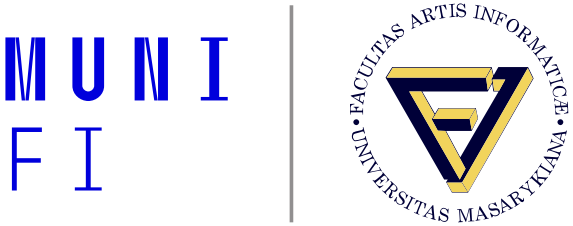


Medical Knowledge Resources for Text-Mining of Health Records in Czech, Polish, and Slovak

Krištof Anetta

xanetta@fi.muni.cz

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

December 10, 2022

What we want

Resekát levá mamma : resekát mléčné žlázy o rozměrech 80-79-45 mm . Ventrálně je volně pohyblivá fascie . Na řezu je mamma prostoupena hrubou fibrózou , v níž se nachází ostře ohraničené suspektní ložisko , obdobného vzhledu jako fibrosa , s patrnými prokrvácenými punkčními kanály , které má přibližné rozměry 19-20 mm . Mediální okraj je cca 5 mm , ventrální a kraniální 12 mm , kaudálně navazuje hrubá fibróza .

(fabricated text closely following the structure of a real health record)

Difficulties

- Related to the state of the art in low-resourced languages
  - No annotated biomedical corpora
  - No machine learning models for this purpose

Interim orientation

- Vocabulary text lookup needs to play a major role
- But the vocabulary concepts need to have a globally recognizable identifier so that UMLS concept connections can be leveraged

Difficulties with vocabulary lookup

- Related to health records
  - Widely varying wordings of concepts
  - Abbreviations
  - Typos
- Related to concepts
  - Specific concepts often longer than 3 words => almost never found literally in health records

Interim orientation

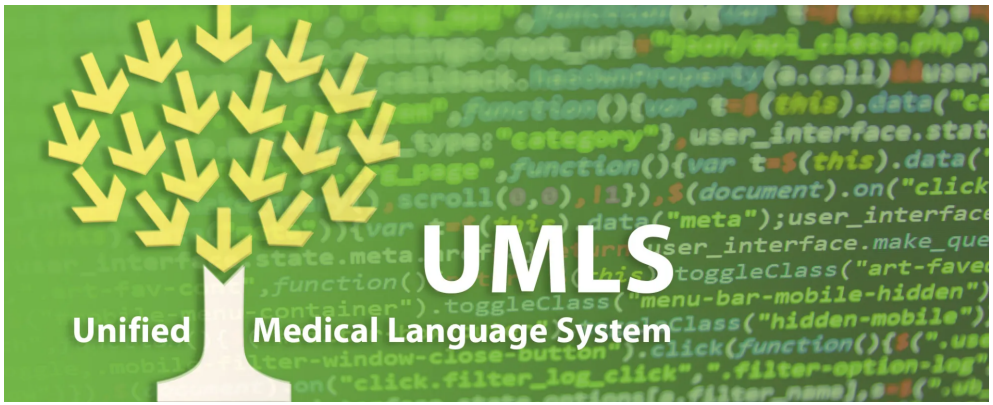
Many vocabularies need major pre-processing to be useful

Long vocabulary entry example

ICD-10: V30.34

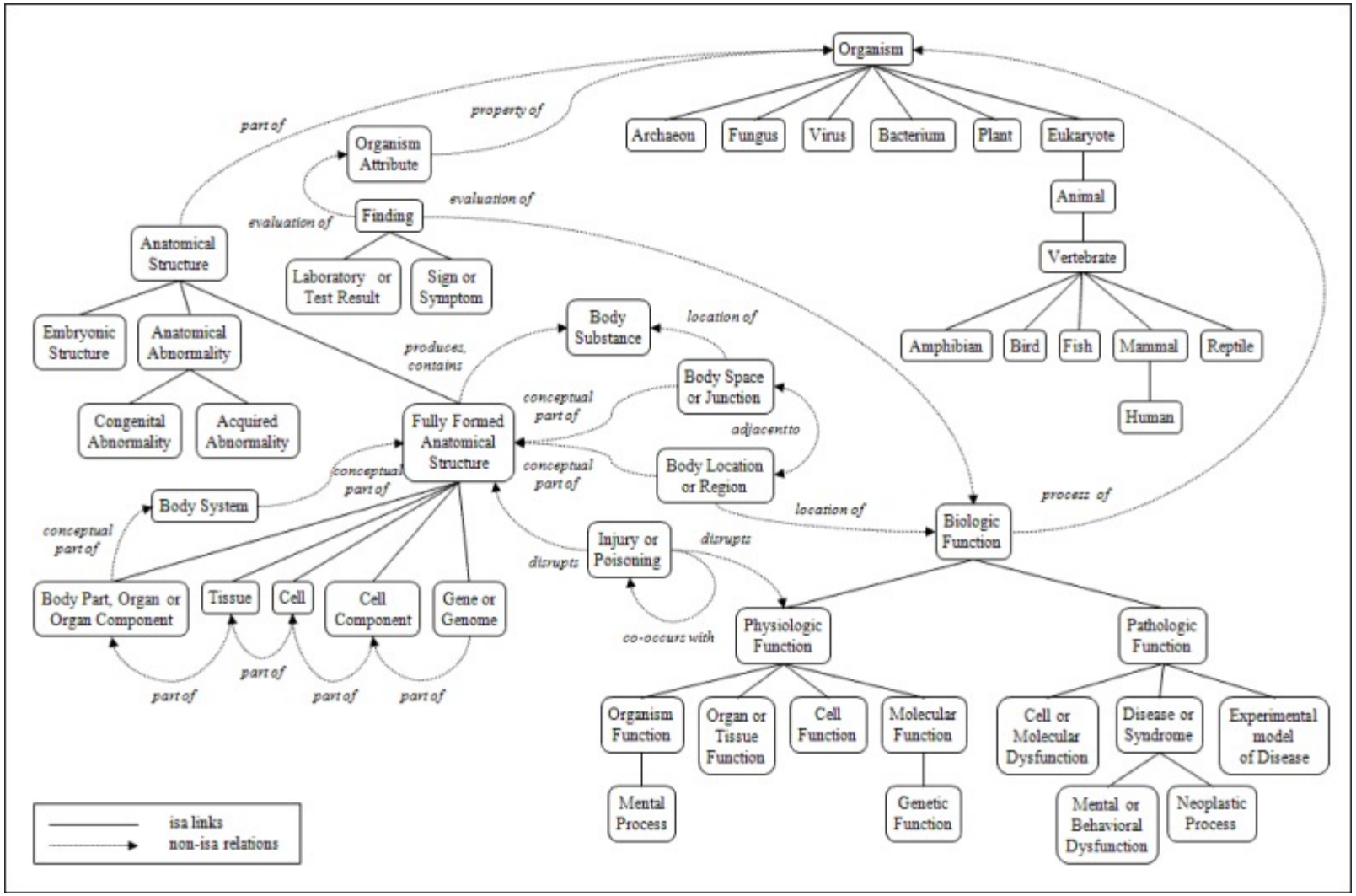
Člen osádky tříkolového motorového vozidla zraněný při srážce s chodcem nebo zvířetem; neurčený člen osádky tříkolového motorového vozidla zraněný při neprovozní (mimosilniční) nehodě; odpočinek, spánek, jídlo

UMLS: Mother of all ontologies



- 10+ million of concepts and synonyms for English and several other languages
- Interconnected through a system of unique identifiers
- Available free of charge after registration
- MRCONSO.RRF - a 2.1 GB text file with all concepts in all included languages

UMLS Semantic network



MRCONSO.RRF

```
C0086741|CZE|S|L8078046|PF|S10105604|Y|A26210537|||10005693|MDRCZE|LLT|10031133|Osmolalita|3|N||
C0086743|CZE|S|L8067284|PF|S10079911|Y|A26392675|||10031161|MDRCZE|LLT|10049594|Deformující osteoartritida|3|N||
C0086768|CZE|S|L15092004|PF|S18357353|Y|A29959395|||10047430|MDRCZE|LLT|10047318|Verner-Morrisonův syndrom|3|N||
C0086769|CZE|S|L8055556|PF|S10107125|N|A26240463|||10033664|MDRCZE|LLT|10033664|Panická ataka|3|N||
C0086769|CZE|S|L8055557|PF|S10107133|Y|A26330883|||10033664|MDRCZE|LLT|10033665|Panické ataky|3|N||
C0086795|CZE|S|L8024981|PF|S10101653|Y|A33548993|||10056886|MDRCZE|LLT|10020470|Nemoc Hurlerové|3|N||
C0086795|CZE|S|L8031938|PF|S10097132|Y|A26360753|||10056886|MDRCZE|LLT|10028048|MPS IH|3|N||
C0086795|CZE|S|L8064840|PF|S10120563|Y|A26299620|||10056886|MDRCZE|LLT|10020471|Syndrom Hurlerové|3|N||
C0086809|CZE|S|L13331630|PF|S16292358|Y|A29335149|||10012426|MDRCZE|LLT|10047907|Trichilemální cysta|3|N||
C0086809|CZE|S|L14153574|PF|S17205929|Y|A28360070|||10012426|MDRCZE|LLT|10035034|Cysta z vlasového folikulu|3|N||
C0086818|CZE|S|L8027570|PF|S10122221|Y|A15802694|||10035543|MDRCZE|PT|10035543|Transfuze krevních destiček|3|N||
C0086839|CZE|S|L8042164|PF|S10119465|N|A26392574|||10048738|MDRCZE|LLT|10048738|Stav po porodu|3|N||
C0086873|CZE|S|L16606122|PF|S20089377|Y|A32203573|||10001764|MDRCZE|LLT|10083950|Jizevnatá alopecie|3|N||
C0086879|CZE|S|L10428486|PF|S13003883|Y|A26305122|||10037325|MDRCZE|LLT|10072345|Kapilární plicní tlak v zaklínění|3|N||
C0086890|CZE|P|L8056545|PF|S10114795|N|A26423043|||10037056|MDRCZE|LLT|10050994|Quickův test|3|N||
C0086904|CZE|S|L8048190|PF|S10108157|N|A26389540|||10061858|MDRCZE|LLT|10016805|Perorální náhrada tekutin|3|N||
C0086904|CZE|S|L8063338|PF|S10108158|Y|A26240373|||10061858|MDRCZE|LLT|10031012|Perorální rehydratace|3|N||
C0086922|CZE|S|L8056741|PF|S10116469|Y|A26242438|||10019617|MDRCZE|LLT|10052589|Revmatická purpura|3|N||
C0086966|CZE|S|L8041909|PF|S10117489|N|A26214435|||10067499|MDRCZE|LLT|10067499|Selektivní potrat|3|N||
C0086981|CZE|S|L8056918|PF|S10117835|N|A26422152|||10040767|MDRCZE|LLT|10042844|Sicca syndrom|3|N||
C0087012|CZE|S|L8057074|PF|S10118996|Y|A32152848|||10062002|MDRCZE|LLT|10057660|Spinocerebelární ataxie|3|N||
C0087031|CZE|S|L8072189|PF|S10119680|N|A28360087|||10042061|MDRCZE|LLT|10042061|Stillova nemoc|3|N||
```

CE Slavic languages in UMLS: Numbers

Table: UMLS Metathesaurus entry counts for relevant languages

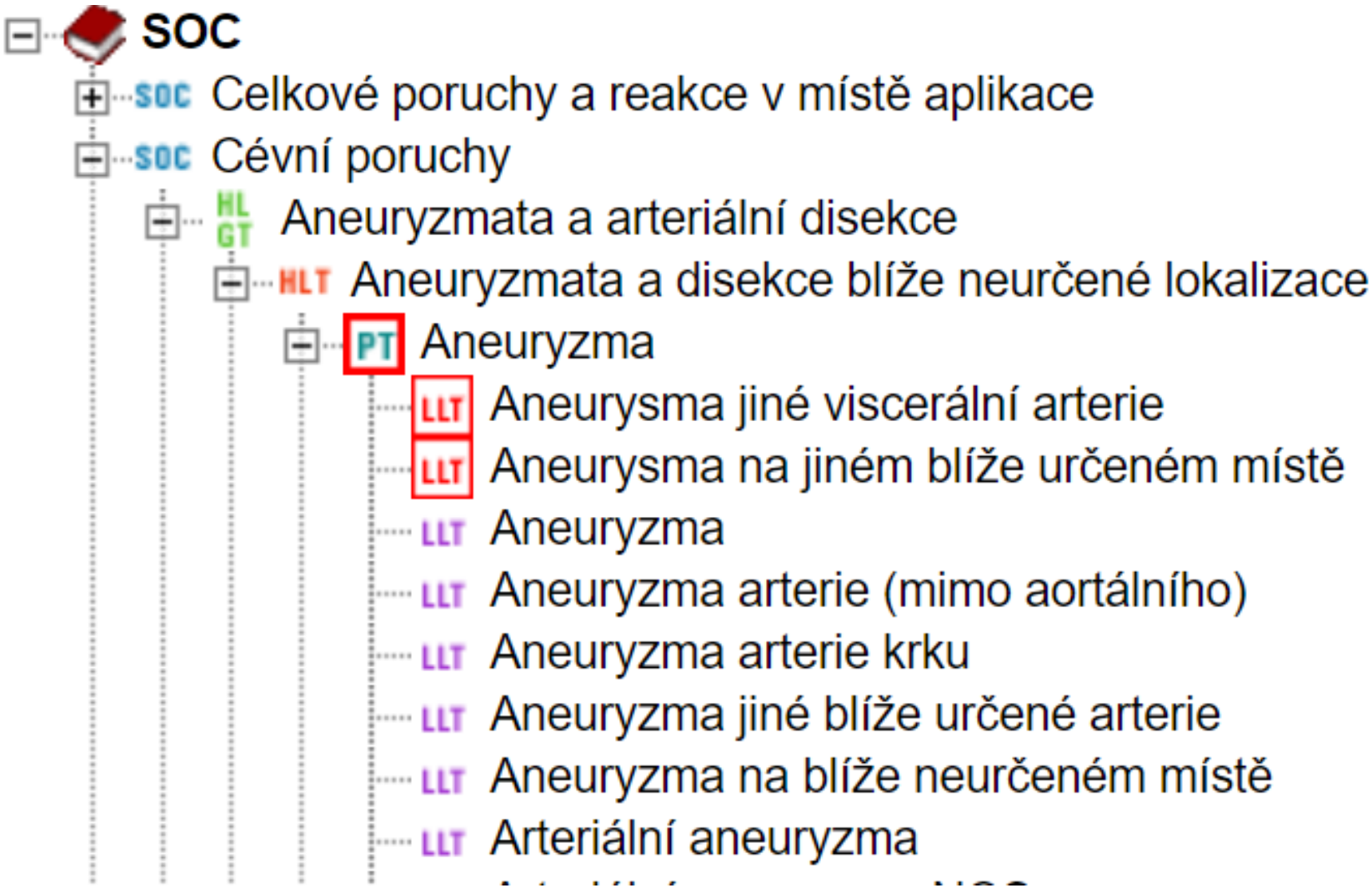
Language	Entry count	Relative size
English	11,855,838	100%
Spanish	1,839,491	15.5%
German	265,553	2.2%
Czech	**212,304**	**1.8%**
Polish	**57,682**	**0.5%**
Slovak	**0**	**0%**

CE Slavic languages in UMLS

- **Available for Czech (76,255 unique strings)**
- Medical terminology dictionary-thesaurus (concepts related to clinical research of medicinal products)
- Avg length: 3.74 words

- **Available for Czech (100,618) and Polish (53,537)**
- Controlled and hierarchical vocabulary for indexing journal articles and books in life sciences
- Avg length: 2.28 words
- Czech MeSH overlap with MedDRA: 5%

Czech MedDRA



Czech MeSH

[A] **Anatomie**

[A01] tělní krajiny

[A01.378] končetiny

[A01.378.610] dolní končetina

[A01.378.610.050] kotník

[A01.378.610.100] hýždě

[A01.378.610.250] noha (od hlezna dolů)...

[A01.378.610.400] kyčel

[A01.378.610.450] **koleno**

[A01.378.610.500] bérec

[A01.378.610.750] stehno

Other resources for Czech

- Most widely known medical categorization system
- 39,000+ entries
- Average length: 4.65 words

SÚKL
Státní ústav pro kontrolu léčiv

- List of registered drugs
- 63,066 entries, ca. 7,500 unique
- Easily discoverable in text, mostly 1 word, if more then in a fixed order

SÚKL databases

8949		EZETROL	10MG	TBL NOB	100	
9709		SOLU-MEDROL	40MG/ML	INJ PSO LC	40MG+1M	
9710		SOLU-MEDROL	62,5MG/M	INJ PSO LC	125MG+2M	
9711		SOLU-MEDROL	62,5MG/M	INJ PSO LC	500MG+7,	
9712		SOLU-MEDROL	62,5MG/M	INJ PSO LC	1000MG+1	
9844		TORECAN	6,5MG	TBL OBD	50	
10032		PIRACETAM AL	800MG	TBL FLM	60	
10033		PIRACETAM AL	800MG	TBL FLM	120	
10045		AGNUCASTON		TBL FLM	30	
10046		AGNUCASTON		TBL FLM	60	
10047		AGNUCASTON		TBL FLM	100	
10052		AGNUCASTON		TBL FLM	300	
10055		TABACUM	31CH-200	GRA	4G	
10063		BROMHEXIN KM	8MG/ML	POR GTT S	1X30ML	
10073		ECHINACEA ANGUSTIFOLIA	31CH-200	GRA	1X4G	
10087		LOBELIA INFLATA	31CH-200	GRA	1X4G	
10111		DHC CONTINUS	120MG	TBL MRL	56	

Czech ICD-10 (MKN-10)

A984	A98.4		Horečka ebola
A985	A98.5		Hemoragická horečka s renálním syndromem
A988	A98.8		Jiná určená virová hemoragická horečka
A99	A99		Neurčená virová hemoragická horečka
B00	B00		Infekce virem Herpes simplex
B000	B00.0		Herpetický ekzém
B001	B00.1		Vezikulární dermatitida, původce: virus Herpes simplex
B002	B00.2		Herpet.gingivostomat.a faryngotonzilit., původce: virus Herpes simplex
B003	B00.3		Herpetická meningitida(G02.0*), původce: virus Herpes simplex
B004	B00.4		Herpetická encefalitida(G05.1*), původce: virus Herpes simplex
B005	B00.5		Oční onemocnění, původce: virus Herpes simplex
B007	B00.7		Diseminované herpetické onemocnění, původce: virus Herpes simplex
B008	B00.8		Jiné formy herpetické infekce, původce: virus Herpes simplex
B009	B00.9		Infekční onemocnění, původce: virus Herpes simplex NS
B01	B01		Plané neštovice [varicella]
B010	B01.0		Varicelová meningitida (G02.0*)
B011	B01.1		Varicelová encefalitida (G05.1*)

Czech ICD-10 (MKN-10): Index

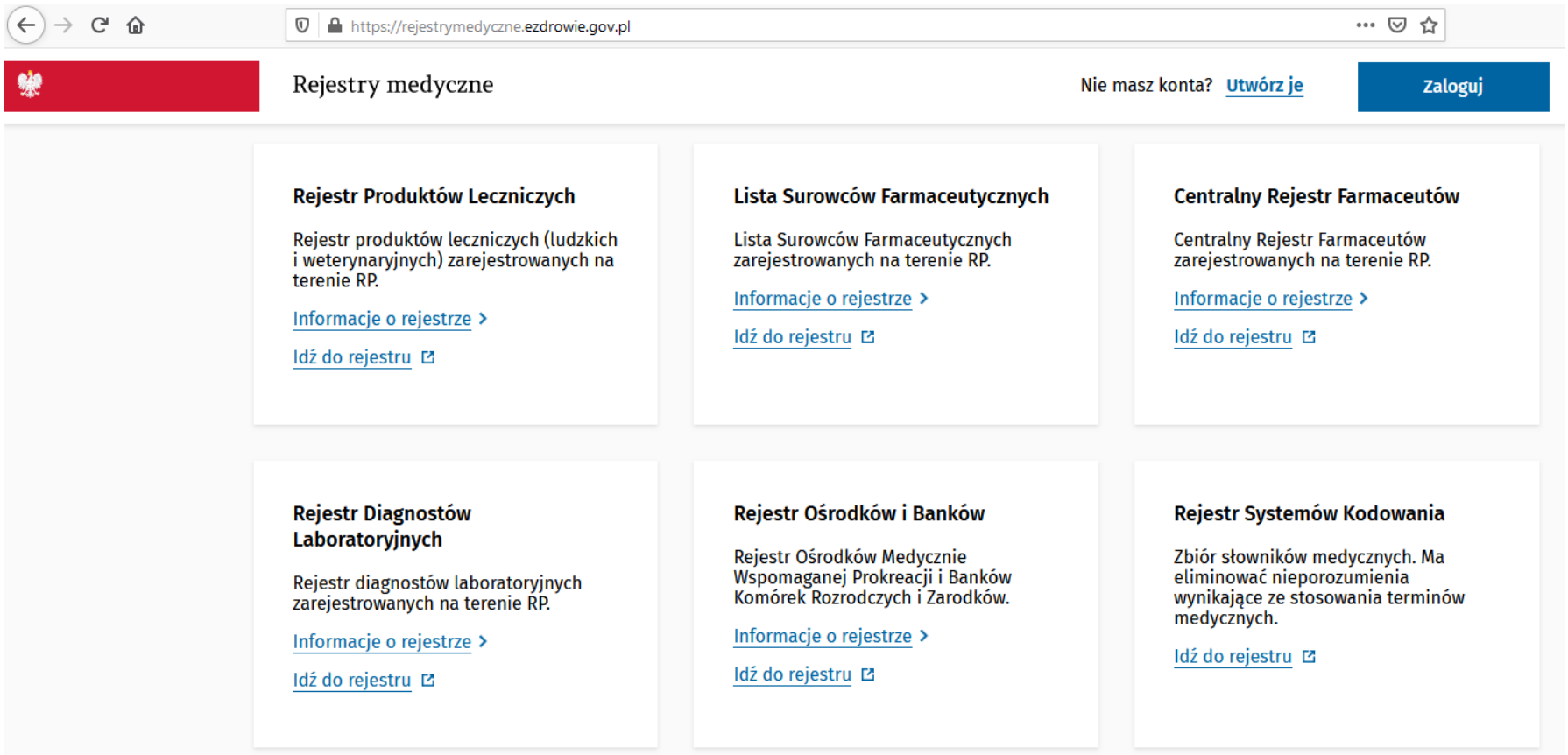
Level 1	Level 2	Level 3	Level 4	ICD-10
Absorpce				
	bílkovin, **porucha**			K90.4
	dusíkatých látek			
	glycidů, sacharidů, porucha			K90.4
	chemikálie			T65.9
		transplacentární		P04.9
			látky z výživy	P04.5

Polish in UMLS

- MeSH translation
- 53,537 unique entries, average length 2.29 words

```
)16606|MSHPOL|MH|D016606|Guzki tarczycy|3|N||
)18828|MSHPOL|MH|D018828|Immunoglobuliny stymulowane przez tarczycę|3|N||
)18828|MSHPOL|SY|D018828|Przeciwciała stymulowane przez tarczycę|3|N||
)13965|MSHPOL|MH|D013965|Tyroidektomia|3|N||
)13965|MSHPOL|SY|D013965|Wycięcie tarczycy|3|N||
)13966|MSHPOL|MH|D013966|Zapalenie tarczycy|3|N||
)13968|MSHPOL|MH|D013968|Zapalenie tarczycy podostre|3|N||
)13969|MSHPOL|MH|D013969|Zapalenie tarczycy ropne|3|N||
)13970|MSHPOL|MH|D013970|Tyroniny|3|N||
)13971|MSHPOL|MH|D013971|Tyreotoksykoza|3|N||
)13972|MSHPOL|MH|D013972|Tyreotropina|3|N||
)13972|MSHPOL|SY|D013972|Hormon tyreotropowy|3|N||
)13972|MSHPOL|SY|D013972|TSH|3|N||
)13973|MSHPOL|MH|D013973|Tyreoliberyna|3|N||
)13973|MSHPOL|SY|D013973|Hormon uwalniający tyreotropinę|3|N||
)13974|MSHPOL|MH|D013974|Tyroksyna|3|N||
)13974|MSHPOL|SY|D013974|Sól sodowa lewotyroksyny|3|N||
)13974|MSHPOL|SY|D013974|Lewotyroksyna|3|N||
```

Polish: centralized e-Health resource



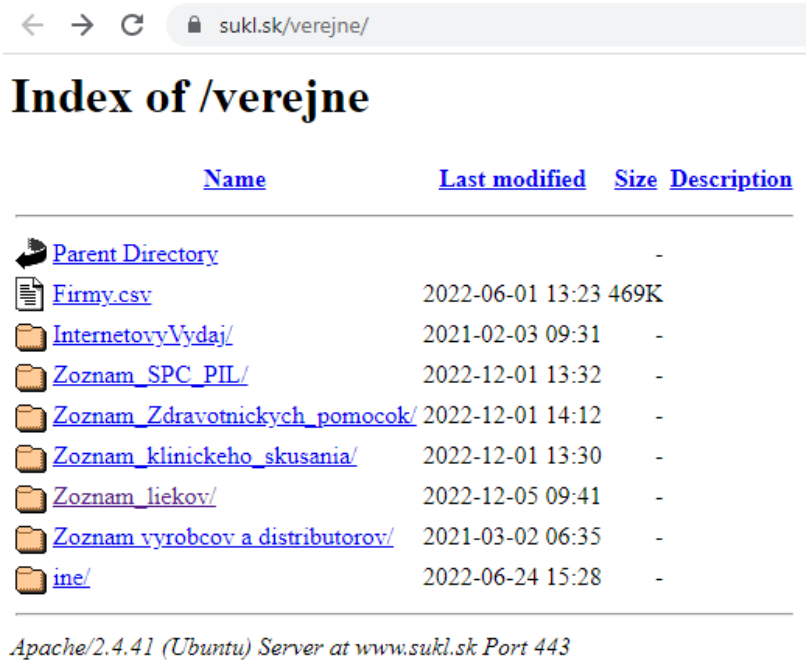
Polish drug list

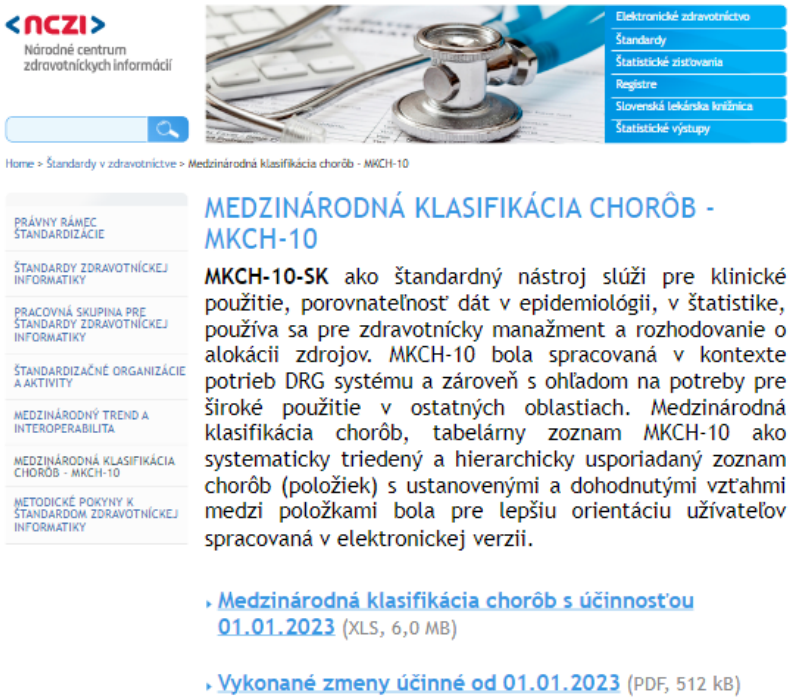
```
  <produktLeczniczy nazwaProduktu="Edelan" rodzajPreparatu="ludzki" nazwaPowszechnieStosowana="Mometasoni furoas" moc="1 mg/g" postac="krem" podmiotOdpowiedzialny="Zakłady
Farmaceutyczne POLPHARMA S.A." typProcedury="NAR" numerPozwolenia="20899" waznoscPozwolenia="Bezterminowy" kodATC="D07AC13" id="100000020">
    <substancjeCzynne>
      <substancjaCzynna>Mometasoni furoas</substancjaCzynna>
    </substancjeCzynne>
    <opakowania>
      <opakowanie wielkosc="1" jednostkaWielkosci="tuba 15 g" kodEAN="05909991023683" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="5" />
      <opakowanie wielkosc="1" jednostkaWielkosci="tuba 30 g" kodEAN="05909991023690" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="6" />
    </opakowania>
  </produktLeczniczy>
  <produktLeczniczy nazwaProduktu="Nalgesin" rodzajPreparatu="ludzki" nazwaPowszechnieStosowana="Naproxenum natricum" moc="275 mg" postac="Tabletki powlekane"
podmiotOdpowiedzialny="Krka, d.d., Novo mesto" typProcedury="DCP" numerPozwolenia="20696" waznoscPozwolenia="Bezterminowy" kodATC="M01AE02" id="100000037">
    <substancjeCzynne>
      <substancjaCzynna>Naproxenum natricum</substancjaCzynna>
    </substancjeCzynne>
    <opakowania>
      <opakowanie wielkosc="10" jednostkaWielkosci="tabl." kodEAN="05909991023744" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="8" />
      <opakowanie wielkosc="20" jednostkaWielkosci="tabl." kodEAN="05909991023751" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="9" />
      <opakowanie wielkosc="30" jednostkaWielkosci="tabl." kodEAN="05909991023768" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="10" />
      <opakowanie wielkosc="40" jednostkaWielkosci="tabl." kodEAN="05909991066178" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="84656" />
      <opakowanie wielkosc="60" jednostkaWielkosci="tabl." kodEAN="05909991023775" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="11" />
    </opakowania>
  </produktLeczniczy>
  <produktLeczniczy nazwaProduktu="Nalgesin Forte" rodzajPreparatu="ludzki" nazwaPowszechnieStosowana="Naproxenum natricum" moc="550 mg" postac="Tabletki powlekane"
podmiotOdpowiedzialny="Krka, d.d., Novo mesto" typProcedury="DCP" numerPozwolenia="20697" waznoscPozwolenia="Bezterminowy" kodATC="M01AE02" id="100000043">
    <substancjeCzynne>
      <substancjaCzynna>Naproxenum natricum</substancjaCzynna>
    </substancjeCzynne>
    <opakowania>
      <opakowanie wielkosc="10" jednostkaWielkosci="tabl." kodEAN="05909991023782" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="12" />
      <opakowanie wielkosc="20" jednostkaWielkosci="tabl." kodEAN="05909991023799" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="13" />
      <opakowanie wielkosc="30" jednostkaWielkosci="tabl." kodEAN="05909991023805" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="14" />
      <opakowanie wielkosc="40" jednostkaWielkosci="tabl." kodEAN="05909991066185" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="84657" />
      <opakowanie wielkosc="50" jednostkaWielkosci="tabl." kodEAN="05909991023829" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="15" />
      <opakowanie wielkosc="60" jednostkaWielkosci="tabl." kodEAN="05909991023836" kategoriaDostepnosci="Rp" skasowane="NIE" numerEu="" dystrybutorRownolegly="" id="16" />
    </opakowania>
  </produktLeczniczy>
```

Polish ICD-10

```
<node code="A80-A89">
    <name>Wirusowe zakażenia ośrodkowego układu nerwowego</name>
    <attributes>
        <attribute name="EN">Viral infections of the central nervous system</attribute>
    </attributes>
    <nodes>
        <node code="">
            <name></name>
            <nodes>
                <node code="">
                    <name></name>
                    <nodes>
                        <node code="A80">
                            <name>Ostre nagminne porażenie dziecięce</name>
                            <attributes>
                                <attribute name="EN">Acute poliomyelitis</attribute>
                            </attributes>
                            <nodes>
                                <node code="A80.0">
                                    <name>Ostre nagminne porażenie poszczepienne</name>
                                    <attributes>
                                        <attribute name="EN">Acute paralytic poliomyelitis, vaccine-associated</attribute>
                                    </attributes>
                                </node>
                                <node code="A80.1">
                                    <name>Ostre nagminne porażenie dziecięce, wirus dziki, importowany</name>
                                    <attributes>
                                        <attribute name="EN">Acute paralytic poliomyelitis, wild virus, imported</attribute>
                                    </attributes>
                                </node>
                                <node code="A80.2">
                                    <name>Ostre nagminne porażenie dziecięce, wirus dziki, tubylczy</name>
                                    <attributes>
                                        <attribute name="EN">Acute paralytic poliomyelitis, wild virus, indigenous</attribute>
                                    </attributes>
```

Slovak





- Drug list
  - Entries: 51,314
  - Unique: 9,090
- List of medical aids and devices

- ICD-10 translation
  - Entries: 20,029
  - Average length: 7.68 words

Slovak drug list

7931C	Caltrate D3 500 mg/1000 IU žuvacie	tbl mnd 60x500 mg/1000 IU (strip	500 mg/1000	60
7932C	Caltrate D3 500 mg/1000 IU žuvacie	tbl mnd 90x500 mg/1000 IU (strip	500 mg/1000	90
7933C	Caltrate D3 500 mg/1000 IU žuvacie	tbl mnd 120x500 mg/1000 IU (strip	500 mg/1000	120
14783	CALTRATE PLUS	tbl flm 15 (fľ.HDPE)		15
34205	CALTRATE PLUS	tbl flm 30 (fľ.HDPE)		30
34228	CALTRATE PLUS	tbl flm 60 (fľ.HDPE)		60
34230	CALTRATE PLUS	tbl flm 2x30 (fľ.HDPE)		60
0372B	CALTRATE PLUS	tbl flm 90 (fľ.HDPE)		90
87814	Calypsol	sol inj 5x10 ml/500 mg (liek.skl.hnedá)	50 mg/1 ml	5
1148E	CAMCEVI 42 mg injekčná suspenzia s	sus ijp 1x42 mg (striek.inj.napl. COC)	42 mg	1
4341A	CAMILIA	sol por 10x1 ml (obal LDPE jednodáv.)	-	10
4342A	CAMILIA	sol por 20x1 ml (obal LDPE jednodáv.)	-	20
4343A	CAMILIA	sol por 30x1 ml (obal LDPE jednodáv.)	-	30
56150	CAMPRAL	tbl ent 84x300 mg (blis.Al/PVC/PVDC)	300 mg	84
40776	CANCIDAS 50 mg prášok na infúzny	plc ifc 1x50 mg (liek.inj.skl.)	50 mg	1
40775	CANCIDAS 70 mg prášok na infúzny	plc ifc 1x70 mg (liek.inj.skl.)	70 mg	1
04784	Candesartan HCT ratiopharm 16	tbl 7x16 mg/12,5 mg (blis.PVC/PVDC/Al)	16 mg, 12,5 mg	7
04785	Candesartan HCT ratiopharm 16	tbl 14x16 mg/12,5 mg (blis.PVC/PVDC/Al)	16 mg, 12,5 mg	14

Slovak ICD-10

D50-D53	Nutričné anémie	
D50.-	Anémia z nedostatku železa	Patrí sem: Anémia: • hypochrómna • sideropenická
D50.0	Anémia z nedostatku železa pri chronických stratách krvi	Posthemoragická anémia (chronická) Nepatrí sem: Anémia zapríčinená akútnym krvácaním (D62) Vrodená anémia zapríčinená fetálnou stratou krvi (P61.3)
D50.1	Sideropenická dysfágia	Kellyho-Patersonov syndróm Plummerov-Vinsonov syndróm
D50.8	Iná anémia z nedostatku železa	
D50.9	Bližšie neurčená anémia z nedostatku železa	
D51.-	Anémia z nedostatku vitamínu B12	Nepatrí sem: Nedostatok vitamínu B12 (E53.8)
D51.0	Anémia z nedostatku vitamínu B12 zapríčinená nedostatkom vnútorného faktora	Anémia: • Addisonova • Biermerova • perniciózna (vrodená) Vrodený nedostatok vnútorného faktora

Conclusion

- Overall, the conceptual range of
  - the Czech resources is good thanks to MedDRA
  - the Polish resources is satisfactory
  - the Slovak resources is only usable for a narrow subset of knowledge extraction scenarios
- Due to string length, intelligent ways of looking for word clusters need to be developed
  - A probability measure based on
    - How many of the full-meaning words from the original concept are near each other? (e.g. 4 out of 6)
    - How far apart are they? (e.g. a maximum of 2 words in between)
    - Are they in the same sentence? Same grammatical function as in the original?

Thank you for your attention!

MUNI
FACULTY
OF INFORMATICS