A review of Spanish corpora annotated with negation

Salud María Jiménez-Zafra[1], Roser Morante[2],
María Teresa Martín-Valdivia[1], L. Alfonso Ureña-López[1]
[1] SINAI, Computer Science Department, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas s/n, E-23071
{sjzafra, maite, laurena}@ujaen.es
[2] CLTL Lab, Computational Linguistics
VU University Amsterdam, De Boelelaan 1105, 1081 HV
r.morantevallejo@vu.nl

Abstract

The availability of corpora annotated with negation information is essential to develop negation processing systems in any language. However, there is a lack of these corpora even for languages like English, and when there are corpora available they are small and the annotations are not always compatible across corpora. In this paper we review the existing corpora annotated with negation in Spanish with the purpose of first, gathering the information to make it available for other researchers and, second, analyzing how compatible are the corpora and how has the linguistic phenomenon of negation been addressed. Our final aim is to develop a supervised negation processing system for Spanish, for which we need training and test data. Our analysis shows that it will not be possible to merge the small corpora existing for Spanish due to lack of compatibility in the annotations.

Title and Abstract in Spanish

Revisión de los corpus españoles anotados con negación

La disponibilidad de corpus anotados con información sobre la negación es esencial para cualquier idioma, ya que son necesarios para poder desarrollar sistemas capaces de procesar este fenómeno lingüístico. Sin embargo, hay una escasez de corpus anotados con negación, incluso para idiomas como el inglés, y cuando hay corpus disponibles, en la mayoría de los casos son pequeños y las anotaciones no siempre son compatibles entre ellos. En este trabajo revisamos los corpus anotados con información sobre la negación en español con el propósito, en primer lugar, de recopilar la información para que esté disponible para otros investigadores y, en segundo lugar, de analizar la compatibilidad de los corpus y las estructuras de negación que se han anotado. Nuestro objetivo final es desarrollar un sistema automático para el procesamiento de la negación en español. Nuestro análisis muestra que no será posible unir los pequeños corpus existentes actualmente en español debido a la falta de compatibilidad en las anotaciones.

1 Introduction

Nowadays, there is a vast amount of information on the Internet. The large number of sources and the high volume of texts make it difficult for users to select information of interest. In order to extract fine-grained information, automatic systems need to be able to process a diversity of linguistic phenomena such as negation, irony or sarcasm that are used to add extra-propositional meaning. In this study, we focus on negation.

Negation is a very relevant linguistic phenomenon for most of Natural Language Procesing (NLP) tasks, such as information extraction, question answering and sentiment analysis, since negation cues act as operators that can change the meaning of the words that are within their scope by changing the truth value of propositions (Horn, 1989). Negation is a main linguistic phenomenon whose computational

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 27th International Conference on Computational Linguistics, pages 915–924
Santa Fe, New Mexico, USA, August 20-26, 2018.

treatment has not been solved yet, not even for English, due to its complexity and the multiple forms in which it can appear (syntactic, lexical and morphological). Because of the low frequency of negations as compared to other phenomena, the impact in the performance of applications is low from a quantitative perspective, but high from a qualitative view, for example, when processing clinical records. Moreover, some systems we use regularly do not treat this phenomenon effectively. Not even Google deals properly with negation in Spanish. The search "películas que no sean de aventuras", returns adventure movies, whereas it should return non adventure movies.

The standard two phases of a negation processing system are: identifying the presence of negation markers and determining their scope. As in NLP tasks, the availability of corpora annotated with information about negation is essential to train algorithms. However, it is not easy to find corpora annotated with negation, especially in languages other than English. In addition, it is not only necessary that corpora exist, but also that corpora are publicly available for the community to use them, that they are well documented, and that they contain annotations of quality. Ideally, they should also be large enough to allow training robust machine learning systems.

There are different catalogs and platforms that provide information about resources and/or access to them, such as LDC catalog[1], ELRA catalog[2], LRE Map[3], META-SHARE[4] and ReTeLe[5] catalog. If we make a query[6] in these repositories for Spanish corpora annotated with negation, we only find 1 resource in ReTeLe catalog, the SFU Review$_{SP}$-NEG corpus (Jiménez-Zafra et al., 2018). Therefore, we decided to perform an exhaustive search of resources. In this way, we facilitate the task for researchers interested in working on this topic by providing a description of the corpora as well as the direct links to get the data when possible.

In this work, we review the existing Spanish corpora annotated with negation. We focus on Spanish because i) there are no negation processing systems available, ii) it is the second language with most native speakers, and iii) it is the third language most used on the Internet. Our final goal is to develop a supervised negation processing system, for which we need to use annotated corpora. The main purposes of our review are to analyze how compatible are the Spanish corpora annotated with negation and to find out whether the annotations account for the complexity of negation in Spanish.

The rest of the paper is organized as follows: in Section 2 we describe the main features of negation in Spanish, in Section 3 we present the corpora annotated with negation, in Section 4 we analyzed them, in Section 5 we propose a solution to the problems found and, finally, we put forward conclusions in Section 6.

2 Negation in Spanish

Processing negation is not as easy as using a list of negation markers and applying look-up methods. They can be used to find out potential negation cues but they are not adequate because the presence of a cue does not imply that it acts as a negation. In the sentence "You bought the car to use it, didn't you?" the cue "not" is not used as a negation but it is used to reinforce the first part of the sentence. Moreover, it is also necessary to identify the scope or part of the sentence affected by the negation and its focus, the part more prominently negated. If we want to advance in the study of this phenomenon, as for most of NLP tasks, the availability of annotated corpora is essential to train algorithms. According to existing resources for English, annotating negation involves the annotation of the following aspects:

- *Negation cue*: lexical item(s) that modify the truth value of the propositions that are within its scope. There are different types of negation according to the type of the negation cue used:

[1] https://catalog.ldc.upenn.edu/

[2] http://catalog.elra.info/en-us/

[3] http://lremap.elra.info/

[4] http://www.meta-share.org/

[5] ReTeLe is a network of resources for language technologies that has as goal the compilation and documentation of linguistic resources created in Spain. ReTele catalog contains metadata describing resources in RDF format. ReteLe catalog: http://linguistic.linkeddata.es/retele-share/sparql-editor/

[6] Query performed on March 6th 2018

  - Syntactic negation, if a syntactically independent negation marker is used to express negation (i.e. *no ['no/not']*, *nunca ['never']*).
  - Lexical negation, if a word is used whose meaning has a negative component (i.e. *negar ['deny']*, *desistir ['desist']*).
  - Morphological negation, if a morpheme is used to express the negation (i.e. *i-* in *ilegal ['illegal']*, *in* in *incoherente ['incoherent']*). It is also known as affixal negation.

- *Scope*: the part of the sentence affected by the negation cue (Vincze et al., 2008). The scope can be continuous or discontinuous.

- *Focus*: the part of the scope that is most prominently or explicitly negated (Blanco and Moldovan, 2011).

- *Negated event*: the event that is directly negated by the negation cue, usually a verb, a noun or an adjective (Kim et al., 2008).

This is just a list of the main aspects that have been annotated for negation. However, each language has specific linguistic resources to express negation and specific negation structures, which should also be reflected in the information annotated in corpora. As we will show in Section 4, most existing annotation schemes for Spanish do not account for the complexity of the linguistic structures used to express negation that are present in texts. This happens mainly because of two reasons: first, annotation of negation started with the annotation of clinical reports in English (Chapman et al., 2001; Goldin and Chapman, 2003; Mutalik et al., 2001a), where there is not too much variation of negation structures. Second, corpora have been created for specific purposes, such as extracting negated clinical events, and not with the intention of accounting for all the linguistic complexity of the negation phenomenon.

An exception to this is the SFU Review$_{SP}$-NEG corpus (Jiménez-Zafra et al., 2018; Martí et al., 2016). The guidelines specify a great variety of negation patterns at the syntactic level that we summarize below. Additionally, the guidelines also specify expressions that involve a negation cue but do not express negation.

On the one hand, patterns that express negation can be divided into three categories:

1. *Simple negation markers*, if they are composed of only one single negation marker (i.e. *no ['no/not']*, *nunca ['never']*).
2. *Complex negation markers*, if negation is expressed using two or more negation markers that can be continuous (i.e. *casi no ['almost not']*, *casi nunca ['hardly ever']*) or discontinuous (i.e. *no ... en absoluto ['not ... at all']*, *no ... mucho ['not ... a lot']*). Complex negation markers are usually used to reinforce negation or to modulate the value of negation (increase or diminish the degree of negation).
3. *Expressions that do not contain any negation marker*: lexicalized complex constructions that express negation in specific contexts even though they do not contain any negation marker (i.e. *en mi vida ['in my life']*).

On the other hand, patterns that not express negation can be categorized as follows:

1. *Rhetorical negation markers*, if a negation marker is used with an emphatic or expletive value, that is, if it is used to make a sentence more full, intense or harmonious, although it is not necessary to understand the meaning of the sentence (i.e. *Viniste a verlo, ¿no? ['You came to see him, didn't you?']*).
2. *Idioms containing negation markers*, if an idiom (i.e. *ni corta ni perezosa ['without thinking twice']*) or a lexicalised cue (i.e. *hasta que no ['until']*) contain a negation marker that does not express negation.

3. *Negation markers in contrastive constructions*, if negation markers are used to counterpose different ideas, to correct something, to introduce new information or to express obligation, rather than to express negation (i.e. *No hay más solución que comprar una lavadora ['There is no other solution than to buy a washing machine']*).
4. *Negation markers in comparative constructions*, if negation markers are used to compare some property with something, that is, negation is used to place an entity below or above another entity on a scale (i.e. *No es tan grande como me lo imaginaba ['It is not as big as I imagined']*).

3 Corpora annotated with negation

In this section, the Spanish corpora annotated with negation are presented[7]. To the best of our knowledge, five corpora exist from different domains, although the clinical domain is the predominant one.

3.1 UAM Spanish Treebank

The first Spanish corpus annotated with negation that we are aware of is the UAM Spanish Treebank (Moreno et al., 2003), which was enriched with the annotation of negation cues and their scopes (Sandoval and Salazar, 2013).

The initial UAM Spanish Treebank consisted of 1,500 sentences extracted from newspaper articles (*El País Digital* and *Compra Maestra*) that were annotated syntactically. Trees were encoded in a nested structure, including syntactic category, syntactic and semantic features, and constituent nodes, following the Penn Treebank model. Later, this version of the corpus was extended with the annotation of negation and 10.67% of the sentences were found to contain negations (160 sentences).

In this corpus, syntactic negation was annotated but not lexical nor morphological negation. It was annotated by two experts in corpus linguistics who followed similar guidelines to those of Bioscope corpus (Szarvas et al., 2008; Vincze, 2010). They included negation cues within the scope as in Bioscope and NegDDI-DrugBank (Bokharaeian et al., 2014). All the arguments of the negated events were also included in the scope of negation, including the subject, which was excluded from the scope in active sentences in Bioscope. There is no information about inter-annotator agreement.

The UAM Spanish Treebank corpus is freely available at `http://www.lllf.uam.es/ESP/Treebank.html`. It is in XML format, negation cues are tagged with the label *Type="NEG"* and the scope of negation is tagged with the label *Neg="YES"* in the syntactic constituent on which negation acts.

3.2 IxaMed-GS

The IxaMed-GS corpus (Oronoz et al., 2015) is composed of 75 real electronic health records from the outpatient consultations of the Galdakao-Usansolo Hospital in Biscay (Spain). It was annotated by two experts in pharmacology and pharmacovigilance with entities related to diseases and drugs, and with the relationships between entities indicating adverse drug reaction events. They defined their own annotation guidelines taken into consideration the issues that should be considered for the design of a corpus according to Ananiadou and McNaught (2006).

The objective of this corpus was not the annotation of negation but the identification of entities and events in clinical reports. However, negation and speculation were taken into account in the annotation process. In the corpus, four entity types were annotated: diseases, allergies, drugs and procedures. For diseases and allergies, they distinguished between negated entity, speculated entity and entity (for non-speculative and non-negated entities). 2,362 diseases were annotated, out of which 490 (20.75%) were tagged as negated diseases and 40 (1.69%) as speculated diseases. 404 allergy entities were identified, of which 273 (67.57%) were negated and 13 (3.22%), speculated. The quality of the annotation process was assessed by measuring the inter-annotator agreement, which was 90.53% for entities and 82.86% for events.

The corpus might be acquired via the EXTRECM project[8] by agreeing to some conditions that include a confidentiality agreement.

[7] `https://github.com/sjzafra/spanish_negation_corpora`
[8] `http://ixa.si.ehu.eus/extrecm`

3.3 SFU Review$_{SP}$-NEG

The SFU Review$_{SP}$-NEG[9] (Jiménez-Zafra et al., 2018) is the first Spanish corpus that includes the event in the annotation of negation and that takes into account discontinuous negation markers. Moreover, it is the first corpus where the effect of the negation on the words that are within its scope is annotated, that is, whether there is a change in the polarity or an increment or reduction of its value. It is an extension of the Spanish part of the SFU Review corpus (Taboada et al., 2006) and it could be considered as the counterpart of the SFU Review Corpus with negation and speculation annotations[10] (Konstantinova et al., 2012).

The Spanish SFU Review corpus consists of 400 reviews extracted from the website *Ciao.es* that belong to 8 different domains: cars, hotels, washing machines, books, cell phones, music, computers, and movies. For each domain there are 50 positive and 50 negative reviews, defined as positive or negative based on the number of stars given by the reviewer (1-2=negative; 4-5=positive; 3-star review were not included). Later, it was extended to the SFU Review$_{SP}$-NEG corpus in which each review was automatically annotated at the token level with POS-tags and lemmas, and manually annotated at the sentence level with negation cues and their corresponding scopes and events. It is composed of 9,455 sentences, out of which 3,022 sentences (31.97%) contain at least one negation marker.

In this corpus, syntactic negation was annotated but not lexical nor morphological negation, as in the UAM Spanish Treebank corpus. Unlike this one, annotations on the event and on how negation affects the polarity of the words within its scope were included. The annotations were performed by two senior researchers with in-depth experience in corpus annotation who supervised the whole process and two trained annotators who carried out the annotation task. The Kappa coefficient for inter-annotator agreement was of 0.97 for negation cues, 0.95 for negated events and 0.94 for scopes.[11] A detailed discussion of the main sources of disagreements can be found in (Jiménez-Zafra et al., 2016).

The guidelines of the Bioscope corpus were taken into account but after a thorough analysis of negation in Spanish, a typology of Spanish negation patterns was defined (Martí et al., 2016). As in Bioscope, NegDDI-DrugBank and UAM Spanish Treebank, negation markers were included within the scope. Moreover, the subject was also included within the scope when the word directly affected by negation is the verb of the sentence, as in ConanDoyle-neg corpus (Morante and Daelemans, 2012). The event was also included in the scope of negation as in ConanDoyle-neg corpus.

The SFU Review$_{SP}$-NEG is publicly available and can be downloaded at `http://sinai.ujaen.es/sfu-review-sp-neg-2/`.

3.4 UHU-HUVR

The UHU-HUVR (Cruz Díaz et al., 2017) is the first Spanish corpus in which affixal negation is annotated. It is composed of 604 clinical reports from the Virgen del Rocío Hospital in Seville (Spain). 276 of this clinical documents correspond to radiology reports and 328 to the personal history of anamnesis reports written in free text.

In this corpus, all types of negation were annotated, syntactic, morphological (affixal negation), and lexical. It was annotated with negation markers and the negated events by two domain expert annotators following closely the Thyme corpus guidelines (Styler IV et al., 2014) with some adaptations. In the anamnesis reports, 1,079 sentences (35.20%) were found to contain negations out of 3,065 sentences. On the other hand, 1,219 sentences (22.80%) out of 5,347 sentences were annotated with negations in the radiology reports. The Dice coefficient for inter-annotator agreement was higher than 0.94 for negation markers and higher than 0.72 for negated events. Most of the disagreements were the result of a human error, i.e., the annotators missed a word or included a word that did not belong either to the event or to the marker. However, other cases of disagreement can be explained by the difficulty of the task and the lack of clear guidance. They encountered the same type of disagreements as Jiménez-Zafra et al. (2016) when annotating the SFU Review$_{SP}$-NEG corpus.

[9]First Online: 22 May 2017 `https://doi.org/10.1007/s10579-017-9391-x`

[10]`https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html`

[11]The inter-annotator agreement values have been corrected with respect to those published in (Jiménez-Zafra et al., 2018) due to the detection of an error in the calculation thereof.

Authors say that the annotated corpus will be made publicly available, but it is not currently available probably because of legal and ethical issues.

3.5 IULA Spanish Clinical Record

The IULA Spanish Clinical Record corpus (Marimon et al., 2017) contains 300 anonymized clinical records from several services of one of the main hospitals in Barcelona (Spain) that was annotated with negation markers and their scopes. It contains 3,194 sentences, out of which 1,093 (34.22%) were annotated with negation cues.

In this corpus, syntactic negation and lexical negation were annotated but not morphological negation. It was annotated with negation cues and their scopes by three computational linguists annotators advised by a clinician. The inter-annotator agreement Kappa rates were 0.85 between annotators 1 and 2, and annotators 1 and 3; and 0.88 between annotators 2 and 3. The authors defined their own annotation guidelines taking into account the currently existing guidelines for corpora in English (Mutalik et al., 2001b; Szarvas et al., 2008; Morante and Daelemans, 2012). Differently from previous work, they did not include the negation cue nor the subject in the scope (except when the subject is located after the verb).

The corpus is publicly available with a CC-BY-SA 3.0 license and it can be downloaded at `http://eines.iula.upf.edu/brat//#/NegationOnCR_IULA/`.

4 Analysis

In order to take an informed decision about which (combination of) corpus can we use to develop a system, we have performed a detailed analysis. An overview of the information annotated can be found in Tables 1 and 2.

Table 1 presents a summary of the negation aspects annotated in each corpus, the domain of the documents, the size in number of sentences and the inter-annotator measure used to estimate the agreement. Table 2 contains the type of negation cues that have been annotated in each corpus and negation types that have been taken into account.

	UAM Spanish Treebank	IxaMed-GS	SFU Review$_{SP}$-NEG	UHU-HUVR	IULA Spanish Clinical Record
Domain	Newspaper articles	Clinical reports	Movies, books, product reviews	Clinical reports	Clinical reports
Total sentences	1,500	NA	9,455	8,412	3,194
Sentences with negation	160 (10.73%)	NA	3,022 (31.97%)	2,298 (27.32%)	1,093 (34.22%)
Negation cue	✓	-	✓	✓	✓
Scope	✓	-	✓	-	✓
Event	-	✓	✓	✓	-
Focus	-	-	-	-	-
IAA measure	NA	%	Kappa	Dice	Kappa

Table 1: Spanish corpora annotated with negation (NA: Non-Available, -: Absent, ✓:Present).

The years of publication of the corpora show the novelty of the task. The first corpus annotated with negation in Spanish appeared in 2013, while the others have been compiled in the last two years. Important aspects of the corpora to be analyzed are the type of documents included, the size, the guidelines applied, the annotation schemes and the inter-annotator agreement.

Three of the five corpora focus on the clinical domain, which reflects the demands for the treatment of negation in this domain. Processing negation in clinical documents is crucial because the health of a patient is at stake, it is not the same to say that *a patient has* or *does not have a disease* or that *he is* or *is not allergic to a compound*. Moreover, we observe that there are other domains of interest, such as product reviews and news. The rating of a film will be totally different if a viewer says *"I liked the*

	UAM Spanish Treebank	SFU Review$_{SP}$-NEG	UHU-HUVR	IULA Spanish Clinical Record
Syntactic	✓	✓	✓	✓
Lexical	-	-	✓	✓
Morphological	-	-	✓	-
Simple	✓	✓	✓	✓
Complex	NA	✓	NA	NA
Expressions not containing negation markers	NA	✓	NA	NA
Rhetorical	NA	✓	NA	NA
Idioms	NA	✓	NA	NA
Contrastive	NA	✓	NA	NA
Comparative	NA	✓	NA	NA

Table 2: Negation types in the Spanish corpora annotated with negation cues (NA: Non-Available, -: Absent, ✓:Present).

movie" or if he says *"I did not like the movie"*. In the case of news, the impact would be totally different if it is said *"A plane crashed"* or if it is said *"Finally the plane did not crash"*.

As for the size, the available corpora are not very large and, although negation is an important phenomenon for NLP tasks, it is relatively unfrequent. In newspaper articles, only 10.73% of the sentences contain negation and, in the case of product reviews and clinical reports this value amounts to 31.97% and 29.22%[12], respectively. These percentages show the need of continuing working on the annotation of negation and its study. Training supervised systems usually relies on the existence of annotated corpora and, consequently, corpus generation is an important part for the development and testing of NLP techniques.

In relation to the guidelines used, it is noteworthy that there is no uniformity. In the first place, there are divergences in the negation aspects being annotated (negation cue, scope, event, focus). None of the corpora contain annotations of the four elements and the focus has been annotated in none of them. Only the SFU Review$_{SP}$-NEG corpus contains annotations of three elements (negation cue, scope and event). The UAM Spanish Treebank and the IULA Spanish Clinical Record corpora have focused on annotating negation cues and their scopes, and the UHU-UVR on the annotation of negation cues and their events. In the second place, these elements have not been annotated in the same way:

- *Negation cue*. As it has been described in Section 2, negation in Spanish is a complex phenomenon. Depending on the negation cue used, it can be syntactic, lexical or morphological. Moreover there are different types of negation patterns that express negation (simple negation markers, complex negation markers and expressions not containing any negation marker) and that do not express negation (rhetorical negation markers, idioms containing negation markers, negation markers in contrastive constructions and negation markers in comparative constructions). Only the UHU-UVR corpus contains annotations about the three types of negation cues (syntactic, lexical and morphological). The UAM Spanish Treebank and the SFU Review$_{SP}$-NEG corpora take only into account syntactic negation, and the IULA Spanish Clinical Record corpus also considered it along with lexical negation. However, in general, it is not specified whether the complexity of negation has been taken into account during the annotation process. An exception to this is the SFU Review$_{SP}$-NEG corpus. The guidelines specify that the different types of negation patterns according to the semantic interpretation have been considered for the annotation of negation at the syntactic level. This

[12]For clinical reports, it has been considered the average corresponding to the sentences annotated in UHU-HUVR and IULA Spanish Clinical Record corpora.

information has been summarized in Table 2.

- *Scope*. In the UAM Spanish Treebank and the SFU Review$_{SP}$-NEG corpora negation cues were included within the scope of negation as in Bioscope (Vincze et al., 2008), but in the IULA Spanish Clinical Record corpus they were not included. On the other hand, in the UAM Spanish Treebank all the arguments of the negated events, including the subject, were included within the scope of negation. However, in the SFU Review$_{SP}$-NEG corpora, the subject was included within the scope of negation when the word directly affected by negation is the verb of the sentence, as in ConanDoyle-neg corpus (Morante and Daelemans, 2012). In the IULA Spanish Clinical Record corpus it was not included, except when the subject is located after the verb.

- *Focus*. This element has not been annotated in the existing corpora.

- *Negated event*. In the SFU Review$_{SP}$-NEG corpus the event is always included within the scope of negation, as in Conan Doyle-neg corpus, and it is usually the head of the phrase in which the negation appears. In the UHU-UVR corpus negated events are annotated if they are clinically relevant, so not all negated events are annotated.

As for the annotation schemes, there is no a standard one. Each project devices the own scheme according to the needs of the project, which has consequences in the compatibility of the annotations across corpora. It is not possible to combine the corpora for machine learning purposes in order to obtain more training data.

Furthermore, the annotated corpora do not use the same coefficient to measure the inter-annotator agreement and there is even a corpus for which this measure is not provided. Providing this measure is very important because it allows to show the reliability of the annotation and the difficulty of the task.

The main purposes of our review was to analyse how compatible are the Spanish corpora annotated with negation and to find out whether the annotations account for the complexity of negation in Spanish. We have found that the existing corpora are not compatible, they have not been annotated with the same purpose and they do not contain annotations for the same negation aspects. Even in those corpora that contain annotations for the same negation aspect, the annotations have been made based on different guidelines. In relation to the complexity of negation, only the SFU Review$_{SP}$-NEG guidelines specify how different negation structures should be annotated. The guidelines of the other corpora do not contain information about the linguistic structures that have been taken into account.

5 A solution

The result of this analysis opens some questions. The existing annotation schemes are not compatible because of differences in genre, annotation guidelines, and the aspects of negation that have been annotated (negation cue, scope, event, focus). Therefore, it would be desirable to define a new scheme that integrates the contents of existing schemes. The scheme should be domain independent and all negation elements should be annotated in the same way. We are currently working on this. We have proposed a workshop (NEGES - Workshop on Negation in Spanish[13]) to advance in the study of this phenomenon and one of the tasks[14] has as goal to reach an agreement on the guidelines to follow. In order to participate in the task, researchers must analyze the existing guidelines and send a document indicating which aspects of the guidelines they agree with and which they do not, all duly justified. This information will be used to discuss the aspects of interest and to try to reach a consensus.

Additionally, we will conduct a study on the feasibility of an automatic conversion of the corpora. It is a pity that after all the work done and the time invested in the annotation of these corpora, it is not possible to merge them. We will explore semi-automatic approaches to re-annotate the most corpora possible.

[13] http://www.sepln.org/workshops/neges/index.php
[14] Task 1 - Annotation guidelines

6 Conclusions

Processing negation is a very important task in NLP because negation can change the truth value of a proposition. Detecting this is crucial in some tasks such as sentiment analysis or question answering. Most existing work on the treatment of negation has been carried out for English, but it is necessary to focus on other languages such as Chinese, Spanish or Arabic.

In this paper, we have presented the existing Spanish corpora annotated with negation information and we have described the main features of each one. Three of them are centered on the clinical domain probably because it is one of the areas where the treatment of negation is crucial in order not to extract false information about clinical conditions. It is worth noting that the first corpus annotated with negation appeared in 2013, whereas the rest have been compiled in the last two years, which shows the novelty of the task.

Our analysis shows that it would not be possible to merge the existing corpora in order to obtain a bigger one to train a machine learning system because of differences in genre, annotation guidelines, and the aspects of negation that have been annotated. The corpora have not been annotated with the same purpose and they do not contain annotations for the same negation aspects. Despite the fact that there have been already several annotation efforts, the community lacks a standard to annotate negation, contrary to what happens with other phenomena such as semantic roles.

As future work we plan to develop annotation standards to annotate negation in Spanish, in such a way that they are applicable to different genres and domains, and to analyze the feasibility of an automatic conversion of the corpora to a common annotation scheme.

Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government. RM is supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

References

Sophia Ananiadou and John McNaught. 2006. *Text mining for biology and biomedicine*. Artech House London.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA, June. Association for Computational Linguistics.

Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco. 2014. Exploring negation annotations in the DrugDDI Corpus. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014)*. Citeseer.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B.G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34:301–310.

Noa P Cruz Díaz, Roser Morante Vallejo, Manuel J Maña López, Jacinto Mata Vázquez, and Carlos L Parra Calderón. 2017. Annotating Negation in Spanish Clinical Texts. *SemBEaR 2017*, page 53.

l. M. Goldin and W.W. Chapman. 2003. Learning to detect negation with 'Not' in medical texts. In *Proceedings of ACM-SIGIR 2003*.

Laurence R. Horn. 1989. *A natural history of negation*. CSLI Publications.

Salud María Jiménez-Zafra, M Teresa Martín-Valdivia, L Alfonso Ureña-López, M Antonia Martí, and Mariona Taulé. 2016. Problematic Cases in the Annotation of Negation in Spanish. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 42–48.

Salud María Jiménez-Zafra, Mariona Taulé, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and M Antónia Martí. 2018. SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9(1):1.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Díaz Cruz, Manuel J Maña López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *LREC*, pages 3190–3195.

Montserrat Marimon, Jorge Vivaldi, Núria Bel, and Roc Boronat. 2017. Annotation of negation in the IULA Spanish Clinical Record Corpus. *SemBEaR 2017*, 5(36.41):43.

M Antónia Martí, M Teresa Martín Valdivia, Mariona Taulé, Salud María Jiménez Zafra, Montserrat Nofre, and Laia Marsó. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57:41–48.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul*. Citeseer.

Antonio Moreno, Susana López, Fernando Sánchez, and Ralph Grishman. 2003. Developing a syntactic annotation scheme and tools for a Spanish treebank. In *Treebanks*, pages 149–163. Springer.

A.G. Mutalik, A. Deshpande, and P.M. Nadkarni. 2001a. Use of general-purpose negation detection to augment concept indexing of medical documents. a quantitative study using the UMLS. *J Am Med Inform Assoc*, 8(6):598–609.

Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. 2001b. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 8(6):598–609.

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza D'ıaz de Ilarraza, and Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

Antonio Moreno Sandoval and Marta Garrote Salazar. 2013. La anotación de la negación en un corpus escrito etiquetado sintácticamente. Annotation of negation in a written treebank. *Revista Iberoamericana de Linguistica*, 8.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45. Association for Computational Linguistics.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 427–432.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1.

Veronika Vincze. 2010. Speculation and negation annotation in natural language texts: what the case of bioscope might (not) reveal. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 28–31. Association for Computational Linguistics.