Annotation of Clinical Narratives in Bulgarian language

Ivaylo Radev **Kiril Simov** **Galia Angelova** **Svetla Boytcheva**
Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences
radev@bultreebank.org, kivs@bultreebank.org,
galia@lml.bas.bg, svetla.boytcheva@gmail.com

Abstract

In this paper we describe annotation process of clinical texts with morphosyntactic and semantic information. The corpus contains 1,300 discharge letters in Bulgarian language for patients with Endocrinology and Metabolic disorders. The annotated corpus will be used as a Gold standard for information extraction evaluation of test corpus of 6,200 discharge letters. The annotation is performed within Clark system — an XML Based System for Corpora Development. It provides mechanism for semi-automatic annotation. First a pipeline for Bulgarian morphosyntactic annotation and a cascaded regular grammar for semantic annotation are run, then rules for cleaning of frequent errors are applied. At the end the obtained result is manually checked. Our goal is to adapt the morphosyntactic tagger to the domain of clinical narratives as well.

1 Introduction

Today the electronic patient records and clinical notes are a fast growing research resource of medical data. These free text documents written by physicians contain a lot of valuable medical information despite the fact that sensitive data makes them hard to work with.

In countries like Sweden, UK and US researchers have started to use the electronic health records (EHR) to create corpora for two main purposes – in order to perform information extraction for medical research and for training domain specific systems to cope with these texts. Related subtasks are: automated de-identification for research work with sensitive data; extraction of medical time-lines in case development, with identification of deceasse and treatment; doing information retrieval and text mining; performing research in order to find relationships between diagnoses, treatments etc.; creation of golden standard corpora for evaluation and training; name entity recognition and annotation.

In this paper we describe annotation with morphosyntactic and semantic information of clinical texts. The corpus contains 1,300 discharge letters in Bulgarian language for patients with Endocrinology and Metabolic disorders. The annotated corpus will be used for information extraction evaluation.

The paper is structured as follows. Section 2 overviews related work with focus on the technological solutions. Section 3 presents the method we use, section 4 – the experiments and results. Section 5 contains the conclusion and discusses future work.

2 Related Work

Relevant references discuss annotation projects for corpora of medical texts in various natural languages. Studying the literature we adapted some principles for our annotation, although the sources are not directly connected to Bulgarian language.

A variety of approaches are described in the literature: e.g. for temporal annotations; pipeline with lexical features to extract time and event mentions; statistical chunking system for annotation; pipeline of tools for automatic processing of clinical texts and tokenization through part-of-speech tagging and dependency parsing; a simplification system, for automated change and adjusting of the text in health records in order to make them easier to understand; biomedical entity recognition dataset using a human-into-the-loop approach. Here we enumerate some annotation approaches correspondingly to language layers.

Proceedings of the Biomedical NLP Workshop associated with RANLP 2017, pages 81–87,
Varna, Bulgaria, 8 September 2017.
https://doi.org/10.26615/978-954-452-044-1_011

Entities. The article (Ogren et al., 2007) reports about the construction of a gold-standard dataset consisting of annotated clinical notes suitable for evaluating a biomedical named entity recognition system. The dataset is the result of consensus between four human annotators and contains 1,556 annotations on 160 clinical notes using 658 unique concept codes from SNOMED-CT corresponding to human disorders. Inter-annotator agreement was calculated on annotations from 100 of the documents for span (90.9%), concept code (81.7%), context (84.8%), and status (86.0%) agreement. Another corpus is designed to support automatic recognition of symptoms in unseen text. It consists of clinical free text records enriched with annotation for symptoms of a particular disease (ovarian cancer). The data (approximately 192K words) was annotated by three clinicians and a procedure was devised to resolve disagreements. The corpus is allows also to investigate the amount of symptom-related information in clinical records that is not coded (Koeling et al., 2011). Recognising entities is related to de-identification of sensitive information; the definitions of annotation classes are not self-evident. The article (Dalianis and Velupillai, 2010) presents two refined variants of an annotated gold standard corpus for de-identification of patient records in Swedish, one created automatically, and one created through discussions among the annotators. These are used for the training and evaluation of an automatic de-identification system based on the Conditional Random Fields algorithm. Promising results are acheived for both Gold Standards: F-score around 0.80 for a number of experiments on 4-6,000 instances, with higher results for certain annotation classes. The construction of three annotated corpora is presented in (Deleger et al., 2012) that serve as gold standards for medical NLP tasks. The annotated narratives are clinical notes from the medical record, clinical trial announcements, and FDA drug labels. High inter-annotator agreements is reported; the corpora are made public to facilitate translational NLP tasks that require cross-corpora interoperability. An annotated corpus (PhenoCHF), focussing on the identification of phenotype information for a specific clinical sub-domain, i.e., congestive heart failure (CHF), is presented in (Alnazzawi et al., 2014). The corpus integrats information from both EHRs (300 discharge summaries) and literature articles (5 full-text papers). The annotation scheme, whose design was guided by a domain expert, includes both entities and relations pertinent to CHF. Two further domain experts performed the annotation with agreement rates up to 0.92 F-Score.

Syntax. The paper (Fan et al., 2013) presents the development of a corpus with syntactic annotation (treebank) with intention to handle ill-formed sentences which are common in clinical text. A supplement to the Penn Treebank II guidelines was developed for annotating clinical sentences. After three iterations of annotation and adjudication on 450 sentences, the annotators reached an F-measure agreement rate of 0.930 (while intra-annotator rate was 0.948) on a final independent set. A total of 1100 sentences from progress notes were annotated that demonstrated domain-specific linguistic features. A statistical parser retrained with combined general English (mainly news text) annotations and our annotations achieved an accuracy of 0.811 (higher than models trained purely with either general or clinical sentences alone). In (Savkov et al., 2016), an approach to training domain specialists with no linguistic background to annotate clinical text is presented. The authors describe a de-identified corpus of free text notes, a shallow syntactic and named entity annotation scheme. A statistical chunking system for such clinical text with a stable learning rate and good accuracy is presented, indicating that the manual annotation is consistent and that the annotation scheme is tractable for machine learning.

Semantics. The Clinical E-Science Framework (CLEF) project aims at the identification of semantic entities and relationships in clinical narratives. The CLEF corpus consists of clinical narratives, histopathology reports and imaging reports from 20 thousand patients. A subset of this corpus was selected for manual annotation of clinical entities and relationships (Roberts et al., 2007). By entity, some real-world thing referred to in the text is meant: the drugs that are mentioned, the tests that were carried out etc. The relationships between entities correspond to the condition indicated by a drug, the result of an investigation etc. Annotation is anchored in the text. Annotators mark spans of text with a type: drug, locus and so on. Annotators may also mark words that modify spans (such as negation), and mark relationships as links between spans. Two or more spans may refer to the same thing in the real world, in which case they co-refer. Each text was annotated by 2 experts independently. In total, 27 annotators are involved in debugging, annotation and review roles. They are drawn from practicing clinicians, medical informaticians, and final year medical students. This corpus was used as a gold standard prividing temporal links (called CTlinks) between TLCs (Temporally Located CLEF entities, which comprise investigations, interventions and conditions) and temporal expressions: dates and times (both absolute and relative), as well as durations, as specified in the TimeML TIMEX3 standard (Roberts et al., 2008). The gold standard is a resource against which to assess the Information Extraction (IE) results of CLEF system. In addition, statistical models of the text may be built by machine learning algorithms. In 2008 the authors write that "the annotated CLEF corpus is the richest resource of semantically marked up clinical text yet created". The semantic annotation scheme, the annotation methodology, and the distribution of annotations in the final corpus are detailed in (Roberts et al., 2009).

Discource and Standardization. The Ontology Development and Information Extraction corpus (ODIE) annotated anaphoric relations in clinical narratives. The gold standard annotations resulted in 7214 markables, 5992 pairs and 1304 chains. These early shared annotation resources revealed the lack of common annotation schemes and community adopted standards and conventions for normalization (Savova, 2017). Recent ambitious projects aim at the annotation of timelines, in order to enable natural language understanding by discovering events and their relations on a timeline. Temporal relations are of prime importance in biomedicine as they are intrinsically linked to diseases, signs and symptoms, and treatments. The annotation guidelines of THYME project (“Temporal Histories of Your Medical Events”) are based on TIMEX3[1].

3 Methodology

The annotation is performed in two steps:

1. Automatic preprocessing and
2. Manual errors checking and correction.

The first step is done by BulTreeBank pipeline for Bulgarian (Savkov et al., 2012) updated with new tools — we substituted previous POS tagger and dependency parser with new ones based on MATE tool[2]. The process starts with simple discharge letter in text format written by the physician (Fig. 1). The text document is converted to XML format. After that we use tokenizer, sentence splitter, POS tagger and lemmatizator to automatically process the raw texts. The result from this processing includes the following information:

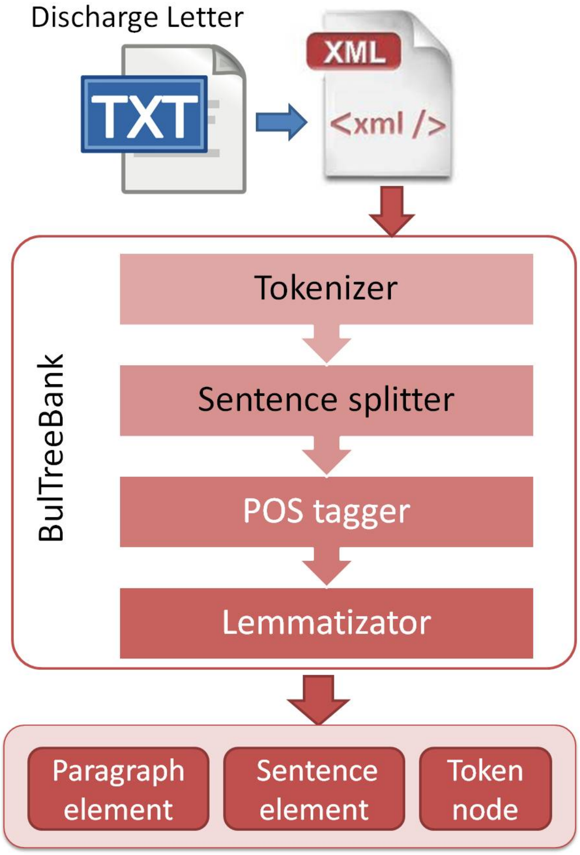


Figure 1: Automatic preprocessing

- ***Paragraph element (p)*** — contains some meta data like age, gender and location of the patient and the main sections of the discharge letter – anamnesis, health status, diagnosis, treatment, clinical exams, consultations, etc.
- ***Sentence element (s)*** — does not have additional information. Very hard to be done because the physicians neglect the punctuation rules.
- ***Token node (tok)*** — the main node of the tree. It has all the linguistic information like

[1] http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

[2] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.en.html

POS and lemmas. Also it has the term attribute.

The overall performance accuracy of the original pipeline droped significantly due to the reach medical terminology included in the texts. The result XML documents are after that checked and annotated further manually. During this process we are using CLaRK system[3] — an XML Based System for Corpora Development (Simov et al., 2001), (Simov et al., 2004). The core of CLaRK is an Unicode XML Editor, which is the main interface to the system. Via it the user could edit, search and process the annotated documents. The system contains several processing tools like XML elements and attributes addition, deletion, and substitution. For navigation over XML documents the system exploit XPath language. Two main tools of the system are (1) Regular Cascaded Grammars; and (2) Constraints over XML Documents.

Regular Grammars in CLaRK System. The regular grammars in CLaRK System work over token and element values generated from the content of an XML document and they incorporate their results back in the document as XML markup (called *return markup*) (Simov et al., 2002). The tokens are determined by the corresponding tokenizer. The element values are defined with the help of XPath expressions, which determine the important information for each element. In the grammars, the token and element values are described by token and element descriptions. These descriptions could contain wildcard symbols and variables. The variables are shared among the token descriptions within a regular expression and can be used for the treatment of phenomena like syntactic agreement. The grammars are applied in a cascaded manner. The general idea underlying the cascaded application is that there is a set of regular grammars. The grammars in the set are in a particular order. The input of a given grammar in the set is either the input string, if the grammar is first in the order, or the output string of the previous grammar. The evaluation of the regular expressions that define the rules, can be guided by the user. We allow the following strategies for evaluation: "longest match", "shortest match" and several backtracking strategies.

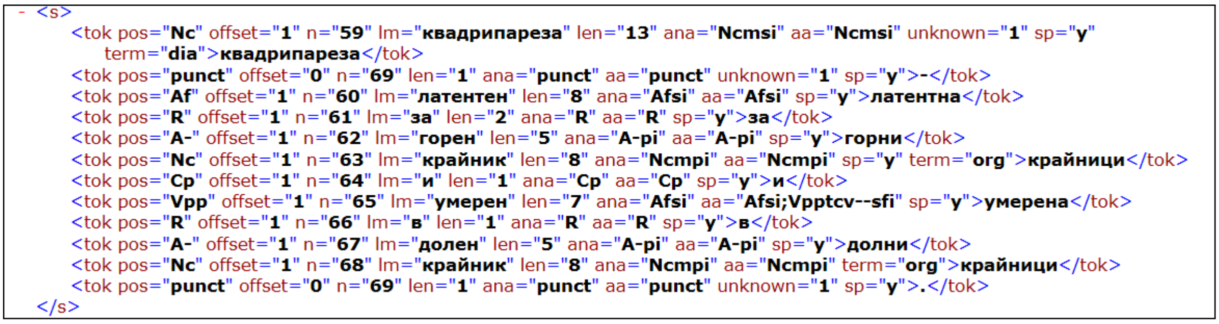
Constraints over XML Documents. The constraints that we have implemented in the CLaRK System are generally based on the XPath language. We use XPath expressions to determine some data within one or several XML documents and thus we evaluate some predicates over the data. Generally, there are two modes of using a constraint. In the first mode **validation**, the constraint is used for a validity check, similar to the validity check, which is based on a DTD or an XML schema. In the second mode **insertion**, the constraint is used to support the change of the document to satisfy the constraint. The constraints in the CLaRK System are defined in the following way: `(Selector, Condition, Event, Action)`, where the selector defines to which node(s) in the document the constraint is applicable; the condition defines the state of the document when the constraint is applied. The condition is stated as an XPath expression, which is evaluated with respect to each node, selected by the selector. If the XPath expression is evaluated as true, then the constraint is applied; the event defines when this constraint is checked for application. Such events can be: selection of a menu item, pressing of a key shortcut, an editing command; the action defines the way of the actual constraint application.

The combination of XLM editor with processing tools is a very powerful tool for minimization of human intervention during the annotation of new corpora. The manual work is inevitable, but many of the mistakes of the automatic processing and also the new annotations are regular. Thus, very quickly the annotator recognizes them. In these cases the system provides necessary support for the annotator to write procedures for automatic repairing or automatic annotation of these regular cases.

At the end a human annotator checks the results and finalizes the annotation. The new information (besides the corrected one) comprises:

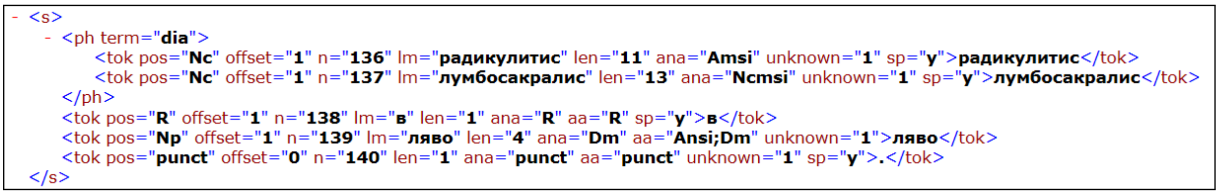
- ***phrase node (ph)*** — subdivision of the sentence with more than one token - bronchial asthma or spine (гръбначен стълб in Bulgarian). It has the term attribute.
- ***time string (ts)*** — subdivision of the sentence with more than one token for dates and time. It has the time attribute.
- ***dosage string (ds)*** — subdivision of the sen-

[3] http://www.bultreebank.org/clark/index.html



```
- <s>
    <tok pos="Nc" offset="1" n="59" lm="квадрипареза" len="13" ana="Ncmsi" aa="Ncmsi" unknown="1" sp="y"
      term="dia">квадрипареза</tok>
    <tok pos="punct" offset="0" n="69" len="1" ana="punct" aa="punct" unknown="1" sp="y">-</tok>
    <tok pos="Af" offset="1" n="60" lm="латентен" len="8" ana="Afsi" aa="Afsi" sp="y">латентна</tok>
    <tok pos="R" offset="1" n="61" lm="за" len="2" ana="R" aa="R" sp="y">за</tok>
    <tok pos="A-" offset="1" n="62" lm="горен" len="5" ana="A-pi" aa="A-pi" sp="y">горни</tok>
    <tok pos="Nc" offset="1" n="63" lm="крайник" len="8" ana="Ncmpi" aa="Ncmpi" sp="y" term="org">крайници</tok>
    <tok pos="Cp" offset="1" n="64" lm="и" len="1" ana="Cp" aa="Cp" sp="y">и</tok>
    <tok pos="Vpp" offset="1" n="65" lm="умерен" len="7" ana="Afsi" aa="Afsi;Vpptcv--sfi" sp="y">умерена</tok>
    <tok pos="R" offset="1" n="66" lm="в" len="1" ana="R" aa="R" sp="y">в</tok>
    <tok pos="A-" offset="1" n="67" lm="долен" len="5" ana="A-pi" aa="A-pi" sp="y">долни</tok>
    <tok pos="Nc" offset="1" n="68" lm="крайник" len="8" ana="Ncmpi" aa="Ncmpi" term="org">крайници</tok>
    <tok pos="punct" offset="0" n="69" len="1" ana="punct" aa="punct" unknown="1" sp="y">.</tok>
  </s>
```

Figure 2: Example 1. Annotation of the upper and lower limbs status



```
- <s>
  - <ph term="dia">
      <tok pos="Nc" offset="1" n="136" lm="радикулитис" len="11" ana="Amsi" unknown="1" sp="y">радикулитис</tok>
      <tok pos="Nc" offset="1" n="137" lm="лумбосакралис" len="13" ana="Ncmsi" unknown="1" sp="y">лумбосакралис</tok>
    </ph>
    <tok pos="R" offset="1" n="138" lm="в" len="1" ana="R" aa="R" sp="y">в</tok>
    <tok pos="Np" offset="1" n="139" lm="ляво" len="4" ana="Dm" aa="Ansi;Dm" unknown="1">ляво</tok>
    <tok pos="punct" offset="0" n="140" len="1" ana="punct" aa="punct" unknown="1" sp="y">.</tok>
  </s>
```

Figure 3: Example 2. Annotation of medical terms in Latin transliterated in Cyrillic

tence with more than one token for doses – 1+1/2 pill or 125 mcg per day.

Information from the attributes:

- ***term attribute*** — marks the medical terms and bears information about their type

- ***term values*** — diagnosis (**DIA**), symptom (**SIM**), status (**STT**), organ (**ORG**), body system (**SIS**), medicament (**MED**), test (**TST**) and index (**POK**). It is likely for more to come up.

- ***time attribute*** — bears information about absolute (**abt** value) time (10.02.1999) or relative (**rtt** value) time (two months ago).

We apply various vocabularies which help us to figure out the semantics of the words in the near context.

The 10 vocabularies are: *(1)* Vocabulary of the 100,000 most frequent Bulgarian terms (Osenova and Simov, 2010); *(2)* Generic medical terms in Bulgarian; *(3)* Anatomical terms in Latin; *(4)* Generic names of drugs for Diabetes Mellitus Treatment; *(5)* Laboratory tests; *(6)* Diseases; *(7)* Treatment; *(8)* Symptoms; *(9)*. Abbreviations; *(10)* Stop words;. These are applied in the specified order and the annotations of the latter ones override the previous ones. The vocabulary coverage is shown on Table 1. In the columns are shown the size of each vocabulary (Size) and the number of tokens matched in the text by this vocabulary (Tokens). The largest coverage has the vocabulary of stop words, then diagnoses, next is the vocabulary of most frequent Bulgarian words followed by the markup words.

Table 1: Lexical Profile Statistics.

Category	Size	Tokens
1. btb	102,730	41,582
2. bg med	3,624	1,545
3. term anat	4,382	3,792
4. drugs	154	12
5. lab test	202	18
6. diagnoses	8,444	54,431
7. treatment	339	4,170
8. symptoms	414	4,180
9. abbrev	477	14,404
10. stop words	805	67,153

4 Experiments and Results

The experiments were done over a set of 1,370 pseudoanonymised discharge letters in Bulgarian for patients with Endocrinology and Metabolic disorders. The discharge letters text contains medical terminology in Latin alphabet (about 1% of all term tokens in our present corpus), sometimes with different transcriptions in Cyrillic alphabet. There are specific term abbreviations both in Bulgarian and Latin (about 3% of the tokens), numerical values (16% of the tokens) and about of 1% of all term tokens are presented as abbreviations.

One of the main problems is that huge groups of out of the vocabulary terms are available in the discharge letters. They are several groups - medical terms in Latin, medical terms in Latin transliterated in Cyrillic; brand names of drugs and medications, abbreviations, etc. There are 7,108 occurrences of drug names in 1,213 of the discharge letters, in average 5.86 drugs per document. These is a quite dynamic information that needs to be updated monthly and the annotation tool also will lack some information.

The problem of Latin written in Cyrillic is about fast and decent annotation by people without knowledge of medical Latin.

статус пост адреналектомиам билатералис(status post adrenalektomiam bilateralis)

аденомектомиам транссфеноидалем ет телегаматерапиам(adenomektomiam transsfenoidalem ет telegamaterapiam)

статус пост тиреоидектомиам про карцинома папиларе лоби синистри (status post thyreoidektomiam pro carcinoma papillari lobby sinistri)

The method is simple: to take every phrase separately and look for attributes and phrasal base and prescribe `Adj` to attributes and `N` for the base (Fig. 3). There is created a grammar in CLaRK for automated phrase (**ph**).

Another problem is that there are many typos in the documents and a variety of abbreviations for same terms.

5 Conclusion and Further Work

We report work in progress about annotation of clinical narratives in Bulgarian.The role of the grammars (phrasal grammar) in quality of the analysis and time-saving in the annotation process. Phrases do not improve the morphological analysis. Good morphological analysis and lemma recognition improves the phrasal grammar and speeds up the work process. One of the main problems is that we did no have yet several annotations for each document and inter-annotation agreement is not evaluated.

Further work include some preprocessing of the corpus for spelling errors correction both for Latin and Cyrillic that will help in the automatic processing. Another direction for further work is the training of a domain specific tokenizer and POS tagger and improving of the general tokenizer and tagger. Iterative enrichment of the vocabularies after the manual correction of the annotation will also help.

Acknowledgments

The research presented here is partially supported by the grant SpecialIZed Data MIning MethoDs Based on Semantic Attributes (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019. The team acknowledges also the support of Medical University – Sofia.

References

Noha Alnazzawi, Paul Thompson, and Sophia Ananiadou. 2014. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In *EACL 2014 Workshop-The Fifth International Workshop on Health Text Mining and Information Analysis, Gothenburg, Sweden, 27 April, 2014, edited by Velupillai, Sumithra and Duneld, Martin and Henriksson, Aron and Kvist, Maria and Skeppstedt, Maria and Dalianis, Hercules*. Association for Computational Linguistics, pages 69–74.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying swedish clinical text-refinement of a gold standard and experiments with conditional random fields. *Journal of biomedical semantics* 1(1):6.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2012, page 144.

Jung-wei Fan, Elly W Yang, Min Jiang, Rashmi Prasad, Richard M Loomis, Daniel S Zisook, Josh C Denny, Hua Xu, and Yang Huang. 2013. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association* 20(6):1168–1177.

Rob Koeling, John Carroll, Rosemary Tate, and Amanda Nicholson. 2011. Annotating a corpus of clinical text records for learning to recognize symptoms automatically .

Philip V Ogren, Guergana K Savova, Christopher G Chute, et al. 2007. Constructing evaluation corpora for automated clinical named entity recognition. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, page 2325.

Petya Osenova and Kiril Simov. 2010. Using the linguistic knowledge in bultreebank for the selection of the correct parses .

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Subbarao Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, et al. 2007. The clef corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2007, page 625.

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.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*. pages 19–26.

Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language resources and evaluation* 50:523.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. 2012. Linguistic analysis processing line for bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Guergana and Savova. 2017. *Annotating the clinical text - MiPACQ, ShARe, SHAPPn and THYME Corpora*.

Kiril Simov, Milen Kouylekov, and Alexander Simov. 2002. Cascaded regular grammars over xml documents. In *Proceedings of the 2Nd Workshop on NLP and XML - Volume 17*. Association for Computational Linguistics, Stroudsburg, PA, USA, NLPXML '02, pages 1–8. https://doi.org/10.3115/1118808.1118820.

Kiril Simov, Petya Osenova, and Milena Slavcheva. 2004. Btb-tr03: Bultreebank morphosyntactic tagset. Technical report, BulTreeBank Project Technical Report.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. 2001. Clark-an xml-based system for corpora development. In *Proc. of the Corpus Linguistics 2001 Conference*. pages 558–560.