Procedia - Social and Behavioral Sciences 95 (2013) 25 October, pp. 211–219

DOI: 10.1016/j.sbspro.2013.10.641

5th International Conference on Corpus Linguistics (CILC2013)

Annotation Issues in Pharmacological Texts

María Herrero-Zazo*, Isabel Segura-Bedmar, Paloma Martínez

University Carlos III of Madrid, Av. Universidad, 30, Leganés 28911, Spain

Abstract

Natural language processing of pharmacological texts includes recognition of drug names and extraction of relationships between them. To this purpose, pharmacological annotated corpora are required. These corpora are usually semantically annotated by domain experts. However, other linguistic aspects should be considered to ensure the quality and consistency of the annotation. This paper introduces several linguistic phenomena affecting the annotation of both drug named entities and drug-drug interaction relationships that arose during the annotation process of the DDI corpus. The detailed documentation of these issues and the decisions on them will improve the quality of the annotated corpus and its usefulness for other researchers or users.

Selection and peer-review under responsibility of CILC2013.

Keywords: linguistics; biomedical corpora; drug nomenclature; drug interactions; annotation.

1. Introduction

With the rapid and exponential growth of biomedical literature in the last decades, Natural Language Processing (NLP) techniques are considered as promising tools for the analysis and management of this huge amount of complex information. Through the automatic processing of texts in natural language, Information Extraction (IE) systems can identify the main concepts in biomedical texts (Named Entity Recognition or NER task) and the relationships between them (Relation Extraction or RE task). However, the development of these automatic systems relies in manually annotated resources, such as corpora, which can be used as gold standards or benchmarks for training and evaluating these systems.

Corpora can be annotated at different levels. They can be annotated with grammatical, semantic or pragmatic functions. Several biomedical corpora have been annotated for biomedical NER, for instance, Genia (Kim, Ohta,

* Corresponding author. Tel.: +34-91-624 -9114.
E-mail address: mhzazo@pa.uc3m.es

Tateisi, and Tsujii, 2003), Yapex (Franzéna, Erikssona, Olssona, Askerb, Lidénb, and Cösterb, 2002), GENETAG (Tanabe, Xie, Thom, Matten, and Wilbur, 2005) or CLEF (Roberts, Gaizauskas, Hepple, Demetriou, Guo, Roberts, et al., 2009) corpus. These corpora have been annotated with semantic classes relevant to the molecular biology domain, such as gene, protein, cell or disease, among others, by experts in the specific field. Although the principal aim of annotation tasks of biomedical corpora usually is annotation at semantic level, "grammar and meaning are intertwined and most annotation efforts combine the two"(Simpson and Demner-Fushman, 2012). Therefore, linguistic phenomena usually arise during the annotation process. For example, term variants such as nested terms or discontinuous names are frequently used to describe biomedical named entities (Alex, Haddow, and Grover, 2007; Kolárik, Hofmann-Apitius, Zimmermann, and Fluck, 2007). In addition, relationships can be expressed in different ways through different syntactic phenomena, such as alternation or coordination (Kulick, Bies, Liberman, Mandel, Mcdonald, Palmer, et al., 2004). Therefore, quality annotated corpora should be annotated taking into account both semantic and grammatical aspects and those decisions made during the annotation process should be documented in form of annotation guidelines (Cohen, Fox, and Ogren, 2005).

Annotation guidelines are the documents defining the annotation task and the annotation conventions (Bird, Klein, and Loper, 2009). The extent and detail of these documents are related with the quality of the annotation process and the agreement between annotators. Moreover, the usefulness of the corpora will depend on the quality of the annotation guidelines (Dipper, Götze, and Skopeteas, 2004). Therefore, it is necessary that the annotation guidelines be accessible to the final user of the corpus (Leech, 1993; Pustejovsky and Stubbs, 2012) and that these documents fulfill requirements such as explicitness and completeness (Dipper, Götze and Skopeteas, 2004). However, some biomedical annotated corpora do not provide annotation guidelines or linguistic aspects are not described with sufficient level of detail (Lu, Bada, Ogren, Cohen, and Hunter, 2006).

In this paper, we describe the main linguistic phenomena identified during the annotation of a pharmacological corpus, the DDI corpus (Segura-Bedmar, Martínez, and Herrero-Zazo, 2013). This is a corpus annotated with pharmacological substances and their pharmacological interactions (drug-drug interactions or DDIs). In this context, we have identified those linguistic phenomena that complicate the manual annotation of drug named entities. In the same way, we describe the main syntactic phenomena that should be considered during the annotation of DDIs to ensure the consistency of the annotation.

The remainder of this paper is structured as follows. The next section describes the main characteristics of the DDI corpus and the annotation process. Section 3 focuses on the description of the main sources of annotation problems. Specific linguistic phenomena affecting the annotation of drug named entities in the DDI corpus are detailed in section 4, while those regarding DDI relationships are described in section 5. The comparison with related works is described in section 6. Finally, section 7 draws the conclusions and future work.

2. Methods

2.1. Outline of the corpus

Automatic extraction of drug related information from pharmacological texts can provide important insights in the Pharmacovigilance domain. In particular, these techniques can be useful in the early detection of unknown DDIs, helping to protect patient safety and to reduce associated healthcare costs. For this goal, manually annotated corpora are valuable and necessary resources. The DDI corpus is a manually annotated corpus developed as gold standard for the competitive challenge DDI Extraction 2013 task (Segura-Bedmar, Martínez, and Herrero-Zazo, 2013) as part of SEMEVAL 2013 (http://www.cs.york.ac.uk/semeval-2013/), which main goal is to provide a common framework for evaluation of information extraction techniques applied to drug named entity recognition and the detection of drug-drug interactions from pharmacological texts. The DDI corpus has been developed in the frame of the MultiMedica Project[1], which main objective is to define and develop information extraction and retrieval techniques based on texts from the medical domain.

[1] http://labda.inf.uc3m.es/multimedica/

This corpus is composed of 1,025 documents from two different databases: DrugBank (Wishart, Knox, Guo, Shrivastava, Hassanali, Stothard, et al., 2006) and MedLine®. DrugBank is a curated database that contains detailed information describing properties, structure and biology of pharmacological substances. The version DrugBank 2.0 contained a specific field describing drug-drug interactions mainly as unstructured text. On the other hand, MedLine is the U.S. National Library of Medicine's® (NLM) bibliographic database. It contains the largest collection of references to journal articles in life sciences. Through the PubMed system[2] it is possible to search in the MedLine database and to obtain abstracts describing drug-drug interactions. These two different text sources were selected to create a pharmacological corpus with different styles of text. Moreover, the documents included in the corpus are representative of different pharmacological information sources: primary literature and curated databases. In this way, DrugBank provides a source of texts focusing in the description of DDIs, written in a less technical form of the language (similar to the language used in package inserts). On the other hand, MedLine abstracts usually describe laboratory or clinical studies in scientific vocabulary and complex syntactic structures are usually used in describing details or making explanations.

The DDI corpus was manually annotated with different types of pharmacological substances: generic drugs, branded drugs, groups of drugs and other active substances that can be involved in a DDI. Moreover, different types of DDI relationships were established regarding the information provided in the sentence (mechanism, effect, advice or interaction). A detailed description of these types of entities and relationships can be found in the annotation guidelines[3].

Manual annotation task was carried out by two experienced pharmacist annotators with expertise in pharmacovigilance. The annotation process was performed following the aforementioned annotation guidelines. These documents were iteratively developed and contain all those rules, conventions and examples on how the annotation task should be carried out. Disagreements between the two annotators were discussed and reconciled during the harmonization process, where a text miner expert helped to make the final decision. These modifications were reflected in the gold standard corpus and in the annotation guidelines, when necessary.

3. Main sources of annotation problems

In this section we describe the main sources of annotation problems identified during the annotation process of the DDI corpus. These issues have been addressed in the framework of a pharmacological corpus. However, due to their general nature, these approaches could be extrapolated to domains others than pharmacology, such as medicine or chemistry.

3.1. Tokenization problems

Corpus texts need to be segmented into words and sentences through tokenization before any further processing can be done. Different tokenizers have been developed for this purpose (He and Kayaalp, 2006). However, pharmacological texts and, specifically, chemical and drug names are a source of ambitious numbers, punctuation marks and parentheses. These ambiguous characters can lead to erroneous sentence splitting and word tokenization.

3.2. Complexity of drug named entities

Drug named entity recognition seems to be a relatively simple task, since the number of possible drug names is limited compared with those of other biomedical entities, such as genes or proteins. Moreover, there are different controlled vocabularies and lists of drugs collecting them, e.g. RxNorm (Nelson, Zeng, Kilbourne, Powell, and Moore, 2011) or the ATC classification system[4]. However, in spite of these facts, current automatic information

[2] http://www.ncbi.nlm.nih.gov/pubmed
[3] http://labda.inf.uc3m.es/ddicorpus
[4] http://www.who.int/classifications/atcddd/en/

extraction systems are not able to properly extract drug names from biomedical texts without human intervention (Jagannathan, Mullett, Arbogast, Halbritter, Yellapragada, Regulapati, et al., 2009). This finding highlights that drug names are complex named entities. There are different characteristics of drug names contributing to this complexity.

First of all, the same drug can have different generic and several trade names in different countries. For example, the drug *paracetamol* is named *acetaminophen* in the USA. Some of its branded names include *Acephen*® (in the USA), *Efferalgan*® (in Spain) or *Ultralief*® (in the UK). Moreover, drug names can have different abbreviations or synonyms. In addition, some drugs are approved to be used in some countries, while they are not in others. Therefore, there is not a comprehensive list of drugs that collects all drug names approved in the world as well as all their synonyms. For example, RxNorm provides normalized names for clinical drugs approved in the USA, linking them to different synonyms, such as branded names. On the other hand, the ATC classification system refers to each pharmacological substance by one official name only, excluding possible synonyms. Another important barrier for the maintenance of an updated controlled list is that new discovered drugs are continuously approved for sale.

Secondly, in pharmacological texts, the mention of terms describing a group of drugs is frequent. These term are complex terms, since they are usually represented by nested multi-word terms, including protein names, numbers, abbreviations or adjectives, as well as punctuation marks such as parentheses or hyphens (Kolárik, Hofmann-Apitius, Zimmermann and Fluck, 2007). Moreover, groups of drugs names are terms with multiple possible variants that usually are not collected in a comprehensive way in a controlled vocabulary or database. For example, the group of drugs *Beta Blocking Agents* (term used in the ATC classification system[5]) can be described as well as *Adrenergic beta-Antagonists, beta-Adrenergic Receptor Blockaders* (terms collected in the MeSH thesaurus[6]) or, simply as *β-blockers,* among others (Paolillo, Pellegrino, Salvioni, Contini, Iorio, Bovis, et al., 2013).

3.3. Complexity of biomedical texts

Manual annotation of biomedical texts should be carried out by domain experts capable of understanding the information described in the corpus. However, not all corpora are equally complex. Complexity varies depending on several facts, such as the source of the documents (e.g. manual curated databases or primary scientific literature), the type of study described in the text (e.g. clinical study or in vitro study) as well as content-related and linguistic aspects, such as the use of technical vocabulary, complex sentences, etc. The level of complexity of texts will determine which annotators should be selected for the annotation task (e.g. a pharmacovigilance expert or a life science bachelor student), the length of training for annotators and the explicit rules that should be described in the annotation guidelines.

3.4. Lack of standard or reference works in the specific domain

A set of standard rules for manual annotation of pharmacological substances or drugs has not been established. This is a difficult task since different corpora are annotated with different final objectives. For example, the DDI corpus has been annotated for the extraction of DDIs, while the aim of the CLEF corpus is to extract clinically significant information from clinical texts (Roberts, Gaizauskas, Hepple, Davis, Demetriou, Guo, et al., 2007). Therefore, different corpora require different annotation schema and annotation guidelines.

However, detailed reference works can improve the objectives achieved by future research groups in the specific domain. Therefore, when research groups create a new manually annotated corpus, comprehensive annotation guidelines should be written. These documents should reflect how the annotation task should be carried out, as well as how annotators should deal with specific or complex linguistic phenomena. For example, in the annotation of a pharmacological corpus, it should be important to specify how annotators should annotate stereoisomers of drugs. These chemical entities are usually described by adding a letter S or R before the drug name (e.g. *S-warfarin*). To ensure consistency between different annotators, a simple rule describing if the annotator should include in the

[5] http://www.whocc.no/atc_ddd_index/?code=C07

[6] http://www.ncbi.nlm.nih.gov/mesh/68000319

annotation span only the drug name (*warfarin*) or the stereoisomer specification (*S-warfarin*) should be described in the annotation guidelines.

Thus, when a new research group creates a related annotated corpus with similar entities, this group could base its decisions on those adopted in the reference work, leading to closest corpora that could be re-used or exchanged in future works.

4. Linguistic aspects of drug names

The manual annotation process of a pharmacological corpus can be a difficult task if, previously, it has not been established what terms should be annotated. As mentioned before in section 3.2, drug names are complex entities and they have several nomenclatures, synonyms and term variants. In this section, we describe some of the main linguistic phenomena regarding drug nomenclature.

4.1. Different nomenclatures

Each drug has a unique and globally recognized name called International Non-proprietary Name (INN) that facilitates the identification of pharmaceutical substances[7]. However, different countries can assign specific nonproprietary names, such us United States Adopted Name (USAN)[8] in the USA or *Denominación Oficial Española* (*DOE*) in Spain[9]. Examples below show some of these different nomenclatures:

i) The effects of ***paracetamol*** *are possibly reduced in patients taking anticonvulsants.*

ii) The absorption of ***acetaminophen*** *may possibly be reduced if colestyramine is given at the same time.*

Sentences i) and ii) describe two different interactions of the same drug. *Paracetamol* is an INN, while *acetaminophen* is the USAN for the same drug. Therefore, there are two different names referring to the same substance and both of them should be annotated.

On the other hand, every drug can have several brand names. That is, a drug marketed under a proprietary, trademark-protected name. There can be several drug brand names for every drug in different countries.

iii) ***Atromid-S*** *may displace acidic drugs such as phenytoin or tolbutamide from their binding sites.*

Sentence iii) describes an interaction of the drug *clofibrate*. However, in this sentence it is named using a brand name: *Atromid-S*. A search in the DrugBank database will show that this drug holds more than 90 different brand names. Therefore, these brand names are different terms referring to the same substance and any mention of them in the text should be annotated.

Drug names usually have different synonyms and abbreviations. Two examples are shown in the following sentences:

iv) HUMIRA has been studied in rheumatoid arthritis patients taking concomitant ***MTX****.*

v) This is typical of the interaction of meperidine and ***MAOIs****.*

Sentence iv) refers to the drug named *methotrexate* using an abbreviation: *MTX*. Sentence v) describes the group of drugs *Monoamine Oxidase Inhibitors* by its abbreviation *MAOIs*. All these terms are synonyms for a drug name or a group of drugs name. Therefore, they should be annotated in the corpus.

4.2. Multi-word terms

Multi-word terms are frequently used to describe drug names and, more often, groups of drugs names. Usually, common nouns such as *drugs*, *agents* or *products*, among others, are preceded by an adjective describing the therapeutic effect, the mechanism or other characteristics of the group of drugs. For example:

[7] http://www.who.int/medicines/services/inn/en/

[8] http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council.page.

[9] Boletín Oficial del Estado. Ley 29/2006, de 26 de julio, de garantías y uso racional de los medicamentos y productos sanitarios.

vi) The treatment of depression in diabetic patients must take into account variations of glycemic levels at different times and a comparison of the available ***antidepressant agents*** *is important.*

However, the term can be shortened and the adjective can be used as a noun.

vii) In the present study we evaluated the interference of ***antidepressants*** *with blood glucose levels of diabetic and non-diabetic rats.*

Two different annotations are possible in sentence vi): just the shorter term <*antidepressant*> or the larger term <*antidepressant agents*>. The first option, the annotation of the shorter term, agrees with the annotation in sentence vii), where there is only one possibility: the annotation of the term <*antidepressants*>. However, information extraction systems or techniques would benefit from the addition of common nouns that could help in the identification of group of drugs names. Therefore, we decided to annotate the longer term, whenever possible.

4.3. Nested terms

A frequent linguistic phenomenon in the pharmacological domain is nested named entities. They are frequently used referring to a specific subgroup of drugs within a group of drugs. Next, some examples of nested named entities are presented:

viii) The concomitant use of allopurinol and ***thiazide diuretics*** *may contribute to the enhancement of allopurinol toxicity.*

These nested terms could be annotated in three different ways: as two independent entities <*thiazide*> and <*diuretics*>; as a unique entity <*thiazide diuretics*>; as three different entities <*thiazide*>, <*diuretics*> and <*thiazide diuretics*>. When the author refers to "*thiazide diuretics*" he or she is alluding to one group of drugs (*diuretics* with a concrete structure defined by the adjective *thiazide*). The annotation of two (option one) or three (option three) different entities would lead to the annotation in the text of more entities than those intended by the author and would complicate the annotation of DDIs between them. Thus, if we annotate more than one entity, we would express that there are two different types of groups: one of them *thiazide* and the other, *diuretics*. Therefore, we decided to annotate a unique entity (option two) <*thiazide diuretics*>.

4.4. Discontinuous names

Another related linguistic phenomenon is discontinuous entities. It is especially common when drug names occur in coordinate structures. For example:

ix) In some patients, the administration of a non-steroidal anti-inflammatory agent can reduce the effects of ***loop, potassium-sparing*** *and* ***thiazide diuretics****.*

In sentence ix) bold terms describe three different groups of drugs. The first one refers to a group of diuretics acting in the loop of Hendle (a specific portion of nephrones in the kidney): <*loop diuretics*>. The second one is a group of diuretics that do not promote the loose of potassium of the body and are denominated <*potassium-sparing diuretics*>. The third one is the abovementioned group of diuretics sharing a common structure, <*thiazide diuretics*>.

The term *thiazide* is used commonly without the term diuretics, preserving its meaning. However, terms <*loop*> or <*potassium-sparing*> always acts as modifiers and do not keep the meaning by themselves. This is the reason why we decided to annotate three different entities <*loop diuretics*>, <*potassium-sparing diuretics*> and <*thiazide diuretics*>.

4.5. Ambiguity

Term ambiguity occurs when the same term refers to many concepts (Ananiadou, Kell, and Tsujii, 2006). As other biomedical entities, drug names can be ambitious. Below, we describe two different examples of ambiguity terms:

x) Therefore, in patients taking ***insulin*** *or oral hypoglycemics, regular monitoring of blood glucose is recommended.*

xi) There is no evidence that EPA supplements have detrimental effects on glucose tolerance, ***insulin*** *secretion or* ***insulin*** *resistance in non-diabetic subjects.*

In sentence x) it is implied by the context that the word *insulin* refers to a drug, since it is stated that it is administered by or to a patient. Therefore, it should be annotated as a drug in the corpus. However, in sentence xi) the same word *insulin* names a substance produced by the own body. In this case, we decided to do not annotate it as an entity, since it does not conform to the previously established definition of drug (this definition can be found in the annotation guidelines[10]).

xii) The ***CNS depressant*** *effect of oxycodone hydrochloride may be additive with that of other* ***CNS depressants****.*

Another source of ambiguity terms are groups of drugs names. In sentence xii) the first term <*CNS depressant*> refers to the depressant effect on the central nervous system by the drug *oxycodone hydrochloride*. Therefore, it is not a group of drugs, but an effect of a drug. However, the second bold term <*CNS depressants*> refers to the group of drugs sharing the common characteristic of having a depressant effect on the central nervous system. Therefore, this second term should be annotated as a group of drugs.

Ambiguity remains an important issue in the development of accurate named entity recognition systems. Therefore, the manual annotation rules established for the annotation of ambiguity terms is a relevant decision in the development of any annotated corpora.

5. Syntactic phenomena in pharmacological texts

As mentioned before, relationships can be expressed in different ways through different syntactic phenomena, such as alternation or coordination. In the DDI corpus, a DDI relationship is a binary relationship annotated at the sentence level with an attribute type (effect, advice, mechanism, int). In this section, we describe some of the main annotation problems identified during the annotation of DDI relationships in the DDI corpus.

5.1. Hypernymic propositions

A hypernymic proposition represents a taxonomic relation between a hyponym and a hypernym. Hypernymic propositions, in particular appositive structures consisting of several entities, are very common in our texts.

xiii) The effects of adenosine are antagonized by ***methylxanthines*** *such as* ***caffeine*** *and* ***theophylline****.*

In sentence xiii) there is an interaction involving the entities described in the appositive structure. In this example, <*methylxanthines*> is the hypernym, while <*caffeine*> and <*theophylline*> are the hyponyms. *Methylxanthines* is a group of drugs, and *caffeine* and *theophylline* are two drugs belonging to this group. Therefore, the sentence states that an interaction can occur between the drug *adenosine* and the members of the group *methylxanthines*, for example, *caffeine* and *theophylline*. Thereby, a DDI relationship for each one of them should be annotated.

However, in some appositive structures, the scope of the interaction only remains the hyponym and not the hypernym. See the example below:

xiv) In addition to this pharmacological interaction, this report describes a novel chemical reaction between ***temazepam (a benzodiazepine)*** *and ethanol.*

Sentence xiv) contains a group, <*benzodiazepine*> and a drug belonging to his group, <*temazepam*>. In this example, the term *benzodiazepine* is describing a characteristic of the drug *temazepam*, and we cannot infer from the sentence that the interaction between *ethanol* and *temazepam* can occurs between *ethanol* and other members of the group *benzodiazepine*, too. Therefore, we decided to annotate one DDI relationship only, between <*ethanol*> and <*temazepam*>.

5.2. Coordinate structures

The same drug can be mentioned several times in the same sentence. In these cases, it could be unclear which one should be included in a DDI relationship.

[10] http://labda.inf.uc3m.es/ddicorpus

xv) The concomitant use of ***nitrofurantoin*** *is not recommended since* ***nitrofurantoin*** *may antagonize the effect of norfloxacin.*

Sentence xv) is a coordinate structure consisting of two coordinate clauses with the conjunction 'since' joining them together. In the first clause, there is just one mention of a drug: <*nitrofurantoin*>. In the second clause, however, there are two mentions of two different interacting drugs: <*nitrofurantoin*> and <*norfloxacin*>. Therefore, we decided to annotate as interacting drugs in the DDI relationships only those drugs mentioned in the second clause.

xvi) The concomitant use of ***nitrofurantoin*** *and* ***norfloxacin*** *is not recommended since* ***nitrofurantoin*** *may antagonize the effect of* ***norfloxacin****.*

In sentence xvi), however, the first clause contains two different interacting drugs, <*nitrofurantoin*> and <*norfloxacin*>, as well as the second does. Therefore, two different DDIs should be annotated: one in the first clause and one in the second clause.

xvii) The concomitant use of ***nitrofurantoin*** *is not recommended since it may antagonize the effect of* ***norfloxacin****.*

Finally, in sentence xvii) the interaction in the second clause is described with an anaphora of the term <*nitrofurantoin*>. Since we did not include anaphora annotation in the DDI corpus, a unique DDI relationship should be annotated between the two mentioned drugs <*nitrofurantoin*> and <*norfloxacin*>.

6. Discussion

For the best of our knowledge, there are only two corpora annotated with DDIs. They are the PK-DDI corpus (Boyce, Gardner, and Harkema, 2012) and the PK corpus (Karnik, Subhadarshini, Wang, Rocha, and Li, 2011). These are manually annotated corpora for pharmacokinetic drug-drug interactions (a type of DDIs caused by a specific mechanism). Although the PK corpus was annotated following hand-made annotation guidelines, they are not publicly available. However, PK-DDI annotation guidelines are available to the research community. They provide instruction for the annotation of drugs and pharmacokinetic DDIs. They include definitions, examples and specific instructions regarding annotation aspects and use of the annotation tool. Rules regarding the annotation of drug named entities differ considerably from those described in the DDI corpus annotation guidelines. Although definitions and examples for all those entities describing drugs in their annotation schema are provided, drug names term variants are not explicitly described. For example, there is not a specific rule regarding the annotation of stereoisomers of drugs (see section 3.4). Regarding DDIs annotation rules, however, explicit and clear instructions are provided. Most of them agree with those established in the DDI corpus annotation guidelines. Therefore, we believe that the previous description of syntactic aspects regarding the annotation of DDIs will be useful for further users for both, the DDI corpus and the PK-DDI corpus.

7. Conclusions

Semantic annotation of pharmacological texts is linked with other linguistic aspects. Their documentation is necessary to ensure the quality of the annotated corpus, to increase its usefulness and to provide a framework for other possible users, such as annotators, language engineers or guideline authors. During the annotation of a pharmacological corpus, the DDI corpus, different linguistic phenomena affecting the annotation of drug names arose. Moreover, the annotation of relationships between these drugs was affected, as well, by syntactic aspects. In this paper, we have reviewed all those annotation issues regarding the annotation of drug names and DDIs. For the best of our knowledge, this is the first work that describes those linguistic phenomena affecting drug named entity and DDI relationship annotation in pharmacological texts.

In our future research work, we plan to enrich the current version of the DDI corpus through the annotation of linguistic phenomena required for a better understanding of the texts, such as negation, modality or anaphora. Moreover, we would also like to study these linguistics aspects in pharmacological corpora in other languages, for example, in Spanish.

The annotation guidelines described in this work, as well as the annotated corpus, are publicly available from http://labda.inf.uc3m.es/ddicorpus.

Acknowledgements

This work was supported by the Regional Government of Madrid under the Research Network MA2VICMR [S2009/TIC-1542] and by the Spanish Ministry of Economy under the project MULTIMEDICA [TIN2010-20644-C03-01].

References

Alex, B., Haddow, B., and Grover, C. (2007). Recognising nested named entities in biomedical text. *Proceedings of BioNLP 2007* (pp. 65–72). Prague, Czech Republic.

Ananiadou, S., Kell, D. B., and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in biotechnology*, *24*(12), 571–9. doi:10.1016/j.tibtech.2006.10.002

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Boyce, R., Gardner, G., and Harkema, H. (2012). Using Natural Language Processing to Identify Pharmacokinetic Drug-Drug Interactions Described in Drug Package Inserts. *Proceedings of the 2012 Workshop on BioNLP* (pp. 206–213).

Cohen, K. B., Fox, L., and Ogren, P. V. (2005). Corpus design for biomedical natural language processing, (June), 38–45.

Dipper, S., Götze, M., and Skopeteas, S. (2004). Towards user-adaptive annotation guidelines. *Proceedings of the COLING 2004 5th International Workshop on Linguistically Interpreted Corpora* (pp. 23–30). Geneva, Switzerland.

Franzéna, K., Erikssona, G., Olssona, F., Askerb, L., Lidénb, P., and Cösterb, J. (2002). Protein names and how to find them. *Int. J. Med. Inf*, *67*(1-3), 49–61.

He, Y. and Kayaalp, M. (2006). A Comparison of 13 Tokenizers on MEDLINE. TECHNICAL REPORT LHNCBC-TR-2006-003. *The Lister Hill National Center for Biomedical Communications*, (December).

Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S., et al. (2009). Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *International journal of medical informatics*, *78*(4), 284–91. doi:10.1016/j.ijmedinf.2008.08.006

Karnik, S., Subhadarshini, A., Wang, Z., Rocha, L. M., and Li, L. (2011). Extraction of drug-drug interactions using all paths graph kernel. *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction* (pp. 83–88). Huelva, Spain.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus--a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(Suppl 1), i180–i182. doi:10.1093/bioinformatics/btg1023

Kolárik, C., Hofmann-Apitius, M., Zimmermann, M., and Fluck, J. (2007). Identification of new drug classification terms in textual resources. *Bioinformatics (Oxford, England)*, *23*(13), i264–72. doi:10.1093/bioinformatics/btm196

Kulick, S., Bies, A., Liberman, M., Mandel, M., Mcdonald, R., Palmer, M., et al. (2004). Integrated Annotation for Biomedical Information Extraction. *Human Language Technology Conf. and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* (pp. 61–68).

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing*, *8*(4), 275–281.

Lu, Z., Bada, M., Ogren, P. V, Cohen, K. B., and Hunter, L. (2006). Improving Biomedical Corpus Annotation Guidelines. *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting* (pp. 89–92). Fortaleza, Brazil.

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., and Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association: JAMIA*, *18*(4), 441–8. doi:10.1136/amiajnl-2011-000116

Paolillo, S., Pellegrino, R., Salvioni, E., Contini, M., Iorio, A., Bovis, F., et al. (2013). Role of Alveolar β2-Adrenergic Receptors on Lung Fluid Clearance and Exercise Ventilation in Healthy Humans. *PloS one*, *8*(4), e61877. doi:10.1371/journal.pone.0061877

Pustejovsky, J., and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning* (p. 342). O'Reilly Media, Inc.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., et al. (2007). The CLEF corpus: semantic annotation of clinical text. *Proceedings of the 2007 American Medical Informatics Association Annual Symposium* (pp. 625–9). Chicago, IL, USA.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., et al. (2009). Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, *42*(5), 950–66. doi:10.1016/j.jbi.2008.12.013

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Simpson, M. S., and Demner-fushman, D. (2012). Biomedical Text Mining: A Survey Of Recent Progress. *Mining Text Data* (pp. 465–517).

Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. (2005). GENETAG: a tagged corpus for gen/protein named entity recognition. *BMC bioinformatics*, *6*(s3).

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, *34*(Database issue), D668–72.