Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 99-109 recibido 09-12-2021 revisado 18-01-2022 aceptado 25-01-2022

A Corpus of Spanish clinical records annotated for abbreviation identification

Un corpus de historias clínicas españolas anotadas para la identificación de abreviaturas

Mercedes Aguado,[1] Núria Bel[2]
[1] Università degli Studi di Milano
[2] Universitat Pompeu Fabra
mercedes.aguado@unimi.it, nuria.bel@upf.edu

Abstract: With the deployment of Electronic Health Records, much effort is being devoted to the development of Natural Language Processing tools that convert information described in these clinical records into structured data to be exploited. Clinical records main characteristic is that they are free text. They are normally written under pressure as memory notes and contain a high number of abbreviations that are an issue for automatic processing. In this article we present the IULA Spanish Clinical Records Corpus annotated for abbreviation identification.
Keywords: Abbreviations, annotated corpus, clinical records, preprocessing.

Resumen: Con la implementación de las historias clínicas electrónicas, se están dedicando muchos esfuerzos al desarrollo de herramientas de procesamiento del lenguaje natural que convierten la información descrita en estos registros clínicos en datos estructurados para ser explotados. La principal característica de las historias clínicas es que son texto libre. Normalmente se escriben deprisa, como notas de memoria y contienen un gran número de abreviaturas que son un problema para su procesamiento automático. En este artículo presentamos el Corpus de historias clínicas españolas del IULA, anotado para la identificación de abreviaturas.
Palabras clave: Abreviaturas, corpus anotado, historias clínicas, normalización, preprocesamiento.

1 Introduction

With the deployment of Electronic Health Records (EHR), much effort is being devoted to the development of Natural Language Processing (NLP) tools that convert information described in these clinical records into structured data that can be exploited. However, clinical records main characteristic is that they are free text, normally written under pressure as memory notes and containing a high number of abbreviations; they result in a telegraphic style that, on the one hand, is quicker for practitioners and experts to write, but on the other hand can be problematic for both human reading and automatic processing by NLP tools. Different methods have been applied to the normalization and analysis of clinical records to make them ready for the most used information extraction tasks like Named Entity Recognition and Classification (NERC), and Relation Identification and Extraction (see for instance Pathak et al., 2013, Gorinsky et al. 2019, Wang et al., 2018). The availability of annotated texts makes the use of machine learning supervised methods possible and allows for a fair comparison among these different methods. Thus, annotated corpora should be made available to support the development and improvement of methods and tools. However, most of the annotated corpora available are in English, as we will see in section 2. Related work, while EHR to be processed are written in many other languages around the world.

In this paper, we describe the IULA Spanish Clinical Record Corpus annotated for abbreviation identification (IULA-SCRC-ABB), a corpus of 3,194 sentences extracted from anonymized clinical records in Spanish annotated with abbreviations and the corresponding annotation guidelines. In the IULA-SCRC-ABB corpus, tokens that are shortened forms of words or phrases (including abbreviations, acronyms and symbols) were identified and tagged according to a linguistically motivated classification. The annotation comprised the identification of the short form, its classification into three classes, and the listing of possible long forms for each. The correct assignment of the long form is a task that requires expert knowledge on medical specialties and their practices and it has been left for future work. We also describe the annotation guidelines and discuss the most problematic cases. To the best of our knowledge, this is the first corpus of Spanish clinical records annotated for abbreviations that is made public and freely accessible at http://eines.iula.upf.edu/brat//#/AcronAbrevOnCR/ .

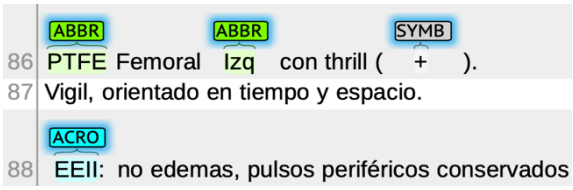

Figure 1: Sample of the IULA-SCRC-ABB corpus displayed with BRAT.

2 *Related work*

Existing medical text corpora annotated with abbreviations have been compiled as datasets of abbreviation identification tools. These corpora mostly contain scientific abstracts and articles in English (see, Islamaj Doğan et al., 2014 for details), although there are some in other languages, but only a few included clinical records. Névéol et al. (2018) and Dallianis (2018) are overviews of research carried out in clinical text mining in languages other than English, and Soto Montalvo et al. (2018), Sánchez and Martínez (2018), Sánchez León (2018), Castaño et al. (2018) and Cuadros et al. (2018) are descriptions of specific abbreviation identification tools for Spanish shown at IberEval2018-BARR, Biomedical Abbreviation Recognition and Resolution, evaluation campaign, whose corpus is described below.

We now describe other corpora consisting of clinical records annotated for abbreviations, and corpus annotated for abbreviations for Spanish, so that the IULA-SCRC-ABB and the methods we used to annotate can be compared with them. (Hua et al., 2007) used 16,949 admission notes from the internal medicine service of The New York Presbyterian Hospital Clinical Data Repository (NYPH-CDR) as the dataset for their machine learning system to be trained to detect abbreviations in clinical notes. For building the dataset, a physician manually reviewed the selected notes, listed all the abbreviations and specified their full forms. The training set consisted of 3,007 tokens of which 418 were abbreviations and the test set contained 2,611 tokens of which 411 were abbreviations. (Hua et al., 2007) also analyzed the abbreviations and further classified them according to the way they are formed. They used three classes: acronyms, shortened words and contractions. Acronyms were short forms, usually associated with multi-word phrases, which were formed by taking the first letter of each word in a phrase. Shortened forms were those which usually are a substring of a long word, although not always. Finally, contractions, considered another type of abbreviation, were those that consisted of an abbreviated contraction of multiple words with a separator (usually "/") between each word, for instance: 't/d/a', whose long form is "tobacco, drugs or alcohol".

Also for English, Wu et al. (2011) built a corpus for testing different machine learning methods for abbreviation detection. Three physicians manually annotated abbreviations in clinical documents randomly taken from the Vanderbilt Medical Center's Synthetic Derivative database, which contains de-identified electronic health records at Vanderbilt University Hospital. A total of 70 documents were annotated first with a pre-processing program that automatically labelled abbreviations using a reliable abbreviation dictionary. The human annotators revised these versions for identifying new abbreviations or removing wrongly labelled words. The developed corpus consisted in a training set with 40 documents of 18,225 tokens which contained 1386 abbreviations, and a test set with 30 documents of 13,913 tokens, containing 12,511 abbreviations.

Kvist and Velupillai (2014) reported about the annotation of two subsets of the Stockholm

Electronic Patient Record Corpus for Swedish (described in Isenius, 2012) with abbreviations and acronyms. The corpus consisted of randomly extracted emergency notes and radiology reports as to amount three sets of about 10,000 words each. Each subset was manually annotated for abbreviations by an expert. In this work, different types of abbreviations were identified: shortened words, *pat* for *patient*; contractions, *ssk* for *sjuksköterska* (nurse); and acronyms, *ECG* for *electrocardiogram*, although these classes were not used for annotation. Kreuzthaler et al. (2016) reported about the creation of a corpus of 1,696 de-identified clinical and outpatient discharge letters in German from the dermatology department of an Austrian university hospital. For training and testing different detection algorithms, instead of annotating the corpus, a list of words ending in period was extracted and manually annotated as whether the period was part of the word or not.

As for Spanish, Rubio-López et al. (2017) reported about having built a corpus to get the data for training and testing a system for abbreviation and acronym identification and disambiguation. The corpus consisted of 150 clinical notes in Spanish about stroke patients. These notes were selected by the high number of potential acronyms found. The notes were cleaned and manually annotated by researchers with a single label. To the best of our knowledge, this corpus is not accessible. Also for Spanish, (Intxaurrondo et al., 2018) described the Spanish corpus created for the BARR shared task held in the framework of IberEval 2017 and 2018 evaluation campaigns. The BARR track objective is to promote the development of biomedical and medical text mining tools together with offering an informed overview of the state of the art techniques and results obtained by the community. In the 2017 edition, the BARR track evaluated systems for detecting mentions of abbreviations-definition pairs: the discovery of abbreviations that were explicitly defined through the corresponding long form in the same sentence. The BARR corpus consisted of 1,050 abstracts for training and 600 abstracts for testing of biomedical articles from different sources. The corpus was manually annotated by biomedical experts. The corpus was annotated with information about abbreviations: short forms and long forms, as well with other relation-related information: derived short forms, global abbreviations, unclear, contextual, etc.; however, no classification of the different types of abbreviations was used. In 2018 BARR2 edition, a new corpus of clinical case studies has been delivered (Intxaurrondo et al., 2019). The corpus is divided into train, development and test sections with 122,594, 56,564 and 90,098 tokens respectively. It has been manually annotated by experts for the task of identifying abbreviations and delivering the corresponding long form. The documentation reports 9,552 annotated abbreviations in the whole corpus. This is the most similar corpus to the one presented here, although there are notable differences, as we will describe in the next section.

The corpus we present here, which is freely available, could be a contribution to BARR for future editions so that the shared task also includes authentic clinical records that present specific characteristics that made them different from scientific literature and clinical case studies.

3 Abbreviations in Spanish clinical records

Clinical records differ from Spanish for general purposes or even from other clinical texts written in Spanish in many linguistic features, such as lexical complexity, word and sentence composition, and sentence structure (see Benavent and Iscla, 2001, for a detailed description of linguistic characteristics of Spanish clinical texts which are very similar to the same genre in other languages as reported in Dallianis, 2018).

Clinical records in Spanish, as well as in other languages, show a higher density of technical terms and in particular of abbreviations (including acronyms, symbols, digits, capitalized letters within words, Roman digits and measurement units).

In our corpus, abbreviations amount 10.2% of the text tokens, while in the most similar corpus to ours, the BARR2 corpus made of clinical notes, abbreviations are about a 3.5% of the tokens. Isenius (2012) and Dallianis (2018) report figures similar to ours in discharge and emergency notes in other languages such as English and Swedish with a 15% of domain abbreviations. Note that the BARR2 corpus has collected clinical notes that are samples of edited text, while IULA-SCR-ABB contains spontaneous writing. The differences are also in

the distribution of the different types. BARR2 corpus is not annotated with types of abbreviations, but according to our annotation guidelines, the abbreviations of the test set would contain a 5.6% of abbreviations, a 67% of acronyms and 38.1% of symbols. In the IULA-SRC-ABB corpus, the distribution of categories is 14.6% abbreviations, 42.1% acronyms, 41.4% symbols and 1.5% unknowns (see section 4.4. Corpus Statistics for more details of the corpus). Moreover, the detection of abbreviations in clinical records is more difficult because of the following issues:

Differently to other general and medical texts, in spontaneous clinical records abbreviations do not occur along with the long form.

Sometimes, practitioners do not use the standard forms of acronyms and abbreviations. For instance, only 11,8 % of abbreviations were marked with a final period. Besides, others are made-up, such as the abbreviation sgto in our texts, which is wrongly used as the short form for segmento (segment), but the standard meaning is sargento (“sergeant”).

It is common an incorrect usage of symbols meant to be international standards, and the texts frequently show wrongly used symbols, such as grs, instead of g (“grams”), seg or min (for segundo “second” and minuto “minute”), which should be s and m, respectively.

Clinical records also exhibit a misuse of capital letters in abbreviations. For example, decilitro ("deciliter") was found written dl and dL; ecografía ("ecography") written ECO and eco; ABD and abd for abdomen.

Moreover, when working with authentic clinical records some other practical issues arise. For instance, we found uppercased full sentences, with uppercased wordforms which became homographs to abbreviations. For instance, we found *SE ADJUNTA* ("annexed"), where *SE* is the reflexive pronoun, but the form also corresponds to the shortened form of *sin especificar* ("unspecified").

4 The corpus

The IULA Spanish Clinical Record Corpus (IULA-SCRC) is a corpus of 3,194 sentences extracted from clinical records and annotated with negation markers and their scope (Marimon et al., 2017). The corpus was conceived as a resource to support clinical text-mining systems, but it is also a useful resource for other NLP systems handling clinical texts: automatic encoding of clinical records, diagnosis support, term extraction, among others, as well as for the study of clinical texts. The corpus was made publicly available with a CC-BY-SA 3.0 license.

This resource was obtained from a set of 300 anonymized clinical reports from several services of one of the main hospitals in Barcelona (Spain). In Table 1 we show the final number of sentences got from different sections of clinical records. Although the corpus was given to us already anonymized (all patient information was removed), the sentences were shuffled to make sure that no traceability of any data is possible.

Section	Sentences	%	Selected
Physical Exploration	5,193	34.61	1,090
Evolution	5,463	36.41	1,147
Radiology	1,751	11.67	367
Current Process	980	6.53	205
Explorations	1,619	10.79	339

Table 1: Statistics of corpus composition.

The texts from the IULA-SCRC corpus were taken as source for the abbreviation annotation, as we explain below. The IULA-SCRC-ABBR corpus is distributed as BRAT files (i.e. raw text and annotations in separate files) in UTF8 encoding and with a CC-BY-SA 3.0 license.

4.1 Pre-processing and pre-annotation system

Following standard practices, the IULA-SCRC-ABB texts were pre-processed to correct misspellings (Lai et al. 2015). Spelling errors are known to be very frequent in medical texts (Dallianis, 2018) and this becomes an issue for automatically processing the texts because misspelled words are not recognized. The texts of our corpus contained about a 2.6% of misspelled words. A 1.1% corresponded to segmentation problems, most typically the last character of a word becomes the first character of the following word (see Table 2 for examples). Missing accents corresponded to 0.86% of the misspellings being the second most frequent error. Other less frequent errors are character inversion, missing spaces, missing letters, unnecessary accents, unnecessary capital letters or wrong characters.

Error type	Error found	Correction
Segmentation	a lingreso	al ingreso (*upon admission*)
Missing accent	simetrica	simétrica (*simetric*)
Character inversion	peirfeira	periferia (*periphery*)
Missing spaces	segundos,sin	segundos, sin (*seconds, without*)

Table 2: Types and examples of misspellings found at the corpus.

Because these errors repeated several times along the different texts, a simple set of regular expressions was used to correct them automatically. Other misspellings were manually corrected, although as we explain in section 4.5. Difficulties and issues, the errors affecting abbreviations were not corrected.

Before annotating abbreviations, sentences were tokenized automatically using Freeling 4.1. (FL, Padró and Stanilovsky, 2012). For speeding up the abbreviation annotation task, we developed a script for identifying abbreviations by accessing a dictionary (like Hua et al., 2007). The dictionary was filled with abbreviations from the dictionary of Spanish medical abbreviations (Yetano Laguna and Alberola Cuñat, 2002) published as a reference for practitioners by the Spanish Health Ministry and other abbreviation lists freely available at the www. The script takes tokens, as found by FL, and makes a look-up at the dictionary database. In case of coincidence, the script retrieves the information about the class and long forms in the database and writes it in a BRAT annotation file. Thus, human annotators are provided with pre-annotated sentences. Annotators had to validate annotations, deleting errors and identifying and annotating new abbreviations missing in the database.

In order to tune FL to clinical text characteristics, we set up the following parameters. For the es.congif file:

- AlwaysFlush = yes, in order FL to process each line as an independent sentence.
- CompoundAnalysis = no. FL can guess compounds by splitting tokens into potential parts of a compound. This option was cancelled.
- QuantitiesDetection = no. FL can also recognize and normalize references to quantities and its measure. For instance, '234 €', is normalized into: 'CUR_EUR:234'. However, the actual spelling of medical measures in our texts showed a significant variation, as well as the use of different abbreviations for the same measure, so we preferred not to normalize them.

The FL named entity recognition module recognizes multiword units relying on uppercased words. However, given that some titles often appear in capital letters, there is a special heuristic to discard long sequences of uppercased words. We set TitleLimit to 1 to prevent the identification of series of uppercased words (a common case in our corpus) as named entities. Finally, since periods are used for sentence segmentation, FL requires the list of period-ending abbreviations to be in the tokenizer.dat file. Therefore, we included here those abbreviations that were at our database.

4.2 Annotation guidelines

In this section, we first introduce the different classes of abbreviations we have used and the motivation and underlying annotation criteria. Secondly, we describe the guidelines given to our two human annotators to identify and annotate abbreviations.

As explained in the Related Work section, most abbreviation annotation practices have not considered subclasses, although Cuadros et al. (2018), Wu et al. (2011) and Kreuzthaler et al. (2016) analyze the differences between short forms demonstrating that they exhibit different formal patterns. Our decision to use three classes for tagging short forms was based on the following differences related to how they are written and how the map to their long form.

In general, shortened forms that we call *abreviaturas* in Spanish usually end in a period; they keep the accented vowel of the long form, if any; they are written mostly with low-case characters and they can be plural forms, such as *págs*. for *páginas* ('pages'). Differently to acronyms, which are read as words (when phonetically possible), *abreviaturas* are read as the corresponding long form[1]. However, the

[1] These reading differences might be the cause of Spanish texts containing many acronyms which are shortened forms of English phrases as, in general,

samples we found in clinical records can deviate from the rules just mentioned and indeed they show high variability with several short forms for the same long form, as we will discuss in section 4.5. Difficulties and issues.

Eventually, the classes used for the annotation were the following:

- Symbols (SYMB): Symbols are tokens consisting of letters (either capital letters or lower cased) and other signs and numbers that are short forms of, mostly, internationally recognized measurement units, chemicals and mathematical terms. The tokens can be alphabetical (e.g., *r*) and non-alphabetical signs (e.g., % -percentage- ; and ∅, which means 'diameter', 'negative' and 'normal') and they never end in period.
- Abbreviations (ABBR): Abbreviations are those terms resulting from the removal of letters from one or more words. An abbreviation can contain letters and numbers, capitalized, lower case or both (starting the word or in other position within the word). Abbreviations also include special characters such as º, ª, er, which contract ordinal numbers (as in 1º, meaning 'first') and other kind of words (as in Hª, meaning 'history'), they can end in a period or not, and they can be hyphenated or not.
- Acronyms (ACRO): Acronyms are those terms that are formed by joining parts of two or more words. Usually, acronyms are composed of the initial part of each word, they can contain capital and/or lower-case letters and they do not end with a period. Besides, sometimes they are composed by one letter of one of the words and more than two letters of the other word, such as "AloTMO", whose long form is "Trasplante alogénico de médula ósea".
- Finally, the *Unknown* label (UNK) was devised to cover misspellings, typographical mistakes and cases where the abbreviation, acronym or symbol could not be attested in any resource.

In section 4.1. Preprocessing and Pre-annotation, we have explained that in order to reduce annotation time and required human resources, we used a simple lexical-lookup tool to pre-annotate the texts. Annotators, who were not practitioners, revised and corrected these pre-annotated texts using the BRAT annotation tool (Stenetorp et al., 2012). Human annotation task was about reviewing and validating the token identified as an abbreviation by the look-up system. In most cases, the pre-annotated class had not to be changed, as the information at the database is correct. However, corrections were required when:

a. A particular shortened form was annotated as belonging to more than one class, for instance: *m* can be both the abbreviation of *mes* (month) and the symbol of *metro* (meter), or *K*, which is the abbreviation of both "Kelvin" and "Karnofsky" and the symbol of *kilo* as well. The annotator had to choose the correct class, according to context.
b. A capitalized word was wrongly annotated as an abbreviation. Annotation has to be deleted.
c. Identifying new abbreviations. For abbreviations in the text but missing in the dictionary, annotators should find the long form and decide the class. The annotators searched for the candidate in different resources (medical dictionaries and databases like SNOMED, MESH, IATE and parallel and comparable corpora) to identify the abbreviation and to collect all possible long forms, if possible. When the annotators could not find the long form of a particular abbreviation, they used the label UNK (unknown), which was reserved for this case.

4.3 Inter-annotator agreement

For validating the annotation guidelines, we followed a two steps procedure. In a first round, the manual annotation task of identifying the short form, its type and possible long forms, as explained before, was performed by two annotators (a specialized translator and a linguist) over the whole set of documents. For the task of identifying the short form and assigning a type, the kappa inter-annotator agreement measure was 0.75. All the mismatches were studied and solved, and the guidelines were refined accordingly. To test the changes in the guidelines, a new round of annotation over 800 sentences was carried out by two new, non-medical expert annotators. The kappa measure was on average 0.75. The major source of disagreement was the distinction between acronyms and abbreviations

acronyms are more likely to become loanwords than their corresponding long forms.

specially for abbreviations in capital letters that were considered acronyms although were listed as abbreviations in different resources.

4.4 Corpus Statistics

The IULA-SCRC-ABBR corpus details are described in Table 3, where number of tokens annotated and number of types for each abbreviations class are presented.

Unit	Number of	Unique forms
Sentences	3,194	
Tokens	38,208	
ABBREVIATIONS	506	163
ACRONYMS	1,460	376
SYMBOLS	1,427	79
UNKNOWN	52	34

Table 3: Details of abbreviation annotation in the corpus.

ABB.	#	ACRO.	#	SYM.	#
U	47	TC	57	mg	188
Tª	27	PAD	54	%	153
mEq	25	PAS	52	mm	139
Rx	19	TP	37	Hg	102
Hb	15	MVC	34	dl	99
ABD	10	EEII	30	L	98
E.	10	FA	29	ºC	53
Dr.	9	VHC	28	h	50
Abd	8	GGT	27	dL	48
Dr	8	NIHSS	25	g	47

Table 4. Ten most frequent abbreviations, acronyms and symbols, and frequency.

Finally, we report about the script to pre-annotate sentences. The script was created just to reduce manual work as the task was to compare tokens in texts with the items of the database created out of an abbreviation dictionary. Table 5 shows the final figures of the corpus after revising the pre-annotated files.

	Manual addition	Final number
ABB.	142	506
ACRO.	190	1460
SYM.	64	1427
TOTAL	396	3393

Table 5: Manual additions after using the script for pre-annotation.

As for the performance of the script, data from the validation exercise with 800 sentences and 475 short forms to be identified and annotated showed that the script initially identified 438 short forms, of which 84, a 19%, had to be manually corrected. As explained in detail in section 4.5. Difficulties and issues, segmentation problems and wrong punctuation that specifically affected abbreviations were not corrected in the source text, thus preventing the script from matching them. We can see an example in Figure 2. Finally, the coverage of the database and the script was 74.5% as 121 forms, mainly abbreviations, were manually added.

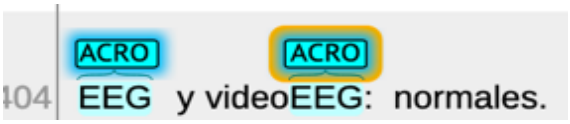

Figure 2. Example of a segmentation misspelling (missing space in videoEEG).

4.5 Difficulties and issues

The processing of health records in Spanish requires correctly identifying the units the text is composed of, including abbreviations. The task of abbreviation identification has been reported to be a challenging task. In Spanish, short forms are considered as belonging to one of three classes: Symbols, Abbreviations and Acronyms. This classification into three different classes should be of interest as it helps to understand the differences and therefore leads to better prediction features. The annotation allowed us to see that Abbreviations suffer of more variability than the other categories. For instance, a quite common term like *hematocrito* ('hematocrit') is written in 4 different ways: 'Hcto' (4)[2], 'Htc' (1), 'hto' (1), 'HTO' (2) and 'Hto' (6). Other samples are *creatinina* ('creatinine') whose abbreviation is 'creat' (2) although variations like 'Crea' (4), 'crea' (1) and 'Cre' (2) were found and *izquierdo* (left) that was found as 'izdo' (2) or 'izqdo' (3). In contrast, variation regarding acronyms is less frequent and the few cases of variations are spelling differences like in *angio-RM* (2) vs. *angioRM* (2) for 'Magnetic Resonance Angiogram', for instance. As for symbols, few cases of variation have been

[2] Frequency of occurrence in parentheses.

found like ‘mmHg’ (34) vs. ‘mmhg’ (2). The explanation of this higher variation for abbreviations could be that, as mentioned before, abbreviations are read as the long form, while acronyms are read as wordforms, what would support a better memorization.

Variations that divert from standard practices were annotated as Unknown, although a proposal for the correct short form and corresponding longform has been provided. As explained in section 4.5.4. Misspellings, the Unknown label was devised to cover misspellings, typographical mistakes and cases of forms looking like abbreviations that, however, were not attested in any of the identified resources (listed in 4.3). Eventually 52 tokens are coded as Unknown. These correspond to 34 unique forms (types) of which only 20 could not be annotated with a longform.

As we mentioned before, a first annotation round was useful to identify the different cases that were not covered by normative descriptions. Despite the apparent simplicity of the task, there were some cases that were difficult to classify and generated some discussions. Now, we list the most significant cases.

4.5.1 Acronyms or abbreviations in other languages, mostly English.

In Spanish practitioner’s reports, there is a surprising abundance of English abbreviations, mostly acronyms, even though there exist a corresponding Spanish term. This was the case of: *SOFA* for English ‘Sequential Organ Failure Assessment’, in Spanish *Evaluación secuencial de fallo orgánico*; DIVAS, for English ‘Digital Intravenous Angiography Subtraction’, in Spanish *Angiografía digital intravenosa de sustracción*; CKD for English ‘Chronic kidney disease’, in Spanish, *Enfermedad Renal Crónica*. For these English acronyms, the annotated long forms are the Spanish ones.

4.5.2 Single characters as type enumerations

Abbreviations that are just one character, usually uppercased, are considered to be a symbol when they have no ending period, e.g., cases such as *A*, in *Virus gripe A* (A influenza virus) because it is an international typing encoding system. We considered their long form to be *Tipo A*, *Tipo T*, etc. The same was done for *ondas T* (T waves).

We did the same for types of vitamins, hepatitis, clusters, etc. since the letters, by themselves do not have a meaning, but are the identifier of a type or a subtype. In some cases, the headword, that is, ‘vitamin’, is missing, for instance *hidroxil B12 B6*. In that case, we included the head in the long form, that is *Vitamina B12*, but we considered the abbreviation a symbol anyway.

Isolated letters in names of medicines and drugs, such as *Gentamicina S* –standing for *Sulfato*--, or *Levofloxacina R* --standing for *Richet*-- are classified as abbreviations, as they are types but not part of an enumeration. Thus, we followed BARR’s annotations for other examples as Proteins C and S, that were identified as abbreviations of ‘peak C’ and ‘Seattle”.

Short forms containing letters and numbers are classified as acronyms when their long form contains more than one word. Short forms containing letters and numbers are classified as ABBR when the letter (or letters) is itself an abbreviation and as symbols when the letters are neither abbreviations nor acronyms (i.e., the elements do not have a long form), but the term as a unit has a long form as in the case *M1 Segmento esfenoidal* (M1 Sphenoidal segment).

4.5.3 Parts of phrases

For terms formed by an abbreviation and a full word, e.g., “E. Coli”, “S. Neumoniae”, “S. Aureus”, “E. Faecium”, “E. Faecalis”, “E. Epidermidis”, only the abbreviation was annotated with its corresponding long form.

Hyphenated words were annotated as a unit, as both parts compose the term. However, if those words lack the hyphen, they are annotated separately.

4.5.4 Misspellings

As already mentioned by (Benavent and Iscla, 2001) incorrect variations of known abbreviations are quite frequent in clinical records. Incorrect forms together with other misspellings such as missing letters and wrong letter order were classified as unknown. However, the correct abbreviation together with the long form, taking the context into account, were suggested in the notes section of annotation. For instance, in Figure 3 we see the proposal for the case of *mgr* instead of *mg* (‘miligram’). Only in 20 cases, they were absolute unknown terms, for instance

"Orientado en espacio y persona, PINR, no refiere diplopía" (Space and person oriented, PINR, does not refer diplopia).

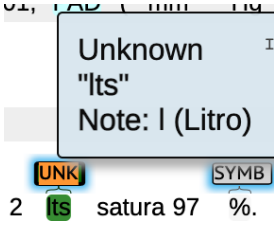

Figure 3. Detail of Unknown annotation: the note might contain the correct short form and corresponding longform.

We have handled the use of commas instead of periods as misspellings. For instance, "E. Coli" was also found in the texts as: "E.coli", "E coli", "E COLI", and "e, coli". They were annotated as abbreviations, whether it is a capital or lower-case character, and not including the comma.The incorrect use of upper and lower-case letters, such as "mmii" instead of "MMII" for *miembros inferiors* (lower limbs) or "eeii" instead of "EEII", for *extremidades inferiores* (lower extremities) is too frequent to be considered an occasional misspelling. We decided to annotate it as a correct form.

Another rather frequent misspelling was when the last part of a word wrongfully joins the next word. In our corpus, very often, the contraction "del" or "al" ('of the' and 'to the', literally) got separated and the letter "l" joins the next word, which can be very confusing and may hinder the identification of the term. In this case, we annotated the abbreviation and ignored the "l".

4.5.5 Mathematical symbols

Mathematical symbols are annotated for consistency. Some mathematical symbols are very ambiguous, but the context helps in deciding about the corresponding long form. Thus, we decided annotating the symbols and coding the appropriate Long Form according to the context. Some symbols deserve special clarification. We classified "+" as a symbol, and the actual long form for each case depends on the context. It can be:

- Positive
- Addition (mathematical symbol). Also, + is commonly used instead of "and".
- Intensity (as in edemas and swellings): + (mild, "leve"), ++ (moderate, "moderado"), +++ (severe, "intenso"), ++++ (serious, "muy severo")
- Levels in mg/dl: it is used in tests to determine the presence of proteins in urine.

We classified the "-" symbol as *negative* or *subtraction*, depending on the context, although, note it could be a hyphen too, but in this case there is no annotation. Other symbols were quotes (' and "), that were classified as symbols of minute and second respectively in the appropriate cases. Finally, we also had to make a distinction between roman numbers (for us, symbols like *IV ventrículo*, fourth ventricle) and acronyms (like *VI*, *ventrículo izquierdo*, left ventricle). We decided whether the term was an acronym or a roman number symbol according to context.

5 Conclusions

In this article, we have introduced the IULA-SCRC–ABBR corpus. It is a dataset of 3,194 sentences extracted from anonymized clinical records and annotated for abbreviation identification, including shortened forms, acronyms and symbols. The corpus was revised and validated by two human annotators. We have also described the annotation guidelines for the annotators, and the underlying criteria that motivated the choice of the three classes: ABBR, ACRO and SYMB. These underlying criteria were based on the characteristics of Spanish abbreviation and in relation with other abbreviation annotated corpora of clinical records already available, although for other languages, that is English, German and Swedish. To our knowledge, the IULA-SCRC-ABBR corpus is the first corpus of Spanish authentic clinical records annotated for abbreviations that is freely accessible under a Creative Commons BY-SA 3.0 license as this resource has been created for supporting the development of natural language processing systems for Spanish and their evaluation.

Acknowledgements

We would like to thank Dr. Pilar Bel, Laura Bernard, Miquel Cornudella, Jorge Vivaldi, and Montserrat Marimon for their collaboration in the annotation task and for their valuable comments and assessment. Research reported in this publication was partially supported by the Project PID2019-104512GB-I00 funded by Ministerio de Ciencia e Innovación (Spain).

References

Benavent, R.A., and A.A. Iscla, A.A. 2001. Problemas del lenguaje médico actual. (ii) abreviaciones y epónimos. *Papeles Med* 10(4), 170–6 (2001).

Castaño, J., P. Ávila, D. Pérez, H. Berinsky, Park, L. Gambarte, and D. Luna. 2018. A Simple Approach to Abbreviation Resolution at BARR2, IberEval 2018. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Cuadros, M., N. Pérez, I. Montoya, and A. García Pablo. 2018. Vicomtech at BARR2: Detecting Biomedical Abbreviations with ML Methods and Dictionary-based Heuristics. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Dallianis, H. 2018. Clinical Text Mining. Secondary Use of Electronic Patient Records. Springer. doi: 10.1007/978-3-319-78503-5.

Gorinski, Ph., H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, and B. Alex. 2019. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. *arXiv preprint arXiv:1903.03985.*

Hua X., P. Stetson, and C. Friedman C. 2007. A Study of Abbreviations in Clinical Notes. *AMIA 2007 Symposium Proceedings*: 821-825.

IATE, Interactive Terminology for Europe. http://iate.europa.eu, accessed 10-07-2021.

Isenius, N. 2012. Abbreviation Detection in Swedish Medical Records. The Development of SCAN, A Swedish Clinical Abbreviation Normalizer. Master's thesis, Department of Computer and Systems Sciences, Stockholm University.

Islamaj Doğan, R., D. C Comeau, L. Yeganova, and W.J. Wilbur. 2014. Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. *Database: The Journal of Biological Databases and Curation*, bau044. doi:10.1093/database/bau044

Intxaurrondo, A, M. Marimon, A. Gonzalez-Agirre, J.A. Lopez-Martin, H.M. Rodriguez, J. Santamaria, M. Villegas, and M. Krallinger, M. 2018. Finding mentions of abbreviations and their definitions in Spanish Clinical Cases: the BARR2 shared task evaluation results. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages* (IberEval 2018). http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Intxaurrondo, A, J. C. de la Torre, H. Rodriguez, M. Marimon, J.A. Lopez-Martin, A. Gonzalez-Agirre, J. Santamaria, M. Villegas, and M. Krallinger. 2018. Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus. Available at http://temu.bsc.es/BARR2/. Accessed 3-12-2021.

Kreuzthaler M., M. Oleynik, A. Avian, S. Schulz. 2016. Unsupervised Abbreviation Detection in Clinical Narratives. In *Proceedings of the Clinical Natural Language Processing Workshop*, 91-98.

Kvist M., and S. Velupillai (2014) SCAN: A Swedish Clinical Abbreviation Normalizer. In Kanoulas et al. (eds.): *CLEF 2014*, LNCS 8685: 62-73.

Lai, K.H., M. Topaz, F.R. Goss, and L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records, *Journal of Biomedical Informatics*, 55.

Marimon, M., J. Vivaldi, and N. Bel. 2017. Annotation of negation in the IULA Spanish Clinical Record Corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles, SemBEaR*.

MESH. Spanish translation of Medical Subject Headings. https://www.nlm.nih.gov/mesh/. Accessed 10-07-2021.

Névéol, A., H. Dallianis, S. Velupillai, G. Savova, and P. Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities

and challenges. *Journal of Biomedical Semantics* 9.

Pathak, J., K.R. Bailey, C.E. Beebe, S. Bethard, D.C. Carrell, P.J. Chen, D. Dligach, C. M. Endle, L.A. Hart, P.J. Haug, S.M. Huff, V.C. Kaggal, D. Li, H. Liu, K. Marchant, J. Masanz, T. Miller, T. Oniki, M. Palmer, K.J. Peterson, and C.G. Chute. 2013. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association: JAMIA*, *20*(e2), e341–e348.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA.

Rubio-López, I., R. Costumero, H. Ambit, C. Gonzalo-Martín, E. Menasalvas, and G.A. Rodríguez. 2017. Acronym disambiguation in Spanish electronic health narratives using machine learning techniques. *Studies in health technology and informatics:* 235-251.

Sánchez, Ch., and P. Martínez. 2018. A Simple Method to Extract Abbreviations Within a Document Using Regular Expressions. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Sánchez León, F. 2018. ARBOREx: Abbreviation Resolution Based on Regular Expressions for BARR2. In Rosso et al. (eds.) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

SNOMED CT Spanish browser, http://browser.ihtsdotools.org/. Accessed 10-07-2021.

Soto Montalvo, S., R. Martínez, M. Almagro, S. Lorenzo. 2018. MAMTRA-MED at Biomedical Abbreviation Recognition and Resolution - IberEval 2018. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (The tool is available at http://brat.nlplab.org/)

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, and H. Liu, 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, *77*: 34-49

Wu, Y., S. T. Rosenbloom, J.C. Denny, R.A. Miller, S. Mani, D.A. Giuse, and H. Xu. 2011. Detecting Abbreviations in Discharge Summaries using Machine Learning Methods. *AMIA Annual Symposium Proceedings,* 1541–1549.

Yetano Laguna J. and V. Alberola Cuñat. 2002. Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias. Madrid: Ministerio de Sanidad y Consumo.