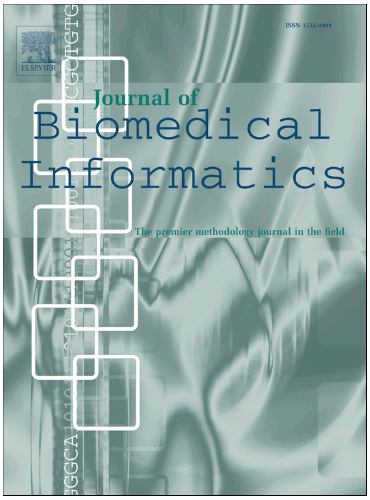


DrugSemantics: a corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics

Isabel Moreno, Ester Boldrini, Paloma Moreda, M. Teresa Romá-Ferri

PII: S1532-0464(17)30136-3
DOI: http://dx.doi.org/10.1016/j.jbi.2017.06.013
Reference: YJBIN 2809

To appear in: *Journal of Biomedical Informatics*

Received Date: 25 January 2017
Accepted Date: 12 June 2017

DrugSemantics: a corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics

Isabel Moreno[a], Ester Boldrini[a], Paloma Moreda[a,*], M. Teresa Romá-Ferri[b]

[a]*Department of Software and Computing Systems, University of Alicante, Alicante, Spain*
[b]*Department of Nursing, University of Alicante, Alicante, Spain*

Abstract

For the healthcare sector, it is critical to exploit the vast amount of textual health-related information. Nevertheless, healthcare providers have difficulties to benefit from such quantity of data during pharmacotherapeutic care. The problem is that such information is stored in different sources and their consultation time is limited. In this context, Natural Language Processing techniques can be applied to efficiently transform textual data into structured information so that it could be used in critical healthcare applications, being of help for physicians in their daily workload, such as: decision support systems, cohort identification, patient management, etc. Any development of these techniques requires annotated corpora. However, there is a lack of such resources in this domain and, in most cases, the few ones available concern English.

This paper presents the definition and creation of DrugSemantics corpus, a collection of Summaries of Product Characteristics in Spanish. It was manually annotated with pharmacotherapeutic named entities, detailed in DrugSemantics annotation scheme. Annotators were a Registered Nurse (RN) and two students from the Degree in Nursing. The quality of DrugSemantics corpus has been assessed by measuring its annotation reliability (overall F=79.33% [95%CI: 78.35-80.31]), as well as its annotation precision (overall $P = 94.65\%$ [95%CI: 94.11-95.19]). Besides, the gold-standard construction process is described in detail. In total, our corpus contains more than 2,000 named entities, 780 sentences and 226,729 tokens. Last, a Named Entity Classification module trained on DrugSemantics is presented aiming at showing the quality of our corpus, as well as an example of how to use it.

Keywords: Corpus, Reliability, Precision, Named Entity Recognition, Spanish, Summary of Product Characteristics

1. Introduction

Nowadays, there is a large amount of information on health and healthcare [1]. Examples of this huge quantity of information available are PubMed [2], a repository that comprises more than 25 million documents on biomedical literature, or the information stored for each patient on its own Electronic Health Record (EHR) during day-to-day care. Due to the high value of such data, exploiting this textual information is critical to: (i) improve healthcare quality; (ii) drive medical innovation research; and (iii) reduce healthcare costs [1]. Nevertheless, healthcare providers have difficulties to use such quantity of information during their professional practice mainly due to two reasons. On the one hand, they have a limited consultation time (i.e. often less than 10 minutes). On the other hand, the required information by them is stored in many and different sources [3].

*Corresponding author at: University of Alicante, Apdo. de correos 99, E-03080 Alicante, Spain. Tel: +34965903400 ext. 2340
Email addresses: imoreno@dlsi.ua.es (Isabel Moreno), eboldrini@dlsi.ua.es (Ester Boldrini), moreda@dlsi.ua.es (Paloma Moreda), mtr.ferri@ua.es (M. Teresa Romá-Ferri)

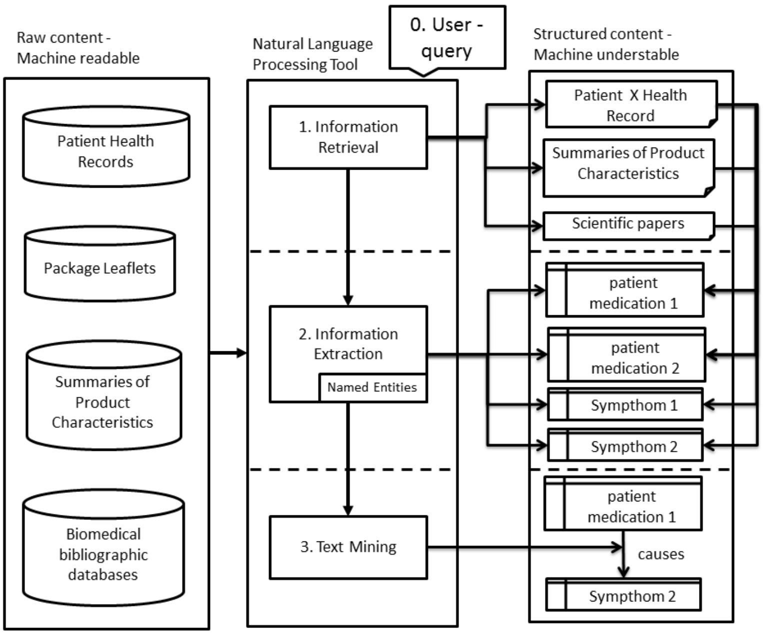


Figure 1: Example of an hypothetical tool (center box) to allow pharmacoterapeutic monitoring using Natural Language Processing techniques, as in the example described in Section 1 Introduction. Its input is raw content (left box) and its output is structured content (right box).

Envision yourself in a primary health care consultation. A general physician attends to a patient that shows several health problems, for instance: overweight, diabetes and hypercholesterolemia. This patient is being monitored with diet, exercise and various medications but, during check-ups monitoring, he/she is not improving. The physician needs to know whether the negative evolution of his/her weight and his/her cholesterol are related or not to the medications employed for his/her treatment (an oral hypoglycemic and a lipid-regulating agents).

Before reaching a conclusion, the physician should analyse a wide range of specialized documents of different sizes and sources. The most relevant ones are: (i) patient EHR, accessible through many and different applications; (ii) Summaries of medicinal Product Characteristics (SPC) or package leaflets for the patients medications, available on medicines agencies web sites at international[1] or national[2] levels; and (iii) scientific papers indexed in biomedical bibliographic databases, such us MEDLINE [2][3], or Scopus [4][4]. For healthcare providers, the analysis of all the information contained in every information source is unmanageable [5–7]. Thus they would need a tool that displays, at a glance, every document relevant to the patient condition with a single query based on their information needs [8].

Natural Language Processing (NLP) is a field of research that addresses the obstacles mentioned above. Its aim is to provide mechanisms to transform unstructured textual information, easy to understand for humans, into structured data that can be exploited by computer processes for different purposes [1, 9]. So, NLP techniques can be employed to achieve the aforementioned tool for healthcare providers. Figure 1 illustrates how to solve this problem using NLP: first, Information Retrieval (IR) techniques

[1] European Medicines Agency has more than 937 authorised medications - January 2017 (http://www.ema.europa.eu/).

[2] Spanish Agency of Medicines and Medical Devices contains more than 13,500 medications on the market - January 2017 (http://www.aemps.gob.es/).

[3] MEDLINE comprises more than 26 million citations (January 2017).

[4] Scopus includes more than 60 million records from journals and books (January 2017).

could be applied. In this way, relevant documents, that satisfy an information need, could be found from a large collection of documents [10, p. 1]. In our pharmacotherapeutic monitoring example, relevant documents would be the patient EHR together with SPCs and scientific papers related to the patient's condition. Afterwards, Information Extraction (IE) techniques could be employed. In this manner, textual and explicit relevant information could be extracted from the retrieved documents in the previous step [11, pp. 94-95][12, pp. 814-815]. In our example, the relevant information could be all the medication that a patient is currently taking and his symptoms (from the EHR) as well as signs commonly associated to these medications (from SPCs and scientific papers), among other relevant information. Then, Text Mining (TM) techniques could be used. This area is in charge of finding information that is not specified explicitly in the document, therefore, further inference is needed [13]. In the case we are dealing with, it would be to discover a reason for the negative evolution of weight and cholesterol level. That is, whether the patients current medication, all of which have been extracted previously from explicit information, is interacting with each other or not. Finally, all the obtained information (both explicit and implicit) would be displayed in an organized and summarized manner to the physician, in order to facilitate reaching a conclusion.

Building such tool for healthcare providers is not a trivial task. This is because IR is a mature area [5, 14] where several IR systems have been developed to retrieve documents that satisfies an user's information need. Some examples are PubMed [2] (professional level) or Google [15] (wide variety of users). However, there is still plenty of work to do, in both IE and TM techniques, to reach suitable results for many and different user profiles [16].

Progress in any of these techniques relies heavily on annotated corpora.This is due to the fact that these resources have mainly two purposes: (i) development - to assist during the creation of rules and statistical models that will control the behaviour of a system; and (ii) evaluation - to provide reference data against which to assess the performance of a system. Nevertheless, annotated corpora for the health domain present two main barriers.

On the one hand, there is a limited number of annotated corpora [17] and existing ones do not consider all relevant information for pharmacotherapeutic care, as Section 2 will show. Therefore, the goal of our research is the construction of DrugSemantics, a pharmacotherapeutic corpus to tackle a part of the IE problem. This resource contains annotations of Named Entities (NE) relevant to the pharmacotherapeutic care. A NE represents a mention of a semantic category in a text [11, 18]. In this field, these NEs categories refer to important information for the prescription and monitoring processes of pharmaceutical products [3, 19] and relates to concepts such as medicines[5] or clinical conditions[6].

On the other hand, most efforts have been focused on English corpora construction [20]. For this reason, DrugSemantics is a resource created using Spanish documents. It represents an attempt to increase the available annotated corpora in this language to be used to detect NEs. This corpus consists of SPCs, a type of document that allows us to overcome limitations regarding patient privacy and access to EHR data for researches outside the healthcare institutions.

This paper has four objectives: (i) to describe the construction of a gold standard, DrugSemantics, which contains pharmacotherapeutic NEs; (ii) to report on agreement between annotators, that is, reliability when a semantic category (i.e. NE) is assigned to a relevant textual fragment from DrugSemantics corpus; (iii) to provide precision of the semantic category assigned that means whether relevant information found in DrugSemantics cor-

[5] For example: trade names of medicines ("*Conacetol®*") or active substances (*"Paracetamol"* - acetaminophen in English).

[6] For instance: therapeutic indications, contraindications or intercurrent illness.

pus is valid or not; and (iv) to demonstrate how to use the DrugSemantics gold standard to deal with NEs.

The rest of the paper is organized as follows. Section 2 describes previous efforts to build annotated corpora. Then, Section 3 presents the materials and methods employed during the construction of DrugSemantics corpus and the quality evaluation of the created gold standard. Next, Section 4 outlines assessment results. Latter, the achieved results are discussed in Section 5. Finally, Section 6 presents our main conclusions.

2. Background

This section reviews existing corpora semantically annotated with NEs that are relevant for pharmacotherapeutic care. Building such resources for English has received considerable attention [21–35]. The most relevant ones are described below:

1. The i2b2 corpus [25–27] consists of 1,243 fully deidentified discharge summaries from Partners Healthcare. This corpus was pre-annotated on the basis of pooled system outputs [18], then 20.19% out of them (i.e. a subset of 251) was manually checked by i2b2 challenge participants (NLP experts) and organizers (domain experts and no experts trained). This corpus includes medications (covering also active substances), dosages, frequencies, durations, routes and reasons of administration.
2. The DrugDDI corpus [24] is made of 1,025 texts from two different sources: MedLine [2] abstracts and DrugBank [36] documents describing drug interactions. All of these (100%) were pre-annotated automatically with UMLS MetaMap Transfer (MMTx) tool [21]. Then two expert pharmacists, with background in pharmacovigilance, reviewed these tags and added new ones, if necessary. This collection contains: generic drug names (i.e. active substances), branded drug names (i.e. medications), drug group names (i.e. pharmatherapeutic group names) and active substances not approved for human use. On average, each document contains 6.63 sentences and 18.05 NEs.
3. The CLEF corpus [23] has 565,000 documents of three types: clinical narratives, histopathology reports and imaging reports. These documents belong to 20,324 deceased patients from Royal Mardsen Hospital. Only 0.27% of these documents (i.e. a subset of 150) were manually labelled by domain experts and NLP experts. The following NEs were included: condition (i.e. disease), drug or device (i.e. medications, active substances, etc.), intervention, investigation, result, and locus. These documents contain 21.63 NEs on average.

In contrast, it was not until recently when Spanish researchers have created corpora related to the pharmacotherapeutic field with semantic information in Spanish. To the best of our knowledge, there are only two available:

1. The IxaMedGS [20] corpus contains 142,154 deidentified Discharge Records of EHR written in Spanish from the Galdakao-Usansolo Hospital. Only 0.01% of all records (i.e. a subset of 75) were pre-annotated using Freeling-Med [20] to include: (i) diseases; and (ii) drugs, namely: active substances and medicines. After this, the result was manually checked by four annotators, who have expertise in the Pharmacology and Pharmacovigilance fields. On average, each document contains 72.13 sentences, 555.11 tokens and 52.76 NEs.
2. The SpanishADRCorpus [37] is composed by 397 comments gathered from a Spanish health forum, ForumClinic [38]. Those comments (100%) were labelled manually by two annotators with expertise in Pharmacovigilance. This collection includes information about the entities: (i) drugs (i.e. medicines, active substances, pharmatherapeutic groups, etc.); and (ii)

adverse drug reactions. These documents contain 77.97 tokens [7] and 2.07 NEs on average.

A comparative summary of the analysed corpora, as well as DrugSemantics gold standard, can be found in Table 1. Regardless of the language, several conclusions can be drawn. First, most efforts produced a semi-gold standard corpus[8], which could facilitate the manual annotation process by reducing time, as shown by [28, 39], and task difficulty. Second, all of them employed domain experts due to their knowledge in this area. Third, all corpora contain information about medications. Nevertheless, none of them include food or excipients, which are also NEs critical when healthcare providers have to choose a pharmacological treatment (e.g. to prevent interaction and allergies problems) [3, 19]. Fourth, most of them incorporate important clinical conditions: (i) condition [23]; (ii) reasons of administration [25–27]; (iii) diseases [20]; and (iv) adverse drug reactions [37]. But other clinical conditions, such as medical contraindications or overdoses, are not explicitly incorporated and all of them are essential in both medicine prescription and monitoring processes during day-to-day care [3, 19].

Focusing on the differences between the two languages, corpora have been created more frequently for English [21–35] than Spanish [20, 37]. The length of the annotated documents is not shared between the languages: English resources have a more balanced number of NE per document (18.05 versus 21.63) than Spanish ones (52.76 versus 2.07). Furthermore, it should be noted that existing Spanish corpora only include information about diseases, adverse drug reactions and drugs, as Table 1 shows.

Given the identified gaps, this paper reports on the construction of the Spanish DrugSemantics corpus, which is designed to include a larger number of NEs for the pharmacotherapeutic process than previous works.

3. Material and Methods

This section describes the methodology employed to build the DrugSemantics corpus. First, the document sampling procedure is presented in Section 3.1. Then, our annotation scheme is outlined in Section 3.2. Latter, the manual annotation guidelines are described in Section 3.3. Next, the assessment of the manual annotation is defined. On the one hand, Section 3.4 presents the metrics of this evaluation. On the other hand, the methodology followed to compute these metrics is described in Section 3.5. Lastly, the construction of the gold standard is presented in Section 3.6.

3.1. DrugSemantics Corpus Sampling Description

This corpus consists of 30 Summaries of medicinal Product Characteristics (SPC), which includes 7,085 sentences and 175,965 tokens. SPC is an standardised official document [40] that includes wealth information about a medicine, approved by health authorities, and its therapeutic indications [41]. This type of document was chosen due to two reasons: (i) SPCs cover appropriately the information needs and they are a priority information source for prescribing and monitoring medications, as shown in [34, 42]; and (ii) health information from citizens has limited access [1, 43–45] and, nowadays, there is no Spanish open-access repository integrating de-identified information on patients equivalent to *Research Patient Data Repository from Partners Healthcare* [46].

The SPCs were selected from a reliable open-access repository called "Medicines Online Information Center" (CIMA[47]) that belongs to the Spanish Agency for Medicines and Health Products (AEMPS). The aim of the sampling was to consider medicines widely used in Spain

[7] Value computed using LingPipe Tokenizer plugin and publicly available data at `http://labda.inf.uc3m.es/doku.php?id=en:labda_spanishadrcorpus` (last access April 16, 2017). When using LingPipe, the term token refers to all document units (e.g. terms and punctuation symbols) except from spaces.

[8] A semi-gold standard corpus is created by a human annotator who manually checks pre-annotated entities.

Corpus\ Features	**i2b2**	**DDI**	**CLEF**	**Ixa-MedGS**	**SpanishADR**	**DrugSemantics**
Document Type	Discharge Summaries	Abstracts, DrugBank texts	Clinical documents	Discharge Summaries	ForumClinic comments	Summaries of Product Characteristics
Tagged documents (Tagged%)	251 (20.19%)	1,025 (100%)	150 (0.27%)	75 (0.01%)	397 (100%)	5 (16%)
Average Document Length	- - -	6.63 sentences - 18.05 NE	- - 21.63 NE	72.13 sentences 555.11 tokens 52.76 NE	- 77.97 tokens* 2.07 NE	156 sentences 4,535.8 tokens* 448.2 NE
Annotation type	Pre-annotated automatically/ manually checked	Pre-annotated automatically/ manually checked	Manual	Pre-annotated automatically/ manually checked	Manual	Manual
Annotator type	Experts (domain experts & trained), not experts (NLP)	Experts (pharmacist)	Experts (clinician & biologist), not experts (NLP)	Experts (pharmacology)	Experts (pharmacist)	Experts (registered nurse & nursing students)
Named Entities (Number of Named Entities)	Medications, dosages, frequencies, durations, routes and reasons of administration	Generic drug (1,164), branded drug (1,866), drug group (4,225) and active substance not approved for humans (765)	Condition (1,056), intervention (254), investigation (431), result (292), drug_device (197), and locus (1,014)	Disease (2,766) and drug (1,191)	Drug (187) and adverse drug reactions (636)	Disease (724), Drug (657), Unit of Measurement (557), Excipient (66), Chemical Composition (62), Pharmaceutical Form (45), Route (42), Medicament (37) Food (31) and Therapeutic Action (20)
Language	English	English	English	Spanish	Spanish	Spanish

(*): Value computed using LingPipe Tokenizer pluggin from Gate Developer and publicly data available

Table 1: Features of the analysed corpora related to pharmacotherapy

to treat elevated cholesterol levels, as well as to deal with minor health problems, such as fever or mild to moderate pain. Thus, a non-probabilistic sampling, using expert judgement, was manually performed to choose 5 active substance, namely: Atorvastatin, Simvastatin, Acetylsalicylic Acid/Aspirin, Paracetamol/Acetaminophen and Ibuprofen.

For each drug, 6 SPCs were chosen considering only commercialized medicines at that time with different brand names (e.g. Gelocatil® or Paracetamol Ratiopharm®) and pharmaceutical forms (e.g. capsules, tablets) to ensure the highest diversity as possible. In the case of medicines for cholesterol (i.e. Atorvastatin and Simavastatin), the variability is lower because only two pharmaceutical forms are used (i.e. tablets and film-coated tablets). Finally, it should be noted that medicines whose active substance is Atorvastatin are mainly generic. Therefore, according to current Spanish regulations, its brand name must include the active substance name [40] (e.g. *Paracetamol Ratiopharm®*).

3.2. DrugSemantics Annotation Scheme

The objective of DrugSemantics annotation scheme is twofold: (a) to identify NEs relevant for pharmacotherapeutic care that allows defining the annotation guidelines; and (b) to annotate SPCs semantically with NEs from the ones previously identified. Our model is based on (i) OntoFIS pharmacotherapeutical ontology [48]; and (ii) common questions about medicines for both healthcare providers [3, 19] and patients [49]. Using these researches as a basis, DrugSemantics is able to capture real information needs of all the actors involved in the pharmacotherapeutic process.

Our model contains 10 different types of NEs (see Table 2): Chemical Composition, Disease, Drug, Excipient, Food, Medicament, Pharmaceutical Form, Route, Therapeutic Action and Unit of Measurement. Examples for each NE can be found in Table 3. These entities, together with their attributes, are described below in alphabetical order:

Chemical Composition is related to chemical group, which represents the third level of Anatomical Therapeutic Chemical (ATC) classification[9] [51];

Disease relates to any clinical condition and, optionally, distinguishes between:

- *Therapeutic Indication* states whether a medicine must be used for a target disease as treatment, diagnosis, or prevention (primary or secondary);
- *Interaction*, when a set of substances are given to a patient, some actions can change producing new clinical conditions and it can be synergistic (when the drug's effect is increased) or antagonistic (when the drug's effect is decreased);
- *Contraindication* specifies clinical conditions when a medication should not be used due to an allergic reaction, a medical problem, some physiological change (e.g. pregnancy) or other treatments/therapies;
- *Desirable Effect* identifies the usages of a medicament to prevent, cure or palliate clinical conditions or a health problem;
- *Side Effect* refers to unexpected clinical manifestations appearance when a medication is given on its usual dosage; and
- *Overdosage* relates to unexpected clinical manifestations appearance when a higher dosage than usual of a medication is given.

Drug is an active substance, which is the designation for the most specific level of ATC [51]. It distinguishes

[9] The ATC classification system is widely used internationally and, in 2003, it was also adapted to be used in the Spanish healthcare system [50].

Table 2: DrugSemantics Scheme Named Entities and attributes: Named Entities ordered alphabetically

Named Entity	**Attributes (={predefined values})**
Chemical Composition	type = {chemical group ATC, chemical group ATC code, Chemical Name, Formula}
Disease	type = {Therapeutic Indication, Interaction, Contraindication, Desirable Effect, Side Effect, Overdosage}
Drug	type = {Drug Name, Drug ATC Code } Strength UnitOfMeasure
Excipient	
Food	type = {Solid, liquid, Supplements, Additives}
Medicament	TradeName Country = {Spain, Others} Strength UnitOfMeasure Pharmaceutical Form
Pharmaceutical Form	
Route	
Therapeutic Action	type = {TherapeuticGroupName, TherapeuticGroupATCCode, DrugNameGroup, DrugGroupATCCode}
Unit of Measurement	amount unitname magnitude

Table 3: Examples from DrugSemantics Scheme: Named Entities ordered alphabetically

Named Entity	***Example in Spanish* (English translation)**
Chemical Composition	*fibrato* (fibrate)
Disease	*insuficiencia renal* (renal failure)
Drug	*Atorvastatina* (Atorvastatin)
Excipient	*maltosa* (maltose)
Food	*zumo de pomelo* (grapefruit juice)
Medicament	*ALCOSIN 10 mg comprimidos recubiertos con película* (ALCOSIN 10 mg film-coated tablets)
Pharmaceutical Form	*comprimidos* (tablets)
Route	*oral* (oral)
Therapeutic Action	*analgésicos y antipiréticos* (analgesic and antipyretic)
Unit of Measurement	*10 mg al día* (10 mg daily)

between name or code of an active substance, and optionally, includes strength and unit of measurement;

Excipient is a substance included in medicines for the purpose of giving shape, consistency, stability, colour, smell, taste or ease its usage. They can sometimes be the cause of allergic reactions or other undesired effects;

Food refers to food taken by patients that can interact increasing or decreasing the effect of a medicament. It distinguishes between solid, liquid, supplement and food additive;

Medicament is constituted by its brand name and, optionally, by its strength, unit of measurement and pharmaceutical form; all of which is equivalent to the Spanish designation for commercialised medicines [40];

Pharmaceutical Form is the possible dosage form in which these substances are marketed;

Route is a method by which a substance is taken into the body;

Therapeutic Action "is the means by which a product achieves an intended therapeutic effect"[52] and it corresponds to an intermediate ATC [51] level. Optionally, it distinguish between name or code of the therapeutic group, as well as name or code of the pharmacological group;

Unit of Measurement identifies name and quantity of a magnitude adopted by convention.

3.3. Study Sample and Manual Annotation Process

Annotators received the same 5 SPCs (one for each drug - see Table 4) to perform the semantic annotation task. These were randomly chosen from the 30 SPC downloaded from CIMA (see Section 3.1). The reasons behind giving complete SPCs were: (i) ease the annotation process and (ii) facilitate control over the information source.

Table 4: DrugSemantics gold standard characterisation (sample)

Drug	**S***	**T****	**T/S*****	**NE+**
Aspirin	123	3,196	25.98	359
Acetaminophen	146	4,172	28.57	412
Ibuprofen	48	1,225	25.52	81
Atorvastatin	261	8,066	30.9	774
Simvastatin	202	6,040	29.90	615
Total	780	22,679	29.08	2,241

Note: (*) Number of Sentences in each annotated SPC;
(**) Number of Tokens in each annotated SPC;
(***) Average Tokens per sentences in each SPC;
(+) Number of Named Entities per annotated SPC

Annotators were a registered nurse (RN) and two students in their final year from the Degree in Nursing, whose native language was Spanish. These 3 healthcare providers (A1, A2, A3) were chosen as annotators due to their pharmaceutical knowledge and their complete understanding of Spanish SPCs.

The annotation process lasted approximately three months, on a part-time basis, and started with an initial joint training session. Both the annotation tool to use, GATE Developer [53], and the annotation task, with particular emphasis on positive examples for each entity and its attributes, were introduced during this initial meeting. Once this session finished, each annotator, separately, received: (i) 5 SPCs, (ii) the annotation guidelines, with positive examples, and (iii) an ATC classification listing [54] in order to help identifying NEs from our annotation scheme. Each annotator worked independently, meaning that there was no contact between them while the annotation process lasted, owing to schedule difficulties. This fact ensured that there was no influence on the decisions adopted by annotators, and it is based on the methodology of experimental designs in healthcare, whose purpose is to control interpretative bias or performance bias [55]. After the training phase, SPCs annotation was performed in three steps. First, annotators carried out an initial annotation round. Then, each annotator participated in a tailor made session to solve doubts and problems that had appeared during the initial round. Last, a final annotation round was carried out by each annotator. Its purpose was that annotators introduce directions given in the previous step. In this manner, each annotator checked her labelling through all her documents before delivering the final version of her 5 annotated SPCs.

3.4. Metrics for Manual Annotation Evaluation

Once the manual annotation process is completed, its quality is determined by a set of quantitative metrics. Reliability and precision are the two main factors on which annotation quality depends. Reliability is usually determined by means of the agreement reached between annotators who worked on the same set of documents [56, 57]. Whereas precision is based on assessing whether a semantic category manually assigned corresponds with a concept from the annotation scheme [58]. Precision is commonly assumed if there is agreement. This strategy has been employed because there is no gold standard that clearly specifies what is or not valid, since the reference dataset is being built [58]. This work does not consider precision is given by reliability, and as a result, this section defines metrics to determine DrugSemantics corpus quality, gauging both reliability and precision at NE category level (ignoring properties defined DrugSemantics scheme).

3.4.1. Reliability

Reliability of an annotated corpus is computed as the agreement among independent annotators [56, 57]. Although Kappa coefficient has been applied before for measuring agreement over NEs annotation [24, 59], in our case is not applicable because this coefficient takes into account the probability of agreement by chance [57]. Given that annotators could label any textual fragment and assign a category to it, the number of fragments not being considered a NE is a large number, since these can overlap as well as vary in length. Consequently, the probability of agreeing by chance would be near to zero. In these cases, F-measure (F) is commonly used [22, 26, 37] to estimate agreement between ratters, since it approaches Kappa [57]. Thus, agreement was calculated for each entity and annotator as a pairwise F-measure between annotators (FA1-A2, FA1-A3, FA2-A3), and then the average among all pairs [57]. The following considerations were made:

a) Agreement is calculated as lenient F to allow partial matches when two annotated NE share a common span of text [53]. A lenient criterion was chosen due to two reasons. On the one hand, encoding issues added extra space characters in our corpus (e.g. an space was added when an annotation precedes a punctuation symbol). On the other hand, two annotators marked the same fragment except for few characters, which usually represented punctuation or a plural ending. For example, A1 marked: "*malformaciones*" (malformations in English'); whereas A2 tagged: " *malformaciones*" including the space before the NE begins.

b) F is macro and micro averaged at sentence level. **Macro-averaged F** (MF) is calculated across NEs separately and then the arithmetic mean is computed. Specifically for macro-average per annotator pair ($A_i - A_j$), F is calculated for all sentences (s) and for each entity (e), e.g. for Drug: $\{MF_{A_i-A_j}(Drug, s_1), ..., MF_{A_i-A_j}(Drug, s_{780})\}$. Then, the arithmetic mean of these sentence results per annotator pair is calculated:

$$MF_{A_i-A_j}(Drug) = \frac{\sum_{s=1}^{780} F_{A_i-A_j}(Drug, s)}{780} \quad (1)$$

Latter, the arithmetic mean of all entities is computed for each pair:

$$MF_{A_i-A_j} = \frac{\sum_{e=1}^{10} MF_{A_i-A_j}(e)}{10} \quad (2)$$

Finally, the arithmetic mean of our three pairs is performed to obtain overall results:

$$MF = \frac{MF_{A_1-A_2} + MF_{A_1-A_3} + MF_{A_2-A_3}}{3} \quad (3)$$

Whereas **micro-averaged F** (mF) is calculated aggregating exact matches[10], partial matches[11] and non-matches[12] counts on the entire corpus before computing F. Specifically for micro-average per annotator pair ($A_i - A_j$), sum of counts to obtain cumulative exact matches (em), partial matches (pm) and non-matches (nm) are computed for all sentences (s) and for each entity (e). For example, for Drug exact matches:

$$em_{A_i-A_j}(Drug) = \sum_{s=1}^{780} em_{A_i-A_j}(Drug, s) \tag{4}$$

Then, the mF applies traditional F1 formula using these cumulative sentence results per annotator pair and entity as follows:

$$mF_{A_i-A_j} = \sum_{e=1}^{10} \frac{2(em_{A_i-A_j}(e) + pm_{A_i-A_j}(e))}{2(em_{A_i-A_j}(e) + pm_{A_i-A_j}(e)) + nm_{A_i-A_j}(e)} \tag{5}$$

Finally, the arithmetic mean of our three pairs is performed to obtain overall results:

$$mF = \frac{mF_{A_1-A_2} + mF_{A_1-A_3} + mF_{A_2-A_3}}{3} \tag{6}$$

c) Taking into account the lack of consensus on how to interpret agreement values [56], this work adapted the Landis and Koch scale [60] for Kappa: $F \in [100, 80]$ signifies almost perfect agreement; $F \in (80, 60]$ means substantial agreement; $F \in (60, 40]$ represents moderate agreement; $F \in (40, 20]$ signifies fair agreement; and $F \in (20, 0]$ means slight agreement.

d) In our opinion, a corpus is reliable when its agreement exceeds 60 (i.e. $F > 60\%$). Consequently, our hypothesis is that agreement (overall and per annotator pairs) in DrugSemantics corpus must have a F-measure greater than 60.

e) F is accompanied by a Confidence Interval (CI) or a Maximum Margin of Error[13] (MME) to provide a more detailed reliability description. CIs give more information indicating a range of values (interval) that is likely to contain the true value, with a probability or confidence level. In our case, 95% CI has been set and a significance level of $\alpha = .05$. In order to verify our agreement hypothesis[14], when the provided F-measure is greater than 60, one proportion $Z - test_{\alpha=.05}$ will be provided. In this case, if such Z value is greater than $Z_\alpha = 1.645$, then the agreement truly exceeds 60% with significance level of $\alpha = .05$.

3.4.2. Precision

Precision measures the consistency of the manual annotation focusing on whether the semantic category assigned is correct, that is, it corresponds with an appropriate concept from the scheme [58]. In our case, correct (or valid) denotes whether the manual NE annotation is commonly used to refer to the selected entity type [15]. In this narrow domain, there is a wide range of knowledge resources that capture common terms for specific concepts. Thus, precision for each NE and annotator is estimated as the percentage of entities manually annotated by an annotator that are present in a knowledge resource (e.g. a dictionary)[16]. Precision is computed as follows:

$$P_a(NE, D) = \frac{match_a(NE, D)}{total_a(NE)} \tag{7}$$

where a is the annotator's identifier; NE is the target named entity from DrugSemantics scheme; D is a knowledge resource, like a dictionary, that can be employed for comparison because it includes common terminology

[10] Two manually tagged NEs share the same type and have exact character offsets (span).

[11] Two manually tagged NEs share the same type and have a common span.

[12] Two manually tagged NEs do not share type nor offsets.

[13] A Maximum Margin of Error is defined as half of the width of a CI.

[14] See item d above.

[15] For instance, whether acetaminophen is manually annotated as Drug and it is commonly used to refer to active substances.

[16] Further details about knowledge resources that are used in this work can be found in Section 3.5.2

for a given NE type; $total_a(NE)$ is the number of annotations that annotator a has included of type NE; and $match_a(NE, D)$ is the number of manual annotations of type NE from annotator a that are present in the resource D. Specifically, the *match* function compares each annotation of type NE from annotator a with all entries in a resource D and employs simple string pattern matching ignoring: accent marks, lower and upper-case letters.

First, precision is being calculated for each entity and annotator following Equation 7. The purpose is to aid in the construction of the gold standard: when disagreements between annotators appear, precision will take a judge role. Then, in order to estimate an overall precision for DrugSemantics, results were macro and micro averaged for all annotators.

Bearing in mind that the health domain requires high-quality NLP resources to avoid errors and its consequences, our hypothesis is that precision in DrugSemantics must have a Precision greater than 80%. Precision is accompanied by CI or MME to provide a more detailed description. Likewise F, 95 % CI and significance level of $\alpha = .05$ has been set. Similarly, one proportion $Z - test_{\alpha=.05}$ is provided to verify our hypothesis when Precision is greater than 80.

3.5. Methodology for Manual Annotation Evaluation

This section describes the procedures followed to assess manual annotation for each semantic category within DrugSemantics scheme. First, Section 3.5.1 outlines the method to compute Reliability applying F (see Section 3.4.1). Last, Section 3.5.2 defines the method to apply Precision (see Section 3.4.2).

3.5.1. Reliability

Once the annotation process finished, agreement between annotators was calculated with Corpus Quality Assurance, from GATE Developer [53]. Before, an initial preprocessing was conducted to detect sentences within our study sample (5 SPCs manually annotated by 3 annotators). To that end: i) points were added at the end of a title in every section (first level heading) and, if necessary, in every subsection (second level heading) aiming at facilitating the identification of sentences; ii) manual annotations on titles were deleted; iii) both headers and footers, in each labelled SPC, were removed; and finally iv) sentences were automatically detected by LingPipe, a GATE Developer [53] plugging.

3.5.2. Precision

Annotation precision was calculated semi-automatically to compute precision between manual annotations and DrugSemantics annotations scheme. This method used an existing dictionary-based NER system, MaNER [61, 62], in order to reduce the expense of the manual review process for correctness. Each entity MaNER recognises has its own dictionary. Each dictionary was gathered from a reliable biomedical knowledge resource and it contains a list of terms representing common and relevant vocabulary for a given NE [63]. The purpose of using a NER system was to manually review only those annotations that are not present in the dictionaries. However a high effort was still required to manually examine these annotations. Thus, it was decided to review by hand all entities from DrugSemantics scheme except from Disease, given that each Disease subtype required a different dictionary to decide if an annotation is correctly tagged.

Initially, MaNER [61, 62] recognized four NE (i.e. Medicament, Drug, Pharmaceutical Form and Route). Hence, such system was extended to include the five missing entities (i.e. Chemical Composition, Excipient, Food, Therapeutic Action and Unit of Measurement). Besides, a new dictionary for Drug was created to replace ActiLex dictionary [61, 62] and overcome issues previously reported [62], such as multi-words inversion. Therefore, 6 new Spanish dictionaries were created and 3 were kept. All of them were acquired by querying several reliable biomedical knowledge resources, namely:

Chemical Composition: its dictionary was built from ATC [51].

Drug: its dictionary was obtained from ATC [51], which replaced the original MaNER dictionary for this entity, ActiLex [61, 62].

Excipient: its dictionary was gathered from the International Numbering System for Food Additives [64].

Food: its dictionary was compiled from BEDCA [65].

Medicament: MePLex dictionary [61, 62] was built from Nomenclator Digitalis [66].

Pharmaceutical Form: its dictionary [61] was obtained from Nomenclator Digitalis [66].

Route: its dictionary [61] was obtained from Nomenclator Digitalis [66].

Therapeutic Action: its dictionary was gathered from ATC [51].

Unit of Measurement: its dictionary [61] was compiled from SNOMED-CT [67].

Precision was estimated in four stages, as follows:

1. Comparison: Text fragments manually annotated were compared automatically against its dictionary with exact matching, meaning that our comparison employed simple string pattern matching between tokens ignoring only accent marks and capitalization. As a result, a list of not matched annotations was generated automatically for each DrugSemantics entity included in this experiment.
2. Analysis: These lists were manually analysed to know how to update our dictionaries. Each annotation not matched was classified as dictionary fault, human error[17] or partial match (i.e. a manual annotation represents an entry in the dictionary, but the former includes the omission or addition of a character, such as a punctuation symbols).
3. Update: As a result, dictionaries were updated with the annotations classified as dictionary fault to perform an exact perfect matching. In this manner, the dictionaries were enhanced to detect more valid annotations. Additionally, manual annotations labelled as a partial match were fixed by means of adding or removing missing characters.
4. Precision: Again, annotations were compared automatically and precision was calculated for each DrugSemantics entity included within this experiment.

3.6. Gold Standard Construction

Once the evaluation finishes, a gold-standard is built. In our case, the final corpus was created by combining annotations from our three annotators to obtain the largest set of annotations of the highest quality possible in terms of precision. To that end, agreements between at least two of three annotators became part of the DrugSemantics corpus, if $F_{A_iA_j}(NE)$ is greater than 60%. It should be noted that Disease entity agreements were strict (character offsets) and considered its type (e.g. Overdosage, etc.) to decide whether an annotation was in our final corpus or not.

Discrepancies were solved using MaNER dictionaries as a judge, if available: when only one annotator a detects an entity $ne1$, this annotation would be included in the gold standard, if it matches its dictionary D and its precision $P_a(NE, D)$ exceeds 80%. This approach would be applied also when $F_{A_iA_j}(NE)$ is less than 60%, but $P_a(NE, D)$ exceeds 80%.

Besides, when two annotators detect the same entity but its scope differs (i.e. inexact character offsets disagreements), the chosen annotation was the one that is more faithful to the DrugSemantics Scheme or more spe-

[17] Errors in manual annotations, e.g.: physiology of administration routes ("*ingestión*" - intake) instead of the route itself ("*oral*").

cific, even though this annotation is not the most frequent. For instance, it was preferred mentions of Drugs that contained their strength (amount and unit of measurement) in the original text, regardless the number of annotators: "*20 mg de simvastatina*" (20 mg of Simvastatin) instead of "*simvastatina*" (Simvastatin). Regarding annotation specificity, MaNER dictionaries were employed to decide in this case. For example, "*Soluciones Orales*" (oral solutions) appear in text and only one annotator included it as Pharmaceutical Form but the others only selected "*Soluciones*" (solutions), the first one was chosen because it's more specific.

3.7. Gold Standard Use Case: Named Entity Classification

In order to show how this corpus cold be used, this section presents a pilot use case. The DrugSemantics corpus is designed to be used in the Named Entity Recognition and Classification (NERC) task. The goal of NERC is to recognize occurrences of NEs in text, which is known as the recognition phase (NER), and assign them a category, which is referred as the classification phase (NEC). Since NERC can implement both phases separately, our use case is focused on the latter (NEC) assuming the output of a "perfect" NER so as to avoid any bias. Specifically, we employed the NEC from [68–70], which is based on Machine Learning (ML) and profiles. A more detailed description of our method can be found in [69].

This NEC uses the DrugSemantics gold standard as training corpus, whereas the SpanishADR [37] corpus is employed for testing purposes. Although these corpora used different annotation schema and entities are not an exact match, some entities are closely related. For instance, our *Disease* entity is a generalization of *AdverseEffect* entity from SpanishADR. Therefore, this use case is applied to the most frequent entities from both data sets: 724 *Disease* (DrugSemantics) versus 545 *AdverseEffect* (SpanishADR).

The performance of this NEC was assessed in terms of traditional Precision (P), Recall (R) and $F-measure_{\beta=1}$ (F1) for the positive class (i.e. is Disease or is AdverseEffect).

4. Results

4.1. Reliability Results

Table 5 presents DrugSemantics corpus reliability, computed for each annotator pair and globally in terms of Lenient F-measure, its MME and its Z-value. These values are based on 95% CI and 0.05 significance level.

The results in Table 5 by pair show that when annotating most entities (6 out 10) agreement is truly substantial (5) or almost perfect (1) for all pairs. A noteworthy agreement is reached by the student pair (A1-A2), in which 9 out of 10 entities confirm our hypothesis are truly substantial (i.e. F values are truly above 60).

Overall agreement between all pairs, considering both macro (F 72.33% [95%CI: 71.25-73.41]) and micro average (F 79.33% [95%CI: 78.35-80.31]), were substantial between annotators ($F \in (80, 60]$). But the latter is very close to the upper limit of this range. Looking closer, the majority of them (8 out of 10) exhibits substantial or almost perfect agreement, that is, most obtained an F higher than 60. Only two of them have a lower agreement: Chemical Composition - $F \in (60, 40]$ - and Therapeutic Action - $F \in (20, 0]$.

Thus, reliability of DrugSemantics manual annotation is satisfactory.

4.2. Precision Results

Annotation Precision between manual annotations and DrugSemantics annotation scheme is presented in Table 6, as well as its MME and its Z-value. These values are based on 95% CI and 0.05 significance level.

Overall precision among annotators, considering both macro (90.05% [95%CI: 89.33-90.77]) and micro (94.74%

Table 5: Named Entity agreement (Lenient F-measure with its Maximum Margin of Error and Z-value) in DrugSemantics corpus per entity and globally, ordered by global agreement

Named Entity	$F_{A1-A2} \pm MME$		Z	$F_{A1-A3} \pm MME$		Z	$F_{A2-A3} \pm MME$		Z	F±MME		Z
Medicament	100			93	± 5.93	**10.9**	93	± 5.93	**10.9**	95.33	± 3.98	**17.4**
Excipient	94	± 4.18	**15.94**	95	± 3.84	**17.88**	88	± 2.99	**9.35**	92.33	± 3.86	**16.43**
Unit Of Measurement	89	± 1.91	**29.69**	89	± 1.89	**30**	90	± 1.81	**32.53**	89.33	± 1.60	**32.31**
Drug	98	± 0.76	**97.37**	80	± 2.32	**16.89**	81	± 2.29	**17.96**	86.33	± 1.60	**32.53**
Pharmaceutical Form	83	± 7.33	**6.15**	88	± 6.22	**8.83**	82	± 7.94	**5.43**	84.33	± 5.86	**8.14**
Food	96	± 5.66	**36**	71	± 12.84	**1.68**	70	± 13.24	1.48^{Ω}	79	± 9.54	**3.94**
Disease	82	± 1.85	**23.30**	65	± 2.45	**4**	69	± 2.48	**7.11**	72	± 1.87	**12.60**
Route	85	± 7.22	**6.79**	57	± 10.78	#	62	± 11.29	0.35^{Ω}	68	± 8.24	**1.90**
Chemical Composition	73	± 6.56	**3.88**	37	± 6.25	#	30	± 6.24	#	46.67	± 5.59	#
Therapeutic Action	14	± 12.85	#	4	± 5.23	#	12	± 7.72	#	10	± 6.79	#
Macro-average	81	± 1.13	**36.35**	68	± 1.39	**11.32**	68	± 1.41	**11.10**	72.33	± 1.08	**22.36**
Micro-average	88	± 0.94	**58.52**	74	± 1.3	**21.06**	76	± 1.29	**24.25**	79.33	± 0.98	**38.72**

A1: Annotator 1, student; A2: Annotator 2, student; A3: Annotator 3, nurse; MME: Maximum Margin Error; Z: one proportion Z-test value; #: F-measure below our hypothesis; Z column bold-faced: agreement truly has an F>60% with a 95% confidence; Ω: There is a lack of evidence to reject our null hypothesis (F=60%) and to accept our hypothesis (F>60%).

Table 6: Annotation Precision between DrugSemantics scheme and the following Named Entities: Medicament, Drug, Unit of Measurement, Route, Therapeutic Action, Pharmaceutical Form, Excipient, Chemical Composition and Food. Ordered as in Table 5.

Named Entity	$P_{A1} \pm MME$		Z	$P_{A2} \pm MME$		Z	$P_{A3} \pm MME$		Z
Medicament	100			100			97.06	± 5.68	**5.89**
Excipient	100			100			100		
Unit Of Measurement	100			99.03	± 0.84	**44.19**	99.07	± 0.83	**45.22**
Drug	99.39	± 0.60	**63.54**	99.21	± 0.69	**54.72**	95.92	± 1.75	**17.81**
Pharmaceutical Form	100			100			93.62	± 6.99	**3.82**
Food	100			100			100		
Route	82.69	± 10.28	0.51^{Ω}	85.71	± 10.58	1.06^{Ω}	79.31	± 14.74	#
Chemical Composition	66.67	± 9.29	#	81.82	± 9.84	0.36^{Ω}	30	± 7.88	#
Therapeutic Action	85.71	± 25.93	0.81^{Ω}	100			36.17	± 13.74	#
Overall Macro-average (A1+A2+A3) ± MME Z						90.05	± 0.72	**27.24**	
Overall Micro-average (A1+A2+A3) ± MME Z						94.65	± 0.54	**52.81**	

A1: Annotator 1, student; A2: Annotator 2, student; A3: Annotator 3, nurse; MME: Maximum Margin Error; Z: one proportion Z-test value #: Precision below our hypothesis; Z column bold-faced: agreement truly has a P>80% with a 95% confidence; Ω: There is a lack of evidence to reject our null hypothesis (P=80%) and to accept our hypothesis (P>80%).

[95%CI: 94.11-95.19]) was almost perfect. Differences between macro and micro-averaging are due to the fact that macro treats all classes equally, while micro-averaging favours bigger categories (such as Drug and Unit of Measurement). Taking into account precision for each annotator:

Annotator 1 obtains a precision lower than 80% for Chemical Composition. Although Route and Therapeutic Action have a precision above this value, its Z-value tell us that our hypothesis can not be accepted. The six remaining ones are truly above this cut-off point.

Annotator 2 gets a precision higher than 80% for all NEs. However, there is no evidence that confirms that

Route and Chemical Composition are truly below this cut-off point.

Annotator 3 obtains a precision lower than 80% for Chemical Composition, Therapeutic Action and Route, whereas the remaining ones are above this cut-off point.

Thus, precision results indicates that manual annotation worked remarkably well for almost all entities and annotators (except Chemical Composition, Route and Therapeutic Action). As a consequence, precision of DrugSemantics manual annotation is adequate.

4.3. Gold Standard Distribution

Finally, our gold standard is built according to the rules specified in Section 3.6. As a result, the frequency for each entity type can be seen in Table 7 - column GS. Besides, details about manual annotations per entity and annotator are also provided (columns A1-A3) to show differences between annotators and the final corpus. In the gold standard, specifically, Disease and Drug entities are the most numerous ones. On the contrary, the more uncommon gold standard NE, with a frequency less than 30, is Therapeutic Action. Gold standard distribution is generally maintained by all annotators across entities, but A3 is the one that introduces more changes (e.g. Drug or Therapeutic Action).

Our gold standard is publicly available for research purposes (more information in [71]).

4.4. Gold Standard Use Case Results

Table 8 presents results when a NEC system is trained to classify the *Diasese* Entity (from DrugSemantics gold standard) versus the *AdverseEffect* (from SpanishADR corpus). It should be noted that training and testing sets are composed of different textual genres. On the one hand, SPCs are formal, longer and normalized documents. On the other hand, text in forum comments are informal, smaller and content free. Besides, different annotation schema are employed.

Despite these facts, Precision (Pr), Recall (Re) and F1 are always higher than 70% and less 80%. As a consequence, these results prove that DrugSemantics gold standard is useful to deal with named entities, even if the textual genre changes.

Table 7: Statistics about DrugSemantics corpus

Named Entity	**Annotations**			
	GS	A1	A2	A3
Disease	724	887	769	567
Drug	657	651	636	490
Unit of Measurement	557	508	518	540
Excipient	66	65	59	59
Chemical Composition	62	99	77	130
Pharmaceutical Form	45	58	43	47
Route	42	52	42	29
Medicament	37	37	37	34
Food	31	24	22	24
Therapeutic Action	20	7	21	47
Total	2,241	2,388	2,224	1,967

GS: gold standard; A1: Annotator 1, student;
A2: Annotator 2, student; A3: Annotator 3, nurse.

Table 8: Example Use Case results: Named Entity Classification

NE Training	**NE Test**	**Pr**	**Re**	**F1**
Disease	AdverseEffect	**78,8**	**70,1**	**74,2**

Note: Pr: Precision; Re: Recall; F1: $F - measure_{\beta=1}$; NE Training: entity from DrugSemantics corpus as training; and NE Test: entity from SpanishADR corpus for testing.

5. Discussion

This section analyses our obtained results and those are compared with previous research. This section is divided in five parts. First, Section 5.1 examines the methodology employed for the annotation of DrugSemantics corpus. Next, Section 5.2 analyses the methodology applied to asses the quality of DrugSemantics gold standard. Latter, Section 5.3 emphasizes DrugSemantics reliability results. This is complemented by a comparison between our reliability results and the ones from other relevant research (Section 5.4). Finally, DrugSemantics precision is highlighted (Section 5.5).

5.1. DrugSemantics annotation methodology

From a critical perspective, several factors may have affected the DrugSemantics gold standard due to the annotation methodology carried out. The first one is related to the knowledge annotators should have on the working domain. A high quality standard annotation is typically expert-driven [20–27, 37] to ensure consistently great quality corpora. However, researchers have shown that non-expert annotators can be as good as experts with appropriate training [72, 73]. In our case, annotators had a priori an adequate level of pharmacotherapeutic knowledge (1 RN and 2 students in their final year from the Degree in Nursing). However, this knowledge varies over time and it seems more consolidated during university than during professional practice, as our results shown (cf. Tables 5-6). In Spain, once the title of RN is obtained, there is no certification to update knowledge periodically, which may be affected by the type of employment carried out. Such limitation was not considered during the annotators selection process and might explain the observed discrepancies. It should be remember that annotator RN obtains the worst results in comparison with students. Specifically, this influence are evinced by our precision analysis. For instance, Therapeutic Action entity is the one that has more detected errors, either this EN is ignored or it is confused with other substances (Drug or Chemical Composition). However, a combination of different perspectives (i.e. two students and one professional) has been considered adequate to ensure that the annotation represents better the entities under investigation.

The second one is related to the number of annotators involved in the process, which varies depending on the available resources: 2 [20, 24, 37], 3-5 [21, 23], $>$ 6 annotators [22]. Although there is no standard number of annotators, there is a consensus to annotate documents by at least two annotators independently. During training, disagreements are usually solved with all annotators present in order to achieve an agreement and update the guidelines accordingly. In order to build the gold standard, differences are traditionally resolved by an experienced annotator (i.e. a judge) or by consensus between annotators. Besides, the number of annotators not only influences the effort required to assure corpus quality, but also the resulting size (i.e. more pairs produce larger corpora). In light of these, two alternative methodologies have emerged lately trying to overcome these limitations: crowdsourcing and translation. Both are cheaper and faster than traditional annotation efforts, but these are not exempt from certain drawbacks.

Crowdsourcing[18] is a collaborative approach for obtaining larger annotated corpora that allows annotators to work independently no matter the distance. Notwithstanding certain researchers have used the crowd to annotate in the healtcare field, such as [74, 75], “a remaining challenge is that the cost to define a single annotation crowdsourcing project can outweigh the benefits” [76]. Furthermore, “there are legitimate concerns that could be raised regarding its use for medical research” [74].

Another option to obtain an annotated corpus would consists in translating automatically existing pharmacotherapuetic corpora. Although it would be possible, the outcome could be inaccurate for various reasons, which are typical of this challenging domain. Most drugs have names that could be translated using the ATC classifica-

[18] For instance, paid-for marketplaces such as Amazon Mechanical Turk or CrowdFlower.

tion system. However, all countries do not accept specific substances in their healthcare systems. Besides, a pharmaceutical company may commercialize a medicine using different names or using a other set of active substances depending on the target country. Hence, it is not possible to transfer all substances between healthcare systems. More problems of machine translation in the medical domain can be found in [77].

Bearing in mind all these standard practices, DrugSemantics annotation methodology was designed by using high level annotation standards with certain adaptations to satisfy the needs and peculiarities of our research framework. Instead of solving training disagreements with all annotators, DrugSemantics provided a tailor made session to solve doubts guided by one of the authors (MT R-F). The reason was to avoid annotators decision' bias (i.e. annotators could influence each other in a joint session), which in turn would have affected precision results. Regarding DrugSemantics gold standard building, two strategies were applied. First, common consensus through majority voting was utilized. Second, disagreements, instead of being discarded, were resolved. To that end, the judge annotator was a semi-automatic process driven by a named entity recognition and classification system based on knowledge [61, 62]. In this manner, DrugSemantics annotation methodology ensured manual annotations of the highest quality, which in turn drove our rigorous gold standard construction process.

5.2. DrugSemantics quality evaluation methodology

Similarly, several aspects may have an effect upon DrugSemantics evaluation. Regarding reliability, attributes of each entity, as well as confidence of the annotator, were not taken into account when computing agreement. Properties were included in DrugSemantics scheme to give hints to annotators; this made the entity identification more effective. In fact, fill the attributes of each entity is a different information extraction task (e.g. association of a given drug to its strength [27]), which is beyond the scope of this work.

Concerning precision, it is important to emphasize that our pipeline is general enough to be applied in other annotated corpora with NEs. To that end, MaNER [61, 62] can be changed for a dictionary-based NE recognition and classification tool. Besides, knowledge resources can also be changed or updated provided that these are carefully selected due to their informative and reliable value. As previously stated, since there is a lack of Spanish resources in this domain, a careful analysis was made. For instance, BEDCA [65], a Spanish database containing terms related to food that do not follow the terminology of SPCs, or Food Additives [64], is a resource that contains a small number of excipients. Both resources were the only ones we found related to these entities, but they needed an update process to represent more accurately language of SPCs. Discarding these entities from precision analysis was not an option owing to their importance for detecting interactions and allergic reactions. On the contrary, larger knowledge sources were identified: (i) SNOMED is an ontology that is the de facto standard for semantic interoperability available for Spanish; and (ii) ATC and Nomenclator digitalis are the common reference source in Spain. Furthermore, our precision methodology could be applied to all NLP tasks that can be resolved with dictionary-based approaches. It is important that those dictionaries are populated with informative and reliable knowledge resources, provided that these include terminology commonly used in the textual genre selected for the resulting corpus [63].

5.3. DrugSemantics Reliability

As we have seen in Section 4.1, the DrugSemantics corpus has evaluated its reliability in terms of agreement as lenient F-measure. Overall agreement, both micro (79.33 95%CI [78.35-80.31]) and macro (72.33 95%CI [71.25-73.41]), are substantial, since their lower limit CI is greater than 60. Besides, our hypothesis is confirmed for all pairs at an overall level (see Table 5). As a result, the manual annotation was reliable. Besides, differences between these averages are due to the unbalanced NEs distribution (they do not have a similar number of entities).

Looking at the results per NE, Medicament, Excipient, Unit Of Measurement, Drug and Pharmaceutical Form had almost perfect agreement ($F \in [80, 100]$ - see Table 5). That leads us to interpret that these entities are easier to annotate. Despite the fact that Disease is the most common entity and the one with a higher variability in terms of occurrences (A1=887, A3=567), it was harder to annotate since it only achieved substantial agreement ($F \in (80, 60]$). This may be due to different disease designations according to the role of a medicament and its drugs (e.g. Therapeutic Indication, Contraindication, Overdosage, Side Effect). Although they received clear instructions in this regard, make that difference was difficult for annotators. Food and Route entities also have substantial agreement in absolute terms. However, not all annotator pairs confirm our hypothesis for these two entities: A2-A3 pair (Food and Route), and A1-A3 pair (Route). Other reasons behind these variations may be: (i) Disease and Food entities are usually formed by several tokens, for instance, "*insuficiencia renal moderada*" (moderate kidney failure), "*colorantes azoicos*" (azo-dyes); and (ii) Route entity often corresponds to one token but is preceded by trigger words, such as "*vía*" (via), that may mislead to incorrect boundaries detection. This fact, in addition to the inclusion of extra spaces and punctuation symbols in the annotations, motivated our decision of using a lenient criterion in order to count them at least as partial-matches.

5.4. Reliability Comparison with the State of the Art

Comparing our reliability results with other Spanish research [20, 37] is not free of certain limitations. First, Ixa-MedGS [20] used a different reliability metric (i.e. Inter Annotator Agreement, a.k.a. IAA). IAA is computed in terms of matches and non-matches, which in turn refer to other variables commonly employed to compute F-measure. On the one hand, matches is also known as True Positives or correct[19]. On the other hand, non-matches refers to errors denoted as False Positive and False Negative or Type I and Type II. Therefore, although the metrics have different names, IAA and F-measure are calculated using the same figures and, as a result, both are equivalent, as shown in [23]. Besides, our work, as well as [20], chose a lenient criterion but nothing is said in SpanishADR corpus [37]. Hence we could assume [37] chose a strict criterion so as to not take into account partial matches.

Second, despite the fact that these corpora gathered Spanish health texts, each effort choose distinct types of documents (Forums comments [37], clinical documents [20] and SPCs) that do not pose identical challenges. For instance, patients tend to use shorter terms and informal language. While professionals often employ abbreviations, short and agrammatical sentences to fill patient health records. Whereas specific terms, formal language and long sentences are utilized to avoid ambiguity in SPC for future references.

Third, entity types across schema are not always a perfect match because there is no consensus within medical NLP community concerning which elements must be considered. The well-known ones barely scratch the surface of NEs that would be useful for all text mining purposes [18]. Finally, only three out of ten NEs from DrugSemantics scheme are present in other Spanish corpora, namely: Medicament, Drug and Disease.

Consequently, the DrugSemantics corpus complements the efforts to build NLP resources for the Spanish pharmacotherapeutic domain. Therefore, a reliability comparison between our corpus and Spanish corpora presented in Section 2 is provided in Table 9. Such comparison is focused on two groups of entities that represent both similar concepts.

[19] Please note that partial matches can also be considered.

Table 9: Annotated Corpora Reliability Comparison, ordered by DrugSemantics agreement. Only shared entities for Spanish efforts.

Named Entity	**DrugSemantics**	**Ixa-MedGS**	**SpanishADR**
Medicament + Drug	F=90.83*[89.53-92.13]$	IIA=92.12	F=89
Disease	F=72[70.13-73.87]$	IIA=89.81	F=59

Note: (*) Macro-averaged F-measure between Medicament (F=95.33%) and Drug (F=86.33%); ($) 95% Confidence Interval

Table 10: Annotated Corpora Reliability Comparison, ordered by DrugSemantics agreement. Only shared entities for English efforts.

Named Entity	**DrugSemantics**	**i2b2**	**DrugDDI**	**CLEF**
Medicament	F=95.33[91.35-99.31]$	-	K=88.53	-
Drug	F=86.33[84.73-87.93]$	-	K=84.67	-
Medicament + Drug	F=90.83[89.53-92.13]$	F>90	-	IIA=85
Unit of Measurement	F=89.33[87.80-90.86]$	F>61.3	-	-
Disease	F=72[70.13-73.87]$	$70.5 \geq F \leq 75.3$	-	IIA=84
Route	F=68[59.76-76.24]$	F>90%	-	-
Therapeutic Action	F=10[3.21-16.79]$	-	K=82.99	-

Note: ($) 95% Confidence Interval

First, Medicament and Drug entities are combined in one single type. For this reason, we calculate a macro-averaged F-measure between our two values (see Table 9 Medicament + Drug). All corpora reported almost perfect agreement for this group with minor variations. In Ixa-MedGS case, their results are slightly higher than ours; whereas SpanishADR F1 is a little less than ours. In our opinion, these minimum differences are related with a combination of two factors. On the one hand, each type of genre has different problems, as noted earlier. On the other hand, the effort required for the annotation task is different. Ixa-MedGS and SpanishADR have less elements in their annotation schema than DrugSemantics. Furthermore, Ixa-MedGS pre-annotate their documents, thus the annotation task is easier than SpanishADR and DrugSemantics annotation.

Second, Disease is present in all efforts but the differences are more evident. For Ixa-MedGS corpus, Oronoz et. al [20] reported an agreement higher than ours. These differences may be due to the fact that their task was less complex, since their annotators had to revise automatically annotated entities, remove incorrect ones and add missed ones. SpanishADR includes a subset of our Disease (i.e. adverse drug reactions) and exhibit moderate agreement (F = 59%). Discrepancies in SpanishADR corpus, could be due to a greater variability and richness in patients comments than SPCs. For instance, patients tend to use shorter terms and informal language, as said earlier, and they could write: "*infarto*" (heart attack) but also "*infart*" or "*1nfart*" or "*nfrt*", while SPCs would employ formal texts and longer words to avoid ambiguity, such as "*infarto de miocardio*" (myocardial infarction).

We can claim that our reliability is in-line with other Spanish corpora and sometimes better, despite that: (i) our annotation scheme includes a higher number of entities which increases obstacles during manual annotation (ratters deal simultaneously with ten NEs across 5 documents with 780 sentences and 22,679 tokens); and (ii) our annotators were free to label any textual fragment on a SPC, while Ixa-MedGS used pre-labelled corpus before annotators intervene.

As regards to English corpora, the reliability among efforts is not directly comparable due to substantial differences among these languages. Still, our agreement results are comparable with what has been shown for other English corpora, as can be seen in Table 10. DrugSemantics annotation scheme is more similar to English than Spanish schema. As in the Spanish case, the documents gathered to create English corpora (EHR, scientific abstracts, DrugBank texts) are different to the one this paper employed (SPCs), even though DrugBank texts and SPCs are the most similar ones. Our agreement results are analogous to other English efforts, since our reliability is almost perfect and substantial for 6 NEs; despite the weak agreement for our Therapeutic Action.

Finally, it should be noted that although DrugSemantics may seem limited in terms of size (only 5 SPCs), our resource present a high level of richness in terms of linguistic and semantic elements. For instance, Ixa-MedGS is the corpus with the highest number of sentences (almost 73 - see Table 1), tokens (555.11 - see Table 1) and entities (around 53 ENs - see Table 1) on average per document. However, DrugSemantics increases all these figures on average per document: sentences raises 2 times (156 - see Table 1), tokens raises 8 times (4,535.8 - see Table 1) and NEs increases 7 times (more than 400 entities - see Table 1). Proportionally, DrugSemantics is bigger because our documents are longer and semantically richer on average than similar corpora in this domain.

5.5. *Precision*

Before computing the precision figures, the lists of not matched annotations for each entity and annotator were analysed manually in depth. As a result, Table 11 presents all types of conflicts identified initially in all our dictionaries before computing Precision. In view of this analysis, precision relies heavily on knowledge resources (i.e. dictionaries), as expected. The good results are due to the manual identification of dictionary issues, which allowed to ignore false errors. These dependencies were overcome by adding the required variations (abbreviations, etc.). The identified types of conflicts in all our dictionaries are:

Table 11: Dictionary gaps in relation to Annotation Precision by entity. Ordered increasingly by number of gaps.

Named Entity	**Gaps**
Medicament	η
Food	$\beta\,\gamma\,\zeta$
Excipient	$\beta\,\gamma\,\zeta\,\eta$
Route	$\alpha\,\gamma\,\zeta\,\eta$
Therapeutic Action	$\gamma\,\epsilon\,\zeta\,\eta$
Unit of Measurement	$\alpha\,\gamma\,\zeta\,\eta$
Drug	$\alpha\,\beta\,\epsilon\,\zeta\,\eta$
Pharmaceutical Form	$\beta\,\gamma\,\epsilon\,\zeta\,\eta$
Chemical Composition	$\alpha\,\beta\,\gamma\,\epsilon\,\zeta\,\eta$

Note: (α): Abbreviation; (β) White-spaces and hyphens; (γ) Specific entries; (ϵ) Lexical variations; (ζ) Lack of synonyms; (η) Human failure and partial match.

(α) *Abbreviations are excluded*: During examination, it was noticed that Drug, Route, Unit of Measurement and Chemical Composition dictionaries, only contain full names. In the case of Unit of Measurement, "*gramos*" (grams) is an entry, but "*g.*" (g.), its acronym, is not. A potential solution would be to include these acronyms in their relevant dictionaries.

(β) *Special characters are considered*: On the one hand, white spaces and hyphens are considered for some entities that can be named after codes. Such codes are a single term in our dictionaries. However, Drugs and Excipients annotations often include hyphens or white spaces that do no produce an exact match. For example, in the case of Excipient NE, "E122" is included in its dictionary but "E-122" is not. On the other hand, break-line character as well as adjectives or articles are considered for names of Drugs, Chemical Compositions, Pharmaceutical Forms, Excipient and Food. For instance, in the case of Chemical Composition entity, "*Inhibidores de Proteasa*" (Protease Inhibitors) is an entry in its dictionary but "*Inhibidores de la Proteasa*" (Inhibitors of the Protease) is not. Modify the matching rule to ignore these characters, could be a possible solution.

(γ) *Lexicons contain terms too specific*: One of the most common problem among different NEs is related to differences in granularity. This happens to Drug, Unit of Measurement, Route, Pharmaceutical Form, Therapeutic Action, Excipient, Chemical Composition and Food. This issue is more obvious for Food, since no matches could be produced when comparing manual annotations to its original dictionary. For instance, "*Zumo de pomelo, envasado*" (grapefruit juice, packaged) belongs to the dictionary but it must be separated in two: "*Zumo de pomelo*" (grapefruit juice) and "*Zumo envasado*" (packaged juice). Hence, the terms within our dictionaries should be generalized to obtain a coarser granularity. This generalization is not trivial and needs to be carefully planned to ensure dictionaries reliability. For example, "*hipoglucemiantes*" (hypoglycemics) is a Therapeutic Action, but this word is included along several entries in Chemical Composition dictionary. For instance, "*Combinaciones de drogas hipoglucemiantes orales*" (Combinations of oral blood glucose lowering drugs) can not be divided to provide new terms in Chemical Composition dictionary.

(ζ) *Lack of synonyms*: Our dictionaries contain a large amount of entries, however all their synonyms are not included or most of them are in only one entry. This is the case of Drug, Therapeutic Action, Excipient and Food. For example, Excipient dictionary has as entry "*azorrubina, carmoisina*" (azorubine, carmoisine), but it must be separated in two: "*azorubina*" (azorubine) and "*carmoisina*" (carmoisine). Hence, entries with several synonyms, need to be splitted to be a match and provide a coarser granularity.

(ϵ) *Lexical variations are considered*: Some disagreements were due to misspellings, but also to gender and number variations of the following entities: Drug, Therapeutic Action, Pharmaceutical Form and Chemical Composition. For instance, Therapeutic Action dictionary has as entry "*Antiácidos*" (antacids) in plural, but in SPCs its singular form also appears "*antiácido*" (antacid). Thus, a possible solution could be to lemmatize entries in dictionaries, this way the matching rule could be done at lemma level, instead of using exact pattern matching.

Despite these dictionaries gaps, we can affirm that manual annotation of DrugSemantics is a success globally: both micro ($P = 94.65\%$ [95%CI: 94.11-95.19]) and macro-average ($P = 90.05\%$ [95%CI: 89.33-90.77]) confirm our hypothesis. Furthermore, it is important to highlight that agreement measures are the only ones reported for other annotated corpora. However, this paper proposes a methodology to provide agreement and precision results, considering that computation of both measures strengthen the quality of corpora.

On the one hand, precision at entity level confirms reliability of DrugSemantic corpus. That is, in the vast majority of entities obtained outstanding results for both indicators regardless annotator or pair, specifically: Medicament, Excipient, Unit of Measurement, Drug and Pharmaceutical Form.

On the other hand, differences between the two suggest that certain entity mentions were wrongly missed by one annotator of the pair. For example, Therapeutic Action obtained the lowest agreement (globally and by pair). However, A2 got an excellent precision for this entity (100%), reason why we include these high quality annotations in DrugSemantics. In these cases, therefore, the statistical hypothesis tests notably assist in the gold-standard construction process by means of restricting which annotations have truly enough quality to be included in the final set.

As a result, DrugSemantics has a high quality with the fewest possible errors so as to not bias a NE recognition and classification algorithm. That is also confirmed by the results presented in the use case (see Table 8).

6. Conclusions

This paper has presented the DrugSemantics corpus, a collection of Spanish SPC. These documents have been manually annotated to include significant NEs for the pharmacotherapeutic process, which are specified in DrugSemantics annotation scheme. To the best of our knowledge, no corpus of annotated SPCs written in Spanish has been created to date. Furthermore, no corpus in this domain has been annotated with an annotation scheme as complete as DrugSemantics to date.

Besides, the quality of this corpus has been assessed by means of measuring annotation reliability (overall agreement, F=79.33% [95%CI: 78.16-80.50]), as well as precision (overall precision, $P = 94.65\%$ [95%CI: 94.11-95.19]). For the latter, a semi-automatic methodology is proposed, which is general enough to be applied on other NLP datasets. To that end, a NE recognition system, which is dictionary-based, has been developed for 9 out of 10 entities.

Given this substantial agreement, this almost perfect precision and the statistics hypothesis testing, a high quality gold-standard has been created. The resulting corpus contains more than 2,000 named entities spread in 5 SPCs, 780 sentences and 22,659 tokens. Our gold standard is publicly available for research purposes [71].

A successful example of how to use the DrugSemantics corpus has been shown, in which a NEC system classifies Disease entity. To that end, the DrugSemantics gold standard has been employed to train this system. In order to prove whether it can serve for training purposes, it has been evaluated on a corpus of different genre (formal versus informal). Therefore, we foresee that DrugSemantic will be useful for the development and testing of Spanish NE recognition tools in the pharmacotherapeutic domain.

As future work, we will study an extrinsic evaluation of a NERC system trained on DrugSemantics and its contribution to other NLP systems (such as question answering or other information extraction tools). Besides, we will consider the extension of DrugSemantics to other NLP tasks (e.g. relation extraction or negation). Finally, we plan to enhance our dictionaries driven by the precision analysis.

Conflict of interest

The authors report that there are no conflicts of interest.

Acknowledgements

Funding: This work was supported by the Spanish Government (grants number TIN2015-65100-R; TIN2015-65136-C02-2-R), the Generalitat Valenciana (grant number PROMETEOII/2014/001) and the BBVA Foundation (award number FUNDACIONBBVA2-16PREMIOI).

The authors would like to thank the team of annotators for their time and work during the annotation of this corpus, as well as Maria Rey Gilabert for her pharmacotherapeutic knowledge. Thanks to the anonymous reviewers for their suggestions which have helped to improve this work.

References

[1] C. Friedman, T. C. Rindflesch, M. Corn, Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine, J. Biomed. Inform. 46 (5) (2013) 765–773. doi:10.1016/j.jbi.2013.06.004.

[2] PubMed.
URL `http://www.ncbi.nlm.nih.gov/pubmed/(accessed:10.01.2017)`

[3] A. González-González, E. Escortell Mayor, T. Hernández Fernández, J. Sánchez Mateos, T. Sanz Cuesta, R. Riesgo Fuertes, Necesidades de información de los medicos de atención

primaria: análisis de preguntas y su resolución, Aten. Prim. 35 (8) (2005) 419–431.

[4] Scopus.
URL http://www.scopus.com/(accessed:10.01.2017)

[5] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, W. Hersh, State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track, Inf. Retr. 19 (1-2) (2016) 113–148. doi:10.1007/s10791-015-9259-x.

[6] G. Del Fiol, T. Workman, P. Gorman, Clinical questions raised by clinicians at the point of care: A systematic review, JAMA Intern. Med. 174 (5) (2014) 710–718. doi:10.1001/jamainternmed.2014.368.

[7] A. I. González-González, J. Sánchez Mateos, T. Sanz Cuesta, R. Riesgo Fuertes, E. Escortell Mayor, T. Hernández Fernández, Estudio de las necesidadesde información generadas por los médicos de atención primaria (proyecto ENIGMA)*, Aten. prim. 38 (4) (2006) 219–224.

[8] G. Del Fiol, A. I. Weber, C. P. Brunker, C. R. Weir, Clinical questions raised by providers in the care of older adults: a prospective observational study, BMJ Open 4 (7) (2014). doi:10.1136/bmjopen-2014-005315.

[9] S. Doan, M. Conway, T. M. Phuong, L. Ohno-Machado, Natural language processing in biomedicine: a unified system architecture overview, Methods in Mol. Biol. (Clifton, N.J.) 1168 (2014) 275–294. doi:10.1007/978-1-4939-0847-9_16.

[10] C. D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008. doi:10.1017/CBO9780511809071.

[11] R. Feldman, J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge University Press, New York, 2007. doi:10.1017/CBO9780511546914.

[12] M. Ben-dov, R. Feldman, Text Mining and Information Extraction, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer US, Boston, MA, 2010, Ch. 42, pp. 809–835. doi:10.1007/978-0-387-09823-4_42.

[13] M. A. Hearst, Untangling text data mining, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, 1999, pp. 3–10. doi:10.3115/1034678.1034679.

[14] A. Singhal, Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (2001) 35–43.

[15] Google.
URL http://www.google.com(accessed:10.01.2017)

[16] W. J. Wilbur, A. Rzhetsky, H. Shatkay, New directions in biomedical text annotation: definitions, guidelines and corpus construction, BMC Bioinform. 7 (2006) 356. doi:10.1186/1471-2105-7-356.

[17] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, O. Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions., JAMIA 18 (5) (2011) 540–543. doi:10.1136/amiajnl-2011-000465.

[18] K. Bretonnel Cohen, D. Demner-Fushman, Biomedical Natural Language Processing, Vol. 11 of Natural Language Processing, John Benjamins Publishing Company, Amsterdam, 2014. doi:10.1075/nlp.11.

[19] J. W. Ely, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. Chambliss, E. R. Evans, Analysis of questions asked by family doctors regarding patient care, Br. Med. J. 319 (1999) 358–361.

[20] M. Oronoz, K. Gojenola, A. Pérez, A. D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions, J. Biomed. Inform. 56 (2015) 318–332. doi:10.1016/j.jbi.2015.06.016.

[21] E. M. van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. a. Kors, L. I. Furlong, The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships, J. Biomed. Inform. 45 (5) (2012) 879–884. doi:10.1016/j.jbi.2012.04.004.

[22] R. I. Doğan Doan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10. doi:10.1016/j.jbi.2013.12.006.

[23] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, A. Setzer, Building a semantically annotated corpus of clinical texts, J. Biomed. Inform. 42 (5) (2009) 950–966. doi:10.1016/j.jbi.2008.12.013.

[24] M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, T. Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920. doi:10.1016/j.jbi.2013.07.011.

[25] O. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, JAMIA 17 (5) (2010) 514–518. doi:10.1136/jamia.2010.003947.

[26] O. Uzuner, I. Solti, F. Xia, E. Cadag, Community annotation experiment for ground truth generation for the i2b2 medication challenge, JAMIA 17 (5) (2010) 519–523. doi:10.1136/jamia.2010.004200.

[27] Ö. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, JAMIA 18 (5) (2011) 552–556. doi:10.1136/amiajnl-2011-000203.

[28] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J. Meinzen-Derr, M. Kaiser, L. Stoutenborough, Q. Li, I. Solti, Evaluating the

impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements, JAMIA 21 (3) (2014) 406–413. doi:10.1136/amiajnl-2013-001837.

[29] Q. Li, L. Deleger, T. Lingren, H. Zhai, M. Kaiser, L. Stoutenborough, A. G. Jegga, K. B. Cohen, I. Solti, Mining FDA drug labels for medical conditions, BMC Med. Inform. Decis. Making 13 (1) (2013). doi:10.1186/1472-6947-13-53.

[30] S. Pradhan, N. Elhadad, W. W. Chapman, S. Manandhar, G. Savova, SemEval-2014 Task 7: Analysis of Clinical Text, Proceedings of the 8th International Workshop on Semantic Evaluation (2014) 54–62.

[31] B. Rosario, M. Hearst, Classifying semantic relations in bioscience texts, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (2004) 430–437doi:10.3115/1218955.1219010.

[32] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, J. Biomed. Inform. 45 (5) (2012) 885–892. doi:10.1016/j.jbi.2012.04.008.

[33] A. Patki, A. Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, Mining Adverse Drug Reaction Signals form Social Media: Going Beyond Extraction, in: Proceedings of BioLinkSig, 2014, pp. 9–19.

[34] K. W. Fung, C. S. Jao, D. Demner-Fushman, Extracting drug indication information from structured product labels using natural language processing, JAMIA 20 (3) (2013) 482–488. doi:10.1136/amiajnl-2012-001291.

[35] N. Elhadad, S. Pradhan, S. L. Gorman, S. Manandhar, W. W. Chapman, G. Savova, SemEval-2015 Task 14 : Analysis of Clinical Text, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 303–310.

[36] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (Database issue) (2006) D668–D672. doi:10.1093/nar/gkj067.

[37] I. Segura-Bedmar, R. Revert, P. Martínez, Detecting drugs and adverse events from Spanish health social media streams, in: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis, 2014, pp. 106–115.

[38] ForumClinic.
URL `http://www.forumclinic.org/(accessed:10.01.2017)`

[39] B. R. South, D. Mowery, Y. Suo, J. Leng, Ó. Ferrández, S. M. Meystre, W. W. Chapman, Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text, J. Biomed. Inform. 50 (2014) 162–172. doi:10.1016/j.jbi.2014.05.002.

[40] Real Decreto 1345/2007, de 11 de octubre, por el que se regula el procedimiento de autorización, registro y condiciones de dispensación de los medicamentos de uso humano fabricados industrialmente (2007).

[41] G. Martín, F. J. Morales-Olivas, Nuevos lenguajes informáticos en la difusion de información sobre medicamentos, Med. Clin. (Barc) 128 (13) (2007) 498–503. doi:10.1157/13100938.

[42] S. Rubrichi, S. Quaglini, Summary of Product Characteristics content extraction for a safe drugs usage, J. Biomed. Inform. 45 (2) (2012) 231–239. doi:10.1016/j.jbi.2011.10.012.

[43] Ley 14/1986, de 25 de abril, General de Sanidad. (1986).

[44] Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. (1999).

[45] Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica (2002).

[46] R. Nalichowski, D. Keogh, H. C. Chueh, S. N. Murphy, Calculating the benefits of a Research Patient Data Repository, in: AMIA Annual Symposium proceedings, 2006, p. 1044.

[47] CIMA (Centro de Información online de Medicamentos de la AEMPS).
URL `http://www.aemps.gob.es/cima/(accessed:10.01.2017)`

[48] M. T. Romá-Ferri, OntoFIS: tecnología ontológica en el dominio farmacoterapéutico, Ph.D. thesis, Universidad de Alicante (2009).

[49] W. Shrank, J. Avorn, C. Rolon, P. Shekelle, Effect of content and format of prescription drug labels on readability, understanding, and medication use: a systematic review, Ann. Pharmacother. 41 (5) (2007) 783–801. doi:10.1345/aph.1H582.

[50] Real Decreto 1348/2003, de 31 de octubre, por el que se adapta la clasificación anatómica de medicamentos al sistema de clasificación ATC. (2003).

[51] WHO Collaborating Center for Drug Statistics Methodology, ATC/DDD.
URL `http://www.whocc.no/atc/(accessed:10.01.2017)`

[52] U. G. P. Office, Food and Drugs, 21CFR3.2 (2016).
URL `https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=3.2(accessed:10.01.2017)`

[53] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, E. Al, Developing Language Processing Components with GATE Version 7 (a User Guide), 2012.

[54] WHO Collaborating Center for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment 2016, Oslo, 2016.

URL http://www.whocc.no/filearchive/publications/2016_guidelines_web.pdf(accessed:10.01.2017)

[55] The Cochrane Collaboration, Assessing risk of bias in included studies, in: Julian PT Higgins and Sally Green (Ed.), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011], 2011, Ch. 8.
URL http://handbook.cochrane.org/chapter_8/8_assessing_risk_of_bias_in_included_studies.htm

[56] R. Artstein, M. Poesio, Inter-Coder Agreement for Computational Linguistics, Comput. Linguist. 34 (4) (2008) 555–596. doi:10.1162/coli.07-034-R2.

[57] G. Hripcsak, A. S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, JAMIA 12 (3) (2005) 296–298. doi:10.1197/jamia.M1733.

[58] F. D. B. Navarro Colorado, Metodología, construcción y explotación de corpus anotados semántica y anafóricamente, Ph.D. thesis, University of Alicante (2007).

[59] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in Chinese clinical text, JAMIA 21 (5) (2014) 808–814. doi:10.1136/amiajnl-2013-002381.

[60] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data., Biom. 33 (1) (1977) 159–174. doi:10.2307/2529310.

[61] I. Moreno, P. Moreda, M. Romá-Ferri, Reconocimiento de entidades nombradas en dominios restringidos, in: Actas del III Workshop en Tecnologías de la Informática, Alicante, Spain, 2012, pp. 41–57.

[62] I. Moreno, P. Moreda, M. Romá-Ferri, MaNER: a MedicAl Named Entity Recogniser for Spanish, in: C. Biemann, S. Handschuh, A. Freitas, F. Meziane, E. Métais (Eds.), 20th International Conference on Applications of Natural Language to Information Systems, Springer International Publishing Switzerland, Passau, 2015, pp. 418–423. doi:10.1007/978-3-319-19581-0_40.

[63] I. Moreno, P. Moreda, M. T. Romá-Ferri, Estudio de fiabilidad y viabilidad de la Web 2.0 y la Web semántica para enriquecer lexicones en el dominio farmacológico, Proces. Leng. Nat. 55 (2015) 65–72.

[64] J. García, EURO-E (2009).
URL http://histolii.ugr.es/euroe/e_index.html(accessed:10.01.2017)

[65] BEDCA (Base de Datos Espaola de Composición de Alimentos).
URL http://www.bedca.net/(accessed:10.01.2017)

[66] Dirección General de Farmacia y Productos Sanitarios. MSSSI. Spanish Government, DIGITALIS (Nomenclátor Digitialis) (2011).
URL http://www.msc.es/profesionales/nomenclator.do(accessed:10.01.2017)

[67] IHTSDO, SNOMED Clinical Terms User Guide, 2014th Edition, 2010.
URL http://www.snomed.org/ug.pdf(accessed:10.01.2017)

[68] Moreno, Isabel and Romá-Ferri, M.T. and Moreda, Paloma, Named Entity Classification based on Profiles: a Domain Independent Approach, in: 22th International Conference on Applications of Natural Language to Information Systems, Liége, 2017 - in press.

[69] I. Moreno, M. T. Romá-Ferri, P. Moreda, Propuesta de un sistema de clasificación de entidades basado en perfiles e independiente del dominio, Proces. Leng. Nat. 59 (2017) – Submitted.

[70] I. Moreno, M. T. Romá-Ferri, P. Moreda, Proposal for a Language Independent Named Entity Classification System based on Profiles, in: Proceedings of the 15th International Conference Text, Speech and Dialogue, 2017 - Submitted.

[71] I. Moreno, E. Boldrini, P. Moreda, M. T. Romá-Ferri, DrugSemantis Gold Standard: an annotated corpus of Spanish Summaries of Product Characteristics annotated for pharmacotherapeutic named entity recognition, https://data.mendeley.com/datasets/fwc7jrc5jr/draft?a=48099c59-ac0d-4366-875e-7ec38b8534b8 (2017). doi:doi:10.17632/fwc7jrc5jr.1.

[72] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, Building gold standard corpora for medical natural language processing tasks., in: Proceedings of AMIA Annual Symposium, Vol. 2012, 2012, pp. 144–153.

[73] W. W. Chapman, J. N. Dowling, G. Hripcsak, Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports, Int. J. Med. Inform. 77 (2) (2008) 107–113. doi:10.1016/j.ijmedinf.2007.01.002.

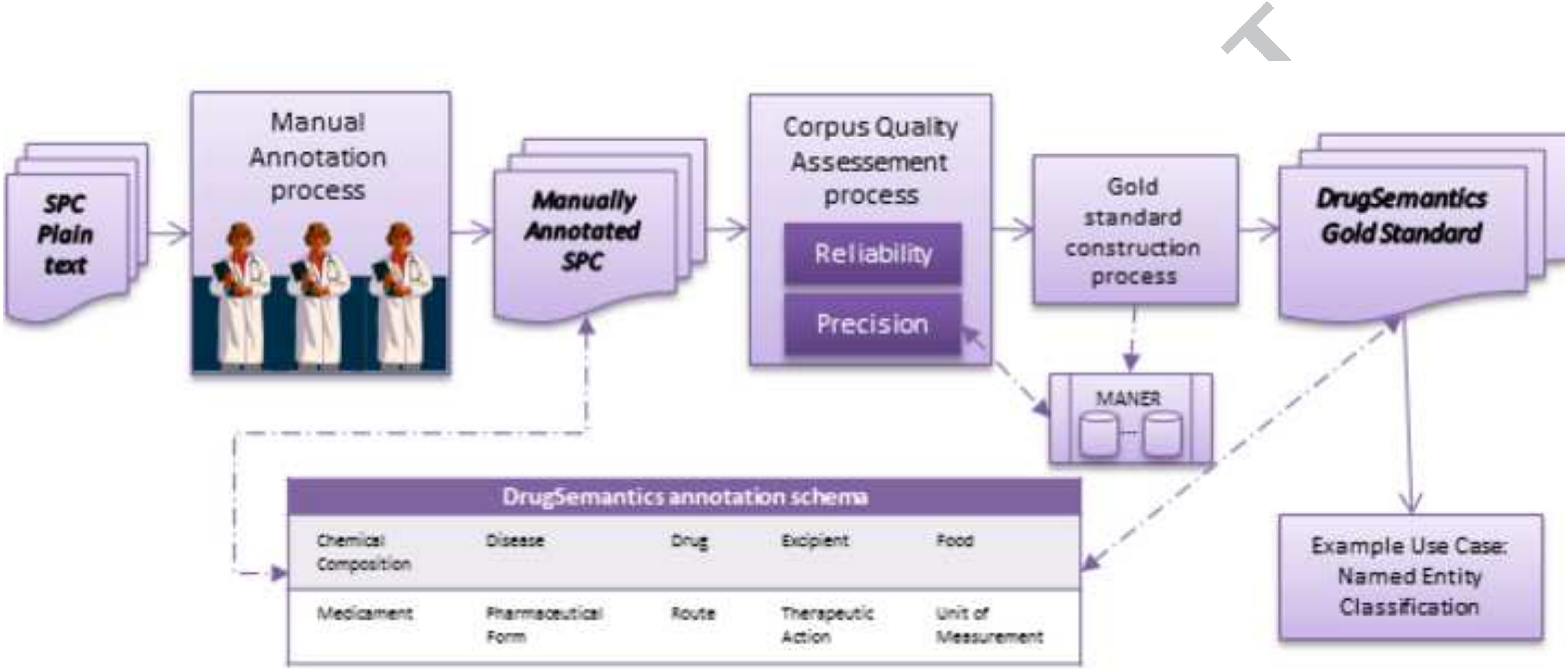
[74] A. Cocos, T. Qian, C. Callison-Burch, A. J. Masino, Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation, J. Biomed. Inform. 69 (2017) 86–92. doi:10.1016/j.jbi.2017.04.003.

[75] Y. Lou, S. W. Tu, C. Nyulas, T. Tudorache, R. J. Chalmers, M. A. Musen, Use of ontology structure and bayesian models to aid the crowdsourcing of icd-11 sanctioning rules, J. Biomed. Inform. 68 (2017) 20 – 34. doi:10.1016/j.jbi.2017.02.004.

[76] M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014, pp. 859–866.

[77] M. R. Costa-Jussà, M. Farrús, J. S. Pons, Machine Translation in Medicine A quality analysis of statistical machine translation in the medical domain, in: Advanced Research in Scientific Areas, 2012, pp. 1995–1998.

Graphical Abstract


SPC Plain text
Manual Annotation process
Manually Annotated SPC
Corpus Quality Assessement process
Reliability
Precision
Gold standard construction process
DrugSemantics Gold Standard
MANER
Example Use Case: Named Entity Classification

DrugSemantics annotation schema

Chemical Composition	Disease	Drug	Excipient	Food
Medicament	Pharmaceutical Form	Route	Therapeutic Action	Unit of Measurement

Highlights

+ DrugSemantics Corpus is a set of Spanish Summary of Product Characteristics
+ 10 pharmacotherapeutic named entity types manually annotated by 3 annotators
+ Corpus Quality: substantial reliability(79.33%), almost perfect precision(94.65%)
+ Quality confirmed through statistical hypothesis testing using Z-test in both cases
+ Precison ensured via semiautomatic method that enhances MaNER, dictionary-based NER