data

MDPI

Article

Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language

Maria Mitrofan [1,†,*], Verginica Barbu Mititelu [1,†,*] and Grigorina Mitrofan [2,*]

[1] Romanian Academy Research Institute for Artificial Intelligence, 13 Calea 13 Septembrie, Bucharest 050711, Romania

[2] National Institute of Diabetes and Metabolic Diseases "N.C. Paulescu", 5-7 Ion Movilă Street, Bucharest 020475, Romania

* Correspondence: maria@racai.ro (M.M.); vergi@racai.ro (V.B.M.); mitrofan.grigorina@gmail.com (G.M.)

† These authors contributed equally to this work.

Received: 29 September 2018; Accepted: 16 November 2018; Published: 23 November 2018

Abstract: Gold standard corpora (GSCs) are essential for the supervised training and evaluation of systems that perform natural language processing (NLP) tasks. Currently, most of the resources used in biomedical NLP tasks are mainly in English. Little effort has been reported for other languages including Romanian and, thus, access to such language resources is poor. In this paper, we present the construction of the first morphologically and terminologically annotated biomedical corpus of the Romanian language (MoNERo), meant to serve as a gold standard for biomedical part-of-speech (POS) tagging and biomedical named entity recognition (bioNER). It contains 14,012 tokens distributed in three medical subdomains: cardiology, diabetes and endocrinology, extracted from books, journals and blogposts. In order to automatically annotate the corpus with POS tags, we used a Romanian tag set which has 715 labels, while diseases, anatomy, procedures and chemicals and drugs labels were manually annotated for bioNER with a Cohen Kappa coefficient of 92.8% and revealed the occurrence of 1877 medical named entities. The automatic annotation of the corpus has been manually checked. The corpus is publicly available and can be used to facilitate the development of NLP algorithms for the Romanian language.

Keywords: corpus; biomedical; Romanian; part-of-speech tags; named entities

1. Introduction

In the field of biomedical sciences, vast quantities of data are generated every year and are available as free texts for natural language processing (NLP) tasks. Substantial efforts were made over the past decades to develop methods and tools to extract useful information from textual records such as full-text journal articles, medical narratives, clinical reports and to organize them in structured data.

In order to automatically process the biomedical literature data, some manually annotated biomedical corpora were developed and used for supervised training of and for evaluating the systems. For example, gold standard corpora have been created for different types of tasks such as part-of-speech tagging [1], named entity recognition [2], relation extraction [3], event extraction [4].

Lately, a slightly increasing number of resources specific to this field have been created for languages other than English. For example, Boytcheva et al. [5] created a biomedical corpus which contains 6400 words, 2000 of them belonging to the Bulgarian medical terminology. For French, Aurélie Névéol et al. created the Quaero French medical corpus [6] that comprises a total of 103,056 words annotated at entity and concept level. Isabel Moreno et al. developed DrugSemantics corpus [7], a collection of Summaries of Product Characteristics in Spanish that contains 226,729 tokens annotated with pharmacotherapeutic named entities. It is necessary to have resources for a variety

of languages in order to develop methods and tools useful for various NLP tasks and also for the biomedical community (decision support systems, cohort identification).

NLP subtasks are chained together to form processing pipelines; therefore, errors produced in these basic subtasks, tokenization, part-of-speech (POS) tagging and named entity recognition (NER), affect the main objectives of the biomedical text mining techniques. The availability of annotated biomedical textual data still remains a barrier in developing state-of-the-art NLP systems, especially for underresourced languages, Romanian among others.

State-of-the-art POS taggers based on statistical approaches achieve an accuracy of between 93–98%. The main problem encountered when using statistical POS taggers is the fact that most of them need corpora annotated with POS labels as training and testing data. In the case of a specialized domain, the main issue is that the accuracy of POS taggers drops on unknown words. For example, the TnT tagger [8] performs at 97% accuracy on known words, but the accuracy decreases to 89% for unknown words. Because of the differences between general language and domain specific vocabulary the accuracy of a tagger decreases when the percentage of unknown words increases. For example, the Romanian TTL POS tagger [9] has an accuracy of 98.23% on newswire domain and 97.83% on biomedical terminology [10]. In order to achieve a good POS tagging accuracy, domain specific annotated gold standard corpora are needed.

Named entity recognition (NER) is another NLP basic task which deals with boundaries identification of domain specific terms such as diseases, procedures, and chemicals for the biomedical domain. Even though multiple efforts have been made to develop domain specific resources useful for biomedical NER (NCBI corpus [11], CHEMDNER corpus [12], BioInfer [13]), the availability of annotated corpora is limited mostly to English. Therefore, in order to evaluate the robustness of NER systems for other languages than English, high quality annotated corpora for the respective languages are needed.

Moreover, it was shown that joint part-of-speech tags and named entities can be used as features to improve the performance of NLP systems. György Móra [14] showed that joint labeling can be used to exploit labels of one task as features in the other task, improving the accuracy for both tasks.

In this paper, we describe the first version of the gold standard morphologically and named entity annotated Romanian medical corpus (MoNERo). In the next section, we describe the corpus: selection of the texts to include in it, annotation levels and guidelines for both morphology and named entities (NEs), and the annotation process. The results are presented in Section 3: descriptive statistics of the corpus, inter-annotator agreement for NER, manual corrections of the POS tagging. The conclusions section closes the paper, with some references to prospective work, too.

The creation of MoNERo is part of a larger effort of developing language resources for Romanian, namely a priority project of the Romanian Academy for creating a large corpus of contemporary Romanian [15].

2. Corpus Description

In this section, we present the effort of collecting the texts, of preprocessing them and of annotating them at the morphological level and with medical terms.

2.1. Corpus Selection

MoNERo contains a selection of biomedical texts which come from three different types of documents: medical books, journal articles and medical blog posts distributed in three different medical subdomains: diabetes, and endocrinology, cardiology. The selection of the corpus and the associated annotation schema closely observe the requirements of the CoRoLa project [15] in the context of which this corpus was created. This is to say that all texts are original, and freely available for search, not for download. However, as we targeted the creation of a freely available resource, the only way not to breach the protocols signed with the texts providers was to select at most two consecutive sentences. There were no implications whatsoever on the morphological annotation, although this had

an impact on the understanding of the named entities by the annotators, who had to invest time in documenting for ensuring the correct annotation of these entities. MoNERo represents a component of the BioRo [16] corpus which is a biomedical corpus for Romanian language, subpart of CoRoLa.

The selection of the sentences included in MoNERo observed the following conditions: presence of Romanian diacritics, coverage of several medical domains (diabetes, endocrinology, cardiology), avoid duplicate sentences. The first condition was necessary for several reasons: (i) observe the norms of writing in Romanian; (ii) avoid ambiguities: many words without diacritics are ambiguous among several forms: e.g., the word without diacritics *fata* can be disambiguated as one of the following nine forms: as the indefinite singular nominative-accusative noun *fată* (En. "girl"), the definite singular nominative-accusative noun *fata* (En. "the girl"), the indefinite singular nominative-accusative noun *față* (En. "face"), the definite singular nominative-accusative noun *fața* (En. "the face"), the indefinite singular nominative-accusative noun *fâță* (a generic name for small fish that move fast), the definite singular nominative-accusative noun *fâța* (referring to a certain such fish), the present verb form *fată* (En. about animals "to give birth"), the imperfective form of the same verb *făta*, the past simple form of the same verb *fătă*; (iii) to obtain better results of the automatic POS tagging phase: the example of ambiguity given above shows that the presence of diacritics reduces ambiguity a lot, although not completely: the form *fată* still has two possible morphologic interpretations: noun or verb; however, the POS tagger very rarely mistakes a noun for a verb or vice versa, as shown in Section 2.2.1 below. The second condition was to have sentences extracted from documents belonging only to the medical domains selected (diabetes, endocrinology, cardiology). The extraction of the sentences was based on the fact that all the files contained in the BioRo corpus have an associated metadata scheme which contains information about the domain. The third condition used in the selection process of the sentences included in MoNERo corpus was to eliminate duplicate sentences, if any. Due to the fact that medical blog posts were included in the corpus, the chances of duplicate sentences to appear increased; therefore, a search for identical sentences was performed.

In Table 1, we present the distribution of the sentences over the three medical domains. It is a roughly equal distribution, with the cardiology domain less well represented. The sentences length is almost the same in all domains, and thus, we can consider such long sentences as a characteristic of the medical domain.

Table 1. Statistics over the corpus.

	#Tokens	#Sentences	Tokens Per Sentence
Cardiology	3428	116	29.55
Diabetes	5504	192	28.66
Endocrinology	5089	187	27.21

2.2. Corpus Annotation Guidelines

The goal of the annotation task was to ensure two levels of annotation (POS tagging and NEs) that would be consistent and suitable for training and evaluating NLP systems.

2.2.1. Annotation Guidelines for Part of Speech Tags

For the morphologic annotation of the corpus, the Romanian M (orpho) S (yntactic) D (escription) tagset developed in the MULTEXT-East project [17] was used. The specifications created in this project identify fourteen classes of words for Romanian: noun, verb, adjective, pronoun, determiner, article, adverb, adposition, conjunction, numeral, interjection, residual, abbreviation, and particle [18]. For each class, a number of attributes are applicable and they are presented in Table 2. Most of them have several possible values. For example, the attribute gender has two possible values in Romanian, masculine and feminine. Although there is also a neuter gender, it is not a separate value, as the neuter borrows the singular form from masculine and the plural from feminine. Thus, neuter nouns have the value *m* for gender when in the singular number and the value *f* when in the plural number.

While all attributes have the same values for different parts of speech, the attribute type has specific values: e.g., for nouns it has the values *common* or *proper*, for verbs it has the values *main* and *auxiliary*, for pronouns-*demonstrative*, *indefinite*, *possessive*, *int-rel*, *personal*, *reflexive*, *negative*, *emphatic* and so on and so forth.

Table 2. The parts of speech and their attributes. N = noun, V = verb, A = adjective, P = pronoun, D = determiner, T = article, R = adverb, S = preposition, C = conjunction, M = numeral, I = interjection, X = residual, Y = abbreviation, Q = particle.

	N	V	A	P	D	T	R	S	C	M	I	X	Y	Q
type	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓
gender	✓	✓	✓	✓	✓	✓				✓			✓	
number	✓	✓	✓	✓	✓	✓				✓			✓	
case	✓		✓	✓	✓	✓		✓		✓			✓	
definiteness	✓		✓							✓			✓	
clitic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓
VForm		✓												
tense		✓												
person		✓		✓	✓									
degree			✓				✓							
owner_number				✓	✓									
owner_gender				✓	✓									
form				✓						✓				
modific_type					✓									
formation								✓	✓					✓
coord_type									✓					
sub_type									✓					
synt_type													✓	
no attribute											✓	✓		

The TTL annotator was used to annotate the corpus with part of speech tags. TTL is a module written in Perl that supports the Romanian, English and French languages, with the following functionalities: sentence segmentation, word segmentation, morphosyntactic labeling, lemmatization, chunking. This tool is a reimplementation of the TnT [8] annotator, but extends it in several ways using the tiered tagging technique where MSDs are recovered from a CTAG (Corpus TAG) annotation [19]. To achieve greater accuracy, it uses a set of 715 labels specific to the Romanian language. In the case of the unknown words TTL assigns the most probable label based on a suffix model.

Compounds that are written as separate words are split. For example, the compound preposition *de la* (En. "from") is analyzed as two tokens: *de* (En. "of") and *la* (En. "at"). An idiom such as *da seama* (En. "realize") is also split into *da* (En. "give") and *seama* (En. "reckoning").

2.2.2. Annotation Guidelines for Named Entities

In the process of annotating biomedical named entities we focused on a subset of four types of entities that are defined based on four semantic groups comprised in the UMLS (Unified Medical Language System) [20].

The four top level entity classes are:

1. **Anatomy (ANAT)**: e.g., *inimă* ("heart"), *țesut epitelial* ("epithelial tissue"), *rinichi* ("kidney");
2. **Chemicals and Drugs (CHEM)**: e.g., *insulină* ("insulin"), *TNF alfa* ("TNF alpha");
3. **Disorders (DISO)**: e.g., *diabet zaharat* ("mellitus diabetes"), *insuficiență cardiacă* ("heart failure");
4. **Procedures (PROC)**: e.g., *coronarografie* ("coronarography"), *plasmaferează* ("plasmaferesis").

In order to prepare our data for future NLP tasks, the IOB2 format was chosen to represent all the named entity chunks contained in the corpus, where "B" denotes the beginning of an entity, "I" represents the continuation of an entity, "O" indicates none of the defined entities [21].

The annotation guidelines were as follows:

1. If a term designating an entity contains in its structure, another term, designating another entity (cascaded constructions [22]), annotate only the longer term. For example, *cancer de rinichi* ("kidney cancer") will be annotated with the DISO label and not as two separate entities, namely *cancer* ("cancer") as DISO and *rinichi* ("kidney") as ANAT.
2. In cases of coordination of two terms sharing (at least) one word, the second term may not contain the common word(s). Romanian is a head-first language and, thus, when two syntactic groups having the same word as head are coordinated, the head of the second term may be omitted: *țesut adipos și [țesut] muscular* ("adipose [tissue] and muscle tissue"). Annotation will be done as: *țesut/B-ANAT adipos/I-ANAT și muscular/I-ANAT*. Here is an example of an annotated construction with disjunctive conjunction: *anemie/B-DISO megaloblastică/I-DISO sau hemolitică/I-DISO* ("megaloblastic or hemolytic anemia").
3. Long distance dependencies must be annotated as contiguous terms. Consider the sentence *Hiposecreția poate fi endocrină (cu produși de secreție în sânge hormoni) sau exocrină (glande care produc și secretă substanțe în lumenul epitelial).* ("Hyposecretion can be endocrine (glands with products secreted in the blood - hormones) or exocrine (glands that produce and secrete substances onto an epithelial surface by way of a duct."). The terms that must be annotated are: *Hiposecreția endocrină* and *Hiposecreția exocrină*. Considering also the treatment of coordinated structures (see above), the sentence will be annotated as: *Hiposecreția/B-DISO poate fi endocrină/I-DISO (glande cu produși de secreție în sânge—hormoni) sau exocrină/I-DISO (glande care produc și secretă substanțe în lumenul epitelial).*
4. In cases of parenthetical constructions, "embedded" terms are annotated: in the sentence *Hiposecreția/B-DISO poate fi endocrină/I-DISO (glande/B-ANAT cu produși de secreție în sânge/B-ANAT—hormoni/B-ANAT) sau exocrină/I-DISO (glande/B-ANAT care produc și secretă substanțe în lumenul/B-ANAT epitelial/I-ANAT).* there are three annotated terms that occur between the B-DISO and the I-DISO of the term *Hiposecreția exocrină*: *glande*, *sânge*, and *hormoni*, all of them annotated as B-ANAT.
5. Anaphoric heads are not annotated. Whenever the syntactic head of a term is replaced by an anaphoric term (usually a pronoun), this is not analysed as part of the term; the annotation is similar to that recommended for coordinated structures: in *valvelor/B-ANAT semilunare/I-ANAT față de cele atrioventriculare/I-ANAT* ("to the semilunar valves as opposed to the atrioventricular ones") the second I-ANAT has the same head as the first one, namely *valvelor*.

Moreover, in order to annotate the corpus with named entities, the annotators were instructed to annotate only the most specific mentions of medical concepts. For example, in the sentence *Pacientul prezintă afecțiuni cardiovasculare aterosclerotice* (En. The patient has atherosclerotic cardiovascular diseases), *afecțiuni/B-DISO cardiovasculare/I-DISO aterosclerotice/I-DISO* was annotated as an entity which belongs to "disorders" entity class, but the more general term *afecțiuni* is left unannotated. The same approach was adopted by [23] for the annotation process of building a medical gold standard corpus annotated with named entities for English. Another common element with this work is that the discontinuous entities were taken into consideration and annotated: e.g., *Anevrismele/B-DISO pot fi fusiforme/I-DISO sau sacciforme/I-DISO* (En. The aneurysms may be fusiform or sacciform).

On the other side, the annotation principles adopted in our work differ from the approach used in the construction of the Quaero corpus [6], a medically named entity annotated corpus for French. The guidelines presented for Quaero corpus imply that a complex entity that can be decomposed into simpler entities which may belong to different entity classes (a specific entity whose span overlaps with that of another, more general, entity) should be annotated at both specific and general level. For example, "infarct de miocard" (en. myocardial infarction) should be annotated as "infarct/B-DISO de/I-DISO miocard/I-DISO" and "miocard" should also belong to the

"anatomy" entity class ("miocard/B-ANAT"). Instead, we annotated this entity only as "infarct/B-DISO de/I-DISO miocard/I-DISO", according to the most specific mention of the concept criterion.

2.2.3. Annotation Process for Part of Speech Tags

The corpus was automatically tokenized, part of speech annotated, and lemmatized with the T(okenizing, part of speech)T(agging and)L(emmatizing) Platform [9] in less than an hour. Each identified token in the corpus is morphologically annotated (i.e., its POS tag is identified) and then lemmatized (i.e., its base form is specified). TTL works with 715 POS tags, resulted from the possible combinations of the values of the attributes enumerated in Section 2.2.1. On the one hand, the high number of POS tags is suggestive of the task difficulty and, on the other hand, it leaves little space for ambiguity. This tagset has been used in all Romanian corpora created at our institute [15,24,25], thus a morphologic comparison among them is always possible.

The automatic tokenization, POS tagging and lemmatization were manually checked by a linguist with a 1000 tokens per hour pace. The percents of corrected errors for each task are included in Table 3. Tokenizing errors were mainly due to a wrong segmentation of intervals, such as periods between two years (e.g., 1992–1999), which were considered a single token by TTL. Another cause was represented by typos in the form of blanks inside a word (*nu măr* instead of *număr*) or inside a number (*1. 5* instead of *1.5*). Obviously, any correction of the tokens implies a correction of the lemma and, most of the times, also of the POS tag. Besides such cases (of tokenization mistakes), there were also cases of wrong lemmatization either in case of unknown words (i.e., medical terms that were not part of the lexicon used by TTL, for example *seroamă* instead of *serom*) or in case of morphologically ambiguous forms, given their wrong POS tagging: e.g., the feminine singular adjective *cronică* from the structure *etapa acută și post-acută, precum și cea cronică* ("the acute and post-acute stage, as well as the chronic one") is considered a noun (given the identity of form) and lemmatized as *cronică* (En. chronicle) instead of the adjectival form *cronic*. The errors in the POS tagging are discussed in Section 3.3 below.

Table 3. The percentages of corrected errors.

Task	Tokenizing	Lemmatizing	Tagging
Percentage	1.65%	2.39%	5.01%

2.2.4. Annotation Process for Named Entities

The annotation process was performed by two human annotators: one expert annotator and one medical domain expert. The annotation started with a training period. During the annotation phase, the annotators collaborated and discussed any annotation issue encountered. The analysis of the inter-annotators disagreement was used to regularly update the annotation guidelines according to the issues that appeared during the consensus sessions. For instance, due to the complexity of the entities and the lack of prior knowledge of biomedical specific vocabulary, the expert annotator sometimes missed out some parts of the entity or the entire entity or mis-categorized it. Whenever the expert annotator encountered difficulties in identifying or distinguishing among different types of entities, the existing medical terminologies were used. For a named entity annotation task, 1000 tokens were annotated in one hour due to the difficulties of the task and the corrections needed. In order to calculate the Inter Annotator Agreement, a fraction of the corpus was annotated separately by each annotator.

3. Results

3.1. Descriptive Statistics of the Corpus

After the annotation process, a corpus of 14,021 tokens (out of which 2018 are punctuation) resulted, distributed into 506 sentences annotated with POS tags and NEs (Table 4). The average

sentence length is quite big, namely 27.69 tokens/sentence. The number of unique lemmas is relevant for calculating the average frequency of each lemma, which is 4.19, irrespective of the type of word (functional or content), although it is widely known that functional words are more frequent than content ones.

Table 4. General statistics over MoNERo corpus.

# Tokens	# Unique Lemmas	# Punctuation	# Sentences	Tokens/Sentence	Punctuation/Sentence
14,012	2856	2018	506	27.69	3.98

The frequency of content words is presented in Table 5 for each part of speech: nouns are the most numerous, but adjectives come second; as also noted by [22], medical terms tend to be descriptive and the morphological reflexion of this fact is the high occurrence of adjectives in medical terms structure. Moreover, some medical terms include even several adjectives: *insuficiență renală cronică* ("chronic renal insufficiency"), *mușchi drepți abdominali* ("rectus abdominis muscles"), *necroze subutanată tisulară* ("subcutaneous tissue necrosis"), etc.

Table 5. Frequencies of content words. N = noun, V = verb, A = adjective, R = adverb.

Tag	N	V	A	R
Number of occurrences	3799	1486	1564	564

Table 6 shows the types and the number of NE annotations in the corpus. The most frequent NE categories are CHEM and DISO, while PROC is rare. This distribution is a consequence of the fact that disorder and medicine names are more frequent in journal articles and medical blog posts. The information provided by this table is very useful in terms of balance if one trains an automatic NER system on this corpus.

Table 6. NEs distribution in MoNERo.

Type	B-DISO	I-DISO	B-ANAT	I-ANAT	B-CHEM	I-CHEM	B-PROC	I-PROC
# of occurrences	399	305	198	127	528	200	67	53

As far as the distribution of NEs over the three medical domains is concerned, we notice in Table 7 the predominance of DISO in cardiology and endocrinology, in both of which the number of PROC is quite low. On the other side, diabetes is dominated by CHEM, while the number of PROC is low as well here. The cardiology field includes many heart diseases (DISO-arrhythmias, cardiac failure, hypertension or myocardial infarction), having different options of pharmacological (CHEM) or interventional treatment (PROC), while most of the anatomical terms are used to explain mainly physiopathological mechanisms. The diabetes category deals particularly with molecular processes (CHEM), such as insulin's mechanisms of action and the effects of different types of treatments, whereas terms tagged as ANAT or DISO are used almost with the same frequency because in diabetes, one deals with the disorder of one anatomical structure (pancreas). Endocrinology follows the same pattern as cardiology, due to the existence of multiple endocrine glands pathologies, with chemical structures (CHEM) that include pharmacological treatment or active molecules, i.e., hormones.

Table 7. NE's distribution differentiated by medical domains.

	ANAT	CHEM	DISO	PROC
Cardiology	43	84	131	36
Diabetes	163	490	187	21
Endocrinology	119	154	386	63

The majority of medical NEs are formed of more than one token, and a few of them even have more than three tokens (see Figure 1).

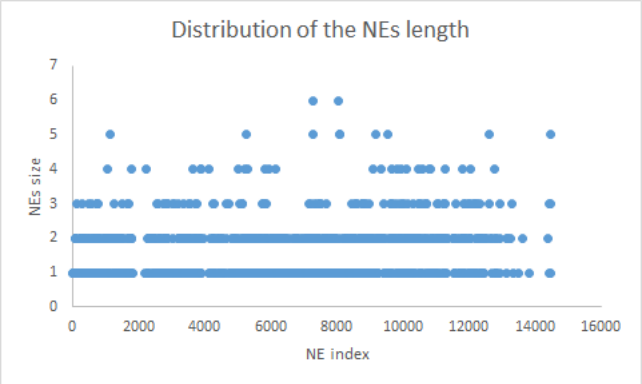

Figure 1. Distribution of the NEs length.

As can be seen in Table 8, the CHEM category has the shortest NEs and the PROC category has the longest ones. The average span of a NE is very relevant in the process of selecting the features for the NER systems due to the fact that the context window size is different in the case of compact NEs compared with the case of long-range NEs.

The statistics reported in the paper were calculated on a previous version of the corpus. Since the corpus is still under development, we uploaded on our website a newer version of the corpus. The reproducibility of some of the statistics will be possible at the end of this process, when we will will release the final version of the corpus.

Table 8. Average size and Standard Deviation of NEs.

Tag	Average	Stdev.
DISO	1.78	0.91
ANAT	1.67	0.69
PROC	1.80	0.97
CHEM	1.39	0.68

3.2. Measurement of Inter-Annotator Agreement

In order to establish the consistency of the annotation, the inter-annotator agreement was computed between the two annotators of NEs on a sample of 1628 tokens (more than 10%) of the corpus. The sample of the corpus used to calculate the inter-annotator agreement was randomly chosen. The measurement of the reliability of the data was done by computing the Cohen Kappa [26,27] coefficient, which is defined as computing the observed agreement $A0$ and the agreement expected by chance Ae:

$$K = \frac{A0 - Ae}{1 - Ae}, \quad (1)$$

The observed agreement $A0$ is the proportion of times the annotators actually agree. The agreement expected by chance Ae is the proportion of times the annotators are expected to agree due to chance. For the sample selected, the Cohen Kappa coefficient was 92.8%, which shows a very high agreement, considering the fact that for most NLP classification tasks a relevant agreement lies between 70% and 80% [27]. For example, for Quaero corpus [6], the authors reported an agreement on entities at the mention and category level for two sets which had 92% agreement.

3.3. Annotation Challenges for Part of Speech Tagging

POS tagging errors were the most numerous (see Table 3). A linguist, with high experience in working with the MULTEXT-East specifications for Romanian, went through the whole corpus and made all the necessary corrections manually. Two major types of errors were found: errors of part of speech and errors of morphological categories. The former implies the wrong identification of the part of speech: for example, a word is analyzed as an adjective instead of an adverb. In Table 9 we show the frequent confusions between parts of speech. Cells are left empty when no confusions of that kind were found in the corpus or the number of cases was very low. *n.a.* stands for *not applicable*; the first column contains the parts of speech that were wrongly identified by TTL, while the first row contains the correct parts of speech, as introduced by the annotator. Most frequently such type of error affects the nouns, then the adjectives and verbs. This type of errors rarely affects adverbs and pronouns, thus they were not added to the table rows. Nouns are usually confused with adjectives (in 10 cases), adjectives with adverbs (13 cases), verbs with adjectives (12 cases) and adverbs with adjectives (52 cases). Here are some examples of these types:

- nouns wrongly annotated as adjectives: *Alterarea* ***podocitelor*** *este implicată în patogeneza...* (En. The alteration of **podocytes** is involved in the pathogenesis...)
- adjectives wrongly annotated as adverbs: *Un studiu* ***recent*** *a arătat că...* (En. A **recent** study showed that...)
- verbs wrongly annotated as adjectives: *... vor trebui* ***upgradate*** *ulterior sau chiar* ***înlocuite*** (En. ... will have to be eventually **updated** or even **replaced**)
- adverbs wrongly annotated as adjectives: *S-a demonstrat recent* ***experimental****...* (En. It has recently been proved **experimentally**...)

The confusion between adjectives and adverbs can be explained through the homonymy between the two parts of speech: their lemma is identical. However, the homonymy is partial: while the adverb does not inflect, the adjective inflects for gender, number, case and may take the definite article and only its masculine singular form is identical with the adverb. The confusion between adjectives and nouns can be explained through the fact that there are frequent cases of zero derivation from adjectives to nouns and the fact that an adjective, especially one unknown to the TTL tool, occurs in prenominal position and takes the definite article, the definiteness being specific to nouns.

The second type of POS tagging errors concerns the morphological categories; the part of speech is correctly identified but some of its characteristics are wrong. This is due to the homonymous forms in the inflection paradigm of the respective words. In the case of nouns there are confusions concerning their gender, number and cases (which are correlated with each other), but also definiteness, although extremely rarely. With verbs the most frequent such type of error concerns the tense: the present of some verbs is mistaken for their past simple, which is a homograph of the present form for some persons. Whenever a verb displays this homonymy of forms, the tagger annotates its forms as being past tense ones, thus offering a consistent annotation. However, we cannot find an explanation for this preference of the tagger.

Most of these types of errors are specific to the language in general, they are independent of the domain to which a text belongs. The only ones that have a higher frequency than in general language corpora are those involving the categories X and Y. The former is attributed to residual words in the corpus, i.e., strings made up of alphanumeric characters or foreign words. Such tokens are quite frequent in the medical corpus: *T1DM*, *Ca2*, *anti-CD38*, *T0*, etc. Sometimes, foreign words (mainly English ones) are used, probably because of the absence of an equivalent Romanian one or because the foreign one is shorter than the existing Romanian one: *tehnicilor tension-free* ("tension-free techniques"), *testarea cognitivă Mini Mental State* ("Mini Mental State examination"), etc. Class Y is the class of abbreviations. In Table 9, one can see that many abbreviations were misanalysed as nouns (more exactly as proper nouns, given the use of capitals in their written form) by TTL and corrected by the human annotator.

Table 9. Confusion matrix for parts of speech.

	Noun (N)	**Adjective (A)**	**Verb (V)**	**Adverb (R)**	**Residual (X)**	**Abbreviation (Y)**
Noun (N)	n.a.	18	2	4	39	**80**
Adjective (A)	10	n.a.	12	**52**	6	
Verb (V)	3	**6**	n.a.			
Adverb (R)		**13**		n.a.		

4. Discussion

MoNERo is a small-sized corpus; the number of tokens is low, as well as the number of medical subdomains represented in it and, consequently, their specific terminology. We are aware that this reduced dimension has an impact on the corpus utility in different bio-NLP tasks, such as named entity recognition, relation extraction, term extraction. We plan to release the final version of the MoNERo corpus in the near future. The dimension of that version will be enough to train and test different systems designed for named entity recognition and different information extraction tasks.

However, as proved by Névéol et al. [28], there is still the need for annotated medical corpora in languages other than English in order to foster development in the biomedical domain. We report here only our first step towards the creation of such a resource for Romanian.

5. Conclusions

This paper describes the construction of a resource that provides gold standards for NLP tasks (POS tagging and NER) in Romanian, thus making it one of the few languages, besides English, for which such a resource exists. The corpus is freely available [29] to those interested. The MoNERo corpus can also contribute to the biomedical terminology development for the Romanian language. The quality of the annotations is very high, with a Kappa coefficient of 92.8% for the manual NEs annotation task.

As future work, we mention its extension in size, so as to make it useful for complex NLP tasks. We also envisage adding a further level of annotation to it, the syntactic one, once the Romanian parser [30] performs well enough. These new versions will also be made freely available.

Author Contributions: Conceptualization, M.M. and V.B.M.; methodology, M.M. and V.B.M.; software, M.M.; validation, M.M., V.B.M. and G.M.; formal analysis, M.M., V.B.M. and G.M.; investigation, M.M., V.B.M. and G.M.; resources, M.M., V.B.M. and G.M.; data curation.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pakhomov, S.; Coden, A.; Chute, C. Creating a test corpus of clinical notes manually tagged for part-of-speech information. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, 28–29 August 2004.
2. Islamaj, D.R.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10.
3. Lee, K.; Lee, S.; Park, S.; Kim, S.; Kim, S.; Choi, K.; Tan, A.C.; Kang, J. BRONCO: Biomedical entity relation oncology corpus for extracting gene-variantdisease-drug relations. *J. Biol. Databases Curation* **2016**. [CrossRef] [PubMed]
4. Verspoor, K.; Yepes, A.J.; Cavedon, L.; McIntosh, T.; Herten-Crabb, A.; Thomas, Z.; Plazzer, J.P. Annotating the biomedical literature for the human variome. *J. Biol. Databases Curation* **2013**. [CrossRef] [PubMed]
5. Boytcheva, S.; Nikolova, I.; Paskaleva, E.; Angelova, G.; Tcharaktchiev, D.; Dimitrova, N. Extraction and exploration of correlations in patient status data. In Proceedings of the Workshop on Biomedical Information Extraction, Borovets, Bulgaria, 14–16 September 2009; pp. 1–7.

6. Névéol, L.A.; Jeremy, C.G.; Rosset, S.; Zweigenbaum, P. The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization. Available online: http://nactem.ac.uk/biotxtm2014/papers/Neveoletal.pdf (accessed on 23 November 2018).
7. Moreno, I.; Moreda, E.; Romá-Ferri, M.T. DrugSemantics: A corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics. *J. Biomed. Inform.* **2017**, *72*, 8–22. [CrossRef] [PubMed]
8. Brants, T. TnT: A statistical part-of-speech tagger. In Proceedings of the Sixth Conference on Association for Computational Linguistics Applied Natural Language, Washington, DC, USA, 29 April–4 May 2000; pp. 224–231.
9. Ion, R. Word Sense Disambiguation Methods Applied to English and Romanian. Ph.D. Thesis, Romanian Academy, Bucharest, Romania, 2007. (In Romanian)
10. Mitrofan, M.; Ion, R. Adapting the TTL Romanian POS Tagger to the Biomedical Domain. In Proceedings of the Biomedical NLP Workshop Associated with RANLP, Varna, Bulgaria, 4–6 September 2017.
11. Doğan, R.I.; Zhiyong, L. An improved corpus of disease mentions in PubMed citations. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montréal, QC, Canada, 8 June 2012.
12. Krallinger, M.; Rabal, O.; Leitner, F.; Vasquez, M.; Salgado, D.; Lu, Z.; Sayle, R.A. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.* **2015**, *7*, S2. [CrossRef] [PubMed]
13. Pyysalo, S.; Ginter, F.; Heimonen, J.; Björne, J.; Boberg, J.; Järvinen, J.; Salakoski, T. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinform.* **2007**, *8*, 50. [CrossRef] [PubMed]
14. György, M.; Vincze, V. Joint Part-of-Speech Tagging and Named Entity Recognition Using Factor Graphs. In Proceedings of the International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 3–7 September 2012.
15. Barbu Mititelu, V.; Tufiș, D.; Irimia, E. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In Proceedings of the 11th Language Resources and Evaluation Conference-LREC, Miyazaki, Japan, 7–12 May 2018.
16. Mitrofan, M.; Tufiș, D. BioRo: The Biomedical Corpus for the Romanian Language. In Proceedings of the 11th edition of the Language Resources and Evaluation Conference, Miyazaki, Japan, 7–12 May 2018.
17. Erjavec, T. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages, Language Resources and Evaluation. *Lang. Resour. Eval.* **2012**, *46*, 131–142. [CrossRef]
18. Tufiș, D.; Barbu, A.M.; Pătrașcu, V.; Rotariu, G.; Popescu, C. Corpora and Corpus-Based Morpho-Lexical Processing. In *Recent Advances in Romanian Language Technology*; Romanian Academy Publishing House: Bucharest, Romania, 1997; pp. 35–56.
19. Tufiş, D. Tiered tagging and combined language models classifiers. In *International Workshop on Text, Speech and Dialogue*; Springer: Berlin, Germany, 1999; pp. 28–33.
20. Available Online: https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt (accessed on 4 April 2018).
21. Sang, E.F.; Veenstra, J. Representing text chunks. In Proceedings of the Ninth Conference on European Chapters of the Association for Computational Linguistics, Bergen, Norway, 8–12 June 1999.
22. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvist. Investig.* **2007**, *30*, 3–26.
23. Deleger, L.; Li, Q.; Lingren, T.; Kaiser, M.; Molnar, K.; Stoutenborough, L.; Kouril, M.; Marsolo, K.; Solti I. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2012; Volume 2012, p. 144.
24. Tufiș, D.; Irimia, E. RoCo-News—A Hand Validated Journalistic Corpus of Romanian. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 22–28 May 2006.
25. Ion, R.; Irimia, E.; Ștefănescu, D.; Tufiș, D. ROMBAC: The Romanian Balanced Annotated Corpus. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey, 23–25 May 2012.
26. Poessio, M.; Vieira, R. A corpus-based investigation of definite description use. *Comput. Linguist.* **1998**, *24*, 183–216.
27. Carletta, J. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.* **1996**, *22*, 249–254.
28. Névéol, A.; Dalianis, H.; Velupillai, S.; Savova, G.; Zweigenbaum, P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *J. Biomed. Semant.* **2018**, *9*, 12. [CrossRef] [PubMed]

29. Available Online: http://slp.racai.ro/index.php/resources/monero-3/ (accessed on 5 April 2018).
30. Ion, R.; Irimia, E.; Barbu Mititelu, V. Ensemble Romanian Dependency Parsing with Neural Networks. In Proceedings of the LREC, Miyazaki, Japan, 7–12 May 2018; ELRA: Miyazaki, Japan, 2018.