Automatic Detection of Contraindications of Medicines in Package Leaflet

Jonas Žalinkevičius
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
jonas.zalinkevicius@hotmail.com

Rita Butkienė
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
rita.butkiene@ktu.lt

***Abstract*— Before physicians prescribe medicines, they must take into consideration the patient's diseases and medicines they use. This is done to avoid complications that may occur. All information about possible contraindications is written in the medicine package leaflet. A system that can automatically detect contraindication mention in the Lithuanian text of leaflet applying natural language parsing is presented. This system gives a possibility to shorten the time needed for medicines prescription decision making. The results of the experiment showed that the created system successfully detected 56 per cent contraindications.**

***Keywords*— medicine contraindications, drug–drug interactions, shallow parsing, morphological analysis, noun phrase detection**

I. Introduction

When a patient is diagnosed with a new disease, additionally physician asks the patient about his allergies, previous health problems, chronic deceases, what medications and food supplements he is using. After taking gathered information into consideration and evaluation of possible contraindications with prescribed medication physician assigns treatment and, if needed, changes previous assignments. Almost all information about contraindications can be found in the medicine package leaflet. According to Lithuania's medicines registration procedure [1], every package must have a leaflet written in Lithuanian. Information in the leaflet must be divided into six sections [2], although the text in a section can be written in not structural manner. So, if a physician needs to find possible contraindications, he must read all text in the second section (Table 1) or search for information on the Internet. Usually, health care information consists of unstructured data and that leads to inaccurate search results that contain hundreds of links to not relevant documents. And the user must read through results to find relevant information.

Automatic information extraction tools can extract biomedical data, save it in a structural way, and minimize information search problem. However, automatic text analysis and information extraction from unstructured text in the medical domain is a challenging task [3]. The aim of this paper is to present a system that gives physicians the possibility of a faster and more accurate way of finding contraindications using automated contraindication detection in the medicine package leaflet.

A system that automates the extraction of contraindications from leaflet text is described is in Section 3. Using this system all leaflets of medicines registered in Lithuania were analyzed. The results of this analysis (contraindications extracted) are used in a commercial medications information system that is used by Lithuanian physicians for prescription of medications. The evaluation of the obtained results is presented in Section 4.

II. Related Work

In Lithuania, it is established that each medicine registered in Lithuania must contain a package leaflet describing therapeutic indications, possible contraindications, safety precautions, and usage information in the Lithuanian language. In order to be sure that the patient does not suffer from possible contraindication, the physician should read through all leaflet text before prescribing the medicine. Usually, the analysis of leaflets is time-consuming, so physicians tend to skip it and rely on the knowledge and experience they have gained.

There are lots of systems developed for analysis and information extraction from the biomedical text in the English language. But there is no solution for the detection of contraindication (i.e. contraindication with disease or contraindication with the pharmacological group) mentions in Lithuanian written text. We have analyzed articles that describe similar problems when analyzing biomedical text. For example, a tool *Semantator* [4] was created for converting biomedical text to linked data. It used ontology-based information extraction using biomedical ontology terms hosted in *BioPortal* and ontology editor *Protégé* for text preprocessing. A semantic annotation and inference platform SENTIENT-MD [3] creates a dependency graph as the first step for dependency parsing which is one of the tasks of semantic annotation of medical knowledge in natural language text. Markus Bundschus [5] used probabilistic graphical models (Conditional Random Fields) to identify semantic relations.

Although all these authors work on texts written in English, we found that common rules and approaches could be applied to Lithuanian texts as well. In order to extract information from text, preprocessing is needed using natural language processing: text segmentation, a morphological analysis should be performed and then a syntactic parse tree or the dependency graph [6]. [7] should be formed. For semantic relations detection, existing ontologies or knowledge bases should be used.

III. System Description

In this section, a system for the detection of contraindication mentions in the medicine leaflet text written in Lithuanian is presented. The system implements a text analysis pipeline of four analysis stages: extraction of contraindication text block, morphological analysis, noun phrase detection, and annotation.

Additionally, all annotated phrases are checked is it in the database of noun phrases to be ignored or not. This database is manually filled and helps to obtain more precise results. The overall pipeline for the detection of contraindication mentions is shown in fig. 1.

Below each stage of text analysis is discussed in more detail.

A. Extraction of contraindication text blocks

In Lithuania, when describing the medicine, a producer must follow a certain template of the package leaflet [2]. This template splits the description of leaflet into 6 sections listed in Table 1

TABLE I. Medicine Package Leaflet Sections

No	Section
1	What X is and what it is used for
2	What you need to know before you <take> <use> X
3	How to <take> <use> X
4	Possible side effects
5	How to store X
6	Contents of the pack and other information

The information which, the patient should be aware of before he or she takes the medicine, is presented in section number two. An example of this section is shown in fig. 2 with highlighted contraindications phrases. So, the first task of our system is to find this section and extract its text for further analysis.

B. Morphological analysis

A morphological analysis forms a background for information extraction about contraindications. In this stage, a given text is split into lexical units (e.g. sentences, lexemes) and analyzed morphologically. For this task, a web service provided by the system "*http://semantika.lt*" [8] is used. The web service returns morphological features for each given lexeme: part of speech, gender, number and so on.

C. Noun phrase detection

Phrases that express a specific contraindication usually are noun phrases, for example, *heart attack*, *type one diabetes*, *pancreatitis*, and so on. Therefore, we chose a phrase structure grammar method because it better fits for noun phrase detection than dependency grammar as it was suggested by Axel Halvoet in his monography [9]. Phrase structure rules are used to split natural language written sentence into its constituent parts: lexical and phrasal categories [9], [10], [11]. For the noun phrase detection in the medicine's leaflet, three phrase structure rules ware specified (see Table 2).

TABLE II. Noun Phrase Structure Rules

No	Rule
1	A lexeme is a part of a noun phrase if it is a noun in the genitive case and follows another noun in the genitive case or adjective or numeral or participle.
2	A lexeme is a part of a noun phrase if it is an attributive adjective in the same case, number, and gender as a base noun and follows noun in the genitive case or adjective or numeral or participle.
3	A lexeme is a part of noun phrase if it is an attributive numeral in the same case, number and gender as the base noun and follows noun in the genitive case, or adjective, or numeral, or participle.

An algorithm implemented for the noun phrase detection checks every lexeme in the sentence for the satisfaction of conditions of at least one rule presents in Table 2. If the condition is satisfied a lexeme is included in the noun phrase. The workflow of analysis of the noun phrase *Lėtinis reumatinis perikarditas* (Chronic rheumatic pericarditis) is shown in Table 3.

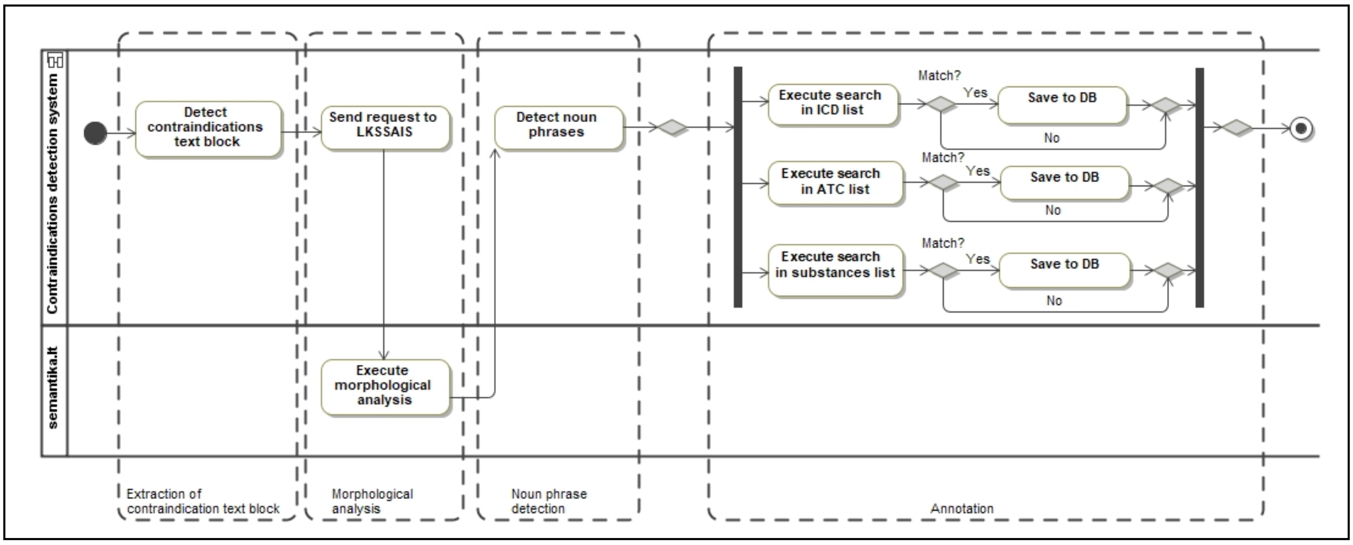
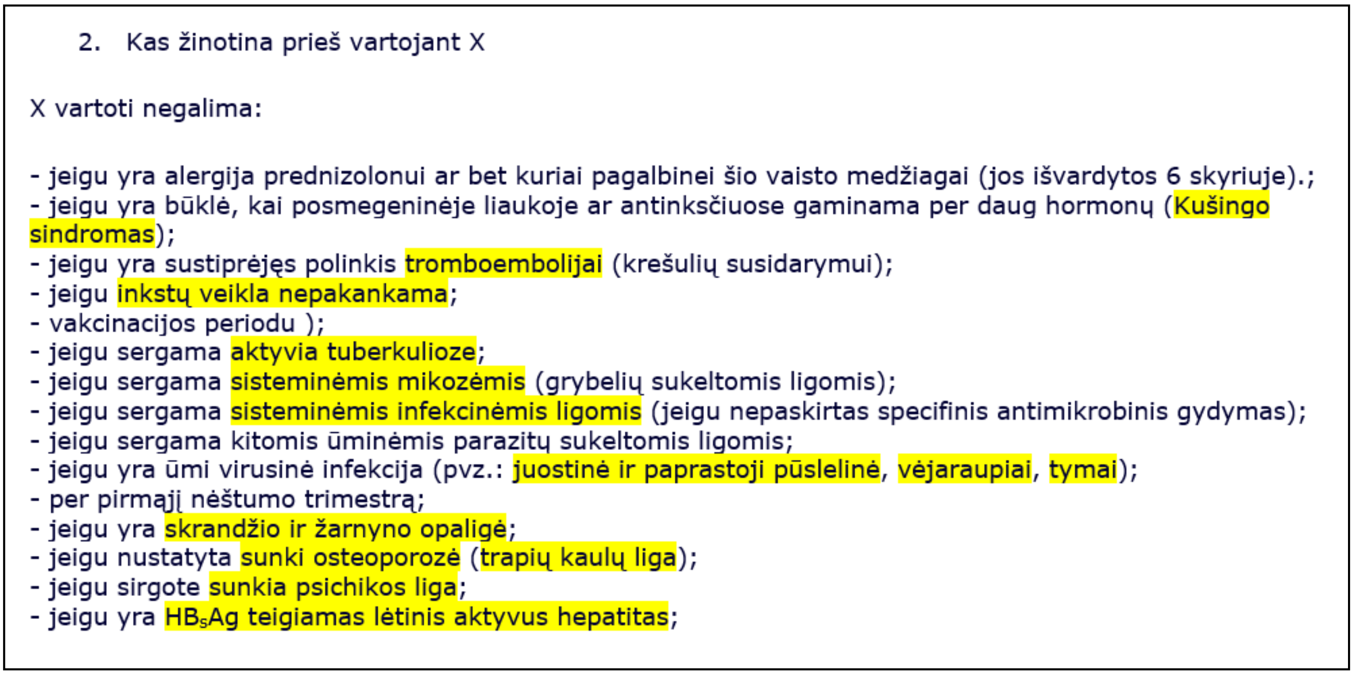

Contraindications detection system
semantika.lt
Detect contraindications text block
Send request to LKSSAIS
Execute morphological analysis
Detect noun phrases
Execute search in ICD list
Execute search in ATC list
Execute search in substances list
Match?
Yes
No
Save to DB
Extraction of contraindication text block
Morphological analysis
Noun phrase detection
Annotation

Fig. 1. Contraindications lookup process activity



2. Kas žinotina prieš vartojant X

X vartoti negalima:

- jeigu yra alergija prednizolonui ar bet kuriai pagalbinei šio vaisto medžiagai (jos išvardytos 6 skyriuje).;
- jeigu yra būklė, kai posmegeninėje liaukoje ar antinksčiuose gaminama per daug hormonų (Kušingo sindromas);
- jeigu yra sustiprėjęs polinkis tromboembolijai (krešulių susidarymui);
- jeigu inkstų veikla nepakankama;
- vakcinacijos periodu);
- jeigu sergama aktyvia tuberkulioze;
- jeigu sergama sisteminėmis mikozėmis (grybelių sukeltomis ligomis);
- jeigu sergama sisteminėmis infekcinėmis ligomis (jeigu nepaskirtas specifinis antimikrobinis gydymas);
- jeigu sergama kitomis ūminėmis parazitų sukeltomis ligomis;
- jeigu yra ūmi virusinė infekcija (pvz.: juostinė ir paprastoji pūslelinė, vėjaraupiai, tymai);
- per pirmąjį nėštumo trimestrą;
- jeigu yra skrandžio ir žarnyno opaligė;
- jeigu nustatyta sunki osteoporozė (trapių kaulų liga);
- jeigu sirgote sunkia psichikos liga;
- jeigu yra HB$_s$Ag teigiamas lėtinis aktyvus hepatitas;

Fig. 2. Example of “What you need to know before use of X” section in the medicine package leaflet

TABLE III. EXAMPLE OF NOUN PHRASE DETECTION WORKFLOW

Step	Action	Rule satisfaction
1	The first lexeme *Lėtinis* (Chronic) is an adjective in the nominative case, singular and of masculine gender	No rule condition is satisfied fully, but according to rule No. 2 the lexeme is a good candidate for the noun phrase.
2	The second word *reumatinis* is an adjective in the nominative case, singular and of masculine gender and follows the adjective *Lėtinis*	No rule condition is satisfied fully, but according to rule No. 2 the lexeme is a good candidate for the noun phrase.
3	The third word *perikarditas* is a noun in the nominative case, singular and of masculine gender It follows the adjectives *lėtinis* and *reumatinis* which are in the same case, number and gender.	The condition of rule No. 2 is satisfied. The noun is a base noun for the first two adjectives. They are attributive adjectives of the noun. So, the condition of rule No. 2 is satisfied as well. The analysis of the third lexeme completes the construction of the noun phrase.

When the construction of the noun phrase is complete the form of the head noun in the phrase is changed to its canonical form (lemma). This is done because the name of item registered in the International Classification of Diseases (ICD) [12], Anatomical Therapeutic Chemical Classification System (ATC) [13] or lists of active substances are in the canonical form, therefore, normalization is required to ensure the correct comparison of values in the next stage of analysis.

D. *Annotation*

All noun phrases identified in the previous stage are reviewed and checked for contraindication. If a contraindication is identified, the phrase is annotated. For annotation three databases are used: ICD, ATC and the lists of active substances. The algorithm compares the noun phrase and name of the item from the database. If the noun phrase matches the name in ICD the phrase is tagged as contraindication with the disease. If the phrase matches the ATC item name, it is tagged as contraindication with a pharmaceutical chemical group, and if the phrase matches the name of the active substance, it is tagged as contraindication with an active substance.

It is worthy to mention that before comparison of the noun phrases all identified phrases are checked against phrases in the database of noun phrases to be ignored. In the text of medicine package leaflet, a lot of words (i.e. illness, hand and so on) that are irrelevant (do not express a contraindication) but are used in ICD, ATC and active substances lists could be found. The database of noun phrases to be ignored was filled manually with the help of a professional pharmacist.

IV. EXPERIMENT

The aim of the experiment is to evaluate the created system and check if a tool can achieve its target - to give physicians the possibility of a faster and more accurate way of finding contraindications. The experiment was done by manually annotating contraindications mentions in the package leaflet text block and comparing results with the system’s results. This was done by a professional pharmacist who works in JSC Skaitos kompiuterių servisas.

A. *Plan*

The experiment was organized as follows. From medicines database ten randomly selected leaflets were analyzed using the system created. The results of the analysis were automatically gathered into the table, which example is presented in Table 4 In the first column the code of item automatically found in the text of leaflet by the system is indicated. The second column represents the database (ATC, ICD or active substances) where the item is registered. The third column was used for the evaluation of annotation correctness.

TABLE IV. AUTOMATICALLY DETECTED CONTRAINDICATIONS RESULTS EVALUATION FOR SINGLE LEAFLET

Code	Domain	Is detection correct
J01CR	ATC	False
J05AE	ATC	True
I09.2	ICD	True

The same randomly selected leaflets were analyzed and annotated manually, and the table of the same structure was filled in with manual annotation results. Manually found contraindications were not interpreted or changed to synonyms. For example, *heart attack* and *myocardial infarction* are the same diseases. But ICD contains only one name of this disease - *myocardial infarction*. The created system is not able to recognize the heart attack as a synonym of *myocardial infarction*.

Additionally, the active substances, mentioned in the leaflet, were translated into the Latin language (nominative and genitive grammatical cases). This was done because the database of active substances, that was provided, has three versions of translation: Lithuanian, Latin in the nominative case and Latin in the genitive case.

B. *Results*

The results of the evaluation are presented in Table 5. The precision, recall and F-Score metrics have been calculated for each leaflet analyzed. Additionally, the ratio between the number of correctly detected contraindications and overall automatically detected contraindications was calculated as well. This metric allows to evaluate how accurate the results are and to use them in further calculations.

Results showed that the system developed is able to correctly detect 56% of relevant contraindications. The average number of links detected automatically is 1482.8 while manually detected links are 197.9. The number of links detected automatically in one leaflet is average four times higher, than detected manually. The average number of erroneous links to ICD is 72%, to ATC - 90%, and links to the list of active substances - 61%.

Calculations show that the system is able to achieve 0.25(±0.23) precision, 0.56(±0.32) recall, and 0.31(±0.19) F-score value. To give a better perspective where the system's failures were and possible reasons for that, Pearson correlation coefficient calculations between various indicators were done (Table 6). The biggest impact on F-Score had incorrectly detected links to ICD, a coefficient was -0.89. The reason why precision was so low is that of the high ratio between automatically and manually detected links.

TABLE V. EXPERIMENT RESULTS

ID	Auto. detected links	Auto. correctly detected links	Man. detected links	Precision	Recall	F-Score	Ratio of links amounts	Err. links to ICD	Err. links to ATC	Err. links to active substances
13092	1906	346	385	0.18	0.90	0.30	4.95	82%	100%	65%
13571	1899	367	444	0.19	0.83	0.31	4.28	81%	100%	58%
859	87	67	162	0.77	0.41	0.54	0.54	17%	100%	100%
1300	400	28	146	0.07	0.19	0.10	2.74	98%	100%	24%
10958	464	14	71	0.03	0.20	0.05	6.54	100%	25%	21%
1872	283	66	68	0.23	0.97	0.38	4.16	77%	100%	43%
5363	473	237	291	0.50	0.81	0.62	1.63	46%	88%	49%
13273	158	51	72	0.32	0.71	0.44	2.19	45%	100%	100%
10744	1199	150	175	0.13	0.29	0.18	6.85	87%	100%	100%
16551	1090	120	204	0.11	0.25	0.15	5.34	90%	87%	51%
Median	468.5	93.5	168.5	0.185	0.56	0.305	4.22	82%	100%	55%
Q1	312.25	54.75	90.5	0.115	0.26	0.158	2.328	54%	91%	45%
Q3	1171.75	215.25	269.25	0.298	0.825	0.425	5.243	89%	100%	91%
Avg	795.9	144.6	201.8	0.25	0.56	0.31	3.92	72%	90%	61%
Std dev	686.52	129.45	132.27	0.23	0.32	0.19	2.10	27%	23%	30%
Min	87	14	68	0.03	0.19	0.05	0.54	17%	25%	21%
Max	1906	367	444	0.77	0.97	0.62	6.85	100%	100%	100%

TABLE VI. CORRELATION OF ESTIMATES AND INDICATORS

		Estimates		
		Precision	*Recall*	*F-Score*
Indicator	Incorrectly detected links to ICD list amount	-0.9655	-0.3114	-0.8939
	Incorrectly detected links to ATC list amount	0.3292	0.4184	0.4382
	Incorrectly detected links to active substances list amount	0.5229	0.1244	0.4523
	Automatically and manually detected contraindications ratio	-0.8119	-0.2583	-0.7682

C. *Conclusions of the experiment*

The experiment shows that the system automatically successfully detected more than half of the relevant contraindication links (56%). But 75% of links were erroneous and the system lacks precision. The reason for that is a high number of incorrect links to ICD (r=-0.9655), this indicator has the most negative impact on the precision and F-Score results. This might be because of commonly used phrases that are not contraindications but used in the ICD list. For example, the word *allergy* does not imply that this is a contraindication and must be ignored. Another reason for low estimates results is, the number of detected contraindications phrases. Calculations show, that the higher is the difference between automatically and manually detected contraindications phrases, the lower are precision and F-Score results. The reason for that is, a high number of noun phrases that are irrelevant to contraindications noun phrases, for example, pill, driving.

Additionally, considering why F-Score is so low (0.31) the assumption that this is because of low precision (0.25) can be done. To raise this indicator the list of phrases to be ignored (common word and phrases) must be used. The most frequent reasons for the incorrect detection of contraindications are:

- the context of the phrase in the sentence is not taken into account;
- Conjunctions are not taken into account and two or more noun phrases (i.e. "…kidney and liver diseases…") are not identified;
- Brackets that are used to specify contraindication are not taken into account ("…liver tumor (malignant or benign)…").

To avoid errors caused by those reasons, users of "*https://gydytojams.vaistai.lt*" IS will be able to mark contraindication as erroneous and if the pharmacist approves that it will be removed from the database.

V. CONCLUSIONS

In this paper, the system which automatically detects contraindications and links them to existing "Skaitos kompiuterių servisas" databases have been introduced. System analyses text of medications leaflets, it extracts noun phrases and links them to corresponding items in ATC, ICD, and active substances list. The system presented was used for the extraction of contraindications from leaflets of all medications registered in Lithuania. Extracted data was used in the pilot project for extending the functionality of the system "*https://gydytojams.vaistai.lt*". The additional function supports physicians in search of possible contraindications that are relevant to patient medical records. Moreover, physicians have the possibility to give feedback about erroneous contraindications presented. In such a way they help in expanding the list of phrases to be ignored and eliminating incorrect contraindication links.

The experiment shows that approximately 56% of contraindications are found but only every fourth is correct. Several changes in the algorithm remain for future work. First, before the noun phrase is looked up in databases, a context must be identified. This would reduce the number of incorrect links. Second, to detect phrases that refer to medication analyzed and to ignore them.

ACKNOWLEDGMENT

Data for this system was provided by JSC Skaitos kompiuterių servisas

REFERENCES

[1] VVKT prie LR SAM, "Įsakymas 2015 m. liepos 3 d. Nr.(1.72E)1A-755 Dėl paraiškų registruoti vaistinį preparatą, perregistruoti vaistinį preparatą, pakeisti registracijos pažymėjimo sąlygas, teisės į vaistinio preparato registraciją perleidimo, nereglamentiniam pakuotės ir (ar," 03 07 2016. [Online].

[2] European Medicines Agency, "European Medicines Agency," 02 2019. [Online].

[3] S. Sahay, E. Agichtein, B. Li, E. V. Garcia and A. Ram, "Semantic Annotation and Inference for Medical Knowledge Discovery," 2007. [Online].

[4] C. Tao, D. Song, D. Sharma and C. G. Chute, "Semantator: Semantic annotator for converting biomedical text to linked data.," Journal of Biomedical Informatics, vol. 46, no. 5, pp. 882-893. 12p., Oct2016.

[5] M. Bundschus, M. Dejori, M. Stetter, V. Tresp and H.-P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields.," BMC Bioinformatics, vol. 9, pp. 1-14, 2008.

[6] Y. Zhang, H.-Y. Wu, J. Xu, J. Wang, S. Ergin, L. Li and H. Xu, "Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature.," BMC Systems Biology, vol. 107, pp. 323-334 12p., 8/26/2016.

[7] R. Frank, Phrase Structure Composition and Syntactic Dependencies, vol. 38, Cambridge, Mass: The MIT Press, 2002, pp. 2-27.

[8] Damaševičius, R., Napoli, C., Sidekerskienė, T. and Woźniak, M., 2017. IMF mode demixing in EMD for jitter analysis. Journal of Computational Science, 22, pp.240-252.

[9] Kaunas University of Technology and Vytautas Magnus University, "Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema," [Online].

[10] A. Holvoet, Bendrosios sintaksės pagrindai, Vilnius: Vilniaus Universitetas, Asociacija „Academia Salensis", 2009.

[11] D. Jurafsky and J. H. Martin, "Formal Grammars of English," in Speech and Language Processing (2Nd Edition), JAV, Prentice-Hall, Inc., 2009, pp. 396-408.

[12] D. Šveikauskienė, "Lietuvių kalbos sintaksinė analizė," Lietuvių kalba, vol. 7, 2013.

[13] Wózniak, M., Połap, D., Nowicki, R.K., Napoli, C., Pappalardo, G. and Tramontana, E., 2015, July. Novel approach toward medical signals classifier. In 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1-7 . IEEE.

[14] Valstybinė ligonių kasa, "TLK-10-AM / ACHI / ACS elektroninis vadovas," [Online].

[15] Norwegian Institute of Public Health, "WHOCC - Structure and principles," [Online].