Medical Knowledge Resources for Text-Mining of Health Records in Czech, Polish, and Slovak

Krištof Anetta

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic
xanetta@fi.muni.cz

Abstract. Knowledge extraction from medical text in small languages like Czech, Polish or Slovak is challenging due to the insufficiency of language-specific medical resources (pretrained models, ontologies, dictionaries). This paper is a survey of noteworthy options for researchers targeting these languages, divided into two sections. First, since the UMLS Metathesaurus for English is by far the most extensive and detailed medical knowledge resource in Western medicine, appreciable results can be achieved by machine-translating the mined text to English – therefore, the relevant English components of UMLS are introduced. Second come the language-specific resources for each language, detailing the publishing institutions, current website locations, contents, and file formats. The contribution of this paper is in collecting and pre-screening widely disparate sources needed for successful medical knowledge extraction in Central European Slavic languages.

Keywords: EHR, electronic health records, healthcare text, UMLS, ICD-10, SNOMED CT, MedDRA, MeSH, NLP, natural language processing, Slavic languages, Polish, Czech, Slovak

1 Introduction

In small, low-resourced languages, producing a well-trained deep learning model for knowledge extraction from medical text is often impossible, the main culprits being limited data availability and even more limited capacity for expert annotation. Therefore, even as medical records in English have been revealing their secrets, small language medical corpora have remained a largely untapped resource of valuable information for both science and medical practice.

In small languages, vocabulary-based knowledge extraction methods are the keystone of all other efforts, identifying the literal occurrences of known medical concepts and, wherever possible, linking them to an existing ontology of medical knowledge.

Since it is highly labor-intensive to create new medical vocabularies that are properly linked to ontologies, the natural first step is to leverage existing resources, optimizing them for the task of medical knowledge extraction. This paper is a survey of major resources available for Czech, Polish, and Slovak and of the opportunities and limitations inherent in them.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 163–171, 2022. © Tribun EU 2022

If not stated otherwise, the focus is on resources that make it possible to locate medical concepts and entities with an unambiguous identifier within existing global coding/classification systems (ICD, ATC) or other major integrative initiatives (UMLS). This is to increase interoperability and enable international comparison in medical knowledge extraction.

2 UMLS: International resources for machine translation approaches

Even though the Unified Medical Language System [2], maintained by the United States National Library of Medicine, does have limited subset translations for individual Slavic languages, the full contents of the English UMLS Metathesaurus with millions of concepts and synonyms is the gold standard of vocabulary-based medical knowledge extraction (also used by Apache cTAKES [5], a leading clinical text analysis system) and its utility transcends the boundaries of English.

As can be seen in Table 1, the sheer advantage English has in concept counts and ready-made synonym permutations is so great that in some cases, instead of piecing the small-language-specific system together from comparatively tiny translated resources, it might well be rational to machine-translate all text to English and annotate it with the UMLS Metathesaurus.

Although a license is required to access the UMLS, there are several identity providers to choose from and its usage is free. Downloads[1] include the `MRCONSO.RRF` file, which contains all concepts, and is the key resource for string-based knowledge extraction (other UMLS files are mostly concerned with concept relationships, attributes, and indexes). Apart from the CUI (Concept Unique Identifier) which secures interconnection among the entire network of meanings, every concept is marked by its language and originating vocabulary, so text-mining researchers can easily filter it for the subset they are looking for.

All of the following resources for English are included in the UMLS release mentioned above.

[1] https://nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

Table 1: UMLS Metathesaurus entry counts for relevant languages

Language	**Entry count**	**Relative size**
English	11,855,838	100%
Czech	212,304	1.8%
Polish	57,682	0.5%
Slovak	0	0%

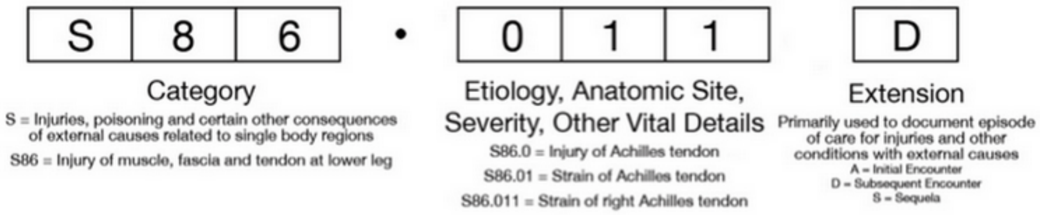

S
8
6
•
0
1
1
D
Category
S = Injuries, poisoning and certain other consequences of external causes related to single body regions
S86 = Injury of muscle, fascia and tendon at lower leg
Etiology, Anatomic Site, Severity, Other Vital Details
S86.0 = Injury of Achilles tendon
S86.01 = Strain of Achilles tendon
S86.011 = Strain of right Achilles tendon
Extension
Primarily used to document episode of care for injuries and other conditions with external causes
A = Initial Encounter
D = Subsequent Encounter
S = Sequela

Fig. 1: ICD-10 code structure [6]

2.1 ICD

The International Classification of Diseases [9], maintained by the World Health Organization, is one of the most widely known classification systems in medicine. This paper deals with ICD-10, its tenth revision, as it has been in use for all or most of the time when electronic health records were being produced in the relevant countries.

ICD uses diagnostic codes (Figure 1) to classify diseases, symptoms, or abnormal findings. Depending on the granularity of representation, the ICD-10 may contain as many as 72,184 codes for different conditions [1]. Czech, Polish and Slovak each have a translation of the ICD-10.

The primary difficulty in using ICD-10 for knowledge extraction is the length of diagnosis names - unlike a traditional dictionary, full ICD-10 names rarely occur in the text and targeted approaches are necessary to atomize the long strings and locate their constituents. This problem will be addressed further in subsections devoted to individual languages.

2.2 SNOMED CT

The Systematized Nomenclature of Medicine Clinical Terms, or SNOMED CT [3], maintained by SNOMED International, a not-for-profit organization, is "the most comprehensive, multilingual clinical healthcare terminology in the world" [7]. As can be seen in Figure 2, it contains a much broader range of concepts than ICD-10, including diagnostic procedures, body structures or substances.

The 2020 international edition of SNOMED CT included 352,567 concepts [7]. Unfortunately, SNOMED CT has not been translated into any of the languages surveyed in this paper.

2.3 MedDRA

The Medical Dictionary for Regulatory Activities (MedDRA), maintained by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), is an international medical terminology dictionary-thesaurus translated into several languages including Czech. Its

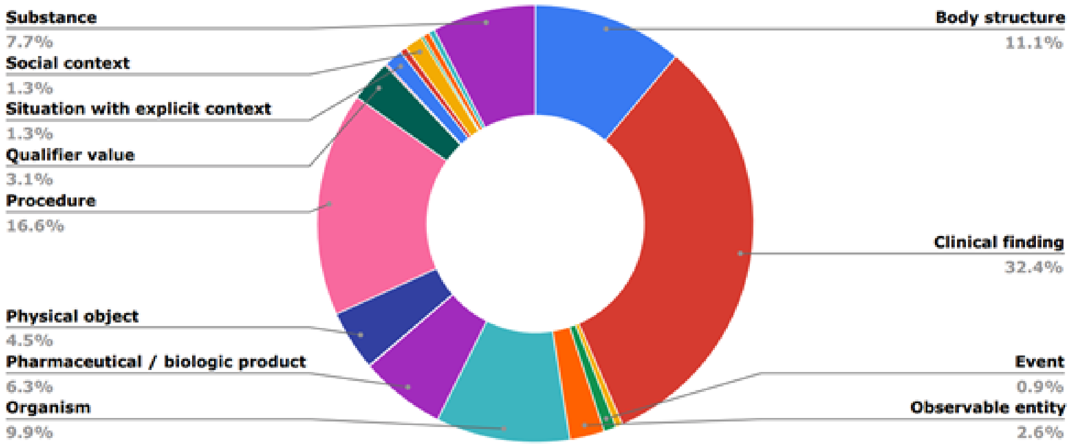


Fig. 2: Breakdown of concepts in SNOMED CT[7]

primary use is in regulatory communication in the biopharmaceutical industry, describing concepts in the clinical research of medicinal products, including adverse events. There are 115,479 English MedDRA concepts in the UMLS. They are organized as a five-level hierarchy:

- SOC (System Organ Class)
- HLGT (High-Level Group Term)
- HLT (High-Level Term)
- PT (Preferred Term)
- LLT (Lowest Level Term)

2.4 MeSH

Medical Subject Headings (MeSH) [4] is a “controlled and hierarchically-organized vocabulary” [8] produced by the United States National Library of Medicine. Its primary use is indexing of journal articles and books in the life sciences, so it can be expected to have less utility in discovering detailed medical information about patients than the previously mentioned vocabularies.

From the languages relevant for this paper, MeSH has been translated into Czech and Polish.

3 Czech

3.1 MedDRA

Unlike Polish or Slovak, Czech provides the option of a MedDRA translation included in the UMLS release listed above. It contains 76,255 unique strings in an interconnected and hierarchical collection of 111,573 entries, the average length of unique entries being 3.74 words. This is relatively high for string lookup, as the probability of unmodified occurrence decreases sharply above 3 words and the Czech MedDRA contains 31,083 unique entries longer than 3 words. Table 2 shows entry counts according to hierarchy level.

3.2 MeSH

Like MedDRA, the Czech MeSH is also included in the complete UMLS release. Czech MeSH is maintained by the Czech National Medical Library. It contains 100,731 entries, 100,618 of them unique, and the average length is 2.28 words, which is fortunate for exact string lookup.

3.3 Registered drug list

The Czech State Institute for Drug Control maintains several drug-related databases[2]. The most relevant resource is the current version of the list of registered drugs[3]. The ZIP file contains several CSV files, of which the most important one, currently named `dlp_lecivepripravky.csv`, contains structured information for each drug organized into 41 columns including the ATC code, which can serve as an international identifier.

There are 63,066 entries in the drug list including different strengths and packagings of the same drug, which boils down to just over 7,500 unique strings.

3.4 ICD-10 translation

The Institute of Health Information and Statistics of the Czech Republic publishes[4] the Czech translation of the ICD-10, referred to as MKN-10.

The files available for download are in a ZIP file referred to as “CSV strukturované podklady” and they are divided into UTF-8 and Windows 1250 versions. The main file (`01_MKN10_5E_2022_cp_w1250.csv` or `01_MKN10_5E_2022_utf8.csv` in the most recent version) contains the concepts ordered by code, starting with A00.

Table 2: Czech MedDRA structure

Hierarchy level	Hierarchy code in UMLS release	Concept count
1 - SOC	OS	27
2 - HLGT	HG	337
3 - HLT	HT	1,737
4 - PT	PT	25,077
5 - LLT	LLT, OL	84,139

[2] https://opendata.sukl.cz

[3] https://opendata.sukl.cz/?q=katalog/databaze-lecivych-pripravku-dlp

[4] https://uzis.cz/index.php?pg=registry-sber-dat--klasifikace--mezinarodni-klasifikace-nemoci-mkn-10#publikace

However, identifying ICD-10 concepts in text is a major challenge due to the long, detailed names of the conditions, extremely unlikely to be found verbatim. As can be seen in Table 3, the average diagnosis name length is 4.65 words, with 19,007 diagnoses (48.6%) of 5 words and more. However, the release of the Czech translation of ICD-10 contains separate files (CSV files with names containing "Abecední seznam") which contain an alphabetically ordered index of concepts split into a hierarchy. Table 4 demonstrates this, showing also that corresponding ICD codes are present in no particular order. With some rule-based processing that removes redundant or impractical text (such as references like "viz" and the contents of parentheses), lemmatizes words, and potentially splits subconcepts into subsubconcepts based on commas, the resulting strings end up much shorter and much more likely to be found in texts on their own. Table 3 shows the dramatic decrease of string length with the index file and with further automated optimization. Together with a cluster searching method that finds collocated subconcepts, this makes the ICD system much better suited for lookup in messy text where diagnoses hide in jumbled or incomplete forms.

Table 3: Czech ICD-10 translation file usability

File	Average item length (words)
Main file ordered by code	4.65
Hierarchical index	2.75
Optimized hierarchical index	1.57

Table 4: Czech ICD-10 alphabetical hierarchy example

Level 1	Level 2	Level 3	Level 4	ICD-10 code
Absorpce				
	bílkovin, porucha			K90.4
	dusíkatých látek – viz Uremie			
	glycidů, sacharidů, porucha			K90.4
	chemikálie			T65.9
		transplacentární (plodem nebo novorozencem)		P04.9
			látky z výživy	P04.5
			látky z životního prostředí	P04.6

4 Polish

4.1 MeSH

The major Polish subset of the UMLS Metathesaurus (available in the UMLS release mentioned above) is MeSH, published by the Polish Central Medical Library. It is smaller than the Czech one with 53,542 entries and an average length of 2.29 words, comparatively suitable for string search.

4.2 Databases published by the Ministry of Health

The Polish Ministry of Health publishes several relevant medical concept databases at [5], including the list of registered drugs and the Polish translation of ICD-10.

Registered drug list The list of registered drugs is available as a XLSX[6] or a CSV[7] file. There are 24 columns containing information about each entry, including extraction-relevant strings like product name, active ingredient, strength, packaging variants, and the internationally recognized ATC code.

There are 22,231 entries in the file, boiling down to just over 20,000 unique names.

The drug names included in this list only need minor automated editing (such as the optional isolation of name alone from manufacturer names, e. g. "GSK" in "Nitrazepam GSK", or quantities, e. g. "150 mg" in "Ranisan 150 mg") and the resulting vocabulary becomes a highly accurate tool for locating drug name mentions.

ICD-10 translation The Polish translation of ICD-10 can be downloaded[8] as an XML file with hierarchically organized nodes based on the granularity of representation (Figure 3).

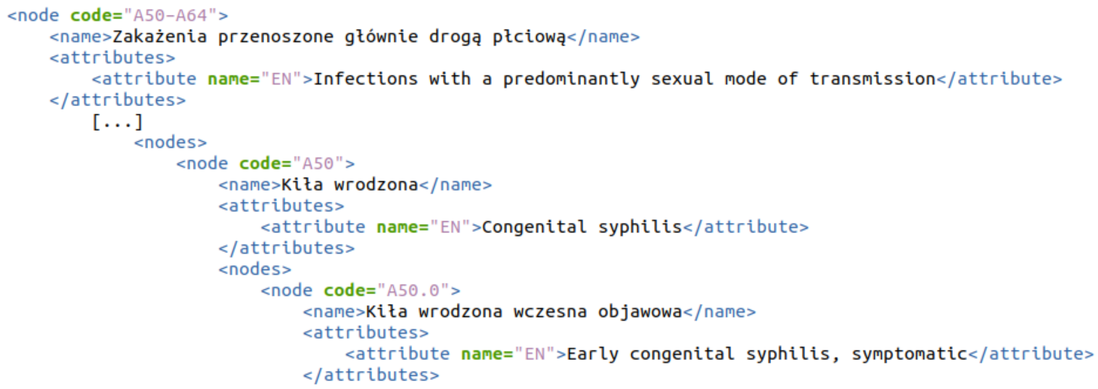
11,314 diagnosis names can be extracted from this file and the average length is 5.33 words, which indicates that most of the strings are not ready to be searched for in medical text. An alphabetically sorted, hierarchical file is not available for Polish, so other automated techniques (such as word separation and cluster search) have to be used to locate ICD-10 concepts in real-world use.

[5] `https://rejestrymedyczne.ezdrowie.gov.pl`

[6] `https://rejestrymedyczne.ezdrowie.gov.pl/api/rpl/medicinal-products/public-pl-report/get-xlsx`

[7] `https://rejestrymedyczne.ezdrowie.gov.pl/api/rpl/medicinal-products/public-pl-report/get-csv`

[8] section "Pliki do pobrania" at `https://rsk3.ezdrowie.gov.pl/resource/structure/icd10/00CD10/011/url`



```
<node code="A50-A64">
    <name>Zakażenia przenoszone głównie drogą płciową</name>
    <attributes>
        <attribute name="EN">Infections with a predominantly sexual mode of transmission</attribute>
    </attributes>
        [...]
            <nodes>
                <node code="A50">
                    <name>Kiła wrodzona</name>
                    <attributes>
                        <attribute name="EN">Congenital syphilis</attribute>
                    </attributes>
                    <nodes>
                        <node code="A50.0">
                            <name>Kiła wrodzona wczesna objawowa</name>
                            <attributes>
                                <attribute name="EN">Early congenital syphilis, symptomatic</attribute>
                            </attributes>
```

Fig. 3: Polish ICD-10 translation data example

5 Slovak

5.1 Registered drug list

The Slovak State Institute for Drug Control maintains several drug-related databases at [9], including an up-to-date list of drugs registered in Slovakia[10]. It can be downloaded as an easily processable XLS file with multiple columns where the structure is similar to that of the Czech registered drug list, including an ATC code which can be used to link the specific drug to all its associated concepts within the UMLS.

There are 51,314 entries in the file, out of which 9,090 are unique strings.

5.2 ICD-10 translation

The Slovak National Health Information Center publishes the Slovak translation of ICD-10, referred to as MKCH-10[11]. Like Polish, there is only one file where diagnoses are ordered by code and no granularized file is available.

The multi-sheet XLS file has 20,029 lines of diagnoses and diagnostic categories, with an average length of 7.68 words, making further automated subdivision necessary for successful string search to be feasible.

6 Conclusion

The available resources for Czech, Polish and Slovak indicate a clear direction for medical knowledge extraction, but they are far from making it straightforward, either due to machine translation issues (when using English resources) or as a result of differences between concepts in dictionaries and actual strings found in medical text. Some information, such as drug names, are easy to extract thanks

[9] www.sukl.sk/verejne/

[10] www.sukl.sk/verejne/Zoznam_liekov/zoznam_liekov.zip

[11] www.nczisk.sk/Standardy-v-zdravotnictve/Pages/Medzinarodna-klasifikacia-chorob-MKCH-10.aspx

to them being mostly one- or two-word strings occurring together. Others, such as ICD diagnoses, are very difficult to find and require dedicated methods of preprocessing of the longer names into subconcepts and special ways of searching for their collocation.

However, if medical informaticians focused on Slavic languages coordinate their efforts, develop pipelines that transform available resources into vocabularies usable for string search, and publish the newly created resources to facilitate an accumulative effect, the resulting systems have the potential to be a major leap in the area of Slavic medical text processing. Poland, the Czech republic, and Slovakia have a combined population of almost 55 million, and this would open up decades of this whole region's health documentation, public or private, to automated analysis, semantic filtering, and statistical research.

Acknowledgements The research in this paper was supported by the Internal Grant Agency of Masaryk University, project MUNI/IGA/1326/2021: *New Horizons of Electronic Health Record Analysis using Deep Learning*, and it was partially carried out within the project MUNI/G/1763/2020: *AIcope - AI support for Clinical Oncology and Patient Empowerment*.

References

1. American Academy of Professional Coders: All about ICD-10 (May 2021), `https://www.aapc.com/icd-10/`
2. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(suppl_1), D267–D270 (01 2004). https://doi.org/10.1093/nar/gkh061, `https://doi.org/10.1093/nar/gkh061`
3. Donnelly, K., et al.: SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics **121**, 279 (2006)
4. Lipscomb, C.E.: Medical subject headings (MeSH). Bulletin of the Medical Library Association **88**(3), 265 (2000)
5. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)
6. Smith, J.: Understanding the ICD-10 code structure (Mar 2022), `https://www.webpt.com/blog/understanding-icd-10-code-structure/`
7. SNOMED International: 5-step briefing, `https://www.snomed.org/snomed-ct/five-step-briefing`
8. U.S. National Library of Medicine: Medical subject headings - home page, `https://www.nlm.nih.gov/mesh/meshhome.html`
9. World Health Organization: ICD-10 : International statistical classification of diseases and related health problems : Tenth revision (2004)