Detecting Latin-Based Medical Terminology in Croatian Texts

Kristina Kocijan[1](✉), Maria Pia di Buono[2], and Linda Mijić[3]

[1] Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
krkocijan@ffzg.hr

[2] TakeLab ZEMRIS, Faculty of Electrical Engineering and Computing,
University of Zagreb, Zagreb, Croatia
mariapia.dibuono@fer.hr

[3] Department of Classical Philology, University of Zadar, Zadar, Croatia
lmijic@unidz.hr

Abstract. No matter what the main language of texts in the medical domain is, there is always an evidence of the usage of Latin-derived words and formative elements in terminology development. Generally speaking, this usage presents language-specific morpho-semantic behaviors in forming both technical-scientific and common-usage words. Nevertheless, this usage of Latin in Croatian medical texts does not seem consistent due to the fact that different mechanisms of word formation may be applied to the same term. In our pursuit to map all the different occurrences of the same concept to only one, we propose a model designed within NooJ and based on dictionaries and morphological grammars. Starting from the manual detection of nouns and their variations, we recognize some word formation mechanisms and develop grammars suitable to recognize Latinisms and Croatinized Latin medical terminology.

Keywords: Medical terminology · Morphological grammars
Latin terms · Latinisms · Croatian · Latin · NooJ

1 Introduction

Health data produced in today's world can easily be classified as big data: it has volume, it has variety and it has velocity. But the main problem we face is that it mostly comes in an unstructured format. NLP can help bring structure to it and with that structure enable learning. This paper will present the first step of a quest in bringing understanding that lies behind unstructured Croatian medical texts.

Before any NLP research on medical data can be started, it is presumed that it exists in the digital format, or as some like to call it the EHR (*Electronic Health Records*). Not all physicians have been eager to transfer to such format, nor happy when it is prescribed to them regardless the benefits such format of

I. M. Mirto et al. (Eds.): NooJ 2018, CCIS 987, pp. 38–49, 2019.
https://doi.org/10.1007/978-3-030-10868-7_4

data enables [1] like easily shared data among physicians but also hospitals and pharmaceutical industry that can help each other learn faster about different treatment results, or why is some drug working in some cases and not in others. This kind of data is usually found in an unstructured format in physicians' and nurses' notes, or CAT scans and MRI readings, that according to research, make up from 50% [2] up to 80% [3] of clinical records. In order to learn from such data and use it to improve patient's care, we need to understand it. Performance of such tools has been demonstrated in [3] and is reported to have more than 90% accuracy in detecting diseases.

There are also other health related data found in the digital format. More and more medical devices are Internet-enabled and are generating our biometric data. There is also metadata about the health terms we search for on the internet.

Still, we are mostly talking about 'privacy regulations' of health data that is present in medical institutions and not as much about how we can learn from this data faster to better suit the needs of each patient. One of the ways is to use the power of NLP to give some structure to the unstructured text and to find the paths to hidden knowledge that lies in it. The importance of morphological processing of biomedical text is seen in more advanced NLP tasks like information extraction [4] and question answering.

One of the potential problems in mining medical texts is diversity of terminology used [5]. The main characteristic of any (English, German, French, etc.) medical language is presence of Latin and Greek. However, in Croatian medical texts, these languages are not solemnly used in its pure original form [6,7]. We have found four types of notations that refer to the same concept: (1) **pure Latin terms** (lat. *diabetes mellitus*); (2) **Croatian translations** (hr. *šećerna bolest*); (3) **Latinisms or Croatian terms** with visible Latin root (hr. *dijabetes melitus*) and (4) **Croatinised Latin words** (Latin root with Croatian case ending) (lat_hr. *diabetes melitusom*). Still, when we search for cases of medicines prescribed, or diagnostics used for *diabetes mellitus*, we would like our results to include the remaining notations as well. Thus, it is important to find a way to link all notations i.e. to normalize them.

In the remaining sections, we will describe our learning corpus, dictionaries of pure Latin [Category 1] and pure Croatian terms [Category 2], and two morphological grammars that recognize the remaining two notations [Categories 3 and 4]. Before we conclude, we will show and explain the results we obtained on our learning corpus.

2 Related Work

The use of neo-classical compounds and morphemes in word formation has been widely analyzed, due to their intense use, especially in some domains with a very long tradition, like medicine. This phenomenon has been studied with regard to different languages [8–11].

The common finding is that a relatively short number of Greek and Latin forms (stems, prefixes and affixes) yields a high number of specialized terms. Further studies aim at extracting semantic information referring to medical entities from raw texts and the identification of the semantic categories that describe the located entities [12].

As regards the first task, many medical lexical databases (e.g., Medical Subject Headings (MeSH), RxNorm, Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine (SNOMED), and Unified Medical Language System (UMLS), which includes all the other sources) can be used as knowledge base for the location of the medical entities. Anyway, the quick evolution of entity naming and the slowness of the manual development and updating of the resources often make it necessary to exploit some word formation strategies, that can be truly helpful in the automatic population of technical-scientific databases. Such strategies concern the morpho-semantic approach and have been successfully applied to the medical domain by [13] on terminal morphemes into an English medical dictionary; by [14] on medical formative elements of Latin and Greek origin; by [15] on the suffix -*itis*; by [16] on suffixes -*ectomy* or -*stomy* and by [17] on the suffix -*osis*. Among the most used tools for the Medical Entity Recognition (MER), we mention MetaMap [18], a reference tool which recognizes and categorizes medical terms by matching noun phrases in free texts to the corresponding UMLS Metathesaurus and Semantic Network, and MEDSYNDIKATE [19], a natural language processor able to automatically acquire data from medical findings reports.

With reference to the second task, we can find in literature rule-based, statistical and hybrid approaches. As regards the contributions that exploit statistical methods for the identification and classification of medical entities, we mention [20], that uses decision trees or SVMs; [21], that uses Hidden Markov Models or CRFs; [22], that presents a machine learning system which makes use of both local and syntactic features of the texts and external resources (gazetteers, web-querying, etc.); and [23], that obtains the nouns of disease, medical condition, treatment and symptom types, by using MQL queries and the Medlineplus Health Topics ontology. Rule-based methods are the ones proposed by [24], who identifies, with a set of graphical patterns, cause-effect information from medical abstracts in the Medline database, and [25], that manages to extract clinical entities disorders, symptoms and body structures from unstructured text in health records, using a rule-based algorithm.

Hybrid approaches have been proposed by [26] for the extraction of gene symbols and names; by [27] for protein name recognition and by [28], which combines terminology resources and statistical methods with sensible improvements in terms of Precision.

3 Corpus

For the preliminary learning phase, we chose two, relatively, small corpora that we have used for the purposes of terminology categorization. The first corpora (**MedNotes**) consists of 100 medical notes regarding the doctor's readings of MR, CT, X-ray and ultrasound images (total of 20.831 tokens). These documents were accessed with great difficulties taking all the necessary steps in protecting the patient's privacy and confidentiality of data (and General Data Protection Regulation (GDPR) as applied in May this year). Thus, the corpora has no mentions of any patient's name or any other personal information except for the gender inferred from the gender of word selections (feminine verbs, nouns and pronouns for female patients and masculine verbs, nouns and pronouns for male patients).

A second corpora (**MedInstruct**) consists of 100 randomly chosen instructions for the use of medicines (total of 213.275 tokens). These documents are not physician's notes but are written for the medical personal. Documents are published by the Agency for Medicines and Medical Products HALMED. Each instruction is written after the more-or-less same pattern and is much longer in length than the medical notes which explains the more numerous tokens then in the first corpora.

Our first assignment was to detect Latin-based nouns in both corpora and define their variations. Our data showed that terminology usage in Croatian medical texts is not consistent. So far, we have been able to detect 4 categories that needed somewhat different approaches for our NLP project:

1. Latin terms (usually written in italics in MedInstruct corpora),
2. Croatian terms,
3. Latinisms or Croatian terms with visible Latin root,
4. Croatinised Latin words (Latin word with Croatian case ending).

Some terms have been found in only two variations, e.g. Category 1 and Category 2: *vertigo* vs. *vrtoglavica*; or Category 1 and Category 3: *urticaria* vs. *urtikarija*; or Category 3 and Category 2: *agitacija* vs. *nemir* and *edem* vs. *otok*. There are also those terms that are found in all four categories. The best example is the term *diabetes mellitus* for which multiple versions are used inconsistently: *diabetes mellitus* (Category 1), *šećerna bolest* (Category 2), *dijabetes* (*melitus*) (Category 3), *diabetes melitusom* (Category 4).

The analysis shows that terminology from Category 1 have been used the least, usually for the names of microorganisms and common expressions like *in vitro*, *in vivo* etc. Some words used in nominative proved to be quite difficult to categorize since they can easily be found in Categories 1 and 4, like *fetus*, *agens*, *gastrosoma*, *lumen*, *tumor*, *uterus* or *retroperitoneum*. The only way to distinguish them was the usage of italics in the MedInstruct corpora. The reason for this ambiguity is that nominative form in both categories is the same.

Category 3 is the most numerous one. The words found in this category are loanwords adapted to the Croatian language by graphic system and appropriate declension. The rules of graphic changes are shown in Table 1.

As the examples from the table show, there are some words that have only one change of either vocal, diphthong or consonant, (sometimes, depending on the position) but there are also words with more changes, e.g. *erythema* to *eritem* and *resistentia* to *rezistencija* that have three changes each.

Some exceptions to these rules are represented by (i) the presence in Latin of two vowels that do not combine into a diphthong, e.g., ae in *aerob* and *aerobilija* and oe in *angioedem*, which in Croatian remain respectively ae and oe; (ii) the

Table 1. Change characteristics for Category 3

Letter	Condition	Latin	Croatian
æ > e	-	gangr**ae**na	gangr**e**na
œ > e	-	**œ**dem	**e**dem
y > i	-	s**y**ncopa	s**i**nkopa
ia > ija	except in words on *a*- w/prefix *anti*-	urticar**ia**	urtikari**ja**
ea > eja	-	diarrho**ea**	dijar**eja**
c > k	w/o -*i*, -*e* behind *c*	syn**c**opa	sin**k**opa
c = c	w/*i*-, *e*- behind *c*	fae**c**es	fe**c**es
ph > f	-	s**ph**incter	s**f**inkter
th > t	-	ery**th**ema	eri**t**em
ch > k	-	**ch**romaturia	**k**romaturia
rrh > r	-	dia**rr**hoea	dija**r**eja
ti > cij	w/vowels or diphthong	resisten**ti**a	rezisten**cij**a
ti = ti	w/*s, t, x* behind *ti*	*conges**tio**	konges**ti**ja
s > z	between vowels	hyperhydro**s**is	hiperhidro**z**a
x > ks	w/o a vowel behind *x*	radi**x**	radi**ks**
x > gz	w/a vowel behind *x*	e**x**anthema	e**gz**antem
dbl cons. > sgl cons.	-	ti**nn**itus	ti**n**itus

sequence *ea* in Latin, which is usually transposed in Croatian like *eja*, in some cases does not undergo any change, e.g., *urea* (and complexes with that word), *kreat(in)in*, *pankreas* (and complexes with that word), *proteaza*, *reapsorpcija* (hence all the words on *a*- with prefix *re*-); (iii) the transformation of double consonant into a single consonant is not applied to all words on *r*- with prefix *hiper*-. e.g., *hiperrefleksija.*

A noun of 2nd declension on *-ium* changes ending to *-ij*, e.g. *delirium* to **delirij*, *cranium* to *kranij* etc. A female noun of 3rd declension on *-tio* changes ending to -*cija*, e.g. *exacerbatio* to *egzacerbacija.* A 3rd declension neuter noun on -*ma* changes ending to -*m* e.g. *oedema*, *atis*, *n* to *edem*; *erythema, atis, n* to *eritem*; **carcinoma* to *karcinom.* Terms with suffix -*oma*, meaning a swelling or tumor, always change -*ma* to -*m.* However, there are more examples where such a noun switches gender in Croatian to become a female noun on -*a* e.g. *coma* > *koma*; *asthma* > *astma*; *plasma* > *plazma*; *stroma* > *stroma*; etc. A 3rd declension female noun ending in -*osis* changes the ending to –*oza*, e.g. *hyperhydrosis* > *hiperhidroza.*

Croatinised Latin words are Latin words with Croatian case endings (Category 4): 2nd declension male nouns on -*us* (e.g. *cryptococcus*, *bacillus*) and 2nd declension male nouns on -*um* (e.g. *sputum*) receive set of case endings characteristic for the Croatian word *stol* (en. table) (a – Gen; u – Dat; / – Acc; e – Voc; u – Loc; om – Inst), and 2nd declension female nouns on –a (e.g. *Candida*) receive set of case endings typical of the word *kuća* (en. house) (e – Gen; i – Dat; u – Acc; o – Voc; i – Loc; om – Inst).

Our goal is to observe the words from different categories as morphological variants that are mapped into the single term. The normalization mapping rules that we use for each category are explained in the following sections.

4 Designing the Dictionary

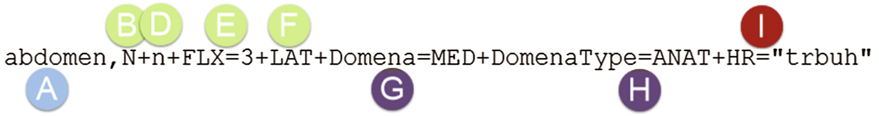
For the purposes of this project, we have designed two separate dictionaries depending on the language the term belongs to: Latin and Croatian. We found it important to keep these two language data separate for both maintenance and cross-language usability purposes. As the new, to us 'UNKNOWN' terms are detected in corpora, they are processed, annotated and added to the main dictionary, either *Lat_MedicalTerms.dic* or *Hr_MedicalTerms.dic.*

4.1 Medical Latin Terminology

Medical Latin terms contain Greek terms that have been Latinized, original Latin terms and artificially created terms, according to the rules of word compilation which combine Latin and Greek stems, prefixes, and suffixes. The development of medicine also develops medical terminology so medical terminology needs to be standardized and taught at the medical faculties. The standardization in medical terminology is a necessity both for the successful work of physicians and for the development of medical informatics (by using computer software to increase the quality of diagnosis and treatment of patients).

Medical Latin encompasses anatomical, clinical and pharmacological nomenclature and is continually revised. The first anatomical nomenclature Basle Nomina Anatomica (BNA), published in 1895, is repetitively revised. The last version of anatomical terminology from 1998 Terminologia anatomica (TA98) is still developing and new terms are introduced every year. It seems that medical terminology now has about one million terms. The development of medicine nomenclature implies medical Latin is not a "dead language", but is still living and developing.

At present, our dictionary of Latin terminology consists of 583 nouns. Each term is marked for word category (N), gender (m—f—n), flective paradigm (FLX = paradigm). Additional semantic annotations are added, where appropriate, describing the language of the term (LAT), the main domain the term belongs to (Domena = MED), one or multiple subdomains where the term is used (DomenaType = ANAT—BACTERIA—DISEASE—DRUG—FUNGUS—KEM—PLANT—PROC—TOOL), Croatian translation of the term (HR = *translation*). Example of an entry *abdomen* [marked with letter A] is visualized in Fig. 1. The word is annotated as a noun [B], neutral in gender [D] using paradigm 3 to build remaining cases for Latin in both singular and plural forms [E] and using LAT as a language marker [F]. Sections marked with letters [G] and [H] show that the word belongs to the medical domain and anatomy as its subdomain field respectfully. The last letter [I] marks Croatian translation that is also provided and further annotated in the Croatian dictionary (Fig. 2).



```
abdomen,N+n+FLX=3+LAT+Domena=MED+DomenaType=ANAT+HR="trbuh"
```

Fig. 1. Example of a dictionary entry for Latin terms used in Lat_MedicalTerms.dic dictionary

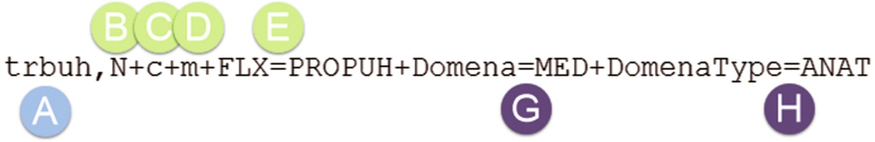
We refer to terms found in this dictionary as Category 1 medical terms. It is to be expected that this dictionary can be used by any other language projects since, except for the last section [Fig. 1.: I], annotations used remain unchanged regardless the language. Thus, sharing this resource with others remains one of our priorities in this project since we believe that it will help us learn faster and be more productive if the similar projects across the globe do not have to start their work from the scratch.

4.2 Medical Croatian Terminology

The medical Latin terms are introduced and adapted to Croatian language or translated to Croatian. This type of terminology is found in the second dictionary, Hr_MedicalTerms.dic, since it is more language specific than is the case with the previously described dictionary. Still, the logic in annotation remains the same.

All the words found in the Latin dictionary, have their partner words (i.e. Croatian translations) in the Croatian dictionary. The opposite is not supported which is evident from the number of terms found in this dictionary (2373). Thus, continuing with our Latin example (Fig. 1) we have *trbuh* in Croatian dictionary (Fig. 2). The term [A] is annotated as a noun [B], of common type [C], masculine in gender [D] using paradigm PROPUH to build remaining 7 cases for Croatian in both singular and plural forms [E]. Sections marked with letters [G] and [H] show that the term belongs to the same domain and subdomain as in the Latin dictionary. Letters [F] and [I] found in the Latin example, are not used for Croatian terms.

Terms found in this dictionary are referred to as Category 2 medical terms. The remaining two categories are words that are produced following certain morphological rules. Thus, we decided to recognize them via grammars and link



```
trbuh,N+c+m+FLX=PROPUH+Domena=MED+DomenaType=ANAT
```

Fig. 2. Example of a dictionary entry for Croatian terms used in Hr_MedicalTerms.dic dictionary

them to Category 1 medical term as their super-lemma [29]. We will explain this in more details in the following section.

5 Grammars

In order to recognize terms belonging to Categories 3 and 4, we have built two separate morphological grammars. We will explain each and discuss the problems we have encountered.

5.1 Grammar for Latinisms

Latinisms are Croatian terms with visible Latin root like *dijabetes melitus* that we recognize from Latin *diabetes mellitus*. These terms are classified as Category 3 and they all use specific morphological rules (see Table 1) to map Latin terms to Latinisms. These rules are consistent with rules for reading Latin. Grammar for recognizing and annotating described patterns (Fig. 3) requires the Latin term, from which Latinism is derived from, to exist in the dictionary. The rules make up a close set of IF-THEN statements such as ***if*** *(ae) in Latin* ***then*** *(e) in Latinism*. Building a grammar for these possibilities is quite straightforward in NooJ.

If each word could have only one change, our grammar would run much faster. However, there are words that undergo more than one change. Let us take the word *syncope* that comes from Latin *syncopa*. It's Latinism is *sinkopa*. In order to recognize it, we need to change *y* to *i* but also *c* to *k*. Another example is the Latin word *hyperhydrosis* that needs to change *y* to *i* twice and *s* to *z* to recognize Latinism *hiperhidroza*.

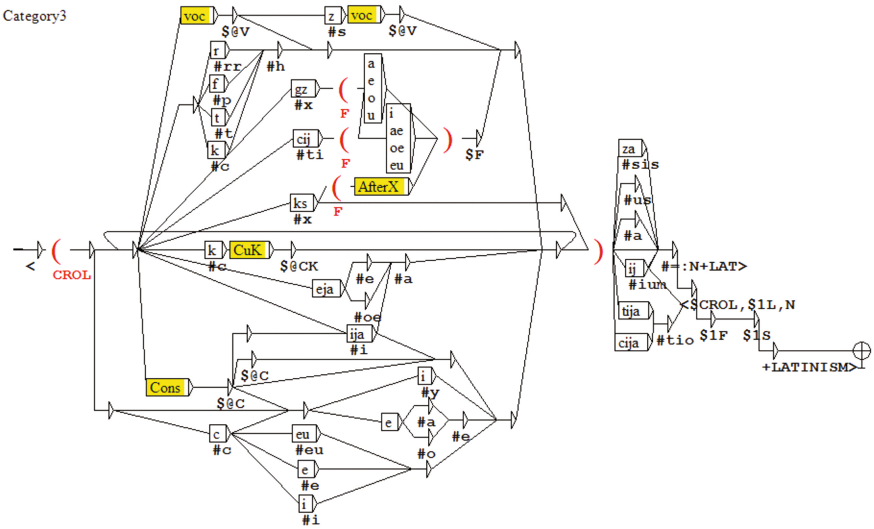


Fig. 3. Grammar for recognizing Latinisms in Category 3

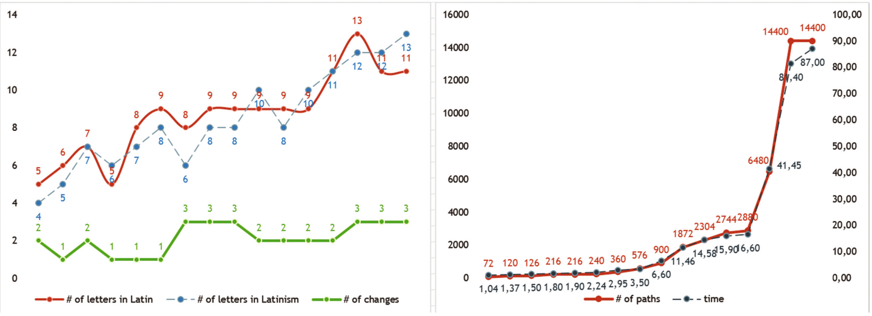


Fig. 4. Correlation of Latin vs Latinism word length vs number of expected changes (left graph) and number of paths vs run-time (right graph)

We have tested 16 example words from Table 1 to check what influences the time required for the grammar to check against all the possible varieties. As the graphs in Fig. 4 show, there is no clear correlation between the number of letters for either Latin words or Latinisms (in most cases, Latin words are longer) and the run time. The same is true for the number of expected changes, as well. There are six 9 letter words in our examples, with 2 or 3 expected changes, and each is taking different time to run, ranging from 2.24 s up to 15.90 s. But, what is correlated with longer run time, is the number of possible paths the grammar can take. This is best seen on a path for double consonants that are present in Latin words, but are not found in Latinisms. However, not all consonants are always doubled. But, since our grammar does not provide the context in which double consonants can occur, it assumes that there are no constraints for this rule and thus, every time it founds a consonant, it checks if it can pass the test that such a word exists in the Latin dictionary.

If we, for example, take the word *tinitus* that the grammar recognizes in the text, it will check for: *ttinitus, ttinnitus, ttinnittus, tinnitus, tinnittus, tinittus* although only *tinnitus* will pass the constraint that it exists as a noun marked with +LAT in the main dictionary. To resolve this problem, we need to find more specific context for such paths in our graph.

5.2 Grammar for Croatinised Latin

Words that are placed in the Category 4 have kept the original Latin spelling in Nominative case. However, all the other case endings belong to the case markers characteristic of Croatian (and not Latin). Thus, instead of finding Latin ablative *diabete mellito* we have *diabetes mellitus****om***.

The grammar for recognizing such Croatinised Latin words uses two variables **$LAT** and **$S** that hold any number of characters that, as such, exist in the Latin dictionary, and Croatian suffix, respectfully. The recognized string is also assigned Latin term as its super-lemma and a semantic label CROLAT (Fig. 5).

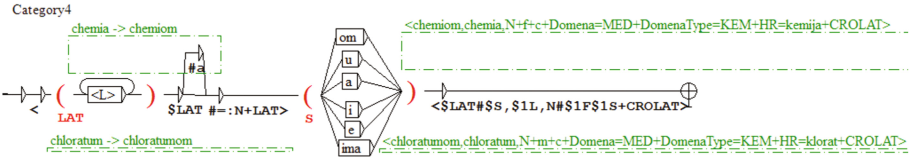

Category4
chemia -> chemiom
<chemiom,chemia,N+f+c+Domena=MED+DomenaType=KEM+HR=kemija+CROLAT>
#a
<
(
<L>
)
LAT
$LAT #=:N+LAT>
(
S
om
u
a
i
e
ima
)
<$LAT#$S,$1L,N#$1F$1S+CROLAT>
chloratum -> chloratumom
<chloratumom,chloratum,N+m+c+Domena=MED+DomenaType=KEM+HR=klorat+CROLAT>

Fig. 5. Graph for recognizing Latinisms in Category 4

6 Results

Latin-based nouns detected by human annotators are distributed across categories in the following manner: Category 1 - 10%; Category 2 - 19%; Category 3 - 70%; Category 4 - 1%. This distribution is only in our test corpus and we expect it to change as the diversity of corpora gets larger.

After our preliminary tests failed to recognize all the terms from Categories 1 and 2, data quality was checked in both dictionaries to fix some erroneous data entries. Also, there were some multy word units that we have decided to deal with at the later stages of the project. Both grammars have correctly recognized and annotated all the tested terms.

Now, when we search for the Latin term *delirium* we recognize all the cases of this term belonging to categories 1 *delirium*, 3 *delirij* and 4 *deliriumom*. In order to recognize Croatian term as well, or all categories when Croatian term *bunilo* is given as a search term, additional grammars will have to be designed to manage such normalization mapping rules.

7 Conclusion

In this paper we present our preliminary work on the the usage of Latin-derived words and formative elements in the development of Croatian medical terminology. We identified four types of this usage: (1) pure Latin terms; (2) Croatian translations; (3) Latinisms or Croatian term with visible Latin root and (4) Croatinised Latin words (Latin root with Croatian case ending). The model we propose here for linking together Latin-Croatian combinations, can also be reused for other languages that are present in medical texts, like Greek, English, or French. Our results give an account of its usefulness and permit to foresee a future fields of work such as (a) establishment of further constraints regarding form combination and formation; (b) analysis in detail of the neoclassical suffixes and prefixes; (c) further study of the combination of a form with general language word, combining the current module with an ontology; (d) test these results in larger corpora.

We also plan to test the model on a bigger and more diverse texts and, if needed, expand the existing grammar with new nodes and rules. In later stages, we plan to expand the research with similar morphological grammars that will recognize other word types such as adjectives and verbs.

Acknowledgement. This research has been partly supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

1. Schneier, B.: The Hidden Battles to Collect your Data and Control your World. Data and Goliath, London (2015)
2. Davenport, T.: Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press, Boston (2014)
3. Simon, P.: Too Big to Ignore: The Business Case for Big Data, vol. 72. Wiley, Hoboken (2013)
4. Liu, H., Christiansen, T., Baumgartner, W.A., Verspoor, K.: Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. J. Biomed. Semant. **3**(1), 3 (2012)
5. di Buono, M.P., Maisto, A., Pelosi, S.: From linguistic resources to medical entity recognition: a supervised morphosyntactic approach. ALLDATA **2015**, 82 (2015)
6. Poljak, Ž.: Quo vadis, Croatian medical terminology-should the diagnoses be written in Croatian, Latin or English? Acta Clinica Croatica **46**(1–Supplement 1), 121–126 (2007)
7. Gjuran-Coha, A., Bosnar-Valković, B.: Lingvistička analiza medicinskoga diskursa. JAHR **4**(7), 107–128 (2013)
8. Estopa, R., Vivaldi, J., Cabre, M.T.: Use of Greek and Latin forms for term detection. In: LREC (2000)
9. Herrero-Zorita, C., Moreno-Sandoval, A.: Medical term formation in English and Japanese. Rev. Cogn. Linguist. **13**(1), 81–105 (2015). Published under the auspices of the Spanish Cognitive Linguistics Association
10. Smith, G.L., Davis, P.E., Soltesz, S.E.: Quick Medical Terminology. In: Smith, G.L., Davis, P.E. (eds.) Consultation with Shirley Soltesz, E. Wiley, Hoboken (1972)
11. Piñero, J.M.L., Terrada, M.L.: Introducción a la terminología médica. Elsevier, España (2005)
12. Abacha, A.B., Zweigenbaum, P.: Medical entity recognition: a comparison of semantic and statistical methods. In: Proceedings of BioNLP 2011 Workshop, pp. 56–64. Association for Computational Linguistics (2011)
13. Pacak, M., Pratt, A.: Identification and transformation of terminal morphemes in medical English Part II. Methods Inf. Med. **17**(02), 95–100 (1978)
14. Wolff, S.: The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. Methods Inf. Med. **23**(04), 195–203 (1984)
15. Pacak, M.G., Norton, L., Dunham, G.S.: Morphosemantic analysis of-itis forms in medical language. Methods Inf. Med. **19**(02), 99–105 (1980)
16. Norton, L., Pacak, M.G.: Morphosemantic analysis of compound word forms denoting surgical procedures. Methods Inf. Med. **22**(01), 29–36 (1983)
17. Dujols, P., Aubas, P., Baylon, C., Grémy, F.: Morpho-semantic analysis and translation of medical compound terms. Methods Inf. Med. **30**(1), 30–35 (1991)
18. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association (2001)
19. Hahn, U., Romacker, M., Schulz, S.: Medsyndikate-a natural language system for the extraction of medical information from findings reports. Int. J. Med. Inform. **67**(1–3), 63–74 (2002)

20. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on Computational linguistics, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
21. He, Y., Kayaalp, M.: Biological entity recognition with conditional random fields. In: AMIA Annual Symposium Proceedings, vol. 2008, p. 293. American Medical Informatics Association (2008)
22. Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., Sinclair, G.: Exploiting context for biomedical entity recognition: from syntax to the web. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 88–91. Association for Computational Linguistics (2004)
23. de la Villa, M., Aparicio, F., Maña, M.J., de Buenaga, M.: A learning support tool with clinical cases based on concept maps and medical entity recognition. In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, pp. 61–70. ACM (2012)
24. Khoo, C.S., Chan, S., Niu, Y.: Extracting causal knowledge from a medical database using graphical patterns. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 336–343. Association for Computational Linguistics (2000)
25. Skeppstedt, M., Kvist, M., Dalianis, H.: Rule-based entity recognition and coverage of snomed ct in swedish clinical text. In: LREC, pp. 1250–1257 (2012)
26. Proux, D., Rechenmann, F., Julliard, L., Pillet, V., Jacq, B.: Detecting gene symbols and names in biological texts. Genome Inform. **9**, 72–80 (1998)
27. Liang, T., Shih, P.-K.: Empirical textual mining to protein entities recognition from PubMed corpus. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 56–66. Springer, Heidelberg (2005). https://doi.org/10.1007/11428817_6
28. Roberts, A., Gaizauskas, R.J., Hepple, M., Guo, Y.: Combining terminology resources and statistical methods for entity recognition: an evaluation. In: LREC (2008)
29. Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley, London (2016)