Kateřina Pořízková
Marek Blahuš
(Centrum jazykového vzdělávání MU)

KORPUS AUTENTICKÝCH KLINICKÝCH DIAGNÓZ V PROSTŘEDÍ SOFTWARU SKETCH ENGINE

Probably every teacher of Latin medical terminology is struggling with a lack of source documents that reflect the real use of Latin in clinical diagnoses, since the possibility of making available the authentic medical documentation is very limited. In cooperation with teaching hospitals in Prague and Brno a corpus of authentic clinical diagnoses has been created. These diagnoses have been excerpted from electronic medical records of the clinics where the frequent use of Latin diagnoses was assumed (i.e. surgery, traumatology as well as gynecology and obstetrics, ORL etc.). In this presentation the authors demonstrate how to use the tools of corpus linguistics (morphological analyzer Morpheus and software Sketch Engine) for innovations in Latin medical terminology teaching and testing. These tools are used to search especially for typical collocations and frequency of terms and their components (e.g. prepositional phrases).

Key words: *Latin Authentic Clinical Diagnoses; Corpus lingustics; Sketch Engine*
Klíčová slova*: latinské autentické klinické diagnózy; korpusová lingvistika; Sketch Engine*

Korpusová lingvistika je relativně mladá jazykovědná disciplína, která nabízí možnosti studia jazyka prostřednictvím autentických, počítačově zpracovávaných textů (písemných i mluvených). U nás je její jméno především spojeno se zpracováváním a studiem Českého národního korpusu (od r. 1994 Ústavem ČNK na FF UK) a s vývojem korpusových nástrojů (Centrum zpracování přirozeného jazyka na FI MU), nicméně i v celosvětovém měřítku jde o velmi progresivní fenomén – vzniká řada různojazyčných korpusů obsahujících texty s různou typologií (synchronní, diachronní, textové žánry, paralelní korpusy apod.).

Pokud jde o latinský jazyk a jeho uplatnění pro zápis diagnóz v současné klinické medicíně, korpus, který by texty tohoto charakteru obsahoval, ještě do nedávné doby neexistoval. Hlavním důvodem je zřejmě to, že klinické diagnózy se nachází v lékařských dokumentech, které nemohou být s ohledem na ochranu osobních dat pacientů i lékařů veřejně přístupné. Na pracovišti Centra jazykového vzdělávání Masarykovy univerzity v Brně se zrodila myšlenka vytvořit ve spolupráci s fakultními nemocnicemi databázi plně anonymizovaných autentických klinických diagnóz. Přes poměrně složitou administrativní proceduru se toto podařilo – jako první poskytla anonymizovaná

data Fakultní nemocnice Brno-Bohunice (2011), následovaly Fakultní Thomayerova nemocnice v Praze (2012) a Fakultní nemocnice u sv. Anny v Brně (2013). Tyto autentické diagnózy jsou v současné době zpracovávány na pracovišti CJV na Lékařské fakultě MU do podoby korpusu.

Ve všech ukázkách z této databáze, které následují, jsou diagnózy zobrazeny v původní podobě, bez jakéhokoli zásahu ve smyslu jazykové korekce.

Krok 1: Databáze autentických klinických diagnóz

Databáze obsahují čistý elektronický výpis autentických klinických diagnóz, které lékaři zaznamenali do nemocničních informačních systémů v průběhu přesně vymezeného časového období (12 měsíců). Jednalo se o záznamy z vybraných klinik, hlavní pozornost byla věnována především chirurgickým oborům (chirurgie, traumatologie, gynekologie-porodnictví, ortopedie apod.).

Tab. 1 Ukázka z databáze autentických klinických diagnóz Fakultní Thomayerovy nemocnice v Praze (Chirurgická klinika)

S7210	Fractura femoris l. dx. pertrochanterica
S4230	Fractura diaphyseos humeri l.dx. spiralis
K509	M. Crohn
S7200	Fractura intracapsularis coli femoris sin. implantatione TEP curata 26.1.11 (FNM)
S700	Contusio coxae l. sin., stp CCEP ante annos.
K359	Irritatio appendicis
K261	Ulcus bulbi duodeni perfor.
J989	Tu lobi sup. pulm. l. sin. - susp. Ca bronchogen.
S0600	Commotio cerebri
S8270	Fractura ATC l. dx. Weber B cum abruptio malleoli med. et margo dorsalis tibiae
K565	Strangulatio int. ilei. Adhaesiones cavi abdominis.
L029	Phlegmona pedis l.dx. st.p. amputationem hallucis.
K573	Diverticulosis colli sinistri
K801	Cholecystitis chronica calculosa.
S7200	Fractura intracapsularis femoris l.sin.
S7210	Fractura pertrochanterica femoris sin. dislocata comminutiva.

Tab. 2 Ukázka z databáze autentických klinických diagnóz Fakultní nemocnice Brno-Bohunice (Gynekologicko-porodnická klinika)

O801	Praesentatio pelvina incompleta natium
O244	Diabetes mellitus gestationis insulinodep.
	Hypotrophia fetus 37/33
	Varicaes uteri
O251	Myoma uteri parvum pendulum

	Primipara vetus
	Obesitas permagna
	Partus medicamentosus non progrediens
	Funiculus umbilicus circum coli fetus semel
	St.p. partuum spontaneus (03.01.2011, 22:10 hod)
	Mater fumator
N994	Adhaesiones pelvs minoris gr II
	Peridnexitis bilat
	Torsio adnextumoris l.dx.
	Pelvalgia chronica
	Metrorrhagia perimenopausalis

Primárním podnětem pro vznik těchto databází byl především nedostatek zdrojových dokumentů, ze kterých by učitelé lékařské terminologie mohli čerpat pro potřeby výuky a testování v rámci klinické a patologické terminologie. Dosud byl latinář odkázán především na existující skripta základů lékařské terminologie, a to s důvěrou, že všechny termíny užité v této literatuře, která nepochybně vycházela i z autentických lékařských zpráv, v praxi klinického lékaře reálně existují a jsou běžně užívané, nebo představují skutečně základní pojem pro daný obor.
Hlavním cílem zpracovávání databází autentických klinických diagnóz je především poskytnout učitelům lékařské terminologie komplexní vhled do reálné podoby využití latiny pro zápis klinických diagnóz. Je třeba upozornit, že současný stav databází je vzhledem k jejich relativně malému rozsahu (z pohledu korpusové lingvistiky) nedostačující pro vyvozování spolehlivých závěrů, nicméně již v této podobě poskytují velmi mnoho zajímavých informací, které jsou využitelné pro výuku lékařské latiny a silně motivační pro studium tohoto předmětu.

Vyhledávání kolokací

Jedním z hlavních důvodů tvorby databází autentických klinických diagnóz a jejich následného přepracování do podoby korpusu je i možnost využití nástrojů korpusové lingvistiky pro vyhledávání typických kolokací, resp. kolokačních termínů. Z formálního, resp. syntaktického hlediska jde především o revizi základních spojení substantiv se shodným přívlastkem (typicky S + A), resp. substantiv ve spojení s neshodným přívlastkem (typicky S +Sgen, resp. předložkové vazby, viz níže). Neméně důležité je i hledisko věcné, resp. sémantické. Jedná se o termíny, které již ze své podstaty představují přesná, jednoznačná pojmenování pro daný klinický jev, a je nezbytně nutné, aby učitel lékařské terminologie měl tyto podklady autentického charakteru k dispozici.

Tab. 3 Ukázka kolokací termínu *ruptura* v databázi autentických klinických diagnóz

S3600	Ruptura lienis inveterata,haemoperitoneum
S3620	Ruptura lienis v.s.post traumatica
S761	Ruptura lig. patelae gen. l. dx. totalis

S761	Ruptura m. quadricipitis femoris l.dx. inveterata.
S863	Ruptura m. tricipitis surae l. dx. susp.
S763	Ruptura partialis musculi bicipitis femoris l.dx.
N139	Ruptura pelvicis renalis l. dx. vs., pyelonephritis obstructiva l. dx.
	Ruptura pseudoaneurysmatis a. hepaticae communis
K863	Ruptura pseudocystae pancreatis
S3700	Ruptura renis l.dx. hemathoma retroperitonei
K359	St.p. APPE, revisionem ca abdominis propter rupt. hepatis
S8260	Fractura ATC l. sin. Weber B cum rupt. lig. deltoidei, syndesmosis et abruptio marg. dorsalis tibiae

Výše uvedená tabulka zobrazuje přehled základních kolokací latinského substantiva *ruptura*. Modře vyznačené jsou lokalizace ruptur, které se typicky týkají měkkých tkání jako např. svalů, vazů, šlach, měkkých orgánů; zeleně jsou vyznačená adjektiva s významem charakteristiky ruptur, a to z hlediska času, doby trvání (*inveterata*), příčiny, resp. okolností vzniku (*traumatica*), rozsahu (*totalis*) a modalitního postoje lékaře k pravděpodobnosti (*suspecta*). S pomocí těchto nástrojů nejde o zpřesnění nebo redefinici významu termínu *ruptura*, ale o jeho typická spojení věcně reflektující to, u kterých anatomických struktur k tomuto klinickému jevu obvykle dochází a ze kterých hledisek se daný jev dále popisuje.

Pro ilustraci uveďme jako další příklad termín *fractura comminutiva* a jeho typické kolokace, především ve vztahu k lokalizaci – jinými slovy, u kterých kostí a ve kterých jejich částech (vyznačeno červeně) typicky dochází k tříštivým zlomeninám:

Tab. 4 Ukázka kolokací termínu *fractura comminutiva* v databázi autentických klinických diagnóz

S5250	Fractura cominutiva partis distalis radii l. sin.
S8230	Fractura cominutiva tibiae dist. l. dx. intraarticul.
	Fractura comminutiva calcanei l. sin.
S4220	Fractura comminutiva colli chirurgicum humeri dx.
S3200	Fractura comminutiva corporsi L2
S4230	Fractura comminutiva diaphyseos humeri l. dx. cum disloc
S6260	Fractura comminutiva dig. III. manus l.dx. cum disloc.
	Fractura comminutiva humeri prox. l. sin
S5250	Fractura comminutiva intraartricularis partis distalis radii sin. dislocata.
S9230	Fractura comminutiva MTT I. et II. pedis l.sin.
S9230	Fractura comminutiva MTT V. pedis l.dx.
S5200	Fractura comminutiva olecrani lat. dx.
S9200	Fractura comminutiva ossis calcanei l. dx.
S3230	Fractura comminutiva ossis ilii l. dx. cum disloc.

Pokud jde o terminologii chirurgických zákroků, i v této oblasti nabízí korpusová lingvistika celou řadu možností, jak využít její nástroje pro vyhledávání typických kolokací. Na následujících konkrétních příkladech je demonstrován rozdíl v souvýskytech u termínů *resectio* a *extractio*.

Tab. 5 Příklady kolokací termínu *resectio* a *extractio* v databázi autentických klinických diagnóz

Z908	Resectio glandulae thyroideae partiale temporo remoto.
K567	Resectio ilei dist. et anast. in toto propter st. ileosus, stenosis ileotreansversoanastomosis
Z904	resectio oesophagei dist. propter empyema thoracis persist.
C19	Resectio sigmoidei sec Hartmann cum ansae int. ilei in toto propter ca

Z470	Extractio cerclage in cursu.
Z480	Extractio corp. metalli, st.p. OS ATC l. dx. (09/2011)
Z470	Extractio osteosynthesis externi in cursu.

Frekvence termínů (zlomeniny)

Dalším, neméně podstatným východiskem pro práci učitele lékařské terminologie je i získání přehledu o frekvenci jednotlivých termínů a jejich kolokacích v praxi klinického lékaře. Práce s korpusem tuto možnost přímo nabízí – viz např. níže uvedenou tabulku s nejčastěji diagnostikovanými zlomeninami. Mimo jiné na přehledu také zjišťujeme, že se při popisu zlomenin často identifikuje přesná anatomická struktura, resp. oblast na dané kosti, jíž se zlomenina týká:

Tab. 6 Přehled nejčastěji diagnostikovaných zlomenin v databázi autentických klinických diagnóz

S8280	Fractura art. talocruralis dx. WB, ZH.
S9230	Fractura baseos metatarsi I. pedis l.dx. comminutiva
S2230	Fractura costae IV l. dx.
S6220	Fractura baseos MTC I. manus l.dx. comminutiva
S9200	Fractura calcanei l. sin .
S5210	Fractura capitis radii l. sin.
S4200	Fractura claviculae l.dx.
S4220	Fractura colli chirurgici humeri l. dx.
S7200	Fractura colli femoris intracapsularis l.sin.
S4220	Fractura diaph humeri l sin partis proximalis
S8270	Fr. diaphyseos tibiae et fibulae l. dx. spiral.
S5200	Fractura olecrani l. dx.
S0221	Fractura osium nasalium, st.p. epistaxim
S5250	Fractura partis dist. radii l. dx cum disloc.
S4240	Fractura partis dist. humeri l. dx. commin. disloc.
S4220	Fractura partis prox humeri l.sin.
S8200	Fractura patellae l dx
S7200	Fractura pertochanterica femoris l.sin.
S6260	Fractura phalangis prox. digit. II. manus dx. dislocata
S3250	Fractura rami inf. ossii pubis lat. dx.

S3250	Fractura rami sup. et inf. ossis pubis l.dx.
S4210	Fractura scapulae l. sin.
S2220	Fractura sterni
S8210	Fractura tibiae part. prox l. dx, comm. sine disl. C3
S5200	Fractura ulnae proximalis l.dx.
S1220	Fractura vertebrae C 5 compressiva

Předložkové vazby

Míra uplatnění předložkových vazeb v klinické a patologické terminologii a tedy i využití akuzativu a ablativu latinských substantiv a adjektiv v lékařské terminologii představuje často diskutované téma. Nicméně otázka by měla směřovat spíše k tomu, které předložky, proč a s jakou primární funkcí se vůbec používají. V souvislosti se vznikajícími databázemi bude možné postihnout nejčastější případy, ve kterých lékař považuje za důležité takový jazykový prostředek uplatnit.

V následující tabulce je uveden příklad diagnóz, ve kterých se objevuje vazba s latinskou předložkou *in*.

Tab. 7 Uplatnění latinské předložky *in* v autentických klinických diagnózách

C341	Carcinoma in situ bronchi sup. pulm. l. sin
	dehiscentio vulneris, sepsis,ileus intestini tenuis, ascites,mors in tabula
	Ca recti in vaginam et sigmoideum increscens, fistula rectovaginalis,St.p. actinotherapiam,Sigmoideostomia
	Ca recti in vaginam et vesicam urinariam increscens, infiltratio coccys,St.p. radiotherapiam,Appendicitis chronica
	Ca pancreatis in venam portam increscens,Pancreatitis chronica,St.p. cholecystectomiam
	Ca recti in vesicam urinariam et prostatam increscens,Infiltratio coccygis tumorosa,Matstasis hepatis S 5
	Ca ovarii l.dx. in vesicam urinariam increscens v.s.,St.p. chemotherapiam,Appendicitis chronica, steatosis hepatis focalis v.s.

Studenti lékařských fakult v České republice se již v úvodních hodinách lékařské terminologie setkávají s výkladem o latinské předložce *in* a základním schématu jejích syntaktických vazeb. Uplatnění předložky *in* s akuzativní vazbou je ovšem v současném rozsahu databází spíše okrajové, v případě ablativní vazby je její použití mnohem častější a relativně běžně se objevuje i v ustálených spojeních typu *in situ, mors in tabula, in anamnesi, in cursu*.

Akuzativní vazba s předložkou *in* je v získané databázi spojená především s adjektivem *increscens*. Sloveso *increscere* má přitom v klasické latině typicky dativní vazbu. Na druhé straně dativ představuje v oblasti současné výuky lékařské latiny marginální jev, v řadě případů již byl ze studijních materiálů zcela vypuštěn. Ve výše uvedených diagnózách je klasická dativní vazba nahrazena předložkovou vazbou s *in* + akuzativ, resp. předložkovou vazbou s *ad*. Lze tedy vyjádřit předpoklad, že lékařská latina je v podstatě živou latinou, která si přizpůsobuje, nebo hledá nové prostředky pro pojmenovávání toho, co se v písemném záznamu lékařské diagnózy ukazuje jako relevantní.

Krok 2: Korpusová lingvistika a její nástroje

Korpus psaného jazyka lze definovat jako rozsáhlou sbírku strukturovaných psaných textů. Zpravidla se přitom jedná o souvislé texty (např. novinové články, knihy) složené z vět. Ve snaze o co nejvěrnější zachycení zkoumaného jazyka bývá do korpusu zařazováno větší množství dokumentů různé provenience (avšak téhož jazyka a u tématicky zaměřených korpusů rovněž téhož žánru), přičemž u každého takového dokumentu jsou evidována metadata informující o jeho původu. Běžně používané nástroje korpusové lingvistiky jsou přizpůsobeny právě práci s takto strukturovanými daty – dokumenty dělenými do vět, případně též odstavců. Příklady takovýchto korpusů standardního typu jsou korpus českých webových stránek nebo literární antologie (tvorba určitého autora či z určitého období). Samostatně zkoumanou jednotkou v těchto korpusech bývá věta, jejíž hranice se obvykle nepřekračuje (výjimkou jsou korpusy s vyznačením anafor).

Klinické diagnózy jako specifický typ korpusu

Získané databáze autentických klinických diagnóz mají sice spíše než literární formu podobu textové databáze (o čemž vypovídá i jejich vstupní formát – typicky tabulka), přesto však mají s tradičními korpusy mnoho společného. Rozhodujícími faktory pro přístup k nim jako ke korpusům jsou uživateli kladné požadavky na generované výstupy (frekvence, kolokace, morfologické značkování) a také rozdílnost datových struktur používaných jednotlivými zdroji (různá pracoviště totiž používají pro evidenci diagnóz různé informační systémy) spolu s nedostatečně vynucovanou jednoznačností vkládaných dat a jejich nízkou zrnitostí (podobně jako u běžného jazyka lze stejnou informaci zapsat více různými způsoby a struktura sdělení není vyjádřena dostatečně explicitně). Z těchto důvodů bylo rozhodnuto přistupovat k databázi diagnóz jako ke korpusu a k jednotlivým diagnózám v ní jako k větám. Pro usnadnění vyhledávání v datech konkrétního pracoviště a lepší zohlednění tamních jazykových specifik je pro každé pracoviště založen samostatný korpus; v budoucnosti je však možné přistoupit k jejich sjednocení do jednoho velkého korpusu. Doplňující informace k diagnóze (především kód MKN) jsou uloženy jako metadata (atributy) příslušné věty, takže lze při vyhledávání provádět filtraci i na základě nich.

Čtyři jazykové roviny

S přihlédnutím k cíli nabídnout uživatelům jednak přístup k plně autentickému znění diagnóz (včetně všech zkratek, překlepů, vynechaných slov apod.), tak i možnost vyhledávání nad základními slovními tvary a podle morfologických značek bylo při tvorbě korpusu přistoupeno k práci s texty diagnóz hned na čtyřech rovinách:

1. **Povrchová rovina** („word“) zachycuje původní text diagnózy, rozdělený do jednotlivých slov (v korpusech se hovoří o *pozicích*) s využitím tokenizéru – nástroje, který pomáhá vyhledávat hranice slov (typicky mezery). V případech, kdy slova následují po sobě bez mezery, je třeba provést rozdělení ručně – jde např. o zkratkové spojení „l.dx.“ zapsané bez mezery nebo o označení obratle „L1“ (což se na druhé rovině rozepisuje jako dvě slova „lumbalis primae“). V takovýchto případech je informace o vynechání mezery

mezi slovy v původním textu zachována prostřednictvím standardní značky „<g/>“ (pro „glue“ = „lepidlo“). Od slov se při tokenizaci odděluje větná interpunkce (typicky tečka za větou, v diagnózách však nejčastěji čárka, pomlčka apod.), která v korpusu vystupuje jako samostatná pozice, ovšem i při jejím oddělování je používána značka „<g/>“, aby bylo z povrchové roviny vždy možné rekonstruovat přesnou podobu původního textu diagnózy. Při manuální kontrole výstupu z tokenizéru je rovněž třeba správně rozlišovat zkratky od tečky oddělující pseudověty tvořící části diagnózy (např. v diagnóze „Fractura pertrochanterica femoris dx. comminutiva. AO 31-A2.“ je první tečka součástí zkratky „dx.“ = „dextri“ a další dvě jsou jakousi tečkou za větou, která součástí předcházejícího výrazu není – výraz „comminutiva“ není zkratka).

2. **Hloubková rovina** („revword“) přiřazuje ke každému výrazu z povrchové roviny jeho hloubkovou, tj. plně rozvitou, pravopisně a gramaticky správnou formu, jakou by měl v nestaženém a jazykově správném latinském textu diagnózy. Hloubková rovina používá stejné pomocné značky jako povrchová a i ona tedy umožňuje rekonstruovat z jednotlivých po sobě jdoucích pozic smysluplný prostý text. Za tvorbu obsahu na této rovině již je však odpovědný především lingvista vytvářející korpus, protože automatickými nástroji lze do ní pouze přichystat povrchové tvary k manuální úpravě a nanejvýš ještě u některých častých a přitom jednoznačných zkratek nabídnout ke kontrole jejich rozvitý tvar. Tvůrce korpusu musí zavést a při zpracování používat určité konvence, kterými zajistí ve všech případech stejné zpracování opakujícího se jevu a nepřímo tak též definuje množinu rysů lékařské latiny, které přes jejich odchylky od klasické latiny připouští (např. „in vaginam increscens“ neopraví na „vaginae increscens“), a naopak množinu rysů pozorovaného jazyka, které považuje za chyby či zápisová zjednodušení a v hloubkové rovině je podle svého uvážení upravuje do konvenční podoby (např. doplnění chybějícího slova „regionis“ do spojení „V. sclopetarium popliteae“). Je tedy zřejmé, že na rozdíl od povrchové roviny je ke zpracování hloubkové roviny korpusu klinických diagnóz již potřeba dobrá znalost latinského jazyka i lékařských reálií.

3. **Rovina lemmat** („lemma“), tj. základních tvarů slov, je tvořena tzv. lemmatizérem, tedy programem, který k libovolnému slovnímu tvaru přiřadí příslušný tvar základní. U substantiv je v našem případě takovým tvarem nominativ singuláru, u adjektiv nominativ singuláru pozitivu, rovněž u zájmen a číslovek je základním tvarem nominativ, u sloves je dle slovníkových zvyklostí základním tvarem první osoba singuláru aktiva prézentu (slovesa se však v klinických diagnózách prakticky nevyskytují), u příslovcí je lemmatem jejich tvar v pozitivu a u dalších slovních druhů pak jejich již obvykle jediný tvar. Přiřazení tvaru k lemmatu však není vždy jednoznačné (např. ke tvaru „fractura“ nachází lemmatizér vedle substantivního lemmatu „fractura“ i slovesné lemma „frango“) a při absenci statistického značkovače (který by pro svou činnost vyžadoval velké množství již jednoznačně označeného textu) je nutné ponechat disambiguaci (tj. zjednoznačnění) lemmatizace na tvůrci korpusu. Ten musí rozhodnout rovněž v případě, kdy použitý lemmatizér určitý slovní tvar nerozpoznal. S výhodou však lze již jednou manuálně určený slovní tvar přidat do databáze lemmatizéru, takže příště již bude rozpoznán automaticky.

4. **Rovina morfologických značek** („tag“) nese informace o gramatických kategoriích příslušných slovních tvarů. Množina určovaných kategorií se liší v závislosti na slovním druhu a morfologické značkování korpusu je tedy úzce svázané s lemmatizací, protože slovní druh je vlastností lemmatu (byť nebývá v korpusech vyznačen

už na úrovni lemmatické, nýbrž až mezi morfologickými značkami). Např. u substantiv je vhodné určovat rod, pád a číslo – substantivum „articulationis" s lemmatem „articulatio" se proto opatří morfologickými značkami „fem gen sg" (s významem „femininum, genitiv singuláru"). Také na rovině morfologických značek dochází k nejednoznačnostem, a to ještě častěji než na rovině lemmat – i proto, že totožné tvary lze nalézt už přímo v mnoha latinských ohýbacích paradigmatech (např. je shodný nominativ, akuzativ a vokativ pro všechna substantiva neutra). Přesné určení gramatických kategorií pro každý slovní tvar vyžaduje dobré porozumění latinskému textu diagnózy, nejvíce odolává pokusům o automatizaci a tvůrce korpusu tak obvykle stráví zpracováním roviny morfologických značek největší část času potřebného k přípravě korpusu.

Souborový formát a velikost

Nejjednodušším souborovým formátem pro zápis korpusu je tzv. vertikál, což není nic jiného než textový soubor, ve kterém každý řádek odpovídá jedné korpusové pozici a na řádku jsou za sebou (s použitím znaku tabulátoru jako oddělovače) vypsány hodnoty pro danou pozici z každé z existujících rovin. K vytvoření takového souboru stačí použít běžný textový editor, případně tabulkový procesor typu Microsoft Excel nebo LibreOffice Calc. Následuje ukázka několika řádků z takového vertikálu:

word	**revword**	**lemma**	**tag**
femzr	*femur*	femur	noun neut nom sg
l.	*lateris*	latus	noun neut gen sg
dx.	*dextri*	dexter	adj neut gen sg

Velikost korpusu se udává v počtech pozic, což odpovídá počtu řádků vertikálu. Platí obecná zásada, že čím více pozic – a tedy obecně textu – korpus obsahuje (samozřejmě za předpokladu vyváženosti výběru a dalších podmínek), tím větší je jeho schopnost podávat kvalitní statistické výpovědi o zkoumaném jazyce. U korpusů obecného jazyka se běžně počty pozic pohybují v řádu milionů, čehož v případě budovaných korpusů latinských klinických diagnóz nedosahujeme – i tak ale pracujeme s poměrně velkými objemy dat, protože např. jen data z chirurgického pracoviště Fakultní Thomayerovy nemocnice v Praze obsahují 7417 hospitalizačních diagnóz, jejichž tokenizací jsme získali více než 60.000 pozic. Úměrně počtu pozic roste samozřejmě i manuální lingvistická práce, kterou je potřeba v rámci přípravy korpusu vykonat.

Použitý software

Z důvodu rozsáhlosti zpracovávaných dat a za účelem automatizace mnoha prováděných úkonů se k usnadnění tvorby a zpracování vertikálů se využívá specializovaný software. Obecně použitelným nástrojem je tzv. korpusový manažer. Jde o program sloužící k práci s korpusy, resp. vertikály, který uživatelé používají především ke kladení dotazů nad korpusy a generování různých statistických ukazatelů a přehledů (frekvence, konkordance, kolokace, slovní profily). Hlavní silou korpusového manažeru je vhodné předzpracování uživatelem vloženého vertikálu (tzv. kompilace korpusu), které je uživateli skryto, ale díky němuž dovede software následně odpovídat na kladené dotazy ve zlomcích sekundy. Korpusový manažer obvykle obsahuje i tokenizér a pomůže tedy rovněž s vytvořením první verze vertikálu z vloženého prostého textu.

Pro implementaci korpusu autentických klinických diagnóz jsme využili korpusový manažer Sketch Engine vyvíjený společností Lexical Computing Ltd. Prostřednictvím tohoto nástroje již byly vytvořeny velké korpusy (nad 1 milion pozic) pro 52 jazyků, především z textů posbíraných z webu. Sketch Engine je vyvíjen od roku 2003 a zajímavostí je, že jeho vývojářem je Pavel Rychlý z Centra zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity. Přístup ke korpusovému manažeru a jeho prostřednictvím též k vytvořeným korpusům je licencovaný (podrobnosti na http://www.sketchengine.co.uk), ale uživatelé s e-mailovou adresou v síti Masarykovy univerzity je mohou po registraci konzultovat bezplatně prostřednictvím adresy https://ske.fi.muni.cz/ Tam je v současnosti nabízeno 169 korpusů pro 50 různých jazyků – vedle větších i menších evropských jazyků jsou zastoupeny i některé „exotické" jazyky africké a asijské (např. telugština či malajština), avšak hotový latinský korpus zatím není k dispozici žádný. To nám však nebrání využít vestavěných funkcí pro tvorbu a sdílení uživatelských korpusů, vytvořit si v prostředí korpusového manažeru latinský korpus vlastní a ten pak sdílet s dalšími uživateli. V současnosti jsou tak vytvářené korpusy autentických klinických diagnóz přístupné na požádání všem uživatelům, kteří se mohou při registraci prokázat e-mailovou adresou v síti Masarykovy univerzity. Výhledově se však počítá i se zpřístupněním korpusů na vlastním volně přístupném webovém serveru s využitím bezplatné verze softwaru Sketch Engine, omezené o některé pokročilejší vyhledávací funkce (především o tvorbu slovních profilů, která jej nejvíc odlišuje od konkurenčních produktů, ale která ostatně poskytuje kvalitní výsledky až pro korpusy větších rozsahů než jsou ty námi budované).

Vedle obecného softwaru, nezávislého na jazyku zpracovávaných dat, se však při tvorbě korpusů používají i softwarové nástroje a databáze zaměřené na konkrétní jazyk. Do této kategorie patří především lemmatizér, resp. morfologický značkovač. Pro nejpopulárnější jazyky korpusů (především angličtinu) existuje velké množství pokročilých nástrojů, které usnadňují tvorbu korpusů např. i s využitím předpočítaného jazykového modelu, který jim umožňuje poskytovat kvalitní odhady správné lemmatizace a značkování na základě kontextu a dosahovat tak vysoké přesnosti i bez nutnosti manuální disambiguace člověkem. Takováto data však zřejmě nejsou dostupná pro latinu a i škála existujících nástrojů pro tento jazyk je zatím poměrně omezená. Alarmující stav veřejně dostupných počítačově-lingvistických zdrojů pro latinu potvrzuje i autorka práce „Analysis of a Potential Latin Treebank" (Jacque 2006, str. 3) a navzdory stáří této práce se zdá, že ani v posledních letech nedošlo k významným posunům k lepšímu. Nejpokročilejší využitelné nástroje jsou tak – k našemu štěstí – právě ty, které jsou k proveditelnému počítačem asistovanému zpracování korpusu středního rozsahu nezbytně potřeba.

K urychlení lemmatizace a morfologického značkování (třetí a čtvrtá rovina korpusu) tak používáme morfologický analyzátor Morpheus. Vznikl v rámci digitální knihovny klasických jazyků Perseus (http://www.perseus.tufts.edu/hopper/) a jeho základem je rozsáhlá databáze latinských slovních tvarů: používaná verze pokrývá 232 732 tvarů přiřazených k 48 816 různým lemmatům a nabízí 720 255 možných morfologických analýz (přechod od slovního tvaru k lemmatu a morfologickému značkování). I přes tuto svou velikost se však ukázalo, že databáze neobsahuje některá lemmata, která se v lékařské latině běžně objevují (např. „talocruralis"), proto je nutné ji průběžně vlastními silami o tato dosud neznámá lemmata rozšiřovat.

Ukázky dotazů nad korpusem

Korpusový manažer Sketch Engine nabízí jednoduché grafické uživatelské rozhraní, jehož prostřednictvím je možno mnohá vyhledávání realizovat prostřednictví několika kliknutí myší. Složitější vyhledávací dotazy je již třeba zapisovat manuálně v jazyce CQL (Corpus Query Language), který poskytuje velkou volnost v kombinování různých rovin a sousedních pozic v dotazu, včetně možnosti využití regulárních výrazů (tzv. „wildcards" – vyhledávání slov podle masky). I s výsledky takovýchto dotazů je pak ale možné dále manipulovat s využitím myši a snadno si tak zobrazovat např. levý či pravý kontext nalezených korpusových pozic.

Nad budovaným korpusem autentických latinských klinických diagnóz lze pokládat např. takovéto dotazy (uvedeny jsou názvy položek menu a vyplňované hodnoty formulářů):

- **Frekvence slovních tvarů (unigramů) v korpusu:**
  Seznamy slov – Hledej atributy: lc – Vytvořit seznam slov
- **Výskyty slovního tvaru *fractura* v korpusu s kontextem:**
  Konkordance – Jednoduchý dotaz: *fractura* – Vytvořit konkordanci
  - **Seřazení nálezů podle pravého kontextu (tj. co následuje za nalezenou pozicí):**
    Třídění – Pravý kontext
  - **Frekvence slovních tvarů na 1. pravé pozici (tj. hned za nalezenou pozicí):**
    Frekvence – První úroveň – Atribut: lc – Pozice: 1R – Vytvořit frenkvenční seznam
- **Frekvence dvojic slovních tvarů (bigramů) v korpusu:**
  Seznamy slov – Hledej atributy: lc – Použít n-gramy. Hodnota n: 2 – Vytvořit seznam slov
- **Výskyty lemmata *dexter* v korpusu s kontextem:**
  Konkordance – Typy dotazů – Typ dotazu: lemma – Lemma: *dexter* – Vytvořit konkordanci
  - **Frekvence nalezených slovních tvarů (na hloubkové rovině):**
    Frekvence – První úroveň – Atribut: lc – Vytvořit frekvenční seznam
- **Výskyty zkratek v korpusu (slovních tvarů na povrchové rovině zakončených na tečku) s kontextem:**
  Konkordance – Typy dotazů – Typ dotazu: CQL – CQL: "..*\." – Implicitní atribut: word – Vytvořit konkordanci
  - **Frekvence nalezených slovních tvarů (na povrchové rovině):**
    Frekvence – První úroveň – Atribut: word – Vytvořit frekvenční seznam
- **Výskyty slov s příponou *-itis* (na hlobukové rovině) v korpusu s kontextem:**
  Konkordance – Typy dotazů – Typ dotazu: CQL – CQL: ".*itis" – Implicitní atribut: revword – Vytvořit konkordanci
  - **Stejný seznam seřazený podle slovního tvaru na nalezené pozici:**
    Třídění – Atribut: revword – Třídící klíč: node – Třídit konkordanci

Bibliografie

Crane, Gregory R. [Ed]. 2014. „Perseus Digital Library.“ [online] Tufts University. Dostupné z: http://www.perseus.tufts.edu/ [cit. 2014–03–30].

Čermák, František [Ed] – Blatná, Renata [Ed]. 2006. *Korpusová lingvistika – stav a modelové přístupy.* Praha: Nakladatelství Lidové noviny (Studie z korpusové lingvistiky, sv. 1).

Jacque, Kristin. 2006. „Analysis of a Potential Latin Treebank. In Natural Language Processing: Fall 2006 Projects.“ [online] Ann Arbor: University of Michigan. Dostupné z: http://web.eecs.umich.edu/~rthomaso/courses/nlp/projects.06/Kris_Jacque.pdf [cit. 2014–03–30].

Kilgarriff, Adam – Rychlý, Pavel – Smrž, Pavel – Tugwell, David. 2006. „The Sketch Engine.“ [online] In *Proc EURALEX 2004,* 2006. Lorient, 105–116. Dostupné z: http://www.sketchengine.co.uk/documentation/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw [cit. 2014–03–30].

RESUMÉ

Učitelé lékařské latiny se obvykle potýkají s nedostatkem zdrojových dokumentů, které by odrážely reálný stav použití latiny v klinickém prostředí, jelikož možnosti zpřístupnění autentické lékařské dokumentace jsou vzhledem k ochraně osobních údajů pacientů i lékařů velmi omezené. Ve spolupráci s fakultními nemocnicemi v Praze a v Brně vzniká korpus autentických klinických diagnóz, které jsou excerpovány z elektronických chorobopisů vybraných klinik (u nichž je předpokládáno frekventované využití latiny pro zápis hospitalizačních a operačních diagnóz, tj. chirurgie, traumatologie, dále i gynekologie-porodnictví, ORL atd.). Ve svém příspěvku autoři demonstrují, jak lze využít nástroje korpusové lingvistiky (morfologický analyzátor Morpheus a software Sketch Engine) pro potřeby učitele lékařské latiny při tvorbě výukových materiálů, resp. testů. Jde zejména o možnosti vyhledávání typických kolokací a přehled z hlediska frekvence jednotlivých termínů nebo jeho komponentů (např. předložkové vazby).