A Pilot Study for Enriching the Romanian WordNet with Medical Terms

Maria Mitrofan
ICIA
Romanian Academy
maria@racai.ro

Verginica Barbu Mititelu
ICIA
Romanian Academy
vergi@racai.ro

Grigorina Mitrofan
National Institute of Diabetes
and Metabolic Diseases
mitrofan.grigorina@gmail.com

Abstract

This paper presents the preliminary investigations in the process of integrating a specialized vocabulary, namely medical terminology, into the Romanian wordnet. We focus here on four classes from this vocabulary: anatomy (or body parts), disorders, medical procedures and chemicals. In this pilot study we selected two large concepts from each class and created the Romanian terminological (sub)trees for each of them, starting from a medical thesaurus (SNOMED CT) and translating the terms, process which raised various challenges, all of them asking for the expertise of a specialist in the health care domain. The integration of these (sub)trees in the Romanian wordnet also required careful decision making, given the structural differences between a wordnet and a terminological thesaurus. They are presented and discussed herein.

1. Introduction

The experiment presented here is to be understood within the larger context of medical terms identification, automatic extraction and classification of relations between them.

For the automatic identification and classification of the relations between terms occurring in a corpus, we need access to a knowledge source. This can be either a corpus annotated with medical terms and relations between them or a resource containing terms and relations between them. For Romanian, a medical corpus (of more than 100,000 tokens) was annotated with medical entities (Mitrofan, 2017). This corpus contains four top level entity classes (Anatomy, Chemicals and Drugs, Disorders, Procedures) corresponding to UMLS (Unified Medical Language System) semantic groups, but no semantic relation is annotated between these entities.

We present here a pilot study of the steps necessary for the development of an ontology-like resource for Romanian medical terminology. Translating SNOMED CT is a huge effort, which takes too long and also requires funding for finding the professionals to do this. However, taking advantage of the existence of a wordnet for Romanian (see below, section 3.1.1.), we decided to expand its set of medical terms with structured information from SNOMED CT (see below, section 3.1.2.) in the form of subtrees containing medical terms of four semantic groups, corresponding to the ones with which the Romanian medical corpus is annotated: anatomical ones, terms designating disorders, medical procedures and chemicals. Given our interest in the diabetes domain, the terms we selected are specific to it, as shown below. Our aims were to establish a work methodology and to discover the difficulties encountered during the translation process and methods to address them. By no means are the selected terms and all their subtrees (when existent) representative for all possible cases of translation equivalents.

2. Related Work

Even though the current version of Princeton WordNet has a broad coverage of the medical terminology, it has certain issues which reflect both the fact that the WordNet was not built for domain-specific ap-

Keywords: WordNet, Romanian Wordnet, medical terminology

plications and the need of specialized expertise when compliance with medical terminologies is needed (Bodenreider et al., 2003).

In the literature several reports on enriching wordnets with medical terms are available. Barry and Fellbaum (2004) presented an important initiative in this direction. They attempted to create a free-standing lexical resource designed for natural language processing applications in the medical domain, Medical WordNet (MWN). This lexical database contains medical terms used by non-expert subjects and two sub-corpora, Medical FactNet (MFN) and Medical BeliefNet (MBN). The former contains validated sentences that illustrate medically relevant vocabulary and the latter contains statements which are believed to be true by non-experts.

Recently, another attempt at integrating medical terms from the National Cancer Institute Thesaurus (NCIt) into Open Multilingual Wordnet (OMW) was presented (Hicks et al., 2018). Two methods were investigated: one which automatically calculates mappings between the literals and their definitions in the two resources, and another one, manual, was the analysis of the Princeton WordNet (PWN) v.3.0 coverage of the thesaurus. Both methods involved literals and glosses alike.

Toumouth et al. (2006) addressed the possibility of mapping a general ontology (PWN) to a specific domain (medicine) by extracting 57887 pairs of nouns separated by a conjunction from Ohsumed corpus (W. Hersh and Hickam, 1994) and automatically introducing them into PWN. For the domain-specific words with multiple senses the accuracy was 46% and the recall 51%.

As far as the Romanian wordnet is concerned, we are not aware of any such initiative.

3. The Experiment

This section contains the description of the resources used and the methodology observed in our pilot study of investigating the necessary steps to follow and decisions to be made for an appropriate integration of the medical terms in wordnet, so as to serve the aim of helping in the automatic identification of terms and of relations between them in a medical corpus.

3.1. Resources

Two types of resources were used: two wordnets and a hierarchical thesaurus of medical terms. One wordnet is the Romanian one, which is to undergo enrichment. The other one was the English wordnet: it is richer than the former and, consequently, it can be referred to when drawing a comparison between common language hierarchies and the domain-specific ones (see below, section 4.2.).

3.1.1. Wordnets

Princeton WordNet

This lexical resource (Fellbaum, 1998) contains words mainly from the general language, although some lexical items specific to various domains do occur in it: consider the literal *oxidized LDL cholesterol*, specific to the medical domain, or *levodopa* from pharmacology. The words are nouns, verbs, adjectives and adverbs, each class with its own internal organization: nouns and verbs are distributed in hierarchies, descriptive adjectives are organized in bipolar clusters, while adverbs and relational adjectives lack an organization. For the present study we are interested only in the noun component of this resource.

PWN is a semantic network in which the synonymy relation between word senses is the condition for grouping them together in the nodes of the network: the nodes are called *synsets*, as they represent sets of synonymous words. The structure of the network of nouns is given by the semantic relation of hyperonymy between word senses in the nodes[1]. One hyperonym can have several hyponyms and a hyponym can also have more than one hyperonym.

The Romanian WordNet

The wordnet for Romanian (RoWN) (Tufiș and Barbu Mititelu, 2014) has been developed following the expand method, which implied its alignment to PWN, that is importing the structure and creating the Romanian equivalents of the English synsets. At the moment, RoWN is aligned to PWN v.3.0 and contains almost 60,000 synsets.

[1] Another relation linking the noun synsets is meronymy, with scarce interest for the present study.

3.1.2. Thesaurus

SNOMED CT

Over the past decades the need for common, controlled, sharable and reusable medical terminologies has been widely recognized. Therefore, there is a growing interest in comprehensive medical terminologies that allow consistent ways of storing, indexing, and retrieving medical data. In order to fulfill these requirements several concept representation systems were created. For example, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a comprehensive clinical terminology covering procedures, clinical findings and diseases, which was created by health-care specialists in order to structure and to reduce the variability of the way medical data is used. SNOMED CT is a multiaxial nomenclature that contains more than 1 million distinct medical terms, 326,734 active concepts (2017 release), 19 upper level hierarchies [2].

A SNOMED CT concept is described by a unique name, a unique numeric code and descriptions (one more frequent term and one or more synonyms). SNOMED CT is officially released in English[3] and Spanish[4]. A UMLS License is required to access SNOMED CT and to use the UMLS SNOMED CT Browser. Free of charge accounts are possible for scientific research in medical informatics.

3.2. Methodology

Making a pilot study, we did not target a complete coverage of the medical domain or at least of the diabetes subdomain. We worked with medical terms from four large categories extracted from the UMLS semantic groups, which are important for our further work: body parts (anatomy), disorders, procedures and chemicals. These terms are: body parts: *pancreas* (En. *pancreas*), *glandă endocrină* (En. *endocrine gland*); disorders: *boală metabolică* (En. *metabolic disorder*), *nefropatie diabetică* (En. *diabetic nephropathy*); procedures: *glicemie* (En. *blood glucose*), *hemoleucogramă* (En. *complete blood count*); chemicals: *colesterol* (En. *cholesterol*), *sânge* (En. *blood*).

We started from SNOMED CT, so from the English terms, and we translated them and all their direct and indirect hyponyms into Romanian. Given that the terms involved are not easily understood by a non-specialist, we had a physician assist the translation. Whenever a series of terms can be used as synonyms, i.e. to refer to the same concept, they are all added to the same node, in random order.

In order to enhance the RoWN with medical terminology, a translation methodology should be taken into consideration (Reynoso et al., 2000), because very often a term-to-term translation is either not possible or inappropriate.

This section is not meant as a thorough presentation of the challenges of terms translations. That is why we discuss only several of the problems we confronted with when translating the SNOMED CT selected terms.

In general, in domains where discoveries and innovations easily spread from one language community to another, specialized lexicons often incorporate new terms by borrowing them from the language of origin, either for a short, transitional period, during which an indigenous word is created and spreads among the specialists, or for ever. Such borrowing are, in Romanian, *by-pass*, *gastric-sleeve*, *pacemaker*, which are used in the medical community. Even though these terms have Romanian equivalents, they need to be included in the lexicon based on the frequency of their usage. Moreover, some of the borrowed terms are usually adapted to the pronunciation and the morphology of the Romanian language *pic* for *peak*, *șuntare* from *shunt*. So, adaptation should also be taken into consideration in the process of enhancing the RoWN with medical terminology.

In the medical domain, the usage of acronyms is a widely accepted practice. As far as English and Romanian are concerned, some acronyms are used in both languages, although they reflect the term structure only in one of the languages: see the use of *HDL* in both the English *HDL cholesterol* and the Romanian *colesterol HDL*. *HDL* is the acronym for *high-density lipoprotein*, so an English structure. Its

[2]https://www.snomed.org/snomed-ct/snomed-ct-worldwide

[3]https://www.snomed.org/about

[4]https://confluence.ihtsdotools.org/display/RMT/2017/10/12/October+2017+SNOMED+CT+Spanish+Edition+Beta+release+available+to+Members

Romanian translation is *lipoproteina cu densitate înaltă*. However, the Romanian medical terminology does not include a term such as *colesterol LDI*, which would be the normal translation of the English term. Instead, only the word is translated and is used with the English acronym, that is *colesterol HDL*. What is more, the most frequently occurring form of the term is, in fact, *HDL colesterol*, which displays a word order that is not specific to Romanian (in which the modifier usually occurs postnominally), but to English (in which the modifiers occur before the modified noun). On the one hand, such examples show that the acronym transfer can be obligatory and must be encoded as a synonym of the term, that is as members of the same synset in wordnet. It can be doubled by the Romanian acronym when existing and being used in the literature. Some other times the terms in each language have their own acronym which is used by specialists and thus the acronym borrowing becomes unnecessary: consider *AIDS* in English and *SIDA* in Romanian, where the English abbreviation is not used. On the other hand, the example shows that structural differences between languages can be overridden in terminology translation.

Turns of phrases occur in term translation: a noun phrase like *120 minute blood glucose measurement* is reorganized: the equivalent of the modifier *blood glucose* becomes the head of the Romanian term and the English head noun *measurement* gets translated as a participle (*măsurată*) modifying this noun: *glucoză măsurată la 120'*.

Such an example is also illustrative of the fact that terms may not always translate in the same way: some of the occurrences of the noun *measurement* in the hyponyms of *blood glucose* are translated by the adjective *măsurată* (e.g., *120 minute blood glucose measurement* translates as *glicemia măsurată la 120 de minute*, etc.), others are translated by the noun *determinarea* (e.g., *capillary blood glucose measurement* translates as *determinarea glicemiei capilare*, etc.).

In conclusion, a well-developed methodology for translating the medical terms is needed mostly because they may be translated incorrectly when following a term-to-term approach: consider also the alteration, be it even slight, of the adjectives order between the corresponding terms *idiopathic transient neonatal hyperinsulinemia* and *hiperinsulinemie idiopatică neonatală tranzitorie*. Usually, more adjectives modifying the same English noun are translated in reverse linear order in Romanian, but it is not the case here.

4. Data Analysis

Data analysis implies discussing to what extent the selected terms and the trees whose roots they are occur in the wordnets and whether their structure is similar or not in the two types of resources, so as to help us to decide on the procedure for their integration in the RoWN.

4.1. Wordnet Coverage of Selected Terms

A first point of interest is the existence of the selected terms in the resource that is to be enriched, namely the RoWN. We notice that three out of the eight terms have not been implemented in RoWN yet: *hemoleucogramă*, *nefropatie diabetică*, *glicemie*. However, for the last two of them their English counterparts do not exist in PWN either: the term *diabetic nephropathy* is not recorded in PWN, while the term *blood glucose* is recorded, but with a different meaning from the one that is of interest for us, namely the medical procedure, not the chemical substance.

For the terms occurring in RoWN, we extracted their hyponymy and mero-part relations and compared them to the hierarchical structure of the corresponding terms in SNOMED CT. The reason for choosing two relations in the wordnet, but only one in the thesaurus will become clear from the trees analysis presented below.

Out of the five terms occurring in RoWN, two are leaves in the network: *colesterol* and *pancreas*. Whereas the former has two children in PWN (*HDL cholesterol* and *LDL cholesterol*[5]), which have not been implemented in RoWN yet, the latter is also a leaf in PWN.

As a consequence, there are only three terms that have subtrees in RoWN: an anatomical one (*glandă endocrină*), one designating a disorder (*boală metabolică*) and one from the category chemicals (*sânge*).

[5]Comparing *cholesterol*, on the one hand, and *HDL cholesterol* and *LDL cholesterol*, on the other hand, from the perspective of terminology formation, we notice that the hyponyms are formed, in this case, by modifying the hyperonym. As other terms in this paper show, this is not the only way of creating specialized terms.

The trees headed by the synsets they belong to were extracted from RoWN and compared to the trees headed by their equivalent synsets in PWN (for an online synchronous visualization of PWN and RoWN synsets visit http://dcl.bas.bg/bulnet/).

For all these terms we notice that the Romanian coverage of the English equivalents is quite high: *endocrine gland* has 16 hyponyms and 13 of them are implemented as hyponyms of *glandă endocrină*, *blood* has 10 hyponyms and 8 of them are implemented as hyponyms of *sânge*, and *metabolic disorder* has 6 hyponyms and 3 of them are implemented as hyponyms of *boală metabolică*.

These (sub)trees depths are quite low and identical for the two wordnets: *endocrine gland* and *glandă endocrină* have a depth of 3 levels, *blood* and *sânge* have 2 levels, and *metabolic disorder* and *boală metabolică* have a depth of 3, respectively 2, levels. The trees for the last two terms are rendered in Figure 1. The identically coloured edges are used for highlighting the hyperonymy relations of the same child. For example, the node *abetalipoproteinemia* has two hyperonyms: *hypobetalipoproteinemia* and *lipidosis*. The edges between the hyponym and each of its hyperonyms are the same colour.

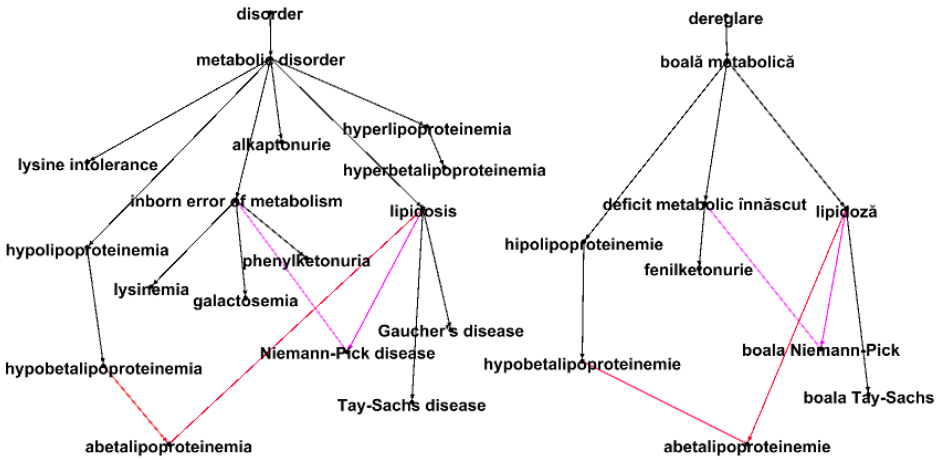

Figure 1: The subtrees of *metabolic disorder* in PWN (left) and of *boală metabolică* in RoWN (right)

Analyzing the direct and indirect hyponyms of these terms, as well as the relations in which these hyponyms are involved in the wordnet we can make several remarks:

- some hyperonyms share a part of their hyponyms: consider the term *boala Niemann-Pick* (En. *Niemann-Pick disease*) which is a hyponym of *deficit metabolic înnăscut* (En. *inborn error of metabolism*) and also of *lipidoză* (En. *lipidosis*). These two hyperonym terms are, in their turn, cohyponyms whose hyperonym is *boală metabolică* (En. *metabolic disorder*) - see Figure 1;
- some cohyponyms are also involved in the mero-part relation: this is the case with several terms in the (sub)tree of *glandă endocrină*: see Figure 2. For example, the term *cortex adrenal* (En. *adrenal cortex*) is both a cohyponym of *glandă suprarenală* (En. *adrenal gland*) and a meronym of it. There is one case when even a part and a subpart of this part are cohyponyms with the whole: *pars intermedia*, *posterior pituitary* and *pituitary gland* are cohyponyms of the hyernym *endocrine gland*. However, *pars intermedia* is a meronym of *posterior pituitary*, and *posterior pituitary* is a meronym of *pituitary gland* - see the same figure.

4.2. Comparison between Wordnet Hierachies and SNOMED CT Hierarchy

A first remark on the corresponding (sub)trees in wordnets and in SNOMED CT concerns their depth: in the case of the terms with hyponyms in wordnet (*endocrine gland*, *blood* and *metabolic disorder*) the difference is rather small: with the exception of *endocrine gland*, which has 3 levels in wordnet and 4 in SNOMED CT, the other (sub)trees have the same depth in both resources: *blood* has 2 levels and *metabolic disorder* has 3.

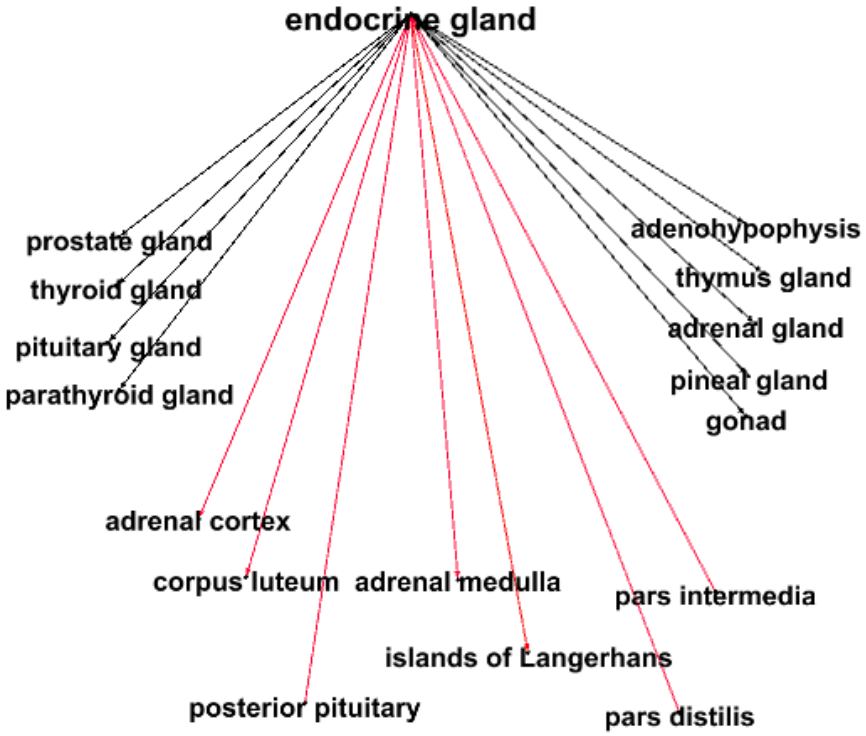

Figure 2: The subtree of *endocrine glands* in PWN

The terms that are not present in wordnets, namely *blood glucose* and *diabetic nephropathy*, have a 3-level depth in SNOMED CT. Moreover, these terms have a large number of children in the SNOMED CT hierarchy: each of them has 11 children. In Figure 3 we illustrate the hierarchy for *blood glucose* in SNOMED CT.

From a medical point of view the assessment of blood glucose is the most important paraclinical measurement, being used for screening, diagnosis and follow up of the patients with diabetic disease. Therefore if one needs to perform named entity recognition (NER) or relation extraction (RE) using PWN or RoWN the annotation of this term is essential. Consequently, because *blood glucose* is one of the most frequently used medical term in the diabetes domain, it should be introduced in both RoWN and PWN.

The terms that are present in PWN but lack any hyponyms are *complete blood count* and *pancreas*. In SNOMED CT, the same terms have different depths: the former has a 2-level depth, while the latter has a 7-level depth. They have three, respectively six children. We notice again the richness of the subtrees headed by these concepts, thus proving their importance in the medical domain.

The case of *cholesterol* is suggestive of the selectivity of wordnets, which, as resources aiming at reflecting the general language rather than the domains of activity, out of the big number of specialized terms encode only those that are accessible to all people: in PWN there are only two children for *cholesterol*: *LDL cholesterol* and *HDL cholesterol*. A further proof of the accessibility of these terms to non-specialists and of their use in the general language is the creation, in the common language, of two synonyms whose meaning is more transparent for patients given the use of very common adjectives in the structure of these terms: they are *colesterol rău* (En. *bad cholesterol*) and, respectively, *colesterol bun* (En. *good cholesterol*). In SNOMED CT, however, the concept *cholesterol* has 8 other children besides these two. Moreover, among all these ten children, the two mentioned above have the highest number of children, in their turn: *HDL cholesterol* has 7 children and *LDL cholesterol* has 5 children.

All these terms discussed so far in this section show that wordnet hierarchies tend to be simpler, thus more accessible to non-specialist people and useful in Natural Language Processing when dealing with

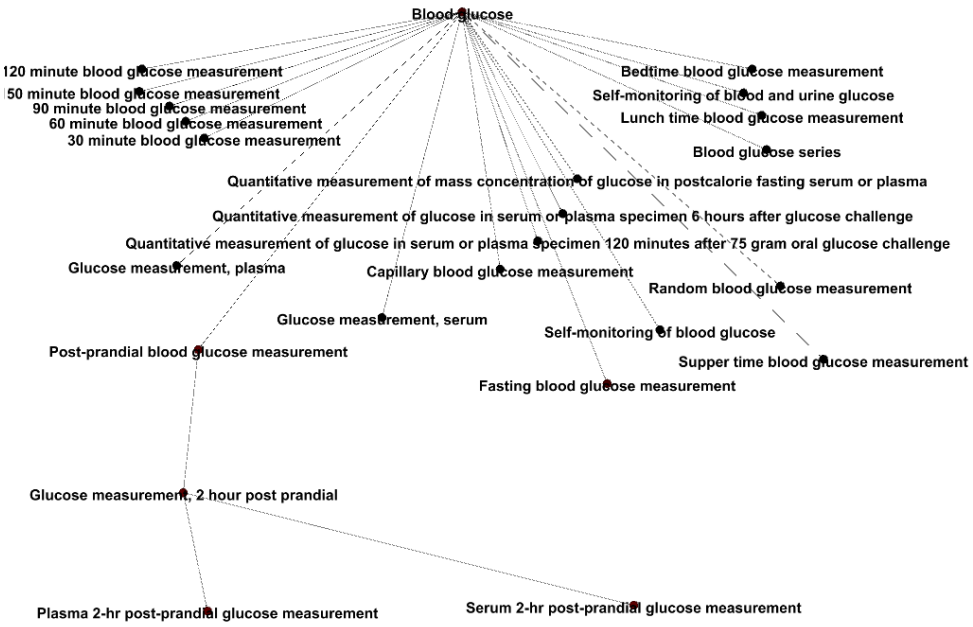

Figure 3: The subtree of *blood glucose* in SNOMED CT

texts from general language, rather than domain-specific ones (Bodenreider et al., 2003). The term that we discuss below is meant to highlight further shortcomings of wordnets when dealing with (at least) medical texts and terms, namely conceptual organization of terms in partial contradiction to medical knowledge.

When comparing the subtrees[6] for *endocrine gland* in PWN (see Figure 2) and in SNOMED CT (see Figure 4), we notice several mismatches of the same kind: several parts of endocrine glands are encoded as such (with the help of the mero-part relation), but also as co-hyponyms of the nouns designating these glands. Consider the *adrenal cortex*. In PWN it is in mero-part relation with *adrenal gland*, but also its co-hyponym, both sharing the hyperonym *endocrine gland*. However, in SNOMED CT it is a hyponym of *layer of adrenal gland*, which has *endocrine gland* as hyperonym; at the same time, *adrenal cortex* is a part of the *adrenal*. A similar, although slightly more complicated case, is that of *pars intermedia*: in PWN it is a part of *posterior pituitary* which, in its turn, is a part of *pituitary gland*; both *posterior pituitary* and *pars intermedia* are co-hyponyms of *pituitary gland*, all having the hyperonym *endocrine gland*. In SNOMED CT *pars intermedia* is a hyponym of *adenohypophysis* which is a hyponym of *pituitary part*. So, we notice the wrong attachment of *pars intermedia* as a part of the *posterior pituitary* instead of the *adenohypophysis* in PWN, alongside its inappropriate attachment as a hyponym of *endocrine gland*. All the red edges in Figure 2 mark (PWN) relations that are non-conformant to the ones in green in Figure 4 (in SNOMED CT).

We only mention now, but leave it unexplained given space constraints, that the hierarchy of *metabolic disorder* in wordnets (as represented in Figure 1) does not conform with the SNOMED CT hierarchy.

5. Integrating Medical Terms in RoWN

Analyzing the eight different cases of diabetes related general terms we dealt with in this pilot study, we can identify several scenarios for their implementation in RoWN, together with all their hyponyms. These scenarios depend on the current status of the implementation of these terms in RoWN and PWN and the cases are: (i) the term is not implemented in RoWN and it is not in PWN either: this is the case

[6]These subtrees are quite dense in both these resources and, given space constraints, we needed to simplify them so as to serve the purpose of the discussion in this section.

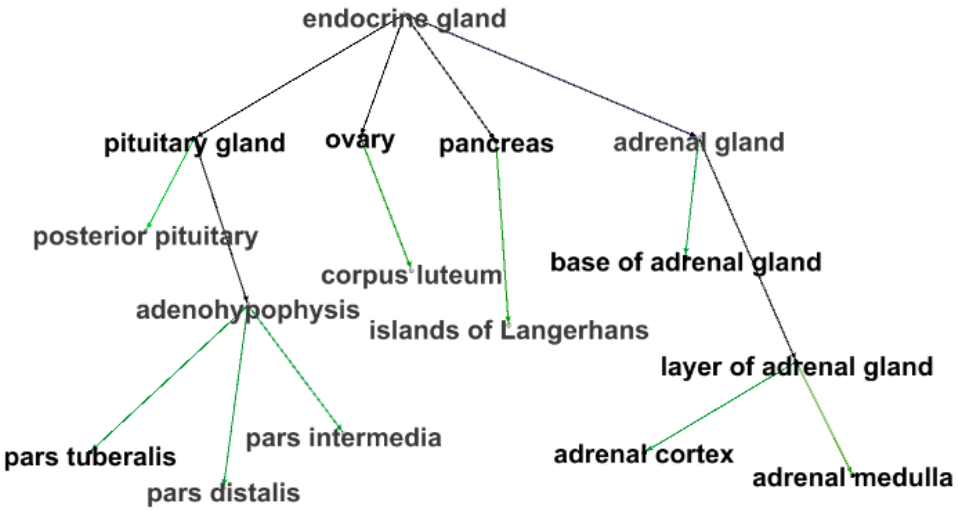

Figure 4: The subtree of *endocrine gland* in SNOMED CT

of *glicemie* (En. *blood glucose*) and of *nefropatie diabetică* (En. *diabetic nephropathy*); (ii) the term is not implemented in RoWN but it is in PWN, where it is a leaf: this is the case of *hemoleucogramă* (En. *complete blood count*); (iii) the term is implemented both in RoWN and in PWN as a leaf: see the case of *pancreas*; (iv) the term is implemented both in RoWN and in PWN, but it has children only in PWN, in RoWN being a leaf: this is the case of *colesterol*; (v) the term is implemented in both RoWN and PWN and has children in both of them: see the cases *boală metabolică* (En. *metabolic disorder*), *sânge* (En. *blood*) and *glandă endocrină* (En. *endocrine gland*).

For these cases we propose the following respective scenarios: (i) go up, in parallel, in the PWN and SNOMED CT hierarchies until corresponding nodes are found; translate and implement in the RoWN the terms and their hyponymic organization in SNOMED CT starting from this common node and going down until the leaves level; keeping track of them is ensured by assigning them a unique distinctive ID[7]; (ii) translate the general term from PWN and implement it in RoWN with the same ID as in PWN; find its equivalent in SNOMED CT and translate and implement all children down to the leaves level, importing also their hyponymic structure; all children get a unique distinctive ID; (iii) the same as at (ii) above, but skipping the implementation of the general term; (iv) if the children that are already in PWN are also children of the same general concept in SNOMED CT, then implement them also in RoWN, with the same ID as in PWN; if their structure is different from the one in SNOMED CT, see the next scenario; translate and implement all the other children (if any) and all children's children down to the leaves level, importing also their hyponymic structure; all children that were not in PWN get a unique distinctive ID; (v) this is the most challenging situation, because it has to deal with the discrepancies between the two types of resources, some of which are presented above. Modifying the existent RoWN structure and completely replacing it with the one from SNOMED CT is excluded because the alignment of RoWN and PWN confers the former (but also the latter) a great value for bi-/multilingual applications involving natural language processing (Tufiș et al., 2004). As a consequence, we can create a parallel medical structure in the RoWN: the nodes that do not conform with the SNOMED CT organization will be doubled and marked distinctively. The result will be the existence of two synsets containing the same literals, but establishing different relations in the network. Applications requiring access to medical data will ignore the wordnet-specific relations and make use of the SNOMED CT-specific ones in such cases.

[7]The synsets in RoWN have an ID identical with the one in PWN, thus ensuring the alignment of the two resources.

6. Conclusions

Resources such as SNOMED CT are expensive both time-wise and money-wise. For languages such as Romanian, lacking such a thesaurus, an alternative solution can be the integration of necessary knowledge from SNOMED CT in the existing wordnet. However, as we showed here, the process involves making important decisions at two levels: finding the Romanian equivalents of the English terms in SNOMED CT (see section 3.2.) and establishing various scenarios for their integration in the wordnet (see section 5.). These are not trivial, given the importance of having aligned wordnets for different languages, thus the inefficiency of a solution that would modify the wordnet structure.

Even though the wordnet was not built for domain-specific applications it can be enriched with specialized terminologies (medical) extracted from already existing specialized ontologies (SNOMED CT) in order to perform terms identification or relations extraction.

References

Barry, S. and Fellbaum, C. (2004). Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*.

Bodenreider, O., Burgun, A., and Joyce, A. M. (2003). *Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study.*

Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hicks, A., Seppälä, S., and Bond, F. (2018). Toward Constructing the National Cancer Institute Thesaurus Derived WordNet (ncitWN). In *Proceedings of the 9th Global WordNet Conference, Singapore*.

Mitrofan, M. (2017). Bootstrapping a romanian corpus for medical named entity recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 501–509.

Reynoso, G., March, A., M Berra, C., P Strobietto, R., Barani, M., Iubatti, M., P Chiaradio, M., Serebrisky, D., Kahn, A., Vaccarezza, O., L Leguiza, J., Ceitlin, M., Luna, D., Quirós, F., I Otegui, M., Puga, M., and Vallejos, M. (2000). Development of the spanish version of the systematized nomenclature of medicine: methodology and main issues. In *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 694–8, 02.

Toumouth, A., Lehireche, A., Widdows, A., and Malki, M. (2006). Adapting wordnet to the medical domain using lexicosyntactic patterns in the ohsumed corpus. In *Computer Systems and Applications, 2006.*, pages 1029–1036.

Tufiș, D. and Barbu Mititelu, V. (2014). *Language Production, Cognition, and the Lexicon*, chapter The Lexical Ontology for Romanian. Springer.

Tufiș, D., Ion, R., and Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva.

W. Hersh, C. Buckley, T. J. L. and Hickam, D. (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *In Proceedings of the 17th annual conference on Research and Development in Information Retrieval (SIGIR-94)*, pages 192–201.