MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language

Maria Mitrofan
RACAI, Bucharest, Romania
maria@racai.ro

Verginica Barbu Mititelu
RACAI, Bucharest, Romania
vergi@racai.ro

Grigorina Mitrofan
NIDMD "N.C. Paulescu"
Bucharest, Romania
grigorinam@gmail.com

Abstract

In an era when large amounts of data are generated daily in various fields, the biomedical field among others, linguistic resources can be exploited for various tasks of Natural Language Processing. Moreover, increasing number of biomedical documents are available in languages other than English. To be able to extract information from natural language free text resources, methods and tools are needed for a variety of languages. This paper presents the creation of the MoNERo corpus, a gold standard biomedical corpus for Romanian, annotated with both part of speech tags and named entities. MoNERo comprises 154,825 morphologically annotated tokens and 23,188 entity annotations belonging to four entity semantic groups corresponding to UMLS Semantic Groups.

1 Introduction

Natural Language Processing (NLP) is a research area that provides methods to convert (human-understandable) unstructured textual information into (machine-readable) structured data and uses it for different objectives. NLP techniques can be used to process and exploit the large amount of biomedical information which is continuously generated. Examples of such repositories are MEDLINE[1], which contains more than 25 million documents belonging to the biomedical domain, or PubMed Central[2], which is an archive of biomedical journal literature and contains more than 5 million full-text articles. These resources can be exploited and used together with different NLP systems previously adapted to the biomedical field to improve the quality of the health care process, to further develop research in the field and benefit both physicians and patients. Information Extraction (IE) tools can be used to extract relevant information from biomedical textual resources (Goeuriot et al., 2017; Li et al., 2017). Reaching suitable results for this NLP subtask is not trivial and there is still room for improvement of results. Advances of these IE tools depend on the existence of annotated resources specific to the field of study (Wilbur et al., 2006; Thompson et al., 2009; Kilicoglu, 2017), annotated corpora being relevant in both phases: development of the models that will determine the behaviour of the system and system performance evaluation. Even though the availability of these resources has increased lately, the main part of the efforts have been directed to the development of annotated corpora for English in different subdomains. However, MoNERo is a resource created for the Romanian language that helps the development of named entity recognition and classification task especially for this language. Romanian benefits from the existence of other corpora created in our institute: the representative corpus of contemporary language (CoRoLa) (Barbu Mititelu et al., 2018), a balanced corpus (ROMBAC) (Ion et al., 2012), the corpus annotated with verbal multiword expressions (Barbu Mititelu et al., 2019). Just like all of these, MoNERo is annotated at the morphological level. However, it stands out given its annotation with four types of Named Entities (NEs) for the medical domain, which are relevant to the identification of: anatomy parts, diseases and disorders, chemicals and drugs, and medical procedures.

This paper has four main objectives: (i) to present the construction of a biomedical gold standard corpus annotated both with part-of-speech tags and named entities; (ii) to present general statistics over the corpus; (iii) to release the final

[1] https://www.nlm.nih.gov/bsd/medline.html

[2] https://www.nlm.nih.gov/bsd/medline.html

Proceedings of the BioNLP 2019 workshop, pages 71–79
Florence, Italy, August 1, 2019.

version of the corpus to the scientific community, (iv) to show the contribution in the development of NLP tools for Romanian language. All the results are discussed in parallel for the two types of annotations.

2 Related Work

This section reviews relevant corpora annotated with NEs specific to the biomedical domain.

1. For English we mention:
   - CLEF corpus (Roberts et al., 2009) – it contains 150 documents of clinical narratives, histopathology reports and imaging reports. It was subtracted from a corpus of 565,000 documents and manually annotated with six types of NEs (condition, intervention, investigation, result, drug or device, locus);
   - i2b2 corpus (Uzuner et al., 2010) – it contains 1243 discharge summaries automatically pre-annotated, out of which a subset of 251 was manually revised. This corpus contains seven types of NEs (medications, dosages, modes, frequencies, durations, reasons of administration, list/narrative);
   - NCBI corpus (Doğan et al., 2014) – a gold-standard corpus for disease mentions and concepts that contains 793 abstracts extracted from PubMed;
   - CHEMDNER corpus (Krallinger et al., 2015) – a corpus of 10,000 abstracts collected from PubMed annotated with two types of NEs: chemicals and drugs.
2. For French there is the Quaero corpus (Névéol et al., 2014) which contains 103,056 words collected from three types of documents: texts with information on drugs extracted from European Medicines Agency (EMEA), titles from research articles comprised in MEDLINE and patents. This corpus was annotated with ten types of NEs defined using UMLS: anatomy, chemical and drugs, devices, disorders, geographic areas, living beings, objects, phenomena, physiology, procedures.
3. For Spanish the following corpora exist:
   - IxaMedGS corpus (Oronoz et al., 2015) – it is composed of 142,154 discharge records out of which 75 were annotated with two types of NEs: diseases and drugs;
   - DrugSemantics corpus (Moreno et al., 2017) – it has 226,729 tokens annotated with ten types of NEs: chemical composition, disease, drug, excipient, food, medicament, pharmaceutical form, route, therapeutic action, and unit of measurement.

All these corpora are available and have had a significant role in information extraction research, especially in named entity recognition (NER) research and were developed for well-established purposes, having in mind the possibility of reusability.

3 Corpus Development Description

3.1 Selection of Corpus Documents

The gold standard morphologically and named entity annotated Romanian medical corpus (MoNERo) was extracted from the BioRo corpus (Mitrofan and Tufiş, 2018), a Romanian biomedical corpus. MoNERo contains texts extracted from three types of documents: scientific medical literature books, scientific medical journal articles and medical blog posts, but predominant are those coming from medical literature. The medical books were chosen as the main source because they contain descriptive materials, full of domain-specific terms. In addition, the texts are of good quality and the use of medical terms is correct. The medical journal[3] from which a part of the texts were extracted is a scientific journal that addresses the specialists, so the language used is specific to the medical domain. In the case of blog posts those collected were texts of popularization and awareness of various medical problems.

The texts were selected so that they belong to three medical subdomains: cardiology, diabetes and endocrinology (see table 3). The main motivation behind choosing these three medical domains is that our textual resources available were centered around the pathology of Diabetes. Since Diabetes is an endocrine disorder it is naturally included in the Endocrinology category. In the same

[3] https://rmj.com.ro/

time because of a very close relation between diabetes and cardiovascular diseases we also obtain a significant category from Cardiology field. Other categories such as neurology, nephrology would have had a very low contribution and we chose not to take them separately but in Diabetes field, because the terms were related to diabetes complications.

The selection was made based on the metadata scheme associated with each document present in the BioRo corpus. The order of the sentences was preserved.

All these texts are Intellectual Property Right (IPR) cleared, thus enabling us to make it available to the community (see section 6).

3.2 MoNERo Annotation Scheme

The annotation scheme of MoNERo has two different levels: (i) a morphologic level at which all part of speech tags were revised by an experienced linguist; and (ii) a named entity level at which NEs were identified and classified in the corresponding semantic group.

3.2.1 Part of Speech Annotation Scheme

The process of the annotation of the corpus with part of speech tags had two phases: automatic annotation (all the texts comprised in this corpus were previously processed when included in BioRo, the source from which MoNERo was extracted) and manual verification of the tags allocated by the tool used (see below section 3.3.1). Here we present the manual verification phase which was done by an expert linguist. The annotation scheme used for morphologic annotation was based on the MSD tagset developed in the Multext-East project (Dimitrova et al., 1998), which contains 715 tags for Romanian. This tagset is very complex and precise, containing fourteen classes of words (noun, verb, adjective, adverb, pronoun, determiner, article, adposition, conjunction, numeral, interjection, abbreviation, residual and particle), each class having a set of attributes such as: type, gender, number, case, definiteness, clitic, verb form, tense, person, degree, etc. (Tufiş et al., 1997).

3.2.2 Named Entities Annotation Scheme

In the case of named entities identification the annotation scheme was based on UMLS[4] (Unified Medical Language System) semantic groups. This resource contains concepts from different terminologies specific to the biomedical domain. Moreover, UMLS is organized as a hierarchical semantic network that comprises semantic types and semantic relations. All the semantic types are grouped in 15 semantic groups (McCray et al., 2001). For this work the annotation scheme contains four semantic groups chosen from the UMLS scheme: anatomy, chemicals and drugs, disorders and procedures. The attributes of each entity type are described below:

1. **Anatomy (ANAT)**: body location or region, body part, organ, or organ component, body substance, body system, cell, fully formed anatomical structure, tissue;

2. **Chemicals and Drugs (CHEM)**: amino acid, peptide, protein, antibiotic, biologically active substance, chemical, clinical drug, hormone, organic chemical, pharmacologic substance, receptor, steroid, vitamin;

3. **Disorders (DISO)**: acquired abnormality, anatomical abnormality, cell or molecular dysfunction, congenital abnormality, disease or syndrome, experimental model of disease, finding, injury or poisoning, sign or symptom;

4. **Procedures (PROC)**: diagnostic procedure, health care activity, laboratory procedure, molecular biology research technique, therapeutic or preventive procedure.

Examples for each type can be seen in Table 1.

Named Entity	Example
Anatomy	*pancreas* (“pancreas”) *nerv optic* (“optic nerve”)
Chemicals and Drugs	*paracetamol* (“paracetamol”) *acid folic* (“folic acid”)
Disorders	*diabet* (“diabetes”) *fibrilaţie* (“fibrillation”)
Procedures	*EKG* (“EKG”) *CT* (“CT”)

Table 1: Examples of named entities extracted from MoNERo.

[4] https://semanticnetwork.nlm.nih.gov/

The main reason for choosing these four types of entities was a trade off between the minimum number of entities (due to an increased complexity of the annotation process) and the maximum relevance for our corpus. However we had some challenges. For example, Physiology was a category that could be included, but due to the fact that the medical texts available were mainly related to pathology, the contribution would have been limited (less than 5%).

Having a tokenized corpus with each token on a separate line, we chose IOB2 (Insid-Outside-Beginning) (Sang and Veenstra, 1999) as the annotation format for named entities. Lately, this format has become popular within the scientific community, being also supported by the CoNLL challenges [5]. The B-tag is used for the first token of every NE, I-tag indicates the token that is inside a named entity and O-tag is used for surrounding tokens that do not belong to a NE (*Ân/O schimb/O, HDL-colesterolul/ B-CHEM, apolipoproteinele/B-CHEM A/I-CHEM şi/O B/I-CHEM sunt/O superiori/O ca/O indicatori/O de/O risc/B-DISO cardiovascular/I-DISO ./O*) ("On the other hand, HDL-cholesterol and lipoproteins A and B are superior as cardiovascular risk indicators."). For ease of reading, in all the examples below we chose not to mention the O-tag, but only the B- and I-tags.

3.3 Annotation Guidelines

3.3.1 Part of Speech

In the initial phase the corpus was automatically preprocessed (sentence split, tokenized, lemmatized) and annotated with POS tags using the TTL annotator (Ion, 2007; Mitrofan and Tufiş, 2018), which was trained on news corpora of about 200,000 tokens with POS labeling checked by trained linguists (Tufiş, 2000). The accuracy for this task was 98.23%. When TTL was trained in order to perform domain adaptation for biomedical domain the accuracy was 97.83% (Mitrofan and Ion, 2017). Therefore, in order to annotate this corpus with POS tags the baseline model was chosen. The second phase of the annotation process, which makes the focus of this paper, was to manually check all the automatically assigned labels. A trained and experienced linguist revised all the tokens included in MoNERo. For this task the guidelines were:

1. correct the token if needed;
2. correct the lemma if needed;
3. correct the POS tag if needed;
4. compounds written as separate words should be split.

3.3.2 Named Entities

The guidelines for named entity annotation were:

1. a complex entity will not be decomposed into simpler entities belonging to different semantic groups; only one semantic group will be associated to the longest entity (*cancer de ficat* ("liver cancer") will be annotated only as a disorder, not as a disorder (*cancer/B-DISO de/I-DISO ficat/I-DISO* ("cancer"/B-DISO) "of"/I-DISO "liver"/I-DISO) and an anatomical part (*ficat/B-ANAT* ("liver"/B-ANAT)); so, there is no embedded annotation;
2. in cases when one head noun is shared by two or more biomedical named entities (coordinations or disjunctions) the annotation will be done as follows: in case of coordinations *ateroscleroza aortei şi a vaselor periferice* ("atherosclerosis of the aorta and peripheral vessels"), should be annotated as *ateroscleroza/B-DISO aortei/I-DISO şi vaselor/I-DISO periferice/I-DISO* or in case of disjunctions *celule beta pancreatice sau hepatice* ("pancreatic beta or hepatic cells") should be annotated as *celule/B-ANAT beta/I-ANAT pancreatice/I-ANAT sau hepatice/I-ANAT*);
3. discontinuous entities will be annotated as contiguous terms and classified in the same semantic group: in the examples *Anevrismele/B-DISO pot fi fusiforme/I-DISO (aspect cilindric al vasului/B-ANAT sangvin/I-ANAT) sau sacciforme/I-DISO* ("Aneurysms/B-DISO may be fusiforms/I-DISO (cylindrical appearance of the blood/B-ANAT vessel/I-ANAT) or sacciforms/I-DISO") the NEs *Anevrismele fusiforme* and *anevrismele sacciforme* are discontinuous;
4. in case of cascaded constructions when one entity is incorporated in another entity (eg. parenthetical constructions) the annotation will be done as: *Anevrismele/B-DISO pot fi fusiforme/I-DISO (aspect cilindric al vasului/B-ANAT sangvin/I-ANAT) sau*

[5] http://www.conll.org/previous-tasks

sacciforme/I-DISO (“Aneurysms/B-DISO may be fusiforms/I-DISO (cylindrical appearance of the blood/B-ANAT vessel/I-ANAT) or sacciforms/I-DISO”). Within the discontinuous NE *Anevrismele sacciforme* there is another NE, *vasului sangvin*.

3.4 Annotation Development

3.4.1 Part of Speech Tags

Even though the accuracy of the automatic annotation with POS tags was very high (subsection 3.3.1), given the high number of POS tags in the Romanian MSD tagset, there was a lot of manual work to be done by the linguist. This task involved manual validation of tokenization, lemmatization, and also correcting the errors of part of speech and errors of morphological categories (see 3.5.1) for each token.

3.4.2 Named Entities

For the named entities annotation task two annotators were employed: one physician and one experienced annotator, both having Romanian as native language. The physician was chosen as annotator due to her capacity of understanding the medical field. Prior to the annotation process there was a training period for both annotators. In this phase they debated issues such as whether or not to annotate overlapping terms, when and if complex terms should be decomposed, how conjunctions should be treated.

Even though the initial guidelines gave them instructions on what should and should not be annotated, they collaborated and discussed throughout the annotation process. Even if the identification of a biomedical entity was a relatively easy task, fitting it into the correct semantic group sometimes required prior knowledge of the biomedical vocabulary. Therefore the experienced annotator has accessed various terminological resources in order to better understand the terms and to categorized them into the correct semantic group. In a post-annotation phase, the two annotators discussed the annotation differences in order to reach agreement.

3.5 Discussion Over the Annotation Process

3.5.1 Part of Speech

During the manual correction process of the part of speech tags the annotator encountered several types of errors generated by the tool used:

1. tokenization errors: wrong segmentation of time intervals (*2000-2001*) was annotated as a single token), typos that led to wrong tokenization of the word (*fi cat* instead of *ficat* (“liver”));

2. lemmatization errors: in case of the unknown words (*adenoamă* instead of *adenom* (“adenoma”)) or in case of morphologically ambiguous forms: the form *copii* can be the plural indefinite of either the masculine noun *copil* (“child”) or of the feminine noun *copie* (“copy”); given this homography, the lemmatizer mistakes one of the words with the other one;

3. tagging errors where classified in two categories:

    - errors of part of speech – wrong automatic identification of the part of speech (nouns as adjectives, adjectives as adverbs and vice versa, verbs as adjectives);
    - wrong identification of the morphological class – the part of speech is correctly identified but some of the specifications are wrongly identified: gender, number, case, etc.

Even though the overall error rate of the tool used was low (1.77% see section 3.3.1) and pre-annotation with POS tags of the corpus was useful, the task of correcting it was a difficult one due to the complexity of the tag set and the laborious manual work needed to determine if the token, lemma and POS tag are correct for each word in the corpus. Annotation time ranges between 17 tokens per minute (at the beginning of the task) and 33 tokens per minute (after the annotator became accustomed with the task and the types of errors). The use of only one annotator for correcting the POS tags is justified, on the one hand, by the low error rate and, on the other, by the expense of the task. However, we are aware of the limitation represented by the lack of inter-annotator agreement measurements (even on a sample) on the morphological annotation.

3.6 Named Entities

The task of annotating the corpus with named entities had an increased difficulty due to several factors such as:

- the need to understand specialized terminology. Several cases can be identified here:
  - completeness of NEs: given the lack of expertise in the biomedical domain, the expert annotator sometimes omitted components of the complex entities, thus attributing the NE a wrong class;
  - ambiguity: both annotators needed to agree upon the cases when to annotate conjunctions present in some entities: for example, although in the vast majority of cases, the conjunction *şi* ("and") is not part of an NE, there are a few cases when it is: one such example is the NE *ocluzia/B-DISO arterelor/I-DISO mici/I-DISO şi/I-DISO mijlocii/I-DISO* ("occlusion of small and medium sized arteries") in which the conjunction *şi* is part of the entity (see its annotation as *I-DISO*) and does not get unannotated as in an example such as *ateroscleroza/B-DISO aortei/I-DISO şi vaselor/I-DISO periferice/I-DISO* ("atherosclerosis of aorta and peripheral vessels");
  - abbreviations: this challenge was encountered especially by the experienced annotator. It is known that biomedical literature is very rich in abbreviations (Federiuk, 1999). Unless their meaning is clear to the annotator, a wrong type can be assigned to it. What is more, many abbreviations are difficult to correctly classify because of their multiple meanings. For example, depending on the context, *ACE* can be *angiotensin convertază* ("angiotensin-converting enzyme") and it belongs to "Chemicals and Drugs" semantic class or *electroforeză capilară de afinitate* ("affinity capillary electrophoresis") and in this case it is correctly labeled as "Procedure"; notice also that the abbreviation is borrowed from English, thus posing challenges to the annotator lacking medical background;
- four different entities types.

  Annotating all relevant entities was itself a challenge. One reason for this is the lack of prior knowledge of biomedical terminologies by the experienced annotator, some of the terms encountered not being covered in the terminological resources used for this task or being present with other senses than the one needed;
- the use of IOB2 format, which is an elaborated type of annotation format.

Estimated annotation time for this task was about 15 tokens per minute (for the experienced annotator) and 30 tokens per minute (for the physician).

The consistency of the annotations was established computing the (Carletta, 1996) coefficient on a sample of 1,628 tokens annotated by the two annotators, especially for this, after they finished the annotation. For this set the Kappa coefficient was 92.8%, denoting high agreement between the two annotators and indicates that the annotation was reliable.

4 Corpus General Statistics

Table 2 presents general corpus statistics offering an overview of the MoNERo corpus. Currently it contains 154,825 tokens (including the punctuation) distributed in 4,989 sentences, all of them annotated with POS tags and NEs. It can be seen that the average sentence length, 31 tokens/sentence, is above 16.06 tokens/sentence, the average sentence length in a balanced Romanian corpus, containing legal, news, medical (i.e. pharmacological), fiction and biographical texts (Ion et al., 2012).

Sentences	4,987
Tokens	154,825
Tokens/Sentence	31.04
Punctuation	20,741
Punctuation/Sentence	4.15

Table 2: MoNERo statistics.

Table 3 presents the distribution of sentences across the domains addressed. As can be seen, the distribution of sentences is not balanced, this being the result of the fact that due to copyright restrictions, the same number of sentences could not be collected for each of the selected domains, especially in the case of endocrinology.

Table 4 presents the distribution of content words. As can be seen, nouns are the most frequent ones, followed by adjectives: medical literature (especially medical literature books) has a

Domain	#tokens	#sentences
Cardiology	63,043	2,028
Diabetes	69,085	2,136
Endocrinology	22,697	823
Total	**154,825**	**4,987**

Table 3: Distribution of corpus sentences corresponding to each medical field.

descriptive structure, there are cases when nouns are modified by two or more adjectives: *bronşită cronică obstructivă* ("chronic obstructive bronchitis"). We notice quite an important number of abbreviations: the scientific subcomponent of the Romanian reference corpus, CoRoLa (Barbu Mititelu et al., 2018), contains 1.16% abbreviations, whereas MoNERo contains 1.9%. In the medical domain, as opposed to other scientific domains, it is common practice to designate concepts by abbreviated forms.

Tag	Percentage
Noun	27.8%
Verb	10.4%
Adjective	11.5%
Adverbs	3.5 %
Abbreviations	1.9 %
Total	55.1 %

Table 4: Percentages of content words.

Table 5 presents the distribution of entity annotation over each of the four semantic groups. This table highlights the fact that the most frequent NE categories are CHEM and DISO, PROC and ANAT being less frequent.

NE type	No. of entities
ANAT	1,964
CHEM	4,156
DISO	6,611
PROC	1,402
Total	**14,133**

Table 5: NEs distribution.

5 Corpus format

The corpus is available in a tabular format that contains four columns, UTF-8 encoded, with LF character as line break. Each line contains annotations of a token in four fields separated by a tab character: word form or punctuation symbol (token), lemma of the word form, NER tag and POS tag. We show below the annotation of the sentence: *Abordul arterei iliace comune se face retroperitoneal, iar grefonul folosit este unul sintetic din Dracon sau PTFE.* ("The access to the common iliac artery is retroperitoneal, and the graft used is a synthetic one from Dracon or PTFE.")

```
Abordul abord B-PROC Ncmsry
arterei arter I-PROC Ncfsoy
iliace iliac I-PROC Afpfson
comune comun I-PROC Afpfson
se sine O Px3--a--------w
face face O Vmip3s
retroperitoneal retroperitoneal O
Rgp
, ,O COMMA
iar iar O Rc
grefonul grefon O Ncmsry
folosit folosit O Afpms-n
este fi O Vmip3s
unul unul O Pi3msr
sintetic sintetic O Afpms-n
din din O Spsa
Dacron dacron O Ncms-n
sau sau O Ccssp
PTFE PTFE O Yn
. . O PERIOD
```

6 Utility of the corpus

There are several reasons for which MoNERo has an important contribution in named entity recognition and information extraction:

- it is the first Romanian gold standard biomedical corpus annotated with both part of speech tags and named entities;
- it was annotated with four types of named entities, making it very useful for training and testing NER systems based on supervised learning;
- it is pre-processed: tokenized, lemmatized and annotated with part of speech tags;
- it has a tabular format that makes it easy to use and the annotations are compliant with IOB2 format standards;

- it is a resource in a language other than English, which can help to train and test NER systems to perform language and domain adaptation;

- it is freely available for download[6] and non-commercial use. The archive contains three files, one for each medical domain, and another file containing all the other ones.

To prove the maturity and utility of this resource we used it to train and test a NER system (Boroş et al., 2018) for biomedical named entity recognition task for Romanian language. The architecture used is based on Bidirectional Long-Short-Term Memory (BDLSTM) networks (Graves, 2012). The system is trained to produce fully connected subgraphs. The feature-set is composed of word embeddings and character-level embeddings. In order to train the system the corpus was split in three sets: training set 80%, development set 10% and test set 10%. The evaluation of the performance of the system was done computing the F1 score and a score of 81.4 was obtained [7]. This experiment represents a starting point for the development/adaptation of NER systems for biomedical domain in Romanian.

7 Conclusions

We presented the MoNERo corpus, a gold standard biomedical corpus for Romanian language enhanced with two types of annotations: morphological and named entities specific to the biomedical field. To our knowledge this is the first biomedical corpus of this type for the Romanian language. This resource has already proven its value and utility, having been used in the development of the NER systems for the Romanian language. The MoNERo corpus is freely available for download and non-commercial use, which makes it even more valuable for the community.

8 Acknowledgements

Part of the work presented here was supported by a grant of the Romanian Ministry of Research and Innovation, PCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/72, within PNCDI III. We would like to thank the three anonymous reviewers for the valuable suggestions and hard work.

[6]The corpus is accessible at http://www.racai.ro/en/tools/text/

[7]The tool trained can be accessed at the following address: http://89.38.230.23/teprolin/index.php?path=teprolin/custom

References

Verginica Barbu Mititelu, Mihaela Cristescu, and Mihaela Onofrei. 2019. The Romanian Corpus Annotated with Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, Florence, Italy.

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1178–1185, Miyazaki, Japan. European Language Resources Association (ELRA).

Tiberiu Boroş, Stefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaalep, and Dan Tufiş. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, pages 315–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Carol S Federiuk. 1999. The effect of abbreviations on medline searching. *Academic emergency medicine*, 6(4):292–296.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, Joao Palotti, and Guido Zuccon. 2017. Clef 2017 ehealth evaluation lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 291–303. Springer.

Alex Graves. 2012. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer.

Radu Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian (in Romanian)*. Ph.D. thesis, Romanian Academy.

Radu Ion, Elena Irimia, Dan Stefanescu, and Dan Tufiş. 2012. Rombac: The romanian balanced annotated corpus. In *LREC*, pages 339–344. Citeseer.

Halil Kilicoglu. 2017. Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Briefings in bioinformatics*, 19(6):1400–1414.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):S2.

Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):198.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

Maria Mitrofan and Radu Ion. 2017. Adapting the ttl romanian pos tagger to the biomedical domain. In *BiomedicalNLP@ RANLP*, pages 8–14.

Maria Mitrofan and Dan Tufiş. 2018. Bioro: The biomedical corpus for the romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Isabel Moreno, Ester Boldrini, Paloma Moreda, and M Teresa Romá-Ferri. 2017. Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of biomedical informatics*, 72:8–22.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization. In *In Proc BioTextM, Reykjavik*. Citeseer.

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz 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.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics.

Paul Thompson, Syed A Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10(1):349.

Dan Tufiş. 2000. Using a large set of eagles-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of LREC*.

D Tufiş, AM Barbu, V Pătraşcu, G Rotariu, and C Popescu. 1997. Corpora and corpus-based morpho-lexical processing. *Recent Advances in Romanian Language Technology, Editura Academiei*, pages 35–56.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):356.