INFL	7,668	0.10
ORTH	26,665	0.34
OTHER	322,858	4.14
PREP	1,146,550	14.71
PRON	9,745	0.13
PUNCT	1,394,900	17.90
SCONJ	933	0.01
SPELL	2,387,712	30.64
VERB	236,658	3.04
VERB:FORM	197	0.00
VERB:SVA	2	0.00
VERB:TENSE	3,437	0.04
WO	352	0.00

Table 7: Augmented corpus' edit and error type distribution.

Altogether, the resulting corpus has a size of over 2.3 million lines and 51 million tokens with almost 8 million annotations. Examples 33 and 34 show some of the resulting sentence pairs[8], while Table 7 describes the error distribution of the corpus. When compared with the original distribution, most categories are similarly balanced, although some like PREP or OTHER have grown and others like SPELL have decreased in size. It is important to keep in mind that, since the generation process was done randomly, re-running the program would result in a similar but different distribution.

33. 'Conlcusiones en escolares LA cobertura VACUNAL sistemática y la antimeningocócica a + C es alta.'

[8]In these examples, the augmented sentence is presented first in roman letters, while the original is presented right after in italics.

System	Precision	Recall	$F_{0.5}$
IMEC +AUG	62.57	35.43	54.26
IMEC +AUG +MEDLINE	62.14	37.11	54.75
IMEC +AUG +NEWS220K	**64.00**	36.14	55.45
IMEC +AUG +NEWSALL	63.89	**38.42**	**56.41**

Table 8: Results of the convolutional neural network trained using additional artificial data.

'Conclusiones: en escolares la cobertura vacunal sistemática y la antimeningocócica A + C es alta.'
'Conclusions: in schoolchildren, the systemic vaccination coverage and the meningococcal A + C are high.'

34. 'Posteriormwente quimioterapia adyuvante carboplatino y paclitaxel finalizando en jumnio de 2011.'
'Posteriormente recibió quimioterapia adyuvante con carboplatino y paclitaxel, finalizando en junio de 2011.'
'Afterwards, (the patient) received adjuvant chemotherapy with carboplatin and paclitaxel, up until June 2011.'

For the experiments, the training set consisted on the augmented corpus concatenated with IMEC's train partition. The results of this model are laid out in Table 8.

5.3.2 Oversampling

Finally, an additional data augmentation technique was used on the IMEC corpus itself. Oversampling is a technique used mostly in classification tasks that consists on duplicating examples from a minority class in order to balance the dataset. It has also seen some use in GEC studies focused on low-resource settings (for instance, see Náplava and Straka (2019)). In these cases, the sentences pairs are duplicated without regarding whether the errors they include are from a minority class or not, as the objective is not to balance the corpus but simply to expand it.

For the experiments, the training section of the corpus was repeated a variable number of times to explore how performance changed by training a different model with each of them. An overview of these tests is provided in Figure 4, which shows that oversampling does indeed improve both precision and recall. There are peaks at various levels and a steep decline after a certain point. Table 9 shows the results of the highest peak and its re-ranking.

Finally, a model using both the oversampled IMEC and the augmented biomedical corpus at the same time was also trained. It used IMEC's training set oversampled 15 times and joint with the augmented corpus. Combining both augmentation techniques with the original corpus actually delivers the best results. Training a model using an

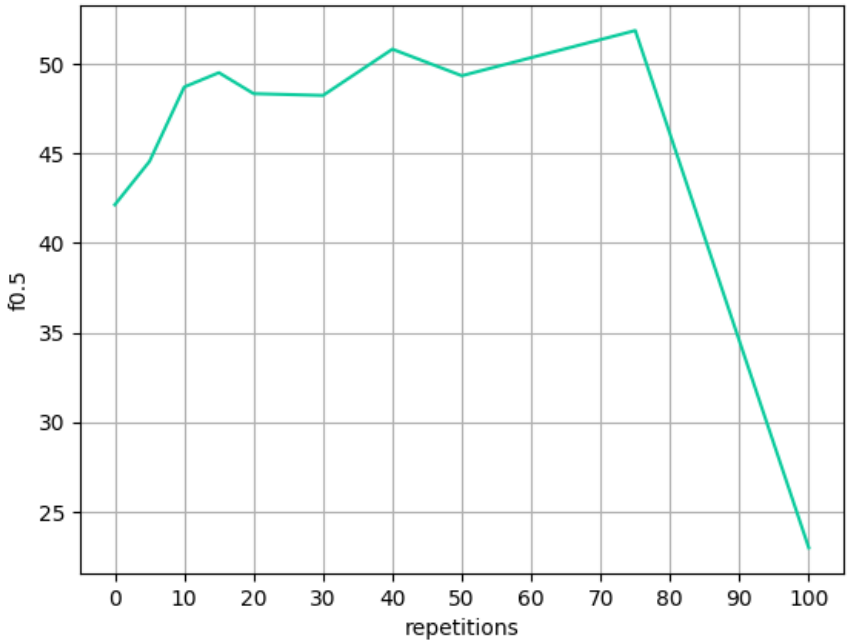

Figure 4: Performance of the convolutional neural network at different oversampling levels.

System	Precision	Recall	$F_{0.5}$
IMEC +OVERS75	53.63	45.81	51.86
IMEC +OVERS75 +MEDLINE	**62.09**	42.71	56.92
IMEC +OVERS75 +NEWS220K	62.01	42.94	**56.95**
IMEC +OVERS75 +NEWSALL	59.85	**45.52**	56.31

Table 9: Results of the best oversampling value with re-ranking.

System	Precision	Recall	$F_{0.5}$
IMEC +AUG +OVERS15	73.71	**59.19**	70.26
IMEC +AUG +OVERS15 +MEDLINE	**76.00**	55.87	**70.89**
IMEC +AUG +OVERS15 +NEWS220K	75.30	55.94	70.43
IMEC +AUG +OVERS15 +NEWSALL	75.81	55.64	70.69

Table 10: Results obtained from the combination of both data augmentation techniques.

oversampled version of IMEC and the augmented corpus as training data elevates the scores up to 70.

It is worth noting that the oversampling size used together with augmentation is different from the size that returned the best results when using oversampling on its own. When experiment with oversampling on its own, bigger numbers seem to generally return better results (up to a point when it declines). However, when training together with the augmented corpus, a small number such as 15 seems to perform better. It could be argued that there's some overfitting going on at higher oversampling levels while lower ones generalize better.

5.4 Discussion

To sum up, the best results from each line of experimentation are shown in Table 11. As we can appreciate, using data augmentation techniques gives us the chance of boosting results without spending a lot on annotating more data or using a bigger architecture. Not only do they increase precision (how many of the corrections that were made were correct), but also recall (how many corrections were made out of the total number of gold corrections). I would argue that, for this task, precision is more important than recall. Even if higher recall indicates that our system is able to correct more diverse errors, correcting the errors that it does detect properly is more relevant for any practical applications.

System	Precision	Recall	$F_{0.5}$
Baseline (Aspell +TRAIN VOCAB)	52.62	14.99	35.03
IMEC +NEWSALL	46.41	41.03	45.22
IMEC +OVERS75 +NEWS220K	62.01	42.94	56.95
IMEC +AUG +NEWSALL	63.89	38.42	56.41
IMEC +AUG +OVERS15 +MEDLINE	**76.00**	**55.87**	**70.89**

Table 11: Best results of each system.

Interestingly enough, the results obtained by the model trained with the augmented corpus are coherent with the theory presented by Felice (2016) about probabilistic generation, presented in Section 2.2.4: a probabilistic generation of errors increases precision while decreasing recall.

Additionally, since each system achieved better results using a different language model for n-best re-ranking, we cannot conclude whether the language model's size and domain

do matter. It is clear, though, that re-ranking is a valuable step that can improve our results.

Next, the strengths and weaknesses of each of the models in Table 11 will be compared in more detail. Fortunately, ERRANT allows an in-depth evaluation based on edit and error type with no extra effort. These detailed evaluations are presented in Table 12 for edit type and Tables 13 and 14. For simplicity's sake, in both of these tables each model has been renamed to its most salient characteristic. In Tables 13 and 14 there are some empty spaces for three categories that were part of the whole corpus but did not make it into the test set: ADV, VERB:TENSE and WO.

Architecture	Edit Type	Precision	Recall	$F_{0.5}$
Baseline	M	100.00	0.00	0.00
	R	54.80	25.02	44.26
	U	100.00	0.00	0.00
IMEC	M	52.76	42.27	50.27
	R	43.98	40.49	43.23
	U	0.00	0.00	0.00
Oversampled	M	66.96	43.08	60.27
	R	61.71	43.39	56.90
	U	0.00	0.00	0.00
Augmented	M	57.14	30.60	48.69
	R	69.27	45.72	62.8
	U	10.00	05.26	08.47
Oversampled + Augmented	M	69.75	47.34	63.72
	R	**80.84**	**64.10**	**76.83**
	U	0.00	0.00	0.00

Table 12: Comparison of each architecture's best model's performance at an edit operation level. M means *missing* (insertion), R means *replacement* and U means *unnecessary* (deletion).

When it comes to edit type corrections, it is striking that most of the models fail at correcting deletions (U), achieving zero points on all three measures. This might be due to the fact that, as Table 3 shows, these errors are by far the least common of all three types. The only two outliers are the baseline and the model that uses the augmented corpus. On the one hand, the baseline achieves a precision score of 100, meaning that it delivered no false positives. However, its recall of 0 indicates that it produced no true positives either. Essentially, the baseline model completely ignored this category. It also seems to ignore insertion (M) errors, which shows that, as expected of a spellchecker, it is only capable of replacing words. On the other hand, the augmented model does correct some deletion errors. However, its results are so poor that it cannot be considered significant.

In regards to the other two edit types, the output for both of them seems to be consistent in all of the models (except for the baseline). It is noteworthy that the oversampled model

Error Type	Architecture								
	Baseline			IMEC			Oversampled		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
ADJ	0.00	0.00	0.00	01.79	25.00	02.19	06.25	25.00	07.35
ADJ:INFL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ADV	-	-	-	-	-	-	-	-	-
AUX	100.00	0.00	0.00	0.00	0.00	0.00	33.33	14.29	**26.32**
CONJ	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DET	0.00	0.00	0.00	60.28	**44.36**	56.25	**65.05**	41.69	58.49
DET:INFL	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MORPH	100.00	0.00	0.00	0.00	0.00	0.00	14.29	12.50	13.89
NOUN	05.88	02.56	04.67	03.96	23.08	04.75	06.25	20.51	07.26
NOUN:INFL	0.00	0.00	0.00	100.00	40.00	76.92	**66.67**	**40.00**	**58.82**
ORTH	100.00	0.00	0.00	70.37	59.38	67.86	76.92	62.50	73.53
OTHER	0.00	0.00	0.00	02.39	13.85	02.86	07.89	18.46	08.92
PREP	100.00	0.00	0.00	46.07	30.60	41.84	68.57	35.82	57.97
PRON	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PUNCT	100.00	0.00	0.00	59.68	38.05	53.58	72.52	41.39	63.04
SCONJ	100.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
SPELL	57.96	29.42	48.54	77.77	43.83	67.30	79.74	46.58	69.81
VERB	100.00	0.00	0.00	47.31	46.32	47.11	68.75	57.89	66.27
VERB:FORM	0.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
VERB:SVA	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
VERB:TENSE	-	-	-	-	-	-	-	-	-
WO	-	-	-	-	-	-	-	-	-

Table 13: Comparison of each architecture's best model's performance at an error type level (1).

Error Type	Architecture								
	Baseline			Augmented			Oversamp+ Augm		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
ADJ	0.00	0.00	0.00	66.67	50.00	62.50	**75.00**	**75.00**	**75.00**
ADJ:INFL	0.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
ADV	-	-	-	-	-	-	-	-	-
AUX	100.00	0.00	0.00	33.33	14.29	**26.32**	16.67	14.29	16.13
CONJ	100.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
DET	0.00	0.00	0.00	53.47	26.26	44.29	64.60	43.32	**58.82**
DET:INFL	100.00	0.00	0.00	20.00	40.00	22.22	**40.00**	**40.00**	**40.00**
MORPH	100.00	0.00	0.00	08.33	25.00	09.62	**20.00**	**25.00**	**20.83**
NOUN	05.88	02.56	04.67	33.33	30.77	32.79	**61.29**	**48.72**	**58.28**
NOUN:INFL	0.00	0.00	0.00	40.00	40.00	40.00	50.00	40.00	47.62
ORTH	100.00	0.00	0.00	**94.74**	56.25	83.33	93.33	**87.50**	**92.11**
OTHER	0.00	0.00	0.00	16.33	12.31	15.33	**35.85**	**29.23**	**34.30**
PREP	100.00	0.00	0.00	46.38	23.88	39.02	75.86	49.25	68.46
PRON	100.00	0.00	0.00	100.00	25.00	62.50	25.00	25.00	25.00
PUNCT	100.00	0.00	0.00	58.40	35.73	51.83	**74.50**	**48.07**	**67.12**
SCONJ	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
SPELL	57.96	29.42	48.54	77.65	50.08	69.95	**85.36**	**68.00**	**81.21**
VERB	100.00	0.00	0.00	67.39	32.63	55.56	**77.03**	**60.00**	**72.89**
VERB:FORM	0.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
VERB:SVA	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
VERB:TENSE	-	-	-	-	-	-	-	-	-
WO	-	-	-	-	-	-	-	-	-

Table 14: Comparison of each architecture's best model's performance at an error type level (2).

is stronger at insertions (M) than replacements (R), while the augmented model does the opposite.

With respect to error type, one of the most remarkable details is the fact there is a clear difference between the performance of some categories. On top of that, out of the five models only the one that uses IMEC on its own seems to be able to correct all different types (even if it completely fails at some of them). It is probably caused by the unbalanced corpus distribution, with more frequent errors being easier to correct than rare ones. This points out to the ever more obvious fact that the quantity of data used for training, as well as its distribution, has a strong influence on the performance of deep learning models. This argument is supported by the fact that the most frequent categories (DET, SPELL, VERB, ...) also have the best results.

Another interesting detail is that the best model overall, the one that uses both oversampling and the augmented corpus, is not always the best at correcting all categories. Some error types, such as ADV or NOUN:INFL, get better results in some of the other models.

To conclude this section, some of the actual outputs of the models are discussed. These examples are shown in Annex B, at the end of this work. The sentences used were handpicked in order to showcase both the parts that the system learnt how to correct properly and the ones where it failed.

Overall, the examples shown are somewhat inconsistent in the sense that, as highlighted earlier when talking error type performance, the best corrections aren't always made by the best system. This can be seen in Example 1 (Augmented + IMEC is capable of correcting 'singo' for 'signo' (*sign*), but the Augmented + Oversampled is not). Still as Examples 3 or 11 show, spelling errors are usually well corrected. This is supported by the data shown in Tables 13 and 14.

Casing errors are somewhat in the middle. Some cases seem to have been corrected very well by most models. For instance, in Example 2 three out of the four models changed 'Amoxicilina' in uppercase to the same word in lowercase, the latter being the correct spelling. Big chunks of text in uppercase also seem to be able to be corrected, as Example 4 highlights. However, too many changes in this respect may lead the systems astray. In Example 15, the original sentence has been greatly changed by all of the systems, causing the loss of a lot of important information.

Orthography is also a strong point of most systems, having the capacity to insert commas (Example 3) or separate distinct sentences using dots (Examples 4 and 10).

Syntactic errors have been generally corrected adequately, specially when it comes to the ellipsis of determiners, preposition and verbs (see Example 2). Gender seems to be more complicated for our systems, as only a few of them choose the correct ending when adding words like determiners (e.g. Examples 6 and 9).

Abbreviations have sometimes been corrected properly, mainly when it comes to their casing (Examples 1 and 9). Sometimes, the system even seems to be able to recognize what is an abbreviation even if it fails at correcting it: in Example 11, the Augmented + IMEC model detects the abbreviation 'FEyVi' and changes it from lower to uppercase.

Surprisingly enough, the model has the ability to develop some simple, general domain

abbreviations even if the Gold Standard does not do it. In Example 2, the Augmented + Oversampled model changes 'gral' to its complete form 'general'.

Again, some examples show that some information may be lost when correcting, this time regarding abbreviations. While some of them have simply not been properly corrected (Example 5), others have been completely changed (Example 11 shows the change from 'FEyVI' to 'fèmr' and 'PSAP' to 'PAIP'). Measurement units' correction are also a mixed bag, with some models correcting them adequately while others simply change them (see Examples 11 and 14).

Symbols seem to be another source of confusion. Although they are sometimes properly corrected (e.g. Example 8 changes '+' to 'más' (*plus*)), since they may have more than one meaning, they may be confused (e.g. Example 9 changes '+', meaning *positive*, to 'más'). Some cases also seem to be particularly difficult (see Example 13), probably because they do not occur too often in the corpus.

Finally, an interesting point is the fact that sometimes the models return correct examples that are not evaluated as such since they differ from the Gold Standard. For instance, in Example 2 the Augmented + IMEC model inserts the verb 'presentar' (*to present*) instead of 'mostrar' (*to show*). This is actually correct, but due to the lack of multiple annotations for each sentence it is evaluated as incorrect.

6 Conclusions and future work

In conclusion, throughout this work I have presented a first approach to Grammatical Error Correction for health records in Spanish. Health records are very important documents, specially for patients. As explained in Section 1, these documents are the main form of communication between specialists and patients. However, their form is a flawed aspect that usually contains multiple orthographic and grammatical problems. As this work has shown, this problem can be lessened using Natural Language Processing techniques. In order to do so, this thesis introduces the IMEC ('Informes Médicos en Español Corregidos') corpus.

IMEC is made up of over 10,000 parallel sentences from health records in Spanish. Orthographic, grammatical and orthotypographic aspects were corrected using the suggestions of multiple sources, such as Aguilar Ruíz (2013) or Bello Gutiérrez (2016). It must be kept in mind that abbreviations and their disambiguation fell out of the scope of this work. A complete tool should include a module dedicated to said task in order to make the text as clear and straightforward as possible. IMEC was manually corrected and automatically annotated using an adaptation to Spanish of the ERRANT toolkit.

The corpus was used to carry out different experiments. First, a baseline was set using the Aspell spellchecker. This spellchecker returned decent result but was not able to correct many error types.

A convolutional neural network was also used to train multiple models with good results. On its own, the model trained using the IMEC corpus returned better results than the baseline, but still had plenty of room for improvement. For this reason, data augmentation techniques were used to try to improve performance and to overcome data sparsity. The training set of the corpus was oversampled and mixed together with a new corpus artificially generated by inducing errors into correct text. These techniques, both on their own and combined, greatly helped the model and improved the results almost by 30 $F_{0.5}$ points.

The experiments also explored how the domain and size of the different components of the training process, namely embeddings and language models, affect the final results. However, the final results seem to be inconclusive as there was no consistent change across all models. A more detailed evaluation of this aspect of the task, mixing multiple embeddings and language models trained on different corpora, would be another possible future development. The same could be said for the data augmentation techniques used in this work. For instance, they could be used to generate a new corpus that balances the corpus distribution.

Due to resources and time restrictions, some of the ideas of this thesis are left as future work. One of the most obvious would be to expand the IMEC corpus even further, allowing it to grow both in size and number of annotators. As a whole, the field of Grammatical Error Correction is more complex than it looks as some of the corrections can be very subjective. Thus, it is important to have datasets annotated by multiple annotators, as it allows a system to consider more than one correct sentence.

It would also be interesting to use the corpus to train new models using some of the most recent and powerful architectures. Again, many of the recent state-of-the-art models

(Omelianchuk et al., 2020; Zhao et al., 2019) use the Transformer architecture, which returns better results but is also more expensive. Big language models like BETO (Cañete et al., 2020) could also be implemented into the pipeline, for instance as part of the re-ranking mechanism as authors like Chollampatt et al. (2019) do.

Although perhaps not as important, evaluating ERRANT in Spanish in the same way it was evaluated in English (Bryant, 2019) would also be another possibility and it would further validate the results of this work.

There are various possible applications for an error correction system such as the one described in this work. Nevertheless, as seen in the examples analyzed in Section 5.4 and compiled in Annex B, the results from the best model are far from perfect, and sometimes unreliable as they might erase information. Because of this, in my opinion, the best practical application for this system would be in any setting where its corrections can be supervised, such as in a post-edition tool or as a real-time suggestion provider. Future iterations of this task that achieve better results may even perform the corrections in an unsupervised manner.

Another possible application would be to normalize text before it is processed by other NLP tools. Using normalized text that is rid of errors might improve said tools' performance. Because of this, it would be interesting to perform an extrinsic evaluation of the system to test its value in settings such as relation extraction or anonymization.

All in all, this work has shown that there are real life problems that have not been explored yet where NLP and artificial intelligence can make a change. As Gutiérrez et al. (2010) note, having computer tools that make the writing process simple and logical is important in a clinical setting. I hope that this first approach will result in more interest in this topic and the eventual development of tools that make both doctors' and patients' lives easier.

References

Manuel José Aguilar Ruíz. 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@: Revista de Medicina, Lenguaje y Traducción*, XIV (37):101–120, 2013.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Not Enough Data? Deep Learning to the Rescue! *ArXiv*, abs/1911.03118, 2019.

Kevin Atkinson. GNU Aspell 0.61 documentation, 2020. URL `http://aspell.net/0.61/man-html/index.html`.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301.

Pablo Bello Gutiérrez. Aprendiendo a redactar mejor tus informes. *Curso de Actualización Pediatría*, pages 391–400, 2016. URL `https://www.aepap.org/sites/default/files/4t2.14_aprendiendo_a_redactar_mejor_tus_informes.pdf`.

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. *CoRR*, abs/1212.0901, 2012.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Boletín Oficial del Estado. Real decreto 9/2015, de 6 de febrero, por el que se regula el registro de actividad de atención sanitaria especializada., 2015.

Adriane Boyd. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6111.

Chris Brockett, William B. Dolan, and Michael Gamon. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220207.

Christopher Bryant. *Automatic annotation of error types for grammatical error correction.* University of Cambridge, 2019. doi: 10.17863/CAM.40832.

Christopher Bryant and Hwee Tou Ng. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1068.

Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1074.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. Robust systems for preposition error correction using Wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517, Atlanta, Georgia, 2013. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*, 2020.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. Improved neural machine translation with a syntax-aware encoder and decoder. *CoRR*, abs/1707.05436, 2017.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014.

Shamil Chollampatt and Hwee Tou Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5755–5762, New Orleans, Louisiana, USA, 2018. AAAI Press.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. Cross-sentence grammatical error correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Spellex Corp. Spellex medicina, corrector ortográfico médico, 2020. URL https://tudiccionariomedico.com.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 7238–7243, Marseille, France, 2020. European Language Resources Association.

Mariano Felice. Artificial error generation for translation-based grammatical error correction. Number 895, 2016.

Mariano Felice and Zheng Yuan. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg, Sweden, 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-3013.

Mariano Felice, Christopher Bryant, and Ted Briscoe. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

Ignacio Mario Morales Flores. Corrector ortográfico para medicina correctm, 2020. URL https://www.cpimario.com/correctm.html.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 449–456. Asia Federation of Natural Language Processing, 2008.

Sylviane Granger. The computerized learner corpus: a versatile new source of data for sla research. In *Learner English on Computer*, pages 3–18, London & New York, 1998. Addison Wesley Longman.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing*, pages 478–490, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-10888-9_47.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4427.

Pedro Gutiérrez, Javier García-Alegría, Ramon Pujol, Inma Alfageme, Sara Menéndez, Raquel Barba, Pedro Cañones, Paloma Pérez, Fernando Alvaro, Luis Royo, Albert Fernández, and Cristóbal León. Consenso para la elaboración del informe de alta hospitalaria en especialidades médicas. *Medicina Clinica - MED CLIN*, 134:505–510, 2010. doi: 10.1016/j.medcli.2009.12.002.

Masato Hagiwara and Masato Mita. GitHub Typo Corpus: A Large-Scale Multilingual Dataset of Misspellings and Grammatical Errors. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6761—6768, Marseille, France, 2020. European Language Resources Association.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, 2011. Association for Computational Linguistics.

Ander Intxaurrondo. SPACCC, 2018. URL `https://doi.org/10.5281/zenodo.2560316`. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1055.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1541.

Satoru Katsumata and Mamoru Komachi. (almost) unsupervised grammatical error correction using synthetic comparable corpus. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 134–138, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4413.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,

pages 1236–1242, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1119.

Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2072.

John Lee and Stephanie Seneff. Correcting misuse of verb forms. In *Proceedings of ACL-08: HLT*, pages 174–182, Columbus, Ohio, USA, 2008. Association for Computational Linguistics.

Jared Lichtarge, Christopher Alberti, Shankar Kumar, Noam Shazeer, and Niki Parmar. Weakly supervised grammatical error correction using iterative decoding. *CoRR*, abs/1811.01710, 2018. URL `http://arxiv.org/abs/1811.01710`.

Salvador Lima López, Naiara Pérez, Montse Cuadros, and German Rigau. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pages 5772–5781, Marseille, France, 2020a. European Language Resources Association.

Salvador Lima López, Naiara Pérez, Laura García-Sardiña, and Montse Cuadros. HitzalMed: Anonymisation of Clinical Text in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pages 7038–7043, Marseille, France, 2020b. European Language Resources Association.

Nina H. Macdonald. Human factors and behavioral science: The UNIX Writer's Workbench software: Rationale and design. *The Bell System Technical Journal*, 62(6):1891–1908, 1983.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. Exploring grammatical error correction with not-so-crummy machine translation. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 44–53, Montréal, Canada, 2012. Association for Computational Linguistics.

Raúl Merino Torre. Editor de textos con corrector ortográfico para textos médicos, 2015. URL `https://addi.ehu.es/handle/10810/15733`.

Ministerio de Sanidad. Recursos físicos, actividad y calidad de los servicios sanitarios. 2018.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India, 2012. The COLING 2012 Organizing Committee.

Jakub Náplava and Milan Straka. Grammatical error correction in low-resource scenarios. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*, pages 346–356, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5545.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel R. Tetreault. GLEU without tuning. *CoRR*, abs/1605.02592, 2016. URL `http://arxiv.org/abs/1605.02592`.

Courtney Napoles, Keisuke Sakaguchi, and Joel R. Tetreault. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain, 2017. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Susanto, and Christopher Bryant. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-1701.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA, 2020. Association for Computational Linguistics.

Lawrence Philips. Hanging on the metaphone. *Computer Language Magazine*, 7(12):39–43, 1990.

Flora Ramírez Bustamante and Fernando Sánchez León. GramCheck: A grammar and style checker. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 175–181, 1996. doi: 10.3115/992628.992661. URL `https://www.aclweb.org/anthology/C96-1031`.

Real Academia Española y Asociación de Academias de la Lengua Española. *Diccionario de la lengua española*. Diccionario de la lengua española. Real Academia Española, 2014. ISBN 9788467041897.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. Artificial error generation with machine translation and syntactic patterns. *CoRR*, abs/1707.05236, 2017. URL `http://arxiv.org/abs/1707.05236`.

Stephen D. Richardson and Lisa C. Braden-Harder. The experience of developing a large-scale natural language text processing system: Critique. In *Second Conference on Applied Natural Language Processing*, pages 195–202, Austin, Texas, USA, 1988. Association for Computational Linguistics. doi: 10.3115/974235.974271.

Alla Rozovskaya and Dan Roth. Grammar error correction in morphologically rich languages: The case of russian. *Transactions of the Association for Computational Linguistics*, 7:1–17, 2019. doi: 10.1162/tacl_a_00251.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162.

Libin Shen, Anoop Sarkar, and Franz Josef Och. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

Connor Shorten and T. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019. doi: 10.1186/s40537-019-0197-0.

Enrique Silver. La información de los sistemas sanitarios y de los pacientes. *Quark: Ciencia, medicina, comunicación y cultura*, 16:19–22, 1999.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Ana Rosa Terroba Reinares. *Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico*. Universidad de La Rioja, 2015. URL `https://dialnet.unirioja.es/servlet/tesis?codigo=46993`.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden, 2010. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. The MeSpEN resource for English-Spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing"*, pages 32–39. European Language Resources Association, 2018.

William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1306.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Shuyao Xu, Jiehao Zhang, Jin Chen, and Long Qin. Erroneous data generation for Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–158, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4415. URL `https://www.aclweb.org/anthology/W19-4415`.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

Helen Yannakoudakis, Marek Rei, Øistein E. Andersen, and Zheng Yuan. Neural sequence-labelling models for grammatical error correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2795–2806, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1297.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267, 2018. doi: 10.1080/08957347.2018.1464447.

Zheng Yuan, Ted Briscoe, and Mariano Felice. Candidate re-ranking for SMT-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 256–266, San Diego, California, USA, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0530.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 649—657, Cambridge, MA, USA, 2015. MIT Press.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1014.

A Annotation Guidelines

This annex compiles the guidelines used to annotate the IMEC (Informes Médicos en Español Corregidos) corpus. In addition, we shortly present some characteristics of the clinical genre in order to justify the decisions taken during the annotation process.

Clinical notes are a recollection of a doctor's observations and thoughts on a patient's visit. They are read by patients themselves, other health professionals and, at times, even used as legal documents. At a linguistic level, Terroba Reinares (2015) shows that this type of documents have many similarities: irregular use of lower and uppercase, common use of nominalizations, use of Latin words and English loanwords, frequent usage of abbreviations, ... Additionally, they also share a wide range of syntactic structures. Due to their various functions, these can also have multiple senders as well as multiple receivers. All of this makes clinical notes a complex text genre and shows a need for correctness.

Bello Gutiérrez (2016) describes three desirable principles that any medical writing should show:

- **Veracity:** what is mentioned in the text should correspond both to reality and to what the author meant. In order to respect this idea, we decided to forfeit annotating any possible lexical or semantic errors. This includes phenomena such as pleonasms (i.e. usage of redundant words or phrases). We consider that this type of errors goes beyond the scope of what we intend to do, and that our expertise in medicine is not sufficient to correct them without unintentionally changing the meaning of the original sentence.

- **Precision:** ambiguous terms should be avoided so there is only one possible interpretation of a given message. The largest case of ambiguity in our corpus are abbreviations. However, even if this principle indicates that we should consider them, we decided not to for multiple reasons. First, disambiguation is a whole different task on itself and treating them would make the scope of our research much bigger. Second, in order to disambiguate them manually, we need certain medical knowledge that we do not possess. Instead, we will normalize them by adapting their spelling following the Real Academia Española's (Real Academia Española y Asociación de Academias de la Lengua Española, 2014) guidelines. This should make treating abbreviations simpler for future works.

- **Clarity:** Texts should be easy to understand for someone with some knowledge of the field. This implies not only using precise terms, but also avoiding rare grammatical structures that difficult the reader's understanding of the text.

Through our corrections, we are able to enhance the clarity principle. This will be performed through proper orthography, proper grammar and a homogeneous use of elements such as punctuation. This document describes some of the most common problems in the corpus related to those three points.

Many of the rules that we describe are based on the proposals by Bello Gutiérrez (2016) and Aguilar Ruíz (2013), as well as some of the recommendations of the Real Academia

Española y Asociación de Academias de la Lengua Española (2014)'s dictionary. Note that some of them are somewhat simplified in order to make the correction process easier and alleviate the number of exceptions, and that not all possible errors are listed here, as some are considered to be obvious to an educated Spanish speaker.

In the examples shown, the first sentence is the original one and is marked with an asterisk (*) to show it is incorrect, while the second one is its corrected version.

A.1 Orthography

Orthographic errors are probably the most common type of error in the corpus. For the sake of brevity, only some of the most relevant phenomena are presented.

Letter case

The use of uppercase and lowercase letters is one of the most irregular aspects of the corpus. There are various cases:

- Some words that are always written in lowercase include weekdays, months, languages, demonyms, job positions, titles, chemical elements, units of measurement, ...

  35. *'Historia referida por su Padre, ya que no habla nada de Castellano el Paciente'
  'Historia referida por su padre, ya que no habla nada de castellano el paciente.'
  'Clinical history explained by the father, since the patient does not speak any Spanish.'

  36. *'Valorada en hospital Dolors Aleu el 4 de Julio'
  'Valorada en el Hospital Dolors Aleu el 4 de julio.'
  'Examined in Dolors Aleu Hospital on July 4th.'

- Any sentence that is written using only uppercase or that mixes both cases is changed into regular case.

  37. *'Intervenciones Secundarias - 53.00, REPARACION UNILATERAL DE HERNIA INGUINAL'
  'Intervenciones secundarias: 53.00, reparación unilateral de una hernia inguinal.'
  'Secondary surgery: 53,00, unilateral reparation of an inguinal hernia'

- After a dash, the next word should start with uppercase; after a colon, it should be in lowercase.

- Diseases' names and drugs referred to using their active ingredient are written with lowercase letters; if drugs are referred to using a brand or commercial name, then they are written with initial uppercase. This is shown in Example 38, where 'ceftriaxona' (a generic name for an antibiotic) is written in lowercase letters, whereas 'Tamiflu' (the commercial name of an antiviral called oseltamivir) is written using

initial uppercase. At times, there might be spelling mistakes in drugs' names, so it is good practice to check whether they are correctly written.

38. *'Iniciamos tratamiento con antibioticos de amplio espectro (ceftriaxona 2gr/24h), esteroides, broncodilatadores, oxigeno y tamiflu'
'Iniciamos un tratamiento con antibióticos de amplio espectro (ceftriaxona 2 g/24 h), esteroides, broncodilatadores, oxígeno y Tamiflu'
'A treatment with broad-spectrum antibiotics (ceftriaxone 2 g/24 h), steroids, bronchodilators, oxygen and Tamiflu.'

- Scientific terms that refer to species are written using initial uppercase only if they are called using their Latin name. If these terms are composed of more than one word, only the first one uses uppercase. Spanish terms are written in lowercase. Compare sentences 39 and 40:

39. 'El 11/08 presentó un cuadro de bacteriemia con HC positivo para Staphylococcus aureus sensible a Augmentine .'
'On 11/08 (the patient) presented bacteremia with a blood culture positive for Staphylococcus aureus sensitive to Augmentin.'

40. 'Se objetiva en el cultivo estafilococo aureus meticilino sensible, iniciándose un tratamiento con cloxacilina (estuvo con él durante 6 semanas).'
'In the culture a Staphylococcus aureus sensitive to methicillin is found, beginning a treatment with cloxacillin (the patient followed it for 6 weeks).'

- Some common nouns can be written using uppercase whenever they refer to specific entities and institutions such as hospitals, universities, departments inside a health centre, ... Then, every word of the name uses initial uppercase. Compare examples 41 and 42: in the former, 'centro de salud' is a generic place and should be written in lowercase letters; in the latter 'Hospital' is part of the name of a specific place, and thus it is written with uppercase initial letter. Notice that in the latter example, 'Neurocirugía' is in uppercase as it refers to a real, specific department within that hospital.

41. *'Retirar puntos en su Centro de salud el día 23 Marzo'
'Retirar los puntos en su centro de salud el día 23 de marzo' 'The stitches will be removed at their health center on March 23rd.'

42. 'Se desestimó tto quirúrgico en su día por parte de Neurocirugia del Hospital Virgen del Palomar'
'Se desestimó un tto. quirúrgico en su día por parte de Neurocirugía del Hospital Virgen del Palomar'
'Surgical treatment was rejected in the past by Virgen del Palomar Hospital's Neurosurgery'

Accentuation

Generally, we follow the guidelines set by the latest edition of the Real Academia Española's dictionary. This includes some changes such as not using an accent for the disjunction 'o' when it is used between numbers or never to accent the adverb 'solo'. For more details, check Real Academia Española y Asociación de Academias de la Lengua Española (2014). Words that have more than one accepted spelling are not corrected (e.g. 'cardiaco' vs. 'cardíaco').

Prefixes spelling

Generally, prefixes are written together with the word they follow. There should not be any spaces between the prefix and the word, nor should they be joined using a dash even if the new word has a double consonant or vowel ('cooficial'). However, prefixes used together with acronyms, numbers or any word that is written with initial uppercase, such as proper nouns, are joint using a dash. This also applies to any scientific term that uses letters from the Greek alphabet (e.g. protein names). There is another exception to this: if a prefix modifies a multi-word expression, it should be written separately. Also, if there is more than one prefix modifying the same word, only the last one follows these established rules. The others should be written separated and with a dash at the end.

Moreover, there are some prefixes that can be written in different ways: *pos-*/*post-* and *tras-*/*trans-*. Both of these forms are usually correct, even if there might be a preferred one [9], so we will not correct them.

Shortenings

Inside this category, we distinguish two different cases: abbreviations and acronyms. We also consider measurement units a special kind of shortening. Following the principles described at the start of this guide, we are generally not concerned with their disambiguation, but rather with whether they are properly written.

- *Abbreviations:* They are the shortened version of one or more words. It is advised not to use them right at the start of the end of a sentence. The main rule is that, whether they refer to a job position, a country, a person's name, ... they must always end with a dot. They retain the gender, casing and accentuation of the original word. That is, if the accented letter also appears in the abbreviation, it is maintained in the abbreviation. Plurals are normally formed by simply adding an *-s* before the dot. If two abbreviations are together, they should separated by a space and not written together (e.g. *'p.ej.' / 'p. ej.').

  Furthermore, depending on how the abbreviation was formed, the rules for feminine construction vary. If the abbreviation is the result of truncation (only the first part of the word is used), then it is formed by adding a superscript *a* after the dot (e.g.

[9] https://www.fundeu.es/recomendacion/pos-y-post-uso-correcto-612/

'Prof.' / 'Prof.ª'). However, if the abbreviation is the result of clipping (the most salient letters are used), it is enough to add an *-a* before the dot.

Additionally, if the abbreviated word is a scientific term, it is recommended to always use the standard abbreviation. It is also not advised to use symbols in the text as abbreviations, such as the percentage symbol (%) for the word 'porcentaje'. An example of this is given in sentences 35 and 36.

35. 'es remitido por su médico por cuadro de fiebre >38ºC'
    'es remitido por su médico por un cuadro de fiebre de más de 38 ºC'
    '(The patient) is referred by their doctor due to fever higher than 38 ºC.'
36. 'en tratamiento actual con furosemida [...] + atorvastatina'
    'en tratamiento actual con furosemida [..] más atorvastatina'
    'Currently in treatment with furosemide [...] and atorvastatin.'

Finally, some authors recommend against using some abbreviations for commonly used words in an excessive manner, but rather limit their use to specific contexts (e.g. citations, tables, lists, ...). For this reason, we decided to always desambiguate abbreviations for the words 'izquierdo' (*left*), 'derecho' (*right*) and their variants.

- *Acronyms:* They are new words created by taking the initial letter of each part of a multi-word expression. In the scientific field, they may also emerge from the different components inside a word (e.g. 'ADN' comes from 'ácido desoxirribonucleico'). Unlike abbreviations, acronyms are usually written using uppercase, with no spaces or dots. They are not pluralized, but rather the same form is used with plural determiners (*'los AINES' / 'los AINE'). Another difference with abbreviations is that acronyms are not written with accents even if any of the original words are.

  37. 'Hace 2 días es valorado por MAP e inicia tto con Amoxicilina 500/125 + AInes i.m. [...]'
      'Hace 2 días es valorado por MAP e inicia un tto. con amoxicilina 500/125 más AINE IM [...]'
      "(The patient) was examined by their PCP two days ago and started a treatment with amoxicillin 500/125 and NSAIDs IM [...]"

- *Measurement units:* They are generally written in lowercase (with some exceptions such as *mmHg*) and without a final dot (unless they happen to be at the end of a sentence). They are always separated from whatever they measure by a space.

  Many units have established abbreviated forms that should be used: for hours the correct form is *h*, not **H* nor **hs*; for minutes, it's *min*, not **m*; for seconds, it's *s*, not **seg* nor **sg*. For grams, the only accepted form is *g*, not **gr* nor **grs*; for kilograms, it's *kg*, not **Kg*. For expressing temperature in the Celsius scale, the correct abbreviation is *ºC* (with the small circle next to the C).

38. *‘TA 100/70mmhg.’
‘TA: 100/70 mmHg.’
‘Blood pressure: 100/70 mmHg.’

A.2 Syntax

This section presents some of the syntactic phenomena that were taken into consideration for our corrections. There are some basic syntactic mistakes in the corpus, such as subject-verb agreement or incorrect preposition use. Those type of mistakes should be fixed but will not be discussed here.

Ellipsis

One of the main syntactic errors is the omission or ellipsis of certain types of words in a generalized way. It happens more commonly with function words, such as determiners and prepositions, but also with content words.

Function words are simple to correct, as the missing word is often obvious given some context and there is usually only one possibility. If there is more than one candidate, we choose any of them arbitrarily. Example 35 shows all of these at the same time.

35. *‘En consulta de control Julio 2017 en Sº Oncologia Medica Residencia Galatea se le practicó un TAC toraco abdominal’
‘En una consulta de control en julio de 2017 en el Sº de Oncología Médica de la Residencia Galatea se le practicó un TAC toracoabdominal’
‘During a follow-up visit on July 2017 at Galatea Residence, (the patient) had a thoracoabdominal CT.’

Missing content words are almost entirely verbs. In many cases, there might be more than one option possible. For this reason, the same subset of verbs must always be used. These are: ‘mostrar’ (*to show*), ‘referir’ (*to refer, to send*), ‘colocar’ (*to put*), ‘ser’ (*to be*), ‘estar’ (*to be*), ‘haber’ (*there is/are*), ‘hacer’ (*to do*) and ‘tener’ (*to have).*

In order to make the corrections more homogeneous, always try to use the most specific word and to respect collocations. The main verb that is used is ‘mostrar’. Even though it is frequently interchangeable with ‘referir’, this verb should only be used when the alternative is not very fluent (see Example 36: ‘viajes largos’ (*long trips*) are not something that can be shown).

36. *‘No viajes largos.’
?*‘No muestra viajes largos.’*
‘No refiere viajes largos.’
‘No long trips.’

A.3 Orthotypography

The correct and regular usage of punctuation is also an important matter. These are some of the main points regarding punctuation and symbol usage:

- Commas between subject and object are incorrect and should be erased.
- At times, sentences on different topics are put together without any punctuation. They should be separated with a dot, as in Example **??**. Long enumerations or multiple bullet points, such as Example 38 and 39, are also separated using dots as the original sentences are often too long.

  37. 'En el dia de hoy el paciente presenta insuficiencia respiratoria con gases Ph: 7,2, Po2: 57, PCO2: 49, Sato2: 82 %, asociada a insuficiencia cardiaca (Radiografia compatible + NT-pro BNP de 2590) y Creatinina de 1,49.'
  'En el día de hoy el paciente presenta insuficiencia respiratoria con gases. PH: 7,2, PO2: 57 , PCO2: 49, sat. de O2: 82 %, asociada a insuficiencia cardiaca (radiografía compatible más NT-proBNP de 2590) y creatinina de 1,49 '
  'Today, the patient shows respiratory insufficiency with gases. PH: 7.2, PO2: 57 , PCO2: 49, oxygen saturation: 82 %, associated to heart failure (the radiography is compatible and NT-proBNP of 2590) and creatinine of 1.49.'

  38. 'No fiebre no clínica constitucional , no otros síntomas añadidos.'
  'No muestra fiebre. No muestra ninguna clínica constitucional. No muestra otros síntomas añadidos.'
  'The patient does not show fever. They do not show any constitutional symptoms nor any other added symptoms.'

  39. 'Por parte del Sº de COT: - Evolución favorable - Profilaxis antibiotica y antitrombótica durante el ingreso - Control radiológico correcto - Alta hospitalaria'
  'Por parte del Sº de COT: - Evolución favorable. - Profilaxis antibiótica y antitrombótica durante el ingreso. - Control radiológico correcto. - Alta hospitalaria.'
  'According to ORTR services: - Positive evolution. - Aantithrombotic and antibiotic prophylaxis during their stay. - Correct radiological control. - Discharge.'

- Gerunds should always be preceded by a comma unless they are used after a conjunction. This is not recommended for all gerund types, but for simplicity's sake we recommend doing it.

  40. 'Levantamiento de colgajo timpano meatal liberando adherencias a pericondrio de anterior intervención.'
  'Levantamiento del colgajo del tímpano meatal, liberando adherencias al pericondrio de una anterior intervención.'
  'Lift of the meatal eardrum flap, freeing adhesions to the perichondrium from a previous surgery.'

- In long sentences, commas were included before the prepositions 'para' and 'por'.
- Prescriptions using numbers are often placed in the middle of the text without any punctuation. For this reason, and following the convention of some of the sentences of the corpus, they should always be inside a parenthesis.

  41. *'Tratamiento actual: Seretide 25/50 2-0-2, Ventolin si necesita, Lactulosa 1-0-0 [...]'
  'Tratamiento actual: Seretide 25/50 (2-0-2), Ventolin si necesita, lactulosa (1-0-0) [...]'
  'Current treatment: Seretide 25/50 (2-0-2), Ventolin if needed, lactulose (1-0-0) [...]'

- Dashes should be used as bullet points in enumeration, not asterisks.

  42. '* Consulta CCEE Traumatología (Dr. Soto del H. 14 de abril, el dia 23 de Octubre a las 12:01h'
  '- Consulta con CCEE de Traumatología (Dr. Soto) del H. 14 de abril, el día 23 de octubre a las 12:01 h.'
  '- Consultation with Traumatology's external consultations (Dr. Soto) at 14 de abril Hospital on the 23rd of October at 12:01 h.'

- Always remember to add a dot at the end of a sentence if there's none.
- Temperature is always written as its own separate word.

  43. *'Acude a urgencias por presentar fiebre de 38,8ºC'
  'Acude a Urgencias por presentar una fiebre de 38,8 ºC'
  '(The patient) comes to the emergency room due to a fever of 38.8 ºC.'

- Percentage symbols are also written separated from the number.

B Example predictions

This section showcases some example outputs from the different models presented during this work.

1. **Source sentence**: "EEII: no edemas ni singos de tvp."
   Gold Standard: "EEII: no muestra edemas ni signos de TVP."
   Translation: "Lower limbs: (the patient) does not show edema nor TVP signs."
   Baseline: "EEII: no edemas ni sirgos de tvp."
   IMEC: "EEII: no muestra edemas ni sensación de tvp."
   Oversampled IMEC: "EEII: no muestra edemas ni sensación de tvp."
   IMEC + Augmented corpus: "EEII: no muestra edemas ni signos de TVP."
   Oversampled IMEC + Augmented corpus: "EEII: no muestra edemas ni singos de TVP."

2. **Source sentence**: "Esta noche empeoramiento importante de su estado gral, con escalofríos , tiritonas y episodios de desorientación, a pesar de que ayer comenzó tto. con Amoxicilina 875/8h (3 dosis)."
   Gold Standard: "Esta noche muestra un empeoramiento importante de su estado gral., con escalofríos, tiritonas y episodios de desorientación, a pesar de que ayer comenzó un tto. con amoxicilina 875/8 h (3 dosis)."
   Translation: "Tonight (the patient) shows serious deterioration of their general state, with chills, shivering and disorientation, even though yesterday they started a treatment with amoxicillin 875/8 h (3 doses)"
   Baseline: "noop"
   IMEC: "Esta noche muestra empeoramiento importante de su estado gral, con escalofríos, ronas y episodios de desorientación, a pesar de que ayer comenzó un tto. con amoxicilina de 875 / 8h (3 dosis)."
   Oversampled IMEC: "Esta noche muestra un empeoramiento importante de su estado gral, con escalofríos, conas y episodios de desorientación, a pesar de que ayer comenzó un tto. con Amoxicilina 875 / 8 h (3 dosis)."
   IMEC + Augmented corpus: "Esta noche presentó un empeoramiento importante de su estado gral, con escalofríos, tiritonas y episodios de desorientación, a pesar de que ayer comenzó el tto. con amoxicilina 875 / 8h (3 dosis)."
   Oversampled IMEC + Augmented corpus: "Esta noche muestra un empeoramiento importante de su estado general, con escalofríos, tiritonas y episodios de desorientación, a pesar de que ayer comenzó un tto. con amoxicilina 875 / 8 h (3 dosis)."

3. **Source sentence**: "Ejn el dia de la fecha se encuentra asintomática y estable por lo que indico el alta a domicilio con control por consultas externas de cardiologia."
   Gold Standard: "En el día de la fecha se encuentra asintomática y estable, por lo que indico el alta al domicilio con control por Consultas Externas de Cardiología."
   Translation: "The mentioned day (the patient) is asyntomatic and stable, because

of which I suggest their discharge and follow-up visits by Cardiology's external consultations"
Baseline: "En el día de la fecha se encuentra asintomática y estable por lo que indico el alta a domicilio con control por consultas externas de cardiología."
IMEC: "Ejn el día de la fecha se encuentra asintomática y estable, por lo que indico el alta a domicilio con control por Consultas Externas de Cardiología."
Oversampled IMEC: "Ejn el día de la fecha se encuentra asintomática y estable, por lo que indico el alta a domicilio con control por Consultas Externas de Cardiología."
IMEC + Augmented corpus: "En el día de la fecha se encuentra asintomática y estable, por lo que indico el alta a domicilio con control por Consultas Externas de Cardiología."
Oversampled IMEC + Augmented corpus: "En el día de la fecha se encuentra asintomática y estable, por lo que indico el alta al domicilio con control por Consultas Externas de Cardiología."

4. **Source sentence**: "INTERVENCIÓN PRINCIPAL: 57.99, OTRA OPERACION SOBRE VEJIGA NCOC Fecha de Procedimiento: 17/06/2016 Técnica: Formolizacion vesical."
Gold Standard: "Intervención principal: 57.99, otra operación sobre la vejiga NCOC. Fecha de procedimiento: 17/06/2016. Técnica: formolización vesical."
Translation: "Main intervention: 57.99, another surgery on NCOC bladder. Date of the procedure: 17/06/2016. Technique: bladder formaldehyde."
Baseline: "INTERVENCIÓN PRINCIPAL: 57.99, OTRA OPERACIÓN SOBRE VEJIGA NCOC Fecha de Procedimiento: 17/06/2016 Técnica: Formolizacion vesical."
IMEC: "Intervención principal: 57.99 , gastrostomía. Fecha de procedimiento: 17/06/2016. Técnica: Fáología vesical."
Oversampled IMEC: "Intervención principal: 57.99, otra operación sobre el aparato musculoesquelético NCOC. Fecha de procedimiento: 17/06/2016. Técnica: enfermedad vesical."
IMEC + Augmented corpus: "Intervención principal: 57.99, otra operación sobre la vejiga NCOC Fecha de procedimiento: 17/06/2016. Técnica: formolizacion vesical."
Oversampled IMEC + Augmented corpus: "Intervención principal: 57.99 , otra operación sobre la vejiga NCOC. Fecha de procedimiento: 17/06/2016. Técnica: formulación vesical."

5. **Source sentence**: "Paciente que ingresa en Hospital Virgen del Palomar por síndrome coronario agudo SCASET ingresa en UCI y posteriormente en el servicio de Medicina Interna."
Gold Standard: "Paciente que ingresa en el Hospital Virgen del Palomar por un síndrome coronario agudo SCASEST. Ingresa en la UCI y posteriormente en el ser-

vicio de Medicina Interna."
Translation: "Patient admitted into Virgen del Palomar Hospital due to an acute coronary syndrome NSTEACS. They are admitted into the ICU and later into the Internal Medicine service."
Baseline: "Paciente que ingresa en Hospital Virgen del Palomar por síndrome coronario agudo SCASEST ingresa en UCI y posteriormente en el servicio de Medicina Interna."
IMEC: "Paciente que ingresa en el Hospital Virgen del Palomar por un síndrome coronario agudo SCASET ingresa en la UCI y posteriormente en el servicio de Medicina Interna."
Oversampled IMEC: "Paciente que ingresa en el Hospital Virgen del Palomar por un síndrome coronario agudo SCASET, ingresa en la UCI y posteriormente en el servicio de Medicina Interna."
IMEC + Augmented corpus: "Paciente que ingresa en el Hospital Virgen del Palomar por un síndrome coronario agudo (SCASET) ingresa en la UCI y posteriormente en el servicio de Medicina Interna."
Oversampled IMEC + Augmented corpus: "Paciente que ingresa en el Hospital Virgen del Palomar por un síndrome coronario agudo SCASET ingresa en la UCI y posteriormente en el servicio de Medicina Interna."

6. **Source sentence**: "Acudirá a la consulta Dra Montero, neumología, como tenía previsto, el dia 13 de mayo a las 07.44 h. en la Clínica 1 de mayo, previa espirometria y gasimetria arterial basa media hora antes."
Gold Standard: "Acudirá a la consulta de la Dra. Montero , Neumología , como tenía previsto, el día 13 de mayo a las 07.44 h en la Clínica 1 de mayo, previa espirometría y gasimetría arterial basal media hora antes."
Translation: "They will visit Dr. Montero's consultation, Neumology, as planned, on May 13th at 07.44 h at 1 de mayo Clinic, after doing a spirometry and baseline arterial blood gas test half an hour earlier."
Baseline: "Acudirá a la consulta Dra Montero, neumología, como tenía previsto, el día 13 de mayo a las 07.44 h. en la Clínica 1 de mayo, previa espirometria y gasimetría arterial basa media hora antes."
IMEC: "Acudirá a la consulta de la Dra. Montero, neumología, como tenía previsto, el día 13 de mayo a las 07.44 h. en la Clínica 1 de mayo, previa espirometría y gasimetría arterial basa media hora antes."
Oversampled IMEC: "Acudirá a la consulta de M. Montero, neumología, como tenía previsto, el día 13 de mayo a las 07.44 h en la Clínica 1 de mayo, previa espirometría y gasimetría arterial basa media hora antes."
IMEC + Augmented corpus: "Acudirá a la consulta Dra. Montero, Neumología, como tenía previsto, el día 13 de mayo a las 07.44 h en la Clínica 1 de mayo , previa espirometría y gasimetria arterial basa media hora antes."
Oversampled IMEC + Augmented corpus: "Acudirá a la consulta del Dra. Montero, Neumología, como tenía previsto, el día 13 de mayo a las 07.44 h en la

Clínica 1 de mayo, previa espirometría y gasimetría arterial basa media hora antes."

7. **Source sentence**: "Evalúo a la paciente en urgencias, y tras hablar con la cirujano y la familia iniciamos tratamiento con Ertepemen 1g endovenoso cada 24h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, primperan y buscapina."
   Gold Standard: "Evalúo a la paciente en Urgencias, y tras hablar con la cirujana y la familia iniciamos un tratamiento con ertapenem 1 g endovenoso cada 24 h con analgesia y sedación a bajas dosis con cl. mórfico, midazolam, Primperán y buscapina."
   Translation: "I evaluate the patient in the emergency room, and after talking to the surgeon and the family, we start a treatment with endovenous ertapenem 1 g every 24 hours with anagelsia and low-dosage sedatives with morphic chloride, midazolam, Primperán and buscapina."
   Baseline: "Evalúo a la paciente en urgencias, y tras hablar con la cirujano y la familia iniciamos tratamiento con Ertepemen 1g endovenoso cada 24h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, Primperan y buscapina."
   IMEC: "Evalúo a la paciente en Urgencias, y tras hablar con la cirujano y la familia iniciamos un tratamiento con Ertepemen 1g endovenoso cada 24 h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, primperan y buscapina."
   Oversampled IMEC: "Solalúo a la paciente en Urgencias, y tras hablar con la cirugía y la familia iniciamos un tratamiento con Ertepemen 1 g endovenoso cada 24 h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, primperan y buscapina."
   IMEC + Augmented corpus: "Evalúó a la paciente en urgencias, y tras hablar con el cirujano y la familia iniciamos tratamiento con Ertepemen 1g endovenoso cada 24 h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, primperan y buscapina."
   Oversampled IMEC + Augmented corpus: "Evalúo a la paciente en Urgencias, y tras hablar con el cirujano y la familia iniciamos un tratamiento con Ertepemem 1 g endovenoso cada 24 h con analgesia y sedación a bajas dosis con cl.morfico, midazolam, primperan y buscapina."

8. **Source sentence**: "Atendida por Unidad Medicalizada le han administrado Actocortina 100 + Polaramine + Nebulización de ventolin + SF 500 cc."
   Gold Standard: "Atendida por la Unidad Medicalizada le han administrado Actocortina 100 más Polaramine más nebulización de Ventolin más SF 500 cc."
   Translation: "(The patient was) helped by the Medicalized Unit who administered Actocortina 100 and Polaramine and Ventolin nebulización and SF 500 cc."
   Baseline: "Atendida por Unidad Medicalizada le han administrado Actocortina 100 + Polaramine + Nebulización de Ventolin + SF 500 cc."
   IMEC: "Atendida por la Unidad Medicalizada le han administrado Actocortina 100 más Polaramine + Nebulización de ventolin más SF 500 cc."
   Oversampled IMEC: "Atendida por la Unidad Medicalizada le han administrado

Actocortina 100 + Polaramine + Nebulización de ventolin más SF 500 cc."
IMEC + Augmented corpus: "Atendida por Unidad Medicalizada le han administrado Actocortina 100 + Polaramine + Nebulización de Ventolin + SF 500 cc."
Oversampled IMEC + Augmented corpus: "Atendida por la Unidad Medicalizada le han administrado Actocortina 100 más polaramine más nebulización de Ventolin más SF 500 cc."

9. **Source sentence**: "Presenta cultivo positivo a E.coli blea + en paciente portadora de sonda urinaria."
Gold Standard: "Presenta cultivo positivo a E. coli BLEA + en paciente portadora de sonda urinaria."
Translation: "Positive culture for E. coli BLEA + in a patient that has a urinary catheter."
Baseline: "Presenta cultivo positivo a E.coli lea + en paciente portadora de sonda urinaria."
IMEC: "Presenta un cultivo positivo a Ecoli blea más en paciente portadora de una sonda urinaria."
Oversampled IMEC: "Presenta un cultivo positivo a E. coli BLEA + en un paciente portadora de una sonda urinaria."
IMEC + Augmented corpus: "Presenta un cultivo positivo a E. coli blea + en una paciente portadora de sonda urinaria."
Oversampled IMEC + Augmented corpus: "Presenta un cultivo positivo a E. coli BLEA más en una paciente portadora de una sonda urinaria."

10. **Source sentence**: "-Furosemida 40 mg, 0.5 comp en el desayuno. (CAMBIO) Su médico valorará más cambios en caso de que precise - Bisoprolol 5 mg, 1comp en el desayuno."
Gold Standard: "- Furosemida 40 mg, 0.5 comp. en el desayuno. (Cambio) Su médico valorará más cambios en caso de que precise. - Bisoprolol 5 mg, 1 comp. en el desayuno."
Translation: "- Furosemide 40 mg, half a pill with breakfast. (Change) Their doctor will make more changes if needed. - Bisoprolol 5 mg, 1 pill with breakfast."
Baseline: "noop"
IMEC: "- Furosemida 40 mg, 0.5 comp. en el desayuno. (CAMBIO) Su médico valorará más cambios en caso de que precise. - Bisoprolol 5 mg, 1 comp. en el desayuno."
Oversampled IMEC: "- Furosemida 40 mg, 0.5 comp. en el desayuno. (CAMBIO). Su médico valorará más cambios en caso de que precise - Bisoprolol 5 mg, 1 comp. en el desayuno."
IMEC + Augmented corpus: "- Furosemida 40 mg, 0.5 comp. en el desayuno. (cambio) Su médico valorará más cambios en caso de que precise - Bisoprolol 5 mg, 1 comp. en el desayuno."
Oversampled IMEC + Augmented corpus: "- Furosemida 40 mg , 0.5 comp.

en el desayuno. (CAMBIO) Su médico valorará más cambios en caso de que precise. - Bisoprolol 5 mg, 1 comp. en el desayuno."

11. **Source sentence**: "Se efectua un ecocardiograma doppler que muestra una aquinesia anterior y septal con severo deteriorto de la feyvi y una PSAp = 50 mmhg."
Gold Standard: "Se efectúa un ecocardiograma Doppler que muestra una aquinesia anterior y septal con un severo deterioro de la FEyVI y una PSAP de 50 mmHg."
Translation: "A Doppler echocardiogram is carried out, showing anterior and septal akinesis with a serious deterioration of the LVEF and a PASP of 50 mmHg."
Baseline: "Se efectuar un ecocardiograma doppler que muestra una aquinesia anterior y septal con severo deterioro de la feyvi y una PSA = 50 mmHg."
IMEC: "Se efectúa un ecocardiograma doppler que muestra una aquinesia anterior y septal con severo deteriorto de la fecha y una PAIP de 50 mm."
Oversampled IMEC: "Se efectúa un ecocardiograma Doppler que muestra una aquinesia anterior y septal con severo deteriorto de la fémr y una PAP de 50 mm."
IMEC + Augmented corpus: "Se efectúa un ecocardiograma doppler que muestra una aquinesia anterior y septal con severo deterioro de la FEYVI y una PSAp = 50 mmHg "
Oversampled IMEC + Augmented corpus: "Se efectúa un ecocardiograma doppler que muestra una aquinesia anterior y septal con severo deterioro de la feyvi y una PSAp = 50 mmHg."

12. **Source sentence**: "OTORRINOLARINGOLOGIA - Interconsulta en Hospitalización Refiere hipoacusia OI ."
Gold Standard: "Otorrinolaringología. - Interconsulta en Hospitalización. Refiere hipoacusia en OI."
Translation: " Otorhinolaryngology. - Consultation in Hospitalization. (The patient) refers hearing loss in the left ear."
Baseline: "OTORRINOLARINGOLOGÍA - Interconsulta en Hospitalización Refiere hipoacusia OI."
IMEC: "Otorratitis - Interconsulta en Hospitalización."
Oversampled IMEC: "Otorrinolaringología. Interconsulta en Hospitalización. Refiere hipoacusia OI."
IMEC + Augmented corpus: "Otorrinolaringología. Interconsulta en Hospitalización Refiere hipoacusia OI."
Oversampled IMEC + Augmented corpus: "Otorrinolaringología. Interconsulta en Hospitalización. Refiere hipoacusia OI."

13. **Source sentence**: "Colección biliar??"
Gold Standard: "¿¿Colección biliar??"
Translation: "Biliary collection??"
Baseline: "noop"
IMEC: "Colección biliar."
Oversampled IMEC: "Colección biliar??."

IMEC + Augmented corpus: “Colección biliar.”
Oversampled IMEC + Augmented corpus: “¿Colección biliar??”

14. **Source sentence**: “Actualmente sobre los 92 Kgrs.”
Gold Standard: “Actualmente sobre los 92 kg.”
Translation: “Currently around 92 kg.”
Baseline: “Actualmente sobre los 92 Kgrs.”
IMEC: “Actualmente sobre los 92 kg.”
Oversampled IMEC: “Actualmente sobre los 92 kg .”
IMEC + Augmented corpus: “Actualmente, sobre los 92 - Kgrs.”
Oversampled IMEC + Augmented corpus: “Actualmente sobre los 92 Kgrs .”

15. **Source sentence**: “- Si la HEMOGLOBINA BAJA por debajo de 8’5g/dl ADEMÁS DE SUSPENDER EL TRATAMIENTO DEBE DE TRANSFUNDIRSE SANGRE EN EL HOSPITAL.”
Gold Standard: “- Si la hemoglobina baja por debajo de 8’5 g/dl , además de suspender el tratamiento debe de transfundirse sangre en el hospital.”
Translation: “- In case (the patient’s) hemoglobin goes under 8’5 g/dl, not only must the treatment be stopped, but blood must also be transfused at the hospital.”
Baseline: “- Si la HEMOGLOBINA BAJA por debajo de 8’5g/dl ADEMÁS DE SUSPENDER EL TRATAMIENTO DEBE DE TRANSFUNDIESE SANGRE EN EL HOSPITAL.”
IMEC: “- Si por debajo de 8’5g/dl y siempre de 8’5 g/dl.”
Oversampled IMEC: “- Si la clínica por debajo de 8’5g/dl.”
IMEC + Augmented corpus: “- Si la hemoglobina baja por debajo de 8’5g/dl además de suspender el tratamiento debe de transfundirse el tratamiento debe de transfundirse sangre EN EL HOSPITAL.”
Oversampled IMEC + Augmented corpus: “- Si la hemoglobina baja por debajo de 8’5g/dl ADEMÁS DE SUSPENDER EL TRATAMIENTO DEBE DE TRANSFUNDIRSE SANGRE EN EL Hospital.”