Univerzita Karlova v Praze

1. lékařská fakulta

Studijní program: Doktorský

Studijní obor: Biomedicínská informatika

Ing. Karel Zvára

Extrakce informací z lékařských textů

Extracting Information from Medical Texts

Disertační práce

Vedoucí závěrečné práce/Školitel: doc. Ing. Vojtěch Svátek, Dr.

Praha, 2017

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem řádně uvedl a citoval všechny použité prameny a literaturu. Současně prohlašuji, že práce nebyla využita k získání jiného nebo stejného titulu

Souhlasím s trvalým uložením elektronické verze mé práce v databázi systému meziuniverzitního projektu Theses.cz za účelem soustavné kontroly podobnosti kvalifikačních prací.

V Praze, 22. června 2017

Ing. Karel Zvára

Podpis

Identifikační záznam

ZVÁRA, Karel, *Extrakce informací z lékařských textů.* Praha, 2017. 78 s., 3 příl. Disertační práce (Ph.D.). Univerzita Karlova v Praze, 1. lékařská fakulta, Ústav hygieny a epidemiologie, Studničkova 2, Praha 2. Školitel: Svátek, Vojtěch.

Identification record

ZVÁRA, Karel, Extracting Information from Medical Texts. Praha, 2017. 78 p., 3 attachments. Dissertation Thesis (Ph.D.). Charles University in Prague, 1st Faculty of Medicine, Institute of Hygiene and Epidemiology, Studničkova 2, Praha 2. Supervisor: Svátek, Vojtěch.

Poděkování

Rád bych poděkoval svému školiteli doc. Ing. Vojtěchu Svátkovi, Dr. za odborné vedení a nepostradatelné rady během doktorského studia.

Velké díky patří MUDr. Marii Tomečkové, CSc. a MUDr. Janu Peleškovi, CSc., kteří svůj čas věnovali zpracovávání původních lékařských zpráv. Bez jejich odborné pomoci by tato práce vůbec nemohla vzniknout.

Můj obrovský dík patří také mé rodině za podporu a každodenní pomoc při mé práci i studiu.

Abstrakt (česky)

Cílem mé práce bylo zjistit specifické vlastnosti českých lékařských zpráv z hlediska možnosti extrahovat z nich konkrétní informace.

Pro svoji práci jsem měl k dispozici celkem 268 anonymizovaných narativních lékařských zpráv ze dvou ambulantních pracovišť. Studoval jsem standardy pro uchování elektronické zdravotnické dokumentace i pro přenos klinických informací mezi informačními systémy ve zdravotnictví. Věnoval jsem se také implementování elektronického zdravotního záznamu v zubním lékařství.

Nejprve jsem se narativní lékařské zprávy snažil zpracovat pomocí nástrojů pro zpracování přirozeného jazyka (Natural Language Processing, NLP). Dospěl jsem k závěru, že narativní lékařské zprávy v českém jazyce jsou typickému českému textu velmi vzdálené zejména pro svoji heslovitost a absenci české větné stavby. Obsahují také velké množství překlepů, zkratek a zkrácených slov. Vzhledem k nedostupnosti hlavních mezinárodních klasifikačních systémů v českém jazyce jsem se rozhodl pokračovat ve výzkumu vývojem metody pro přípravu vstupního textu pro překlad a jeho sémantickou anotaci.

Hlavním cílem této části výzkumu bylo navrhnout metodu a podpůrný software pro interaktivní korekci a sémantickou anotaci narativních lékařských zpráv, které by umožnily jejich snadnější použití, s menším množstvím chyb i mimo jejich původní kontext.

Vyvinul jsem třífázovou metoda předběžného zpracování s cílem podpořit druhotné využití lékařských zpráv. Metoda třífázového předzpracování narativních klinických zpráv byla ověřena na 49 anonymních českých lékařských zprávách z oblasti kardiologie.

Klíčová slova

Lékařské zprávy, zpracování přirozeného textu, extrakce informací, elektronický zdravotní záznam.

Abstract

The aim of my work was to find out the specific features of Czech medical reports in terms of the possibility of extracting specific information from them.

For my work, I had a total of 268 anonymized narrative medical reports from two outpatient departments. I have studied standards for preserving electronic health records and for transferring clinical information between healthcare information systems. I have also participated in the process of implementing electronic medical record in the field of dentistry.

First of all, I tried to process narrative medical reports using natural language processing (NLP) tools. I came to the conclusion that narrative medical reports in the Czech language are very different than a typical Czech text, especially because it mostly contains short telegraphic phrases and the texts lack typical Czech sentence structure. It also contains many misspellings, acronyms and abbreviations. Another problem was the absence of existence of the Czech translation of the main international classification systems. Therefore I decided to continue the research by developing the method for pro-processing the input text for translation and its semantic annotation.

The main objective of this part of the research was to propose a method and support software for interactive correction and semantic annotation of narrative medical reports that would allow their easier use with fewer errors outside of their original context.

I have developed a three-phase pre-processing method to encourage secondary use of medical reports. The method of three-phase pre-processing of narrative clinical reports was verified on 49 anonymous Czech medical reports in the field of cardiology.

Keywords

Clinical reports, natural language processing, information extraction, electronic health record.

Obsah

1. Úvod ... 3
1.1 Cíl práce ... 4
1.2 Struktura této práce ... 5
1.3 Stav výzkumu ... 5
1.3.1 České lékařské zprávy ... 5
1.3.2 Výzkum v zahraničí ... 7
1.3.3 Souhrn poznatků ... 9
2. Legislativní rámec a standardy ... 11
2.1 Legislativní rámec ... 11
2.2 Standardy pro uchování elektronické zdravotnické dokumentace ... 12
2.2.1 Standardy pro předávání zpráv (messaging) ... 13
2.2.2 Standardy pro zaznamenání zdravotnické dokumentace (clinical document) ... 14
2.2.3 Standardy kombinující předávání zpráv a zaznamenání zdravotnické dokumentace ... 15
2.2.4 Klasifikační systémy a číselníky ... 17
2.2.5 Shrnutí cílových standardů ... 20
3. Automatická lingvistická analýza textů lékařských zpráv ... 21
3.1 Podklady a metody ... 21
3.1.1 Příprava volného textu ... 21
3.1.2 Provedení ... 22
3.2 Výsledky ... 25
3.3 Diskuse ... 28
4. Elektronický zdravotní záznam pro orální medicínu s interaktivní komponentou zubního kříže ... 29
4.1 Metody ... 30
4.2 Výsledky ... 30
4.3 Diskuse ... 40
5. Třífázová metoda předběžného zpracování ... 43
5.1 Podklady a metody ... 43
5.1.1 Tokenizace lékařské zprávy ... 45
5.1.2 Normalizace tokenizované klinické zprávy ... 45
5.1.3 Sémantická anotace normalizované klinické zprávy ... 46
5.1.4 Normalizační databáze ... 46
5.1.5 Softwarový nástroj TOCESA ... 48

5.2 Výsledky 49
5.2.1 Fáze I: Automatizovaná tokenizace volnotextových lékařských zpráv 49
5.2.2 Fáze II: Normalizace tokenizovaných klinických zpráv 49
5.2.3 Fáze III: sémantická anotace normalizované klinické zprávy 61
5.3 Diskuse 63
6. Závěry 67
8. Citovaná literatura 71
Výkladový slovník 78
Přílohy 79

1. Úvod

Lékařské texty, kterými se v práci zabývám, jsou texty, které vznikají v průběhu poskytované zdravotní péče v diagnosticko-terapeutickém procesu. Patří k ním zejména lékařské zprávy, které jsou významnou součástí zdravotnické dokumentace. Slouží pro uchování zjištěných informací o zdraví, o provedených úkonech, léčbě a souvisejících administrativních údajích. V České republice mají lékařské zprávy obvykle povahu volného textu, který je formátován jen pomocí mezer, tabulátorů a nových řádků. Lékařské zprávy, jako součást zdravotnické dokumentace obsahují zejména identifikační a administrativní údaje, údaje o zdravotním stavu či okolnostech úmrtí, rozhodnutí a pokyny ošetřujícího lékaře, případně další zákonem vyžadované údaje.

Zdravotnická dokumentace v evropských systémech veřejného zdravotnictví slouží k několika různým účelům:

- zaznamenání údajů pro další poskytování zdravotní péče (ze strany poskytovatele péče, ev. systému veřejného zdravotnictví) či zajištění si další zdravotní péče (ze strany pacienta či jeho zástupce) a pro zajištění kontinuity sdílené péče o zdraví;
- vyhovění zákonným požadavkům, na jejichž plnění se váže i prokazování splnění povinností vůči státní autoritě, a to jak po stránce oprávnění pro výkon činnosti, dodržení stanovených postupů, využití schválených prostředků, léčivých přípravků a léčiv;
- prokázání provedení výkonů ve vztahu k systému veřejného zdravotního pojištění, a to i v případech, kdy jsou úhrady nasmlouvány jako kapitační platba u praktických lékařů, paušálem (u poskytovatelů lůžkové péče) nebo jde o přímé platby (i bez vazby na systém veřejného zdravotního pojištění);

- zaznamenání údajů jako důkazu pro případné budoucí trestní či občanskoprávní řízení, například pro obranu proti žalobě na náhradu škody.

Je nezbytné si vždy uvědomit celou šíři účelů tvorby zdravotnické dokumentace, neboť jednotlivé výše uvedené účely se liší ve své motivaci a forma i konečný obsah zdravotnické dokumentace je ovlivněna všemi těmito motivy. Nelze se proto divit potřebě lékařů zachovat možnost volného textového vyjádření při tvorbě lékařských zpráv.

1.1 Cíl práce

Hlavní cíl práce uvedený v pojednání o disertační práci je *„zjištění specifických vlastností českých lékařských zpráv z hlediska možnosti extrahovat z nich konkrétní informace“*. Pro realizaci tohoto cíle je třeba zajistit splnění těchto dílčích cílů:

1. Zodpovědět otázku: „Které vlastnosti českých lékařských zpráv působí největší problémy v nestatistických fázích zpracování přirozeného jazyka?“

2. Navrhnout základní postup pro analýzu česky psaných lékařských zpráv.

3. Pomocí vlastní implementace s možností využití externích nástrojů ověřit navržený postup pro analýzu lékařských zpráv založených na češtině a základní postup i výsledky publikovat.

4. Ověřit možnosti extrakce strukturované informace z lékařských zpráv a jejího vložení do elektronického zdravotního záznamu.

Hlavní cíl i dílčí cíle se podařilo v rámci výzkumu dosáhnout. Z odborných lékařských zpráv psaných v českém jazyce lze pod supervizí odborníka a za použití technologie pro zpracování přirozeného jazyka (natural language processing) získávat potřebné odborné

informace, například seznam známých alergických reakcí či výsledky biochemických vyšetření.

1.2 Struktura této práce

Během studia jsem se postupně věnoval zpracování narativních lékařských zpráv nástroji pro zpracování přirozeného jazyka, strukturalizaci informací o zdraví do elektronické formy a třífázové metodě předzpracování textu a podpory sémantické anotace. Tomu odpovídá i struktura této práce.

Druhá kapitola popisuje prostředí, ze kterého české narativní lékařské zprávy vycházejí a jejich cílové struktury, klasifikační systémy a základní technologii pro zpracování textů.

Třetí kapitola popisuje výzkum, ve kterém jsem se zaměřil na zpracování narativních lékařských zpráv pomocí nástrojů pro zpracování přirozeného jazyka.

Ve čtvrté kapitole popisuji související výzkumnou činnost zaměřenou na návrh elektronického zdravotního záznamu., který umožňuje ukládat jak strukturovanou informaci, tak i nestrukturovaný lékařský text.

Pátá kapitola se věnuje třífázové metodě pro předzpracování a sémantickou anotaci lékařských zpráv („3PP (three phase preprocessing) metoda“) včetně jejího ověření.

Šestá kapitola diskutuje výsledky výzkumu a sedmá kapitola je shrnuje do závěru.

1.3 Stav výzkumu

1.3.1 České lékařské zprávy

Snaha o částečně automatizovanou extrakci informací z narativních lékařských zpráv je velmi úzkou specializací. Tyto lékařské zprávy lze vnímat jako texty v českém jazyce,

výsledky vlastního i dřívějšího výzkumu však ukázaly, že české narativní lékařské zprávy se od běžného českého textu významně liší. Ve světě se o automatizaci při extrakci informací z narativních lékařských zpráv snaží více skupin, ovšem obvykle se zajímají o aplikaci na anglicky psané texty a nikoliv na jazyk příbuzný češtině.

Tématu zpracování lékařských zpráv v češtině či slovenštině se ve svých diplomových a disertačních pracích věnovali především Semecký v [1], Smatana v [2] a Přečková v [3].

Jiří Semecký navrhl vstupní zprávy zpracovávat pomocí lingvistické analýzy a pomocí regulární analýzy. Ve své práci věnoval také extrakci informací, kterou popsal jako způsob nahlížení na vstupní obsah v rámci znalostní domény. V tradičním pojetí zpracování přirozeného jazyka (Natural Language Processing) zahrnuje například určování postavení tokenu ve větě (Part of Speech Tagging – „PoS Tagging“).

V případě lékařských zpráv jsou tak možné v zásadě dva hlavní přístupy. První přístup předpokládá, že lékařské zprávy jsou řádnými texty v českém jazyce a zpracování by tak vycházelo z tradičního pojetí zpracování přirozeného textu. Druhý možný přístup nepředpokládá, že text lékařských zpráv odpovídá přirozenému psanému projevu a soustředí se proto na identifikaci termínů, případně částí zprávy. Oba přístupy je možné kombinovat.

V práci uvedl dvě definice sémantické analýzy volného lékařského textu. V první definici jde o nalezení algoritmu pro vyhledání fragmentů lékařské zprávy, které mají význam popsaný znalostní bází. V druhé definici jde o nalezení algoritmu pro sémantickou analýzu a označení nalezených fragmentů ve zprávě značkami. Semecký uvádí důvody, pro které se zdá, že lingvistická analýza lékařských zpráv nemůže být úspěšná: *„V lékařských zprávách se objevuje velmi málo souvislých vět a nebývají vždy regulérně odděleny interpunkcí. Jak bylo již výše zmíněno, informačně nejbohatší jsou právě heslovité úseky zpráv, kde nám*

syntaktická dokonce ani lexikální analýza nepomůže, neboť tyto úseky nemají charakter konkrétních vět českého jazyka.“ Jako druhý způsob analýzy volného lékařského textu Semecký uvádí regulární analýzu, tedy ověřování, zda jednotlivé úseky zprávy splňují předem určená, často na jazyku nezávislá pravidla.

Peter Smatana se podobně jako Jiří Semecký ve své práci zabýval lingvistickou i regulární analýzou. Oproti práci Semeckého doplnil slovníky pro lingvistickou analýzu. Podobně jako Semecký došel k závěru, že *„analýza na základe viet v takýchto dokumentoch nie je možná“*.

Smatana obdobně jako Semecký nepoužil žádné databáze možných termínů ani obecné jazykové slovníky. Při rozšíření o použití lingvistické analýzy dochází k mírně lepším výsledkům než při využití jen analýzy pomocí regulárních výrazů.

Petra Přečková se ve své práci mimo jiné věnovala jazykové analýze česky psaných narativních lékařských zpráv. Ve své analýze zmiňuje občasné pořizování českých textů bez diakritiky, velmi časté překlepy, chybějící mezery (zejména mezi číselnými hodnotami a jednotkami) a zaměňování číslice 0 (nula) a písmena O. Zmiňuje také časté používání zkratek a zkrácených tvarů, přičemž i stejný lékař v jedné zprávě může použít více variant zkrácení. V závěru uvádí, že narativní záznam českých lékařských zpráv je velmi nehomogenní a nestandardizovaný.

1.3.2 Výzkum v zahraničí

Lékařskými zprávami a jejich formou ve srovnání s běžným textem se zabýval například Tsung O. Cheng, který v [4] uvedl, že 90 % začínajících lékařů jiných odborností nedokáže porozumět termínům ze zpráv oboru otorhinolaryngologie. Výborný přehled na téma lékařských zpráv sepsala Van Ginneken v [5].

Van Ginneken se v [6] zabývala obsahem narativních lékařských zpráv a jejich možnou strukturalizací. V [7] se věnovala datovému modelu elektronického zdravotního záznamu systému Orca, který byl založený na třech vrstvách: dotazovací vrstvy, funkční vrstvy a vrstvy rozhraní. Porovnání Orca a HL7 se věnoval také Miroslav Nagy ve své disertační práci [8].

Záznamu strukturovaných dat a extrakci informací do elektronického zdravotního záznamu v kardiologii, pediatrii a dalších oborech se dále věnují publikace [9-12].

García-Remesal se v [13] věnoval možnosti integrovat klinická data z různých zdrojů – z narativních zpráv a ze strukturovaných datových zdrojů. Výsledky dotazníkového šetření mezi lékaři vyhodnotil jako srovnatelné.

Možné přístupy k extrakci informací z narativních lékařských zpráv pomocí nástrojů NLP studoval například Blaschke v [14], který se pokoušel extrahovat informace z volných textů s využitím bibliografické databáze PubMed. Johnson se v [15] věnoval možnosti strukturovat narativní lékařské zprávy už v okamžiku jejich tvorby právě s využitím nástrojů pro zpracování přirozeného jazyka. Hui se v [16] věnoval výsledkům, které dosáhl v soutěži extrahování údajů z anglicky psaných narativních lékařských zpráv i2b2. Srovnání výsledku různých metod se věnovali také Meystre a Haug v [17]. Přístupy k extrakci pomocí nástrojů NLP byly dále studovány také v [18-19].

Meystre a další se v [20] věnovali anglicky psaným lékařským narativním zprávám. Popisují, že některé klinické texty mají formu krátkých telegrafických sdělení, zatímco jiné, zejména propouštěcí zprávy, jsou často formulovány tak, aby byly jasné. Všímají si, že narativní lékařské zprávy jsou plné zkratek a zkrácených slov a zhruba třetina z takových zkratek či zkrácených slov je využívána ve více významech tak, že jsou nejednoznačné i při znalosti

kontextu. Obdobně si všímají vysokého počtu překlepů ve zprávách. Popisují techniky snah o extrakci informací, a to jak pomocí nástrojů pro zpracování přirozeného textu, tak pomocí nástrojů pro předzpracování / vyčištění vstupních textů. Obsah shrnujícího článku se vzhledem ke kontextu autorů vztahuje na anglicky psané zprávy.

Automatické extrahování strukturované informace z anglicky psaných narativních lékařských zpráv podporoval systém MedLEE (Medical Language Encoding and Extraction) [21]. Další metody lze najít v článcích a patentech [22-23]. U jiných jazyků již bylo dosaženo částečných úspěchů, ovšem vždy u jazyků s mnohem lépe dostupnými nástroji pro zpracování textů a často i s dostupnými klasifikačními systémy v daném jazyce (např. část SNOMED CT existuje i ve španělské verzi) [24]. Význam jednotlivých termínů v lékařských textech často závisí na kontextu, vlivu místní legislativy a na zvyklostech daných zejména vzděláváním lékařů[25-27].

1.3.3 Souhrn poznatků

Badatelé, kteří zkoumali možnosti extrahovat informace z narativních lékařských zpráv v zahraničí, se zpravidla věnovali anglicky psaným lékařským zprávám. Věnovali se jak metodám využívající lingvistickou analýzu, tak snahám o předzpracování těchto zpráv. Zahraniční výzkum se týkal převážně anglicky psaných narativních lékařských zpráv. To je významné v souvislosti s tím, že cílové klasifikační systémy (zejména SNOMED CT a LOINC) jsou k dispozici v anglickém jazyce. Termíny nalezené v textu tak lze vyhledávat přímo v těchto klasifikačních systémech.

Možnostem extrahovat informace z českých narativních lékařských zpráv se věnovali především Semecký, Smatana a Přečková, kteří se věnovali lingvistické a regulární analýze.

Dospěli k závěrům, že české narativní lékařské zprávy mají českou větnou stavbu, obsahují velké množství zkratek, zkrácených slov, překlepů a jiných chyb.

2. Legislativní rámec a standardy

2.1 Legislativní rámec

Náležitosti zdravotnické dokumentace, dokonce ani povinnost ji vést, nebyly dlouhou dobu v Československu a poté v České republice vůbec upraveny zákonem. Právní otázky v souvislosti se zdravotnickou dokumentací upřesnil například Roman Žďárek v [28].

Náležitosti vedení zdravotnické dokumentace byly zákonem upraveny novelizací zákona o péči o zdraví lidu č. 20/1966 Sb. [29] ve znění novely 260/2001 Sb. [30] Zákon výslovně stanovil, že zdravotnická dokumentace musí obsahovat osobní údaje pacienta v rozsahu nezbytném pro jeho identifikaci a zjištění anamnézy, informace o onemocnění, o průběhu a výsledcích vyšetření, léčení a dalších významných okolnostech souvisejících se zdravotním stavem pacienta a s postupem při poskytování zdravotní péče. Minimální rozsah zdravotnické dokumentace byl stanoven vyhláškou Ministerstva zdravotnictví č. 385/2006 Sb. [31]

Úpravu vedení zdravotnické dokumentace převzal v roce 2011 zákon 372/2011 Sb. „o poskytování zdravotních služeb“ [32], který, podobně jako předchozí úprava, předpokládá možnost jejího uchování na elektronickém nosiči.

České lékařské zprávy obvykle mají formu volného textu. Skutečnost, že lékaři jsou organizováni v profesní organizaci a mají obdobné vzdělání, vede k tomu, že zprávy psané různými lékaři mají obdobnou strukturu, především pořadí jednotlivých částí.

Text bývá formátován pomocí uspořádání mezer, nových řádků a obvykle nepoužívá žádné značkování (markup). Obvyklým způsobem je tvoření nových lékařských zpráv zkopírováním obsahu minulé zprávy a její úpravou. Tím zapisující lékař ušetří velké

množství času a nezapomene do zprávy uvést povinné administrativní údaje a údaje o dlouhodobých diagnózách a dříve prodělaných operacích. Tento způsob vytváření spolu však nese také riziko, že dojde ke zkopírování údajů, které již nejsou platné. Podobné problémy s kopírováním a aktualizací starších zpráv se netýkají jen České republiky, ale i jiných zemích, viz např. [33].

2.2 Standardy pro uchování elektronické zdravotnické dokumentace

Standardům pro předávání a uchovávání údajů zdravotnické dokumentace se dlouhodobě věnuje několik organizací, komunit a implementuje je řada dodavatelů software i hardware pro zdravotnictví. Z hlediska formy užití je nutné rozlišovat mezi předáváním zpráv (tzv. *messaging*) a zaznamenáváním zdravotnické dokumentace v elektronické podobě (tzv. *clinical document*).

Extrakce informací z lékařských textů, kterou jsem se během svého studia zabýval, se týká zdravotnické dokumentace v elektronické podobě (*clinical document)*. Přesto uvádím základní informace i k předávání zpráv, neboť jde o úzce související standardy.

Předávání zpráv lze vnímat jako přenášení části elektronické zdravotnické dokumentace v daném čase. Standardy pro předávání zpráv a pro uchování elektronické zdravotnické dokumentace od stejného původce proto obvykle sdílejí stejné struktury.

To však neznamená, že by bylo možné elektronickou zdravotnickou dokumentaci vnímat jako „množinu předávaných zpráv“. To proto, že elektronická zdravotnická dokumentace má jiné vlastnosti, než předávání zpráv. Elektronická zdravotnická dokumentace se zaměřuje na pacienta, může obsahovat informace ve vztahu k času, je zpravidla dlouhodobá a je sdílena (v českém prostředí zejména prostřednictvím výpisů ze zdravotnické dokumentace) různými poskytovateli zdravotních služeb. Oproti tomu se předávání zpráv omezuje na jejich

účel, je vázaná ke kontextu jejich předání (point-to-point) a často obsahuje jen informaci o změně a nikoliv celkovou stavovou informaci.

2.2.1 Standardy pro předávání zpráv (messaging)

Úkolem standardů pro předávání zpráv je zajistit standardní předávání údajů o péči o zdraví elektronickou cestou. Tento způsob se běžně využívá pro komunikaci informačních systémů s přístroji, pro předávání údajů mezi informačními systémy v rámci zařízení i pro předávání údajů mezi informačními systémy mezi různými organizacemi.

Ve světě zřejmě nejrozšířenějšími takovými standardy jsou HL7 [34] verze 2 a DICOM [35] (Digital Imaging and Communications in Medicine). DICOM je využíván pro předávání obrazových údajů (RTG snímků či záznamů ze sonografie, angiografie a podobně).

HL7 verze 2 se využívá pro komunikaci s jinými přístroji, např. s přístroji provádějícími chemické rozbory či mezi různými systémy v rámci jedné nemocnice. Využití HL7 pro automatizované předávání dat v rámci poskytovatele zdravotní péče či mezi poskytovatelem zdravotní péče a jeho partnery při poskytování péče (např. lékárnami, laboratořemi a podobně) často zahrnuje synchronizaci databáze pacientů (automatické předávání změn a nově vložených záznamů) či předávání údajů od lékaře pro lékárnu.

HL7 verze 3 je již dlouhou dobu ve vývoji. Referenční informační model HL7 verze 3 (RIM) je standardem ISO/HL7 21731 [36]. Jde o velmi obecný a robustní model, jehož implementace je náročná a samotný standard nedává jednoznačné pokyny pro způsob implementace. Pro propojení systémů pomocí protokolů HL7 verze 2 či HL7 verze 3 je zpravidla zapotřebí shoda na způsobu implementace. To zajišťuje sdružení Integrating the Health Enterprise (IHE) [37], jehož členy jsou standardizační organizace (mj. Health Level

7 Inc.), poskytovatelé zdravotních služeb, výrobci software pro zdravotnictví a vládní organizace podílející se na organizaci systému veřejného zdravotnictví.

V České republice vznikl v 90. letech národní standard pro předávání zpráv, který byl pojmenován prostě *Datový standard (Ministerstva zdravotnictví) [38]*[10]. Z tohoto názvu vzniklo zkratkové slovo *DASTA*, které se dnes pro označení tohoto standardu běžně používá. *DASTA* je úzce provázaný s *Národním číselníkem laboratorních položek* (NČLP) [38], přičemž tento „číselník" je národním klasifikačním systémem, který obsahuje množství různých číselníků, a který není začleněný do systému mezinárodně užívaných klasifikačních systémů, zejména do UMLS [39].

V roce 2008 byl v Evropské unii spuštěn projekt epSOS [40], který kombinuje předávání zpráv (messaging) a zaznamenání zdravotnické dokumentace (clinical document). Tento standard je popsán níže v návaznosti na další standardy a klasifikační systémy.

2.2.2 Standardy pro zaznamenání zdravotnické dokumentace (clinical document)

Hlavním, především v USA využívaným, systémem pro uchování elektronické zdravotnické dokumentace, je Health Level 7 Continuity of Care Document (CCD) [34]. Tento standard vznikl ze dvou, původně velmi odlišných, standardů – Health Level 7 Clinical Document Architecture (CDA) [34] a Continuity of Care Record (CCR) [41].

CDA vzniklo jako robustní standard, podobně jako ostatní Health Level 7 standardy. CCR oproti tomu vznikalo z konkrétních komunikačních potřeb („odspodu"), tedy především z klinické praxe. CCR se stalo rychle použitelným, ovšem s absencí robustních mechanismů. Standard CCD převzal výhody obou standardů.

Detailnímu přehledu literatury k tématu elektronických zdravotních záznamů se věnuje [42].

2.2.3 Standardy kombinující předávání zpráv a zaznamenání zdravotnické dokumentace

Zajímavým standardem, který je aktivně využíván zejména v Austrálii, ve Slovinsku a zřejmě bude využíván i v Brazílii, je openEHR [43][13], který vyvíjí stejnojmenná australská nadace. Projekt vznikl z původně evropského projektu Good European Health Records (GEHR) [44] a následných implementačních snah v rámci projektu Synapses [45]. Po nepřijetí v západní Evropě se těžiště projektu přesunulo do Austrálie a vznikla nadace openEHR. Celý standard OpenEHR stojí na definici „částí dokumentace" pomocí *archetypů*, které jsou zapisované pomocí *Archetype Definition Language* (ADL) [46]. Tento způsob přejal CEN ve standardu 13606 [47].

Na přelomu tisíciletí byl příslušnými pracovními skupinami CEN (TC 251) a ISO (TC 215) navržen a následně schválen standard ISO/EN 13606 označovaný *EHRcom*. Tento standard obecně specifikuje požadavky na strukturalizaci elektronického zdravotního záznamu, požadavky na řízení přístupu k němu a obecně stanovuje požadavky na předávání tohoto záznamu jako celku nebo jen vybraných částí. Pojímá tedy zdravotnickou dokumentaci tak, že za základ považuje elektronický zdravotní záznam (clinical document), ze kterého se jeho filtrováním (podle požadavku či přístupových práv) stává zpráva, která je následně předána jinému systému. Kombinuje tak oba výše zmíněné přístupy – elektronický zdravotní záznam i předávání zpráv.

Standard EHRcom navíc kombinuje také standardy Health Level 7 a nadace OpenEHR, ovšem způsob implementace ponechává v informativní části standardu, standard jej tedy nenařizuje.

Na EHRcom navazuje standardizace konceptů při kontinuální péči o zdraví (ISO EN 13940, ContSys) [48].

Účelu zkoumaného tématu je nejblíže výstup projektu epSOS, který probíhal v Evropské unii v letech 2008 až 2014. Jedním z jeho výsledků je epSOS Patient Summary [49], které bylo definováno tak, aby obsahovalo nejdůležitější údaje pro poskytování bezpečné zdravotní péče v případě neočekávané a neplánované situace vyžadující lékařský zásah. EpSOS Patient Summary obsahuje tyto skupiny informací [50]:

- administrativní údaje (např. jméno, datum narození, pohlaví pacienta),
- nejdůležitější klinické údaje (alergie, současné diagnózy, implantáty, nedávné operace),
- aktuální preskripci,
- informace o samotném záznamu (kdy a kým byl vytvořen, údaje nutné pro jeho důvěryhodnost).

EpSOS Patient Summary je zařazen do kombinovaných standardů z důvodu, že projekt předpokládá zařazení vytváření dokumentů epSOS PS v pilotních projektech jednotlivých členských států a standardizaci jeho přenosu.

V současné době probíhá v Evropě pilotní nasazení. V České republice jde o projekt Kraje Vysočina nazvaný NIX-ZD, který je financovaný z programu CEF Telecom [51].

Oproti tomu je v České republice využíván národní standard DASTA. Ten je využíván jak pro předávání zdravotnických informací mezi poskytovateli zdravotních služeb, tak pro hlášení resortu zdravotnictví prostřednictvím systému zdravotnických registrů [52].

2.2.4 Klasifikační systémy a číselníky

Mezinárodní standardy pro uchování a přenos zdravotnické dokumentace se shodují na využití klasifikačních systémů SNOMED CT, LOINC a MKN 10. Všechny tyto klasifikační systémy jsou indexovány v meta-klasifikačním systému Unified Medical Language System (UMLS) [39]. Podrobněji se využití klasifikačních systémů v českých lékařských zprávách věnovala Petra Přečková v disertační práci [53].

2.2.4.1 LOINC

LOINC (Logical Observation Identifiers Names and Codes) je jedním z nejvíce používaných klasifikačních systémů v medicíně. Jeho účelem je umožnit vyžádání a výměnu výsledků v klinické praxi. Jeho položky označují laboratorní a další klinická pozorování, obvykle ta, která je možné exaktně měřit.

Každá položka číselníku může být definována některými nebo všemi z následujících druhů položek:

1. předmět měření – to, co je zjišťováno – například sodík, antigen hepatitis C, hemoglobin a podobně,
2. měřená vlastnost – například hustota, teplota, …,
3. časování – např. zda jde o hodnotu v daném čase nebo například o průměr za 24 hodin,
4. typ vzorku – například moč, krev
5. typ škály – zda je měření kvantitativní (skutečné), ordinální, nominální (např. přítomnost konkrétních bakterií) či popisné (např. popis RTG snímku)
6. metoda použitá pro zjištění výsledku nebo další upřesnění nutné pro interpretaci (například měření výšky vleže / vstoje)

Číselné identifikátory LOINC jsou užívány v mnoha standardech, například v LIM modelech a zprávách HL7 verze 3. [34][35] LOINC je k dispozici ve formě textového souboru s hodnotami oddělenými tabulátorem. [54][52]

2.2.4.2 SNOMED CT

SNOMED CT je kombinací zkratkového slova SNOMED a zkratky CT, kde SNOMED znamená „Systematized Nomenclature of Medicine“ a CT znamená „Clinical Terms“. Jde o kombinaci původně dvou číselníků: SNOMED a Clinical Terms.

SNOMED CT obsahuje komplexní klinickou terminologii poskytující dostatečné vyjadřovací schopnosti pro zaznamenání klinických informací. Další důležitou vlastností SNOMED CT je jeho mapování do UMLS [39], které umožňuje automatické mapování do jiných klasifikačních systémů. Jde o jeden z nejrozsáhlejších klasifikačních systémů ve zdravotnictví. V současnosti SNOMED CT obsahuje více než 300 tisíc aktivních konceptů, které jsou navzájem provázány více než milionem vztahů.

Koncepty SNOMED CT jsou organizovány v hierarchiích s různými úrovněmi podrobnosti. Základní rozdělení termínů ve SNOMED CT je:

- klinický nález (clinical finding / disorder)
- procedura / intervence (procedure / intervention)
- pozorovatelná entita (observable entity)
- část těla (body structure)
- organizmus (organism)
- látka (substance)
- farmaceutický / biologický produkt (pharmaceutical/biologic product)
- vzorek (specimen)

- zvláštní koncept (special concept)
- hmotný objekt (physical object)
- síla (physical force)
- událost (event)
- prostředí či umístění (environment or geographical location)
- sociální kontext (social context)
- fáze nebo jiné škály (staging and scales)

Hlavní část údajů je ve SNOMED CT uložena v základních tabulkách (Core Tables). Ty mají tři části:

- koncepty (concepts)
- popisy (descriptions)
- vztahy (relationships)

SNOMED také obsahuje sadu mapování vlastních kódů přímo do dalších klasifikačních systémů jako je MKN 10. [55]

2.2.4.3 Mezinárodní klasifikace nemocí

Mezinárodní klasifikace nemocí (International Classification of Diseases) [55][55] je ve své desáté verzi vydávána každým rokem. Je vydávána i v češtině, proto pro její označení používám českou zkratku MKN. Lze ji definovat jako systém kategorií, kterým přísluší jednotlivé skupiny patologických stavů. Tento klasifikační systém vytvořila a udržuje Světová zdravotnická organizace. Prvotním účelem MKN je umožnit systematické

zaznamenávání, analýzu, interpretaci a srovnání údajů o nemocnosti a úmrtnosti v čase a v různých zemích.

Přes původní účel se MKN stalo mezinárodním standardem pro kódování diagnóz při péči o zdraví jednotlivce i v epidemiologii.

Z MKN (anglicky ICD) vycházejí další specializované klasifikační systémy, zejména

- ICD-O se zaměřením na onkologii,
- ICD-DA se zaměřením na zubní lékařství a stomatologii a
- ICD-NA se zaměřením na neurologii.

2.2.4.4 LÉKY (SÚKL)

Databáze léčiv a léčivých přípravků Státního ústavu pro kontrolu léčiv (SÚKL) je vytvářena a udržována převážně za účelem jejich registrace, sledování jejich distribuce a preskripce. Léky jsou rozděleny dle ATC skupin [56].

2.2.5 Shrnutí cílových standardů

Pro praktickou použitelnost zejména při přeshraniční péči, zejména s ohledem na využití klasifikačních systémů v epSOS, jsem jako cílová kódování zvolil SNOMED CT, LOINC a MKN 10 (Mezinárodní klasifikace nemocí verze 10). Strukturu a přenositelnost kódování využitého v epSOS diskutovali Estelrich, Chronaki, Cangioli a Melgara v [57]. Jen v případě léčiv a léčebných přípravků jsem zvolil národní databázi LÉKY vedenou Státním ústavem pro kontrolu léčiv (SÚKL), a to z důvodu jeho dobré dostupnosti a především znalosti v České republice užívaných léčiv ze strany spolupracujících lékařů.

3. Automatická lingvistická analýza textů lékařských zpráv

3.1 Podklady a metody

Pro svoji práci jsem měl k dispozici celkem 268 anonymizovaných lékařských zpráv ve formě volného textu. Tyto zprávy byly zbaveny identifikačních údajů a pocházely ze dvou ambulantních kardiologických pracovišť Městské nemocnice Čáslav z období let 2000 až 2004. Tyto zprávy byly získány s informovaným souhlasem pacientů.

Zpracování metodami přirozeného jazyka (natural language processing – NLP) zahrnuje několik základních fází. Jde o tyto fáze:

1. strukturální analýza (izolace jednotlivých částí zpráv),
2. lexikální analýza (identifikace zpracovávaného slova a zjištění jeho významů),
3. slovní rozbor (sestavení variant hierarchií v rámci věty).

V první části výzkumu možnosti extrahovat informace z lékařských zpráv jsem se soustředil na lexikální analýzu zpráv. V jazyce Java SE jsem postupně vytvářel nástroje pro jejich zpracování.

3.1.1 Příprava volného textu

Volný text je nejprve zapotřebí připravit pro další zpracování. Základní metodou je tokenizace vstupního textu, tedy jeho rozdělení do řetězce strojově oddělených součástí. Každá taková součást obvykle sestává z posloupnosti znaků, jednotlivé součásti (tokeny) jsou odděleny tzv. stop-znaky, kterými bývají mezery, pevné mezery, tabulátory či konce řádků.

Tokenizace může být doplněna o vyhledávání pokročilejších vzorů, například odlišením numerických a alfanumerických tokenů či automatizovaným označováním kombinací tokenů dle stanovených pravidel. Takové označování lze zařadit jak do oblasti prosté přípravy volného textu v případě generických pravidel jako je např. vyhledávání vzorů, např. „lomítky oddělená čísla“ či do oblasti extrakce informací, například v případě vyhledávání údajů o krevním tlaku, kde jde o dvě lomítky oddělená čísla se splněním dalších parametrů, tedy především rozsahu hodnot a vztahu mezi hodnotami (např. první číslo je vyšší než druhé).

Na tokenizaci často navazuje stemizace či lemmatizace, tedy identifikace jednotlivých tokenů se základními pojmy (od toho název „stemizace“ – nalezení kmene slova). Tím dojde k automatickému anotování (tagování) jednotlivých slov, které však nemusí být jednoznačné; k jednomu tokenu může být přiřazen i více než jeden možný kořen slova. Tato část již přesahuje do oblasti extrakce informací.

Stemizace a lemmatizace se liší přístupem. Zatímco u stemizace se zpracovává token tak, že se identifikací koncovek, přípon a předpon zjišťuje kmen (kořen) slova. V případě lemmatizace se vychází ze základního tvaru, ze kterého jsou automaticky generovány všechny jeho tvary a ty jsou následně identifikovány s tokenem.

3.1.2 Provedení

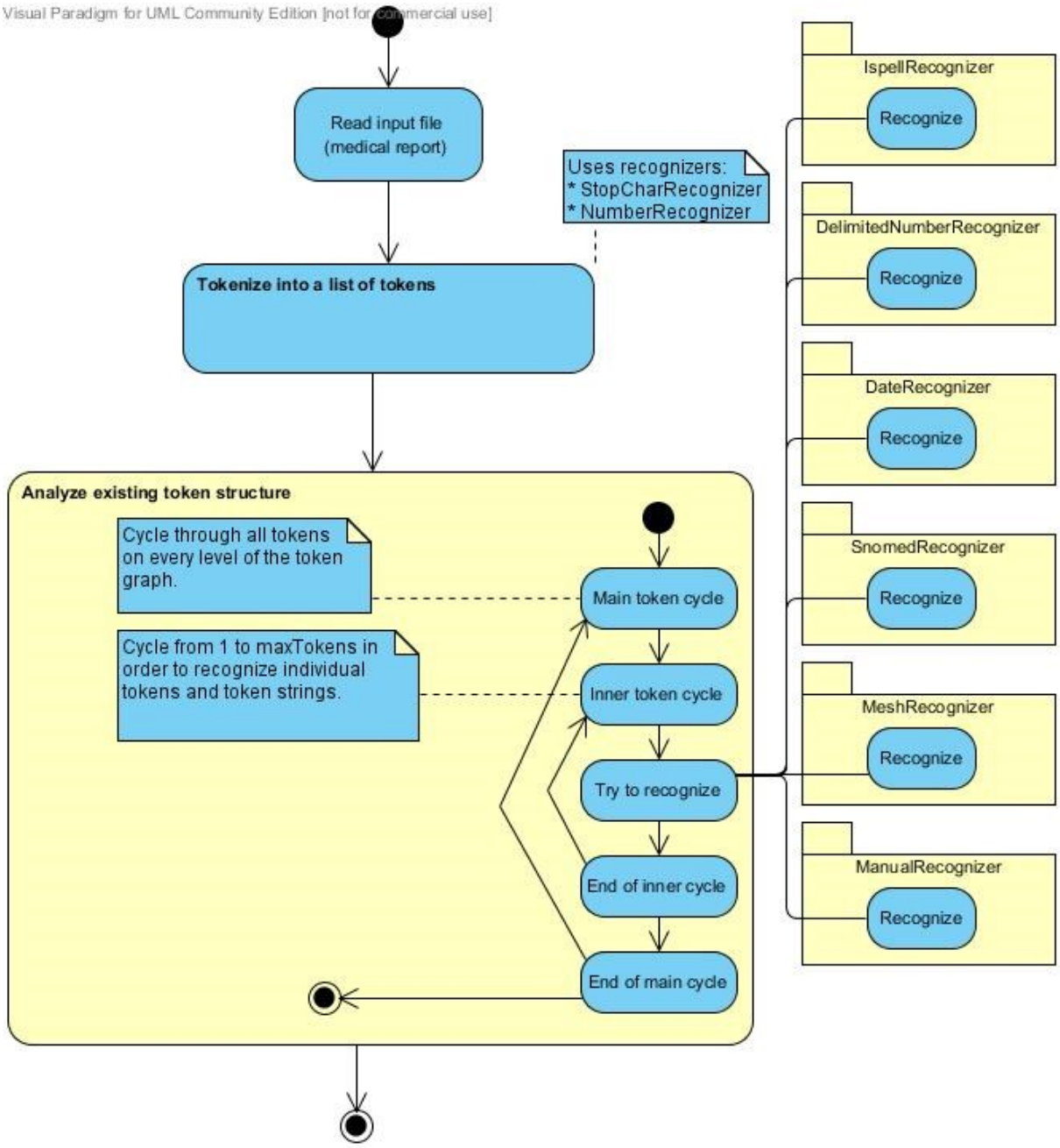
Zprávy byly nejprve automaticky tokenizovány do řetězce tokenů, přičemž obecný alfanumerický token byl označen jako *Container*, obsah s jen numerickým obsahem *Number* a token představující stop-znak *SpecialChar*. V případě tokenů typu *SpecialChar* byla zároveň zaznamenána informace o počtu stejných znaků následujících za sebou. Obrázek 1 uvádí ukázku tokenů reprezentujících informaci o krevním tlaku.

Container	SpecialChar	Container	SpecialChar	Number	SpecialChar	Number
„Krevní“	„ “ (1x)	„tlak“	„ “ (1x)	„145“	„/“ (1x)	„90“

Obrázek 1: Ukázka tokenů reprezentujících informaci o krevní tlaku

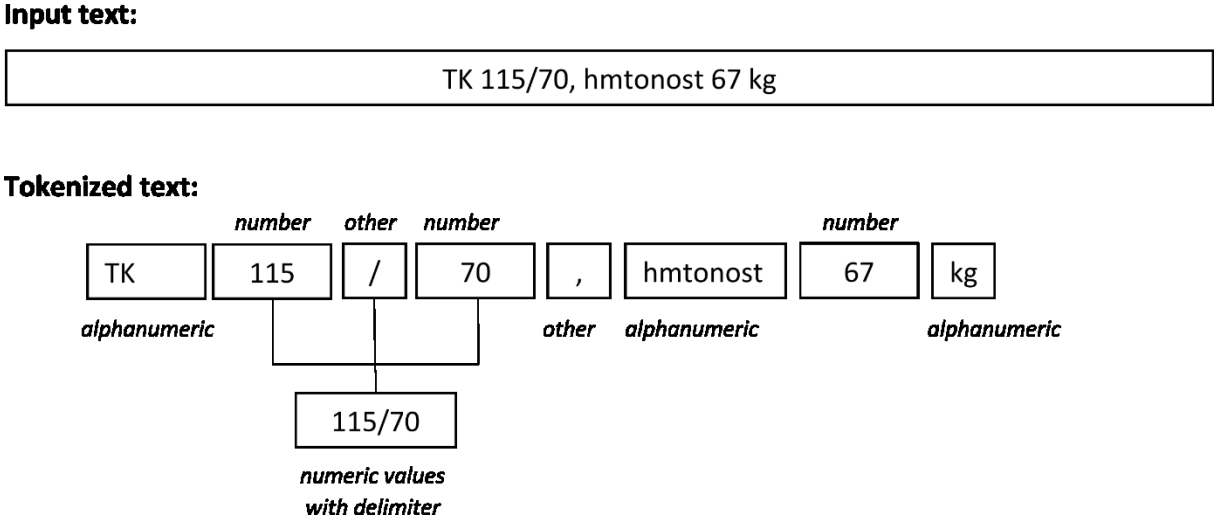
Řetězec tokenů byl následně zpracováván dalšími automatickými nástroji, z nichž některé prováděly sloučení do tokenů vyšší úrovně rozpoznáním s využitím regulárních výrazů. Tímto způsobem byly automaticky do tokenů vyšší úrovně sjednoceny podřetězce tokenů představující čísla oddělená definovaným stop-znakem či datum. Dále jsem vytvořil rozpoznávací nástroje, které obsah tokenů vyhledávaly v databázi MeSH [58], v klasifikačním systému SNOMED CT [59] a v databázi využívající slovník iSpell pro češtinu [60]. Nalezené termíny byly seskupeny v tokenech označených *MedicalTerm*, česká slova nalezená s využitím databáze založené na databázi iSpell byla převedena na tokeny typu *Dictionary*.

Obsah tokenů byl porovnáván s hodnotami položek jednotlivých klasifikačních systémů. V případě, že hodnota nebyla nalezena, ale tokeny obsahovaly stop-znak tečku („.“), došlo k vyhledávání termínů, které se shodovaly v části textu před tečkou. Tečka tedy představovala alespoň jeden libovolný znak slova (ekvivalent označení „.+“ u regulárního výrazu). Obrázek 2 zobrazuje postup rozpoznávání tokenů.



Obrázek 2: Postup rozpoznávání tokenů

Obrázek 3 ukazuje příklad vstupního textu v českém jazyce, jeho tokenizaci a automatickou identifikaci hodnot. Tokeny představující mezery jsou vynechány pro větší přehlednost obrázku.



Obrázek 3: Tokenizace narativní klinické zprávy

Vidíme, že tokeny vyjádřené alfanumerickými znaky mohou být slova použitého cílového jazyka, slova s zapsaná chybně (např. "hmtonost"), zkratky nebo zkratky (např. TK), čísla a jiné znaky (např. "/").

3.2 Výsledky

Výsledky mé snahy o zpracování lékařských zpráv pomocí nástrojů zpracování přirozeného textu (NLP) jsem publikoval především v [61] a v [62]

Rozpoznávací mechanismus pro vyhledání kombinací odpovídajících krevnímu tlaku nalezl celkem 434 případů, přičemž minimum počtu nalezených kombinací ve zprávě bylo 0, nejvyšší počet rozpoznaných zápisů krevního tlaku bylo 12.

Pomocí slovníku iSpell [60] a s využitím nových pravidel pro odvozování tvarů slov pomocí ohýbání (lemmatizace) se podařilo identifikovat počty slov jednotlivých slovních druhů v počtech, které jsou uvedené v tabulce 1. Jak je vidět, jako slovo českého jazyka podařilo identifikovat celkem jen 47,7 % tokenů.

Tabulka 1: Identifikace počtu slov jednotlivých slovních druhů v počtech anotací

	Počet anotací (průměr/zpráva)	**Anotací / slov celkem**
Podstatné jméno	75	30,32 %
Přídavné jméno	23	9,3 %
Zájmeno	0	0 %
Číslovka	0	0 %
Sloveso	17	6,87 %
Příslovce	3	1,21 %
Předložka	0	0 %
Spojka	0	0 %
Částice	0	0 %
Citoslovce	0	0 %
CELKEM	**118**	**47,7 %**

Číselníkové termíny se podařilo nalézt v celkem 107 zprávách. Těchto 107 zpráv bylo automatizovaně rozděleno na celkem 66 376 tokenů. Až na výjimky se podařilo rozpoznat

jen termíny z české verze bibliografického klasifikačního systému MeSH [58]. Termíny ze SNOMED CT se nedařilo rozpoznat, neboť nebyla k dispozici česká verze SNOMED CT. Tabulka 2 uvádí počty rozpoznání nejčastěji rozpoznaných deskriptorů MeSH.

Tabulka 2: Počty rozpoznání nejčastěji rozpoznaných deskriptorů MeSH

Kód	Deskriptor	Počet
H01.671.691	tlak	63
A07.541	srdce	62
C14.907.489	hypertenze	39
A01.047	břicho	39
C14.280.067	arytmie	37
A01.456	hlava	34
C23.550.288	nemoc	33
A01.598	krk	30
D04.808.247.222.284	cholesterol	29
E01.370.370.380.650	puls	28

Snaha o rozpoznání položek z číselníku mezinárodní klasifikace nemocí MKN 10 [55] nevedla k velkému počtu rozpoznání. Byl využit číselník MKN 10 publikovaný Ústavem zdravotnických informací a statistiky (ÚZIS). Tento číselník však obsahoval velké množství zkrácených slov a stejné významy popisuje různými způsoby („diabetes mellitus“ vs „DM“). Samotná česká verze číselníku vykazuje podobné vlastnosti jako narativní lékařské texty.

3.3 Diskuse

Narativní lékařské zprávy, které jsem měl pro účely výzkumu k dispozici, obsahovaly velké množství zkratek a zkrácených slov. Výsledky mého výzkumu tím jen potvrzují předchozí výsledky, kterých dosáhli Petra Přečková a další v [63]. Podařilo se rozpoznat jen malou část lékařských termínů a to téměř výhradně jen s využitím bibliografického, nikoliv klinického, klasifikačního systému. S pomocí techniky lemmatizace se navíc podařilo jako česká slova identifikovat méně než polovinu alfanumerických tokenů.

Ukázalo se, že narativní české lékařské zprávy nejsou typickým českým narativním textem. Jsou textem velmi specifickým, který je z velké části složen z odborných termínů a vyznačuje se intenzivním využitím zkratek, zkrácených slov a absencí explicitního vyjádření větných členů.

Výsledek této části výzkumu potvrzuje domněnky Jiřího Semeckého [1] a Petera Smatany [2], že lingvistická analýza nemůže být úspěšná.

4. Elektronický zdravotní záznam pro orální medicínu s interaktivní komponentou zubního kříže

V roce 2004 byl v Evropském centru pro medicínskou informatiku, statistiku a epidemiologii (EuroMISE centrum) dokončen vývoj prototypu elektronického zdravotního záznamu (EHR), který byl označen jako Multimedia Universal Distributed Electronic Health Record (MUDR). Ten poskytuje způsob pro ukládání strukturovaných dat, který je založený na ontologiích zdravotnických disciplín a umožňuje též připojit samostatný narativní záznam. MUDR není statickým, ale dynamickým modelem, umožňuje tedy rozšířit či modifikovat jeho vlastnosti bez nutnosti zasáhnout do struktury databáze. Podrobné informace lze nalézt například v disertační práci Miroslava Nagyho [8].

Kromě dalších aplikací byl v projektech aplikovaného výzkumu EuroMISE centra vyvinut softwarový nástroj MUDRLite. Ten se skládal z několika součástí. Jednou takovou součástí byl MUDRLite interpreter. Ten vytváří uživatelské rozhraní definované pomocí jazyka MLL a poskytuje rozhraní pro připojení dalších uživatelských grafických komponentů. Komponenta zubního kříže využívala tohoto rozhraní pro propojení s MUDRLite. Tuto technologii komponenty zubního kříže se Ústav informatiky AV ČR pokusil též patentovat. Touto technologií je model ontologie základních zubních struktur člověka, kterým lze popsat všechny situace a neztratit žádnou podstatnou informaci pro obor zubního lékařství. Tato ontologie je pohledem na klinickou informaci o pacientově chrupu v čase a vychází z klasifikace chrupu, která identifikuje zuby pomocí jejich pořadí v kvadrantech chrupu.

4.1 Metody

Stomatologický elektronický zdravotní záznam zachycuje údaje o pacientovi: Jeho osobní anamnézu (historii a předpoklady), provedená vyšetření, léčební úkony i preskripci. Zároveň je podmínkou pro další zpracování těchto údajů, například pro podporu rozhodování lékaře při stanovování diagnózy onemocnění orofaciální soustavy [64].

Jednou z možností je založit stomatologický elektronický zdravotní záznam na interaktivní komponentě zubního kříže, neboť takové zobrazení je názorné a blízké klinické praxi.

Datový model nové implementace byl založen na datovém modelu MUDRLite, avšak nově bylo využito UML modelování včetně dědičnosti tříd.

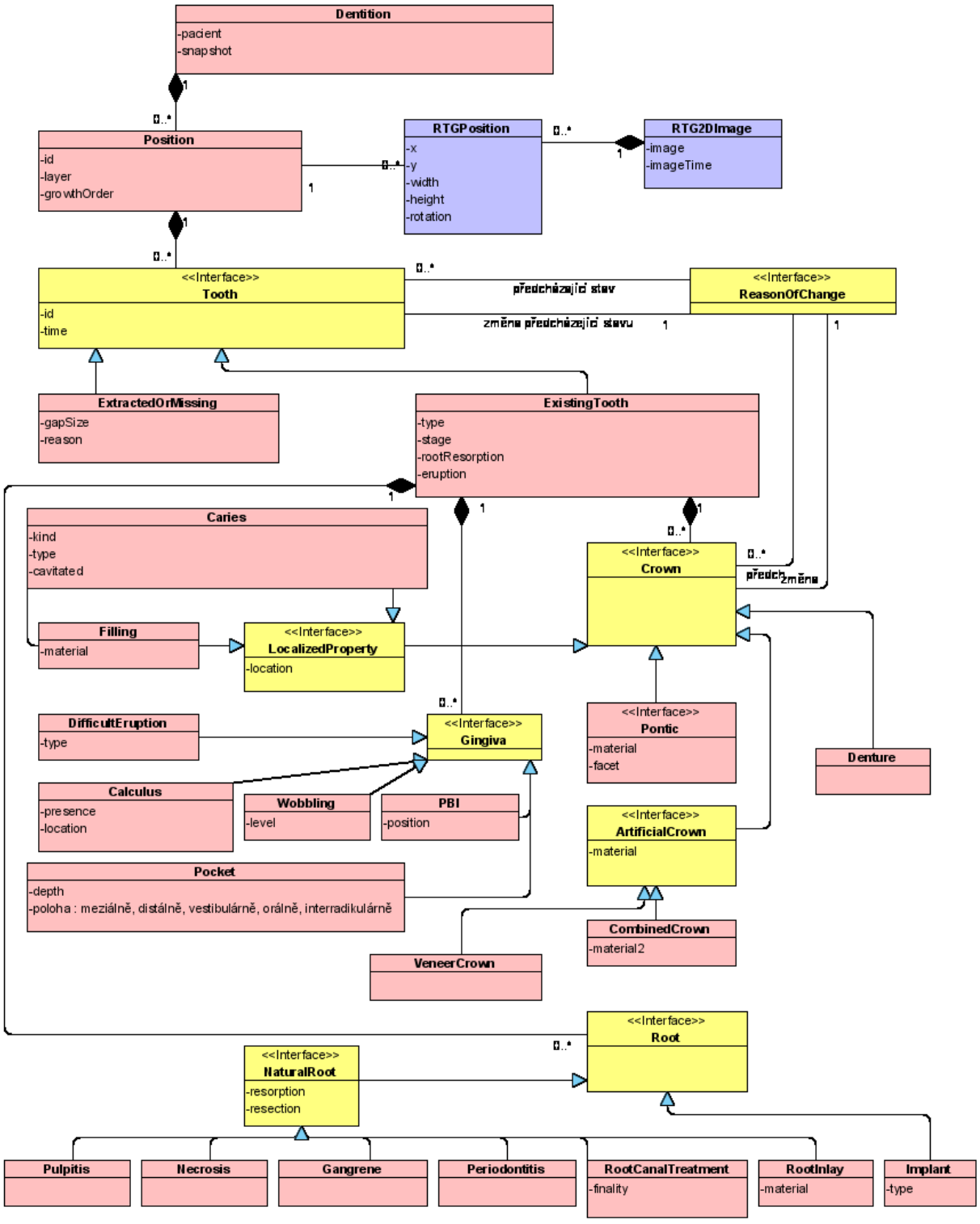
4.2 Výsledky

Interaktivní software s komponentou zubního kříže v první verzi vznikl v rámci projektu Informační technologie pro rozvoj kontinuální sdílené péče o zdraví 1ET200300413 grantové agentury Akademie věd ČR. Tento software podporoval zobrazení pouze stálého chrupu a byl založen na výše popsaném modelu ontologie. Software byl vytvořen pro prostředí Microsoft Windows s .NET Framework jako samostatní knihovna DentCross.dll. Systém byl vyvinut pomocí Microsoft Visual Studie .NET 2003. Komponenta podporovala souběžné ukládání dvourozměrných RTG snímků a fotodokumentace. Zubní lékař mohl využít cca 60 různých druhů zapsání (a zobrazení) informací o vyšetření či ošetření. Software podporoval také stanovení léčebného plánu včetně plánu návštěv a dokázal zobrazit jednotlivé kroky ošetření znázorněním stavu chrupu. Tento model i software byly použity také ve forenzní stomatologii [65].

4. Elektronický zdravotní záznam pro orální medicínu s interaktivní komponentou zubního kříže

V rámci výzkumu jsem navrhl metodu a vytvořil softwarové řešení zahrnující práci s interaktivní komponentou zubního kříže Lifetime DentCross pro celoživotní elektronický zdravotní záznam ve stomatologii, který zahrnuje možnost vkládání údajů nejen pro stálý, ale i pro smíšený a dočasný chrup. Pro tyto situace byl rozšířen i model ontologie stomatology 2. lékařské fakulty Univerzity Karlovy v Praze a Fakultní nemocnice v Motole a je popsán v disertační práci K. Chleboráda [66].

Na základě navržené metody jsem vytvořil datový model pro uložení dat v objektově-relační přístup k databázi, uživatelské rozhraní pro stomatologii a novou softwarovou verzi interaktivní komponenty Lifetime Dent Cross. Použitý datový model je znázorněn na obrázku 4.



Obrázek 4: Objektový model použitý v software Lifetime Dental Cross

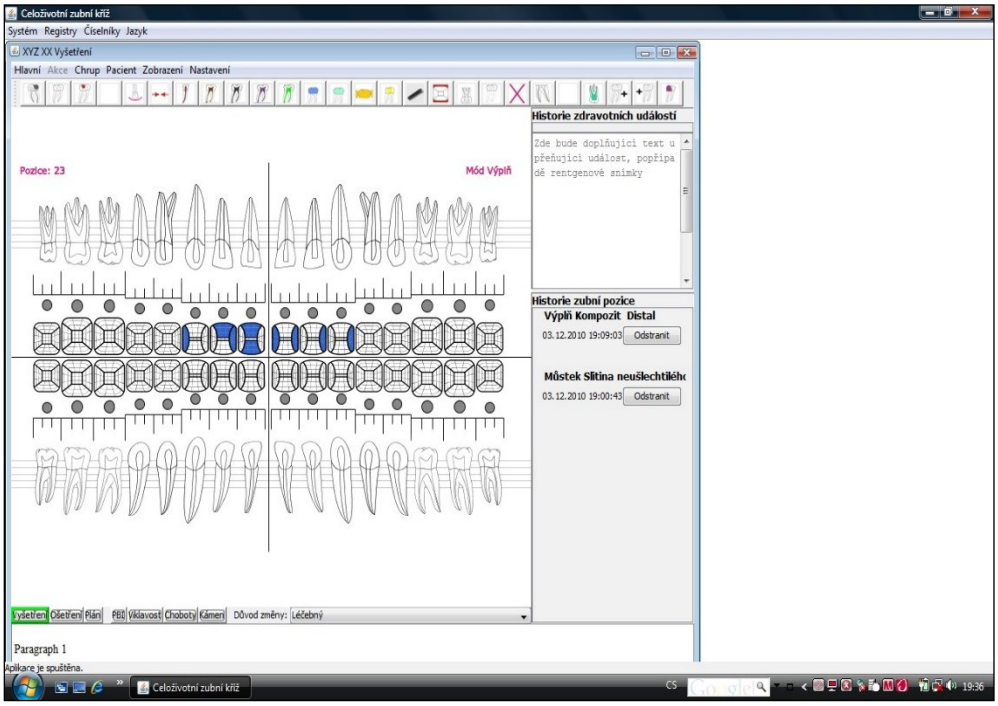
Základní třídou modelu je *Dentice* (chrup), který obsahuje jednotlivé zubní pozice. Tyto zubní pozice představuje obecná třída *Tooth* (Zub), ze které děděním vycházejí třídy *ExtractedOrMissing* (chybějící či extrahovaný zub) a *ExistingTooth* (existující zub).

Třída *ExistingTooth* slouží pro připojení objektu tříd *Gingiva* (dáseň), *Root* (kořen) a *Crown* (korunka), přičemž navázaných objektů všech typů může být více, než jedna. Tyto tři třídy (*Gingiva, Root* a *Crown*) představují jednotlivé vlastnosti příslušných součástí zubního aparátu, a to jak přirozených (např. údaj o zubním kameni u dásně, údaj o snekróze kořene či údaj o stavu jednotlivých plošek u korunky), tak u způsobených ošetřením (implantát, umělá korunka, výplň).

Každá zubní pozice je popsána pomocí základních anatomických struktur, jak je uvedeno výše v popisu datového modelu. Těmito strukturami jsou korunka, kořen a závěsný aparát zubu. Základním prvkem uživatelského rozhraní je graficky zobrazený zubní kříž (obrázek 5). Zobrazení obsahuje jak základní administrativní údaje o pacientovi, jako jsou jméno, příjmení a rodné číslo, tak nástroje pro zapisování údajů do zubního kříže. Tyto nástroje byly navrženy a seřazeny po konzultaci se zubními lékaři a obsahují také přechody do dalších režimů umožňujících například zapsání údajů o výsledcích parodontologického vyšetření, například o přítomnosti zubního kamene, o hloubce parodontálních chobotů, viklavosti zubů a stavu dásní (pomocí PBI – papila bleeding index). Pravá část okna obsahuje historii ošetření dané zubní pozice a legendu. Legenda zobrazuje především barevné kódování materiálů, které uživateli usnadňuje rychlou orientaci v záznamu. Jednotlivé zubní pozice jsou označené dvouciferným číslem podle kvadrantů.

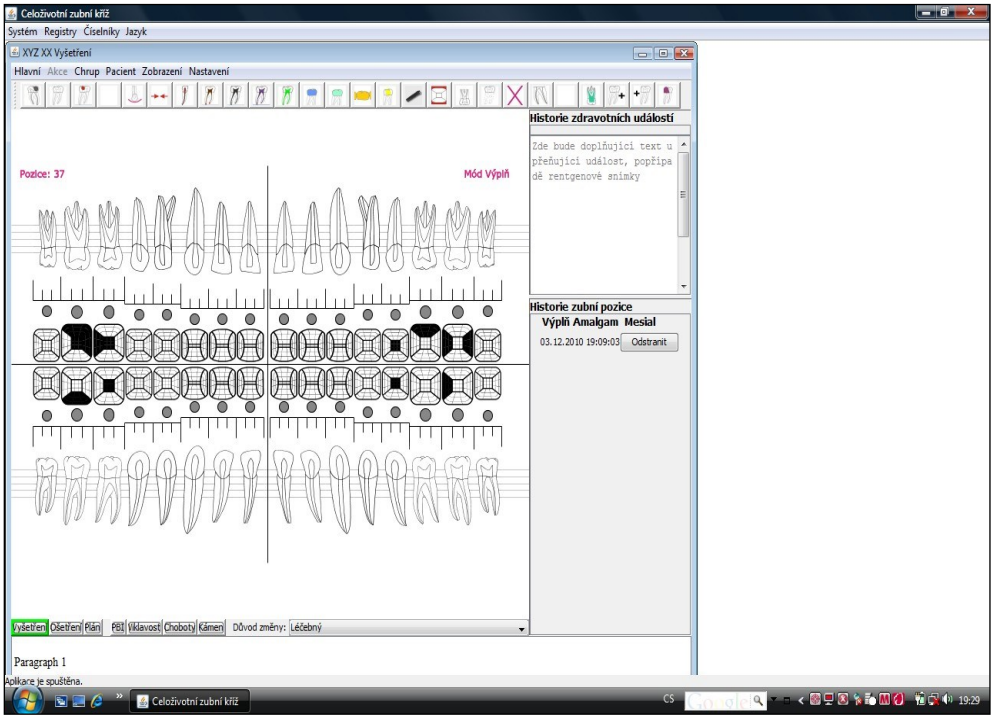
Databázová struktura implementace obsahuje také informaci o pacientech a umožňuje ke každému ošetření vložit textovou část zprávy. Na obrázku 4 je zobrazena jen část modelu zachycující strukturovanou informaci o stavu chrupu, která umožňuje také zachycení změny

v čase propojením objekty třídy *ReasonOfChange* (ten v obrázku modelu již není rozveden na třídy, které tento interface dědí).



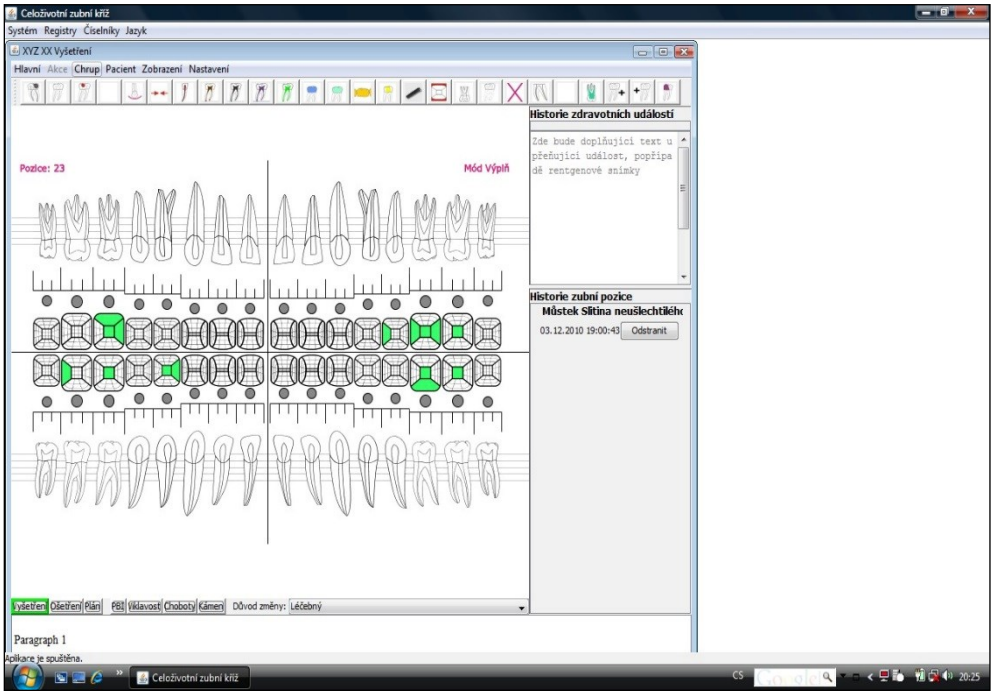
Obrázek 5: Fotokopozitní výplň

Tato implementace umožňuje větší podrobnost pro zapsání pozice kazivé léze či výplně. Korunky zubů jsou rozdělené na 7 polí pro lokalizaci, která jsou označena podle anatomických zvyklostí: M (mesiální), D (distální), O (okluzální), I (incizální), R (orální), V (vestibulární) a C (cervikální). Rozsah kazivé léze lze zaznamenat dle Mountovy klasifikace kazivých defektů (číslem od 1 do 4). Výplně jsou označeny barevným kódováním dle klíče: fotokompozit – modrá (obrázek 5), amalgám – černá (obrázek 6) a skoionomer – zelená (obrázek 7). Díky rozdělení plošek zubu lze souběžně zapsat ošetření jednoho zubu různými materiály či souběžnou přítomnost kazivé léze i výplně na jednom zubu (např. meziokluzní výplň, obrázek 8) či kaz v dočasném chrupu (obrázek 9).

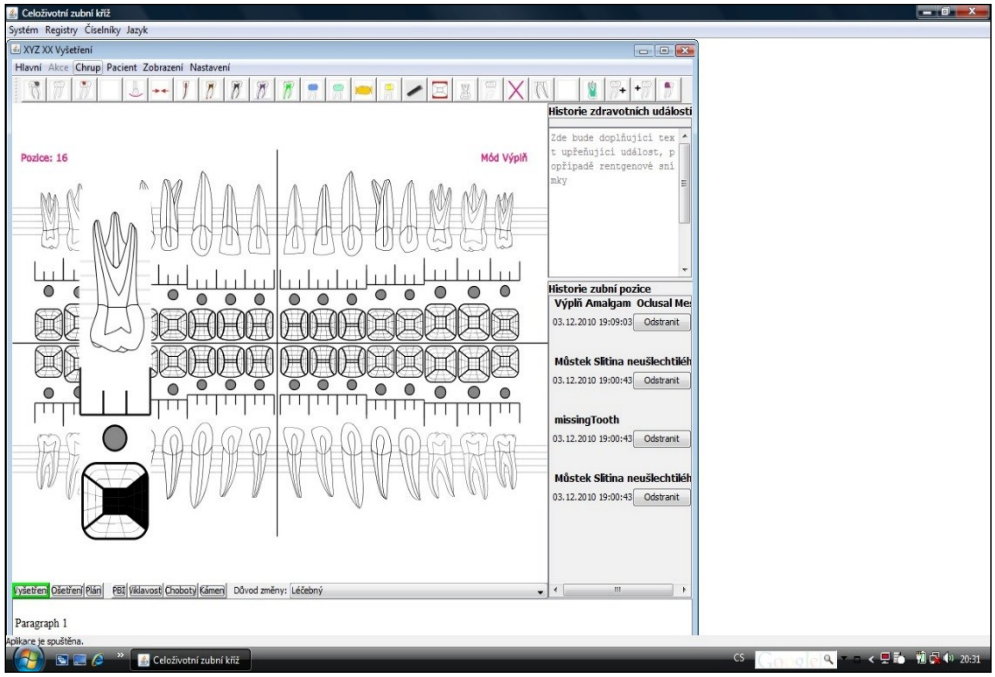


Obrázek 6: Amalgámová výplň

Při zaznamenání ošetření jednotlivých zubních plošek objektu typu *ExistingTooth* přísluší více objektů dědících třídu *LocalizedProperty*, kde každý obsahuje informaci o stavu plošky.

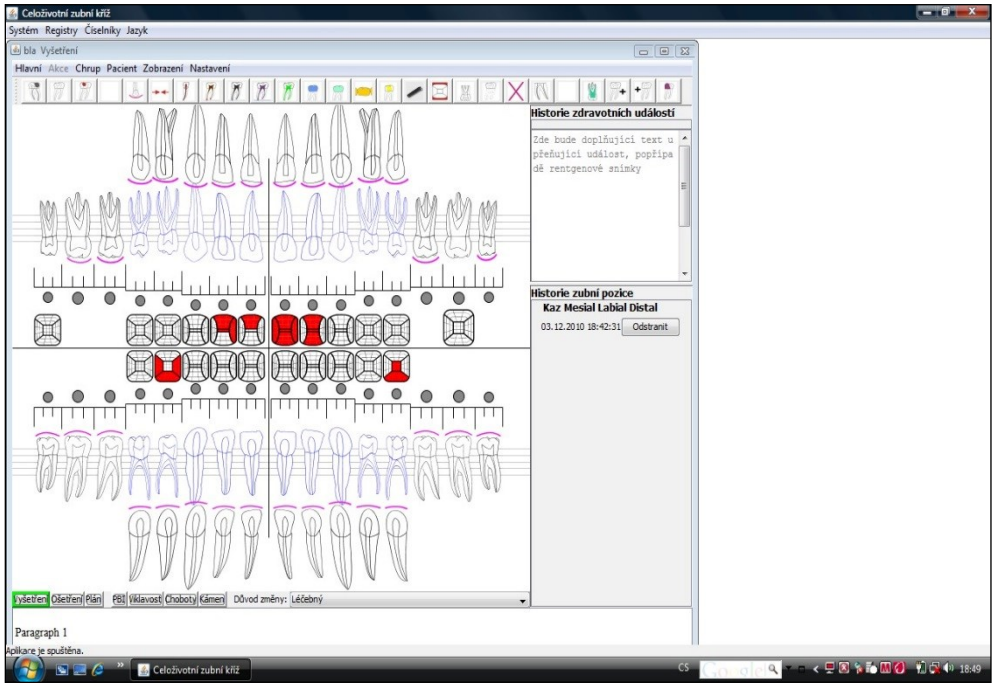


Obrázek 7: Výplň skloionomerem



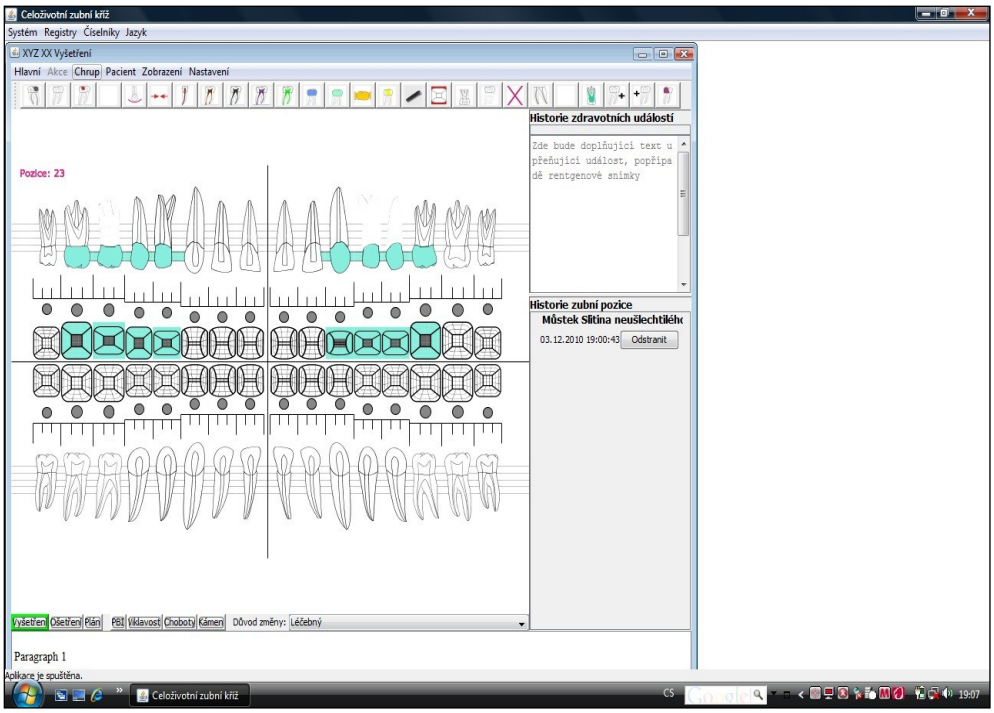
Obrázek 8: Meziokluzní výplň

Kombinaci korunkových náhrad a mezičlenů můstků lze sestavit z nabízených komponent a současně označit i materiálové složení.

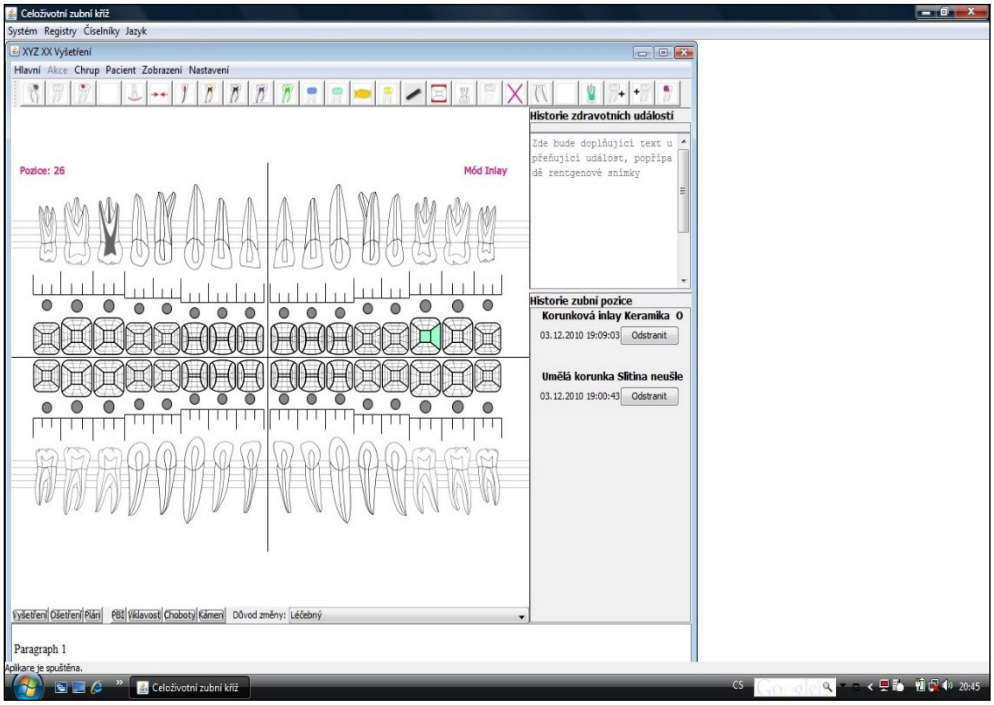


Obrázek 9: Zubní kaz v dočasném chrupu

Z korunek lze vybrat plášťové, fazetované či kombinované, totéž platí u mezičlenů (obrázek 10). Software umožňuje zapsat také speciální náhrady částí zubu, například fazety, polokorunky, korunkové inleje, onleje a overleje včetně kořenových inlejí (obrázek 11). Model i software umožňují také zachycení zvláštních stavů, například nepřítomnost zubu s uzavřením jemu příslušející mezery (obrázek 12) nebo zaznamenání nadpočetného zubu. Lze také zachytit informaci o snímatelných náhradách a protetických ošetření.

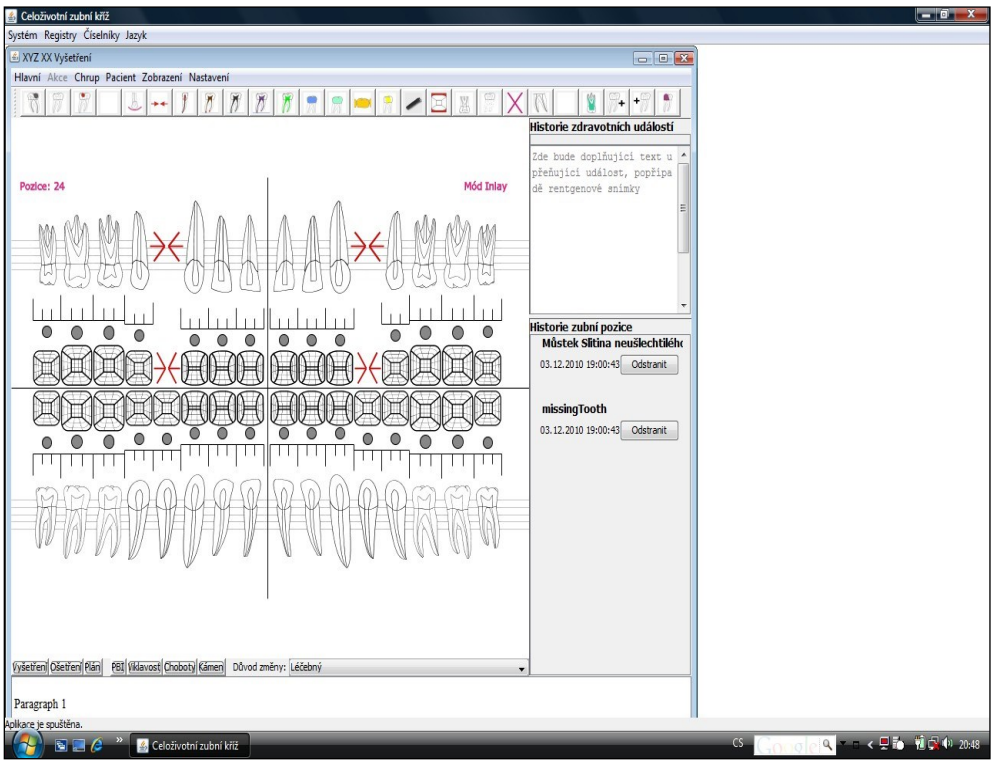


Obrázek 10: Fazetované korunky můstku včetně mezičlenů



Obrázek 11: Kořenová inlej

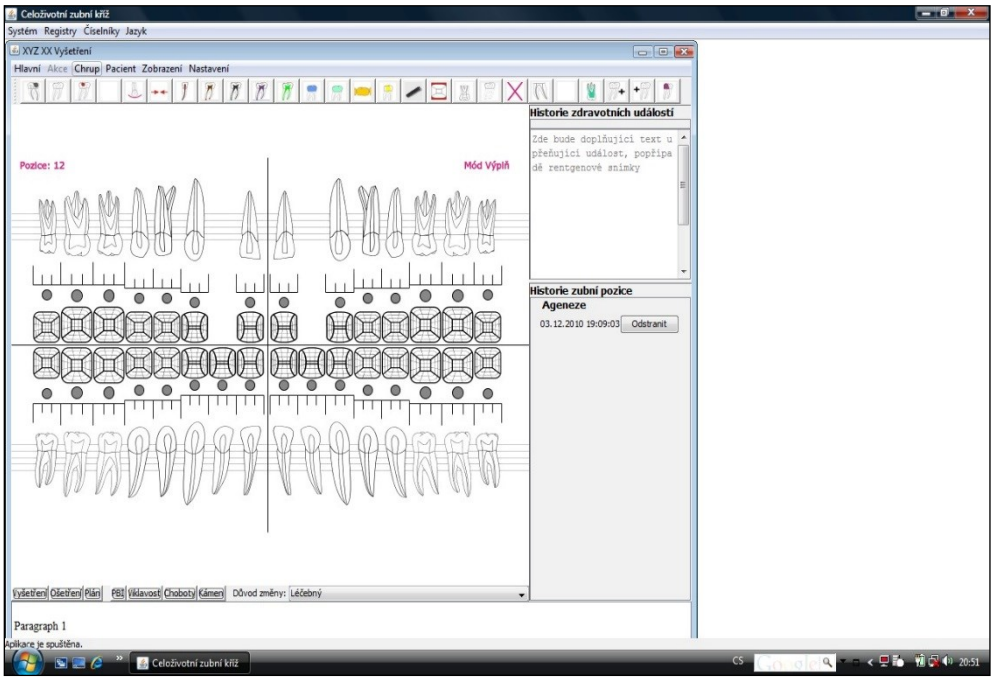
V případě přítomnosti kořenové inleje je k objektu třídy *ExistingTooth* připojen objekt *RootInlay*, který dědí obecnější třídu *Root*. Představuje vlastnosti kořene zubu.



Obrázek 12: Chybějící mezera

4. Elektronický zdravotní záznam pro orální medicínu s interaktivní komponentou zubního kříže

Rozvoj a rozšiřování využívání ortodoncie vede k tomu, že zvláštní situace je zapotřebí zachycovat stále častěji. Viz např. uzávěr mezery po extrakci a andodoncii (obrázek 13).

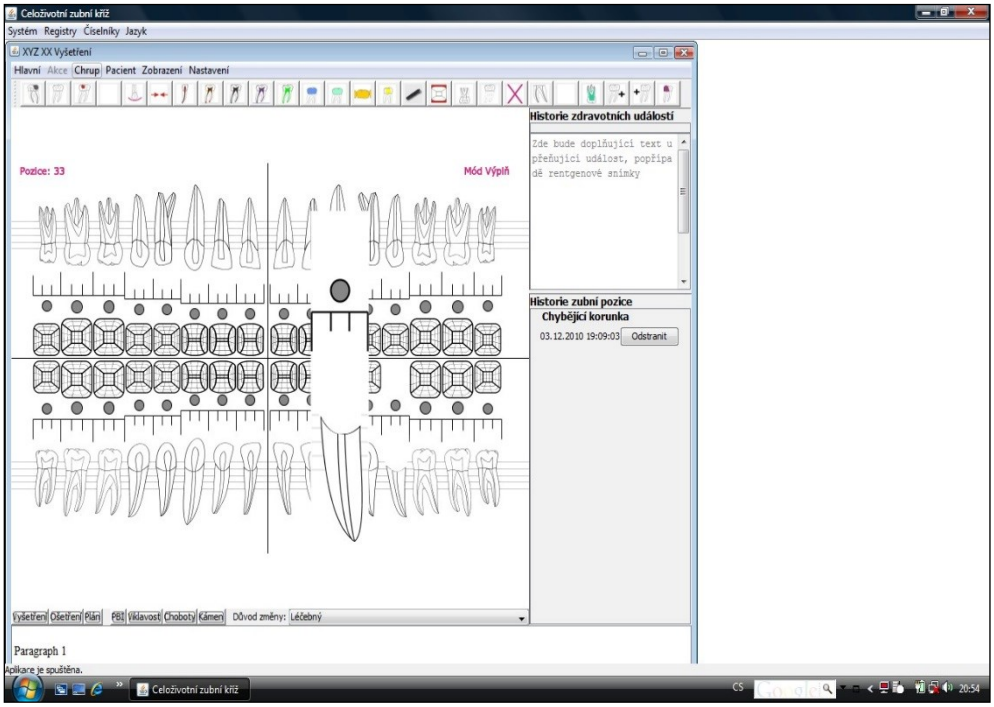


Obrázek 13: Andodoncie zubů na pozicích 12 a 22

Stále častější jsou také implantáty. U implantátů se rozlišují dva základní typy: implantát s fixní nástavbou a implantát pro snímatelnou náhradu.

Údaje se načítají pro každý zub zvlášť, systém zároveň vede informaci o datu a času zápisu a tak lze procházet historii jednotlivých zubních pozic i chrupu jako celku.

Z pohledu stomato-chirurgického ošetření zubní kříž umožňuje zobrazit také prořezávání zubů, jejich zánětlivé komplikace (až po nekrózu) či například ztrátu korunkové části zubu a zobrazení jen zbývajícího kořene (obrázek 14).



Obrázek 14: Samostatný kořen zubu

I když je obor zubního lékařství velmi vhodný pro zapisování strukturované informace a právě strukturalizace v co největším rozsahu byla jedním z hlavních cílů výzkumu, ukázalo se, že není možné opomenout možnost, aby ošetřující lékař do dokumentace vložil svůj volně formulovaný text. Strukturovaná informace je sice nezbytná pro jakékoliv hromadné či automatizované zpracování, nemůže však nahradit expresivní možnosti, které má lékař ve volném narativním textu. Lékař totiž musí mít možnost zapsat informace o podkladech i nejistotě, které vedly k jeho rozhodnutí, o instrukcích, které předal pacientovi a podobně. Zdravotnická dokumentace totiž slouží nejen pro další péči o pacientovo zdraví, ale také jako podklad pro vyúčtování s institucemi zdravotního pojištění a jako důkaz při řízeních v rámci profesních organizací, v rámci státního dohledu nebo před soudem.

Software je možné využít také v rámci forenzní stomatologie. Klasickou úlohou forenzní stomatologie je identifikace zemřelých. Takové rozpoznávání se provádí tehdy, je-li nalezené tělo poškozeno natolik, že nelze provést vizuální identifikaci. Tradičními druhy

událostí, kdy se ve velkém měřítku využívá identifikace podle zubního záznamu, jsou hromadné katastrofy, jako jsou hromadné nehody, letecké katastrofy, války či přírodní katastrofy jakými byly vlny cunami v oblasti Indonésie či v Japonsku.

Software Lifetime DentCross podporuje práci v různých jazycích. V současné době podporuje český jazyk, anglický jazyk, německý jazyk a španělský jazyk. Další jazyk lze přidat přeložením databáze obsahující cca 300 výrazů.

Součástí řešení bylo také zabudované hlasové ovládání. Výzkumný tým výsledky publikoval v [67].

4.3 Diskuse

Model, který byl původně vytvořen pro systém MUDRLite, jsem rozšířil pomocí UML do struktury podporující dědičnost. To umožnilo zjednodušení a zpřehlednění datové struktury i v samotném kódu programu.

Podílel jsem se také na softwarové implementaci nové implementace v programovacím jazyku Java. Využití hlasového ovládání ukázalo, že pro klinickou praxi je důležité nejen zapsání samotných údajů tak, aby bylo možné je efektivně následně využít, ale důležitý je také způsob jejich zadání do systému.

Za důležitostí způsobu zadávání údajů stojí skutečnost, že ošetřující lékař během vyšetřování a ošetřování pacienta přichází přes rukavice do kontaktu s ústní dutinou pacienta. Během vyšetřování či ošetřování proto lékař nemůže své ruce využívat pro interakci s výpočetní technikou, aniž by si po takovém jejich znečištění musel vyměnit. Bezkontaktní hlasové zadávání údajů do elektronického zdravotního záznamu umožňuje lékaři pořizovat záznam

bez nutnosti měnit rukavice nebo využít asistenci další osoby. Tím může dojít k úspoře nákladů.

Implementace byla úspěšně otestována ve Fakultní nemocnici v Motole.

5. Třífázová metoda předběžného zpracování

5.1 Podklady a metody

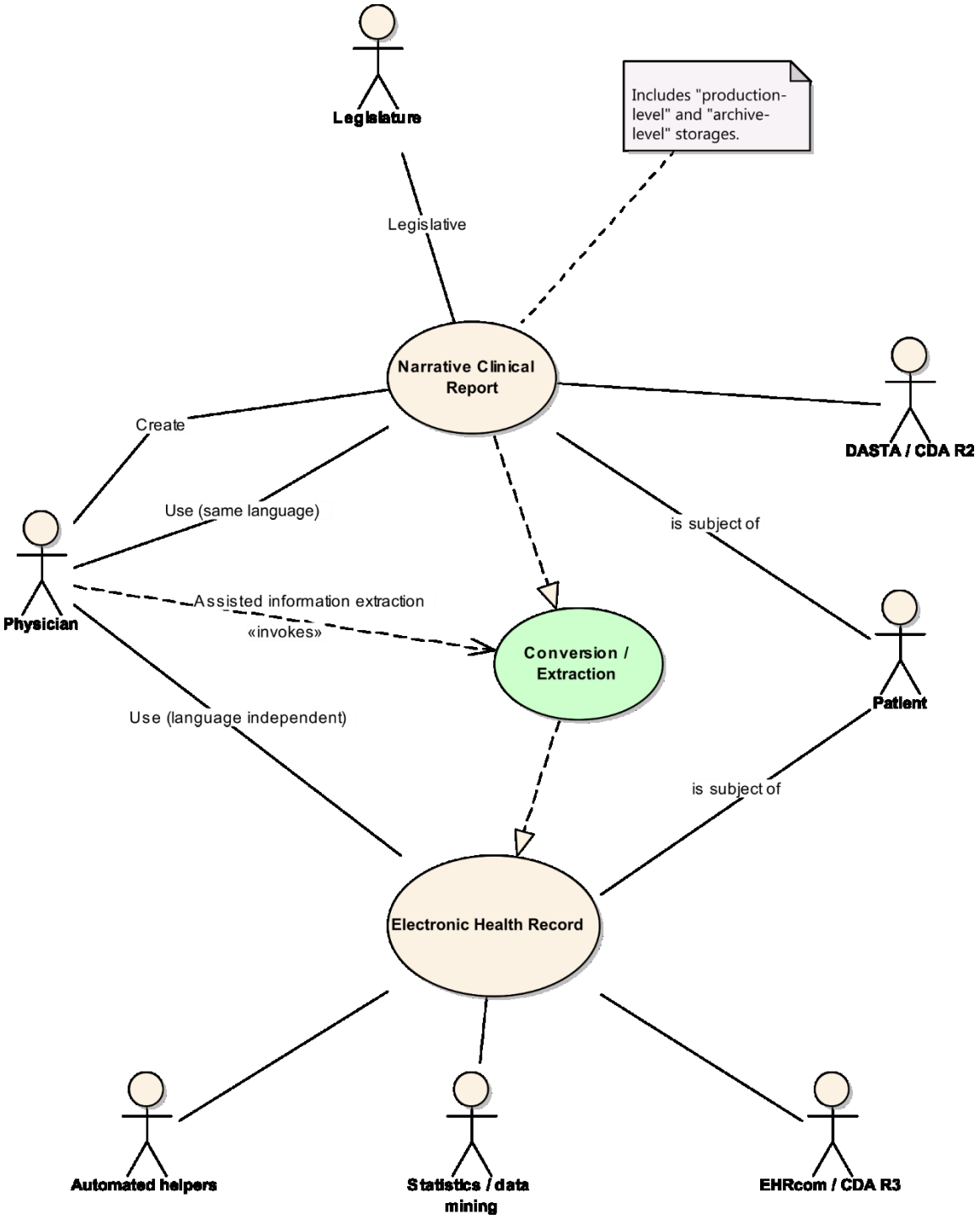
Pro tuto část výzkumu byly využity stejné podkladové narativní lékařské zprávy jako pro výzkum popsaný v kapitole 3. Z těchto zpráv bylo použito jen prvních 49, a to z důvodu omezené kapacity spolupracujících lékařů.

Navrhl jsem třífázovou metodu předběžného zpracování (metoda 3PP) narativních klinických zpráv pro další automatizované zpracování, přičemž dalším zpracováním míním strojový překlad a další analýzu pro extrakci strukturovaných dat. Způsob zapsání strukturovaných informací z narativní klinické zprávy do elektronického zdravotního záznamu (EZZ) je znázorněna na obrázku č. 15 a je dlouhodobým cílem výzkumu. Současné použití a hodnocení metody se soustřeďuje jen na předzpracování narativních zpráv, které by umožnilo strojový překlad.

Výsledky této metody jsem s ostatními spoluautory publikoval v [68].

Navrhovaná metoda má tři hlavní fáze:

1. **Tokenizace** zajišťuje segmentaci textu do slov, frází, symbolů nebo jiných významných prvků nazývaných tokeny. Seznam tokenů se stává vstupem pro další zpracování.
2. **Normalizace** se skládá z korekce pravopisu a z rozvinutí zkratek a akronymů.
3. **Sémantická anotace** poskytuje vazby na strojově čitelné nomenklatury. Dále se předpokládá, že metoda bude použita v souvislosti s automatizovaným strojovým překladem (vytvořením předzpracované zprávy), který se bude opírat o volně dostupné nástroje pro strojový překlad.



Obrázek 15: Schéma získávání informací z narativní klinické zprávy a implementace strukturovaných informací do elektronického zdravotního záznamu

5.1.1 Tokenizace lékařské zprávy

Narativní lékařská zpráva je tokenizována na objekty typu "číslo", "alfanumerické znaky" a "jiný znak". Výstup tokenizace označuji jako **tokenizovanou klinickou zprávu**. Dále budu výrazem „slovo“ označovat také tokeny typu číslo či jiné znaky. Tokenizace vychází z předchozího výzkumu, snahy zpracovávat narativní lékařské zpráv pomocí metod pro zpracování přirozeného jazyka. Získané tokeny tak mohou být kombinovány do specifických typů tokenů podle pravidel založených na regulárních výrazech. V této metodě se však zaměřuji na předzpracování textu pro jejich strojový překlad.

Výstupem první fáze metody 3PP je tokenizovaná klinická zpráva.

5.1.2 Normalizace tokenizované klinické zprávy

Normalizace tokenizované lékařské zprávy spočívá v opravě překlepů a rozšíření zkrácených výrazů do úplné slovní podoby. Korekce pravopisu a rozšíření zkratek či zkrácených slov na plnou formu provádí lékař nebo jiný zdravotnický odborník. Výsledek této druhé fáze nazývám **normalizovanou klinickou zprávou**.

Lékař v tokenizované narativní zprávě provádí úpravy, které označuji jako „transformace“. Jde o následující druhy úprav:

1. dělením slov, která ve zprávě byla nesprávně spojena (chybějící mezera),
2. spojením přilehlých slov, která byly nesprávně oddělena (přebývající mezera),
3. mazáním slov, která do lékařské zprávy vůbec nepatří,
4. změnou obsahu slov, která se dělí na dva druhy:
    a. oprava překlepů a
    b. rozvinutí zkratky nebo zkráceného slova na plný tvar.

Typografické chyby jsou opravovány v samostatném kroku před převodem tokenů na úplné slovní vyjádření. Této druhé fáze 3PP se netýká sémantická anotace. Úpravy se do databáze zaznamenávají v pořadí, ve kterém je lékař provedl. Informace o tom, kde se vyskytla chyba nebo jaký token byl rozšířen do plného tvaru, zůstává v databázi pro budoucí použití. Z technického hlediska je reprezentací výstupu (normalizovaná lékařská zpráva) les, kde kořeny jsou tokeny tokenizované klinické zprávy a listy představují výrazy získané transformací (opravou / rozšiřováním) tokenů, tedy finální výstup.

5.1.3 Sémantická anotace normalizované klinické zprávy

V této fázi měli spolupracující lékaři za úkol pořádně pročíst normalizované lékařské zprávy a sémanticky v ní anotovat výrazy představující klinické pojmy. Pro sémantickou anotaci byly použity mezinárodní klasifikační systémy LOINC **[69]**[3], SNOMED-CT **[59]**[2] a MKN 10 **[55]**[1]. Tyto tři klasifikační systémy jsou mapovány do systému UMLS (Unified Medical Language System) **[39]**[26] a jejich položky tak je možné převádět i do jiných do UMLS mapovaných systémů. Kromě těchto tří systémů byl použit také národní klasifikační systém LÉKY [70], který je databází léčiv a léčivých přípravků registrovaných Státním ústavem pro kontrolu léčiv České republiky.

Výstupem této třetí fáze metody 3PP **je semistrukturovaná normalizovaná klinická zpráva.** Je třeba poznamenat, že zatímco prezentovaný výzkum postupuje zdola nahoru, v závislosti na zkušenostech lékařů, kteří komentují zprávy, v budoucnu by se postup anotace mohl vylepšit sofistikovanými metodami sémantické anotace.

5.1.4 Normalizační databáze

Shromážděná databáze dvojic "nezpracovaný termín“ a „opravený / rozšířený výraz" (normalizační databáze) může mít potenciál pro učení automatizované transformační procedury. Aby bylo zřejmé, že zde takový potenciál je, že musíme zkontrolovat, zda se značný počet nezpracovaných termínů vždy koriguje / rozšiřuje (obecně: transformuje) buď na stejný termín, nebo na skupinu termínů, které mají jen nepatrné gramatické rozdíly; např. jednotné/množné číslo, jiný tvar s ohledem na ohýbání slova, pohlaví nebo sémanticky nevýznamný rozdíl, jako je „laboratorní vyšetření“ vs. „laboratoř“ (vzhledem k jazyku lékařské zprávy je význam shodný). Situace, kdy byl stejný termín (obvykle zkratka) přeměněn na úplně sémanticky odlišné pojmy, např. "LS" na "lumbosakrální" v jednom případě a "levá síň" (levé atrium) v jiném případě, jsou nežádoucí. Jejich příspěvek k chybovosti by však měl být zvážen s ohledem na poměr frekvencí různých alternativ za předpokladu, že automatizovaná transformační procedura by vždy zvolila nejčastější transformaci. Mít takovou míru chybovosti blízkou 0 % znamená, že automaticky transformovaný text (předložený k následnému automatickému překladu) může být srozumitelný, byť ne vždy gramaticky správný, a mít šanci zachovat srozumitelnost i po strojovém překladu do jiného jazyka.

Lepší míry aproximace automatizované transformace lze dosáhnout využitím jen opakovaných transformací daného vstupu, tedy úplným vynecháním případů, kdy jde o transformace bez znovupoužití). Míru chybovosti pak lze spočítat jako relativní podíl opakovaných transformací vážený pravděpodobností, že daná vybraná transformace je skutečně heterogenní.

Odhad míry chyb lze vyjádřit jako:

$$Err = \frac{\sum_r (n_r - 1) \cdot w_r}{\sum_r (n_r - 1)}$$

Kde n_r je počet výskytů nezpracovaného výrazu *r* a w_r je váha *r* pro výpočet chyby.

Váha se vypočítá jako

$$w_r = \frac{|\{(i,j)\};\ i,j \in tsf(r), i \neq j|}{P(|tsf(r)|, 2)}$$

Kde *tsf (r)* je množina transformovaných pojmů, které jsou výsledkem nezpracovaného výrazu *r* v databázi, *i* a *j* jsou výrazy z této množiny (jsou jeho různé prvky, ačkoli možná stejné termíny) a *P (n, k)* označuje *k*-permutaci *n*.

5.1.5 Softwarový nástroj TOCESA

Pro použití metody 3PP narativních klinických zpráv v praxi jsem vyvinul softwarový nástroj TOCESA pro tokenizaci, korekci, expanzi a sémantickou anotaci. Software TOCESA jsem postupně vyvíjel od roku 2010. TOCESA je samostatná PHP webová aplikace s částečným využitím JavaScriptu. Jako úložiště dat využívá databázi mySQL. Aplikace podporuje všechny tři fáze zpracování od plně automatizované tokenizace (první fáze) přes podporu normalizace (druhá fáze) po anotování a jeho ověřování (třetí fáze).

V tomto výzkumu se softwarem TOCESA pracovali dva kardiologové, kteří provedli normalizaci a sémantickou anotaci 49 lékařských zpráv z oboru kardiologie.

Online ukázka nástroje s jednou předem vyplněnou anonymizovanou narativní lékařskou zprávou je zpřístupněna na http://ie-demo.zvara.cz (uživatelské jméno a heslo: "demo"). Tyto přihlašovací údaje již byly zveřejněny v [68].

5.2 Výsledky

Metoda 3PP byla aplikována na údaje 49 anonymních českých lékařských zpráv z oblasti kardiologie. Dva kardiologové metodu 3PP v software TOCESA ověřili v letech 2012 až 2015 na celkem 49 lékařských zprávách z oblasti kardiologie.

5.2.1 Fáze I: Automatizovaná tokenizace volnotextových lékařských zpráv

Proces tokenizace spočívá v jednoduchém lineárním zpracování narativní lékařské zprávy, které znaky stejné třídy (číselné, alfanumerické a jiné) transformuje do sekvence tokenů. White-space znaky (mezera, tabulátor, carriage-return a nový řádek) byly zachovány jako zvláštní typy tokenů, aby se zachovala informace o struktuře. 49 anonymizovaných volnotextových lékařských zpráv z oblasti kardiologie bylo rozděleno na 3 324 tokenů (včetně mezer). Tímto způsobem bylo vytvořeno 49 tokenizovaných klinických zpráv.

5.2.2 Fáze II: Normalizace tokenizovaných klinických zpráv

Cílem této fáze bylo normalizovat tokenizované lékařské zprávy opravou chyb a rozšířením zkrácených termínů. Konkrétně šlo o:

- rozdělení nesprávně spojených tokenů představující slova (chybějící mezera nebo jiný rozdělující znak),
- spojení dvou vedlejších tokenů, které ve skutečnosti mají tvořit jedno slovo (přebývající mezera nebo jiný rozdělující znak),
- odstranění tokenů, které byly zřejmě zadány omylem (např. překlep obklopený mezerami),
- oprava překlepů (formou úpravy slova – tokenu),

- rozšíření zkratky na plný tvar (jen u zkratek či zkrácených slov, která se běžně nepoužívají ve formě zkratky).

Transformace, které lékaři nezávisle na sobě provedli, pochopitelně nebyly identické, neboť musely odrazit lékařské i jazykové znalosti, zkušenosti i běžnou praxi každého z těchto lékařů. Mapování bylo do databáze zaznamenáno ve formě další vrstvy. Provedením každé mapovací operace tedy byly zachovány informace o dříve provedených operacích stejného lékaře. Lékaři byli instruováni, aby v případě, kdy je zapotřebí provést více transformací na jednom tokenu, postupovali v pořadí, ve kterém jsou transformace nabízeny (tedy od rozdělování slov po rozšíření zkratek a zkrácených slov).

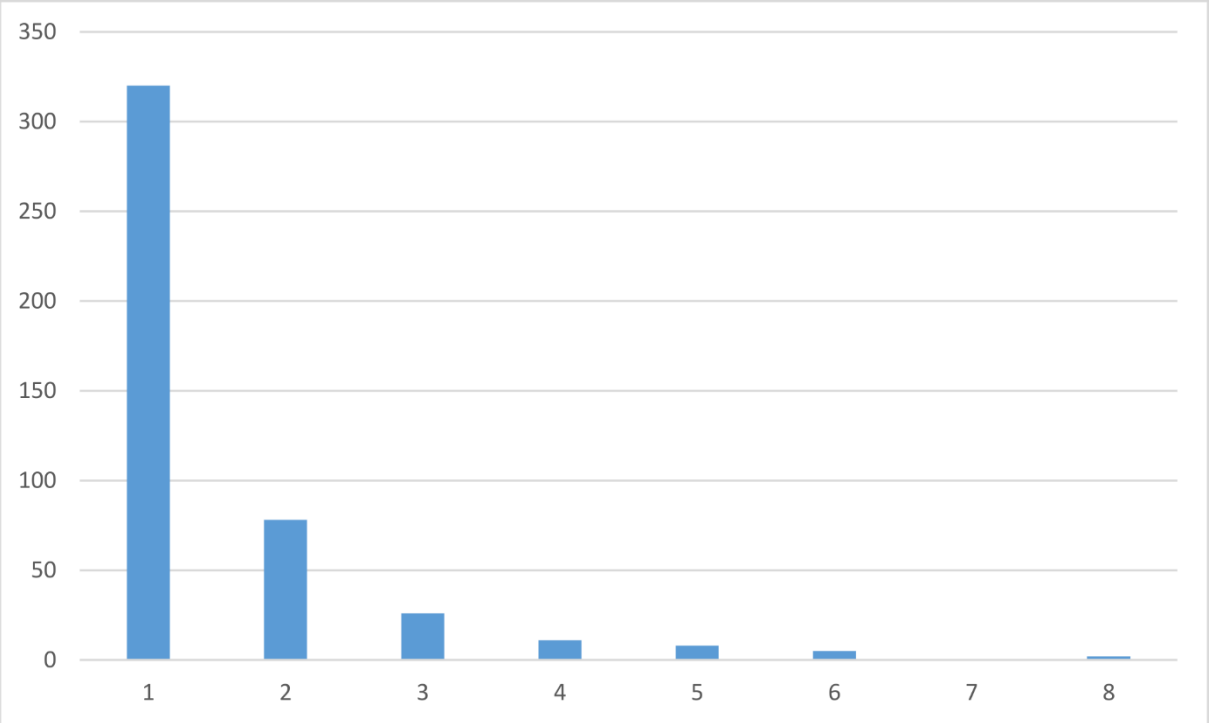
5.2.2.1 Údaje o provedených změnách

První kardioložka prošla normalizační fází pomocí software TOCESA ve všech 49 tokenizovaných klinických hlášeních a provedla opravy, expanze a další výše popsané transformace. Z celkem 127 různých znění tokenů s překlepy vytvořila celkem 129 různých slov, přičemž provedla 148 oprav v celkem 49 lékařských zprávách.

Ze 450 znění zkrácených slov vytvořila 684 slov či slovních spojení v celkem 1 411 případech. Z 92 různých znění zkratek vytvořila celkem 136 různých plných znění v celkem 267 případech. Tabulka 3 a obrázek 16 ukazují počty obsahu tokenů podle počtu slov po rozšíření zkrácených slov na plný tvar.

Tabulka 3: Počty obsahu tokenů podle počtu slov po rozšíření zkrácených slov

počet slov po rozšíření	1	2	3	4	5	6	7	8
počet zkrácených slov	320	78	26	11	8	5	0	2

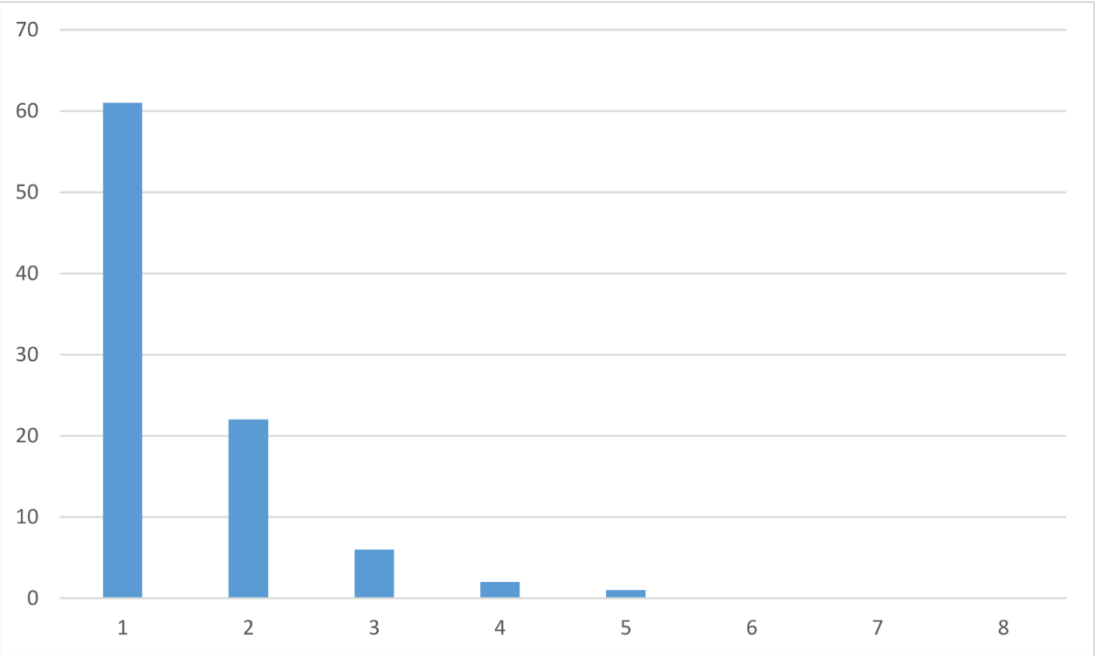


Obrázek 16: Počty obsahu tokenů podle počtu slov po rozšíření zkrácených slov

Tabulka 4 a obrázek 17 ukazují počty obsahu tokenů podle počtu slov po rozšíření zkratek na plný tvar.

Tabulka 4: Počty obsahu tokenů podle počtu slov po rozšíření zkratek na plný tvar

počet slov po rozšíření	1	2	3	4	5
počet zkratek slov	61	22	6	2	1



Obrázek 17: Počty obsahu tokenů podle počtu slov po rozšíření zkratek na plný tvar

Bližší pohled ukázal, že vyšší počet shod byl způsoben potřebou použít různé tvary stejného slova (zejména kvůli skloňování).

5.2.2.2 Analýza normalizační databáze a potenciál opětovného použití

Tabulka 5 uvádí souhrnné informace o provedených transformacích. Čtvrtý řádek tabulky ukazuje, že více, než 80 procent transformací rozšíření zkratek či zkrácených slov, bylo použito více než jednou. V těchto případech je zapotřebí vědět, zda transformace vždy ukázala na stejný cílový tvar. Transformace se stejným cílovým významem označuji za *homogenní*, transformace odkazující na alespoň dva různé cílové významy označuji za *heterogenní*. To je vidět ve střední části tabulky. Přes 90 procent rozšíření zkratek a přes 80 procent rozšíření zkrácených slov jsou sémanticky homogenní, odkazují tedy na stejný význam. To zahrnuje i transformace, které byly zaznamenány jen jedenkrát – ty označuji jako *unikátní*. Poměr mezi počtem unikátních a ostatních (vícenásobných) transformací je v případě zkratek a zkrácených slov zhruba stejný. V případě zkratek šlo o 175 takových

opakovaných transformací, v případě zkrácených slov o 955 opakovaných transformací (dopočteno).

Tabulka 5: Souhrnná informace o provedených transformacích

	Rozvinutí zkratek	**Rozvinutí zkrácených slov**	**Oprava chyb**
Celkem transformací	267	1405	144
Počet původních tokenů	92	450	123
Unikátní transformace	44 (16 %)	246 (18 %)	116 (81 %)
Neunikátní transformace	223 (84 %)	1159 (82 %)	28 (19 %)
Homogenní transformace	246 (92 %)	1159 (82 %)	144 (100 %)
- unikátní	110	510	139
- neunikátní	136	649	5
Heterogenní transformace	21 (8 %)	246 (18 %)	0 (0 %)
Počet opakování stejných transformací	175	955	21
… spočtená míra chybovosti	**5.17 %**	**11.16 %**	**0 %**

Spočtená míra spolehlivosti (5% míra chybovosti) pro zkratky může být považována za přijatelnou. Horší je výsledek u zkrácených slov (11% míra chybovosti). To je nejspíš způsobeno tím, že zkrácená slova jsou často srozumitelná jen v kontextu původní zprávy a

jejich znění se může shodovat se zněním jiných zkrácených slov. Příkladem takového zkráceného slova může být „zj.“, které podle kontextu odpovídá významům „zjištěný“ nebo „zjevný“ či „zřejmý“.

Jak je vidět, výsledky pro opravu chyb jsou zcela odlišné od výsledků pro rozšiřování. V případě oprav chyb nebyly zjištěny žádné heterogenní transformace, míra chybovosti je tedy 0 %. To odpovídá předpokladům, neboť zjevně není příliš pravděpodobné, že dva sémanticky odlišné termíny budou chybně zapsány tak, že chybné znění bude totožné. Také bylo zjištěno jen relativně málo opakovaných korekcí stejné chyby (jen 19 %), takže poměr transformace opětovného použití (bez použití nástrojů jako je počítání lexikografické vzdálenosti) je nízký.

Tabulka 6 uvádí 15 nejčastějších transformací při rozšíření zkráceného slova na plný tvar. Transformace rozlišují velká a malá písmena.

Tabulka 7 uvádí 15 nejčastějších transformací při rozšíření zkratky na plný tvar.

Tabulka 6: 15 nejčastějších zjištěných transformací při rozšíření zkráceného slova

původní tvar	**cílový tvar**	**počet transformací**	**druh**
Obj.	Objektivně	38	homogenní
vyš.	vyšetření	31	homogenní
Subj.	Subjektivně	30	homogenní
norm.	normální	28	homogenní
r.	roku	22	homogenní
nebol.	nebolestivé	19	homogenní
Pac.	Pacient	17	homogenní
bpn	bez patologického nálezu	16	homogenní
neg.	negativní	16	homogenní
Pac.	Pacientka	14	homogenní
frekv.	frekvence	14	homogenní
pravid.	pravidelný	14	homogenní
th.	therapie	14	homogenní
pac.	pacient	13	homogenní
palp.	palpačně	13	homogenní

Tabulka 7: 15 nejčastějších zjištěných transformací při rozšíření zkratky

původní tvar	**cílový tvar**	**počet transformací**	**druh**
OA	Osobní anamnéza	15	homogenní
RA	Rodinná anamnéza	11	homogenní
DK	dolní končetiny	10	homogenní
DM	diabetes mellitus	9	homogenní
DK	dolních končetin	8	homogenní
ES	extrasystol	7	homogenní
IM	infarkt myokardu	7	homogenní
AV	atrioventrikulární	6	homogenní
TK	krevního tlaku	6	homogenní
AA	alergická anamnéza	6	homogenní
ICHS	Ischemická choroba srdeční	6	homogenní
LK	levé komory	6	homogenní
AP	anginy pectoris	5	homogenní
NO	nynější onemocnění	4	homogenní
AV	atrio-ventrikulární	4	homogenní

5.2.2.3 Dopad normalizace na strojový překlad

Na původní (neupravené) i na normalizované lékařské zprávy jsme společně se spolupracujícími kardiology aplikovali strojový překlad z češtiny do angličtiny pomocí překladače Google. Obrázky **18A, 18B a 18C** ukazují výsledky strojového překladu, přičemž obrázek 18A ukazuje výsledek strojového překladu původní zprávy, obrázek 18B ukazuje strojově přeloženou normalizovanou zprávu a obrázek 18C ukazuje rozdíly mezi překlady.

Patients with implanted pacemaker Biotronik Pikos mode VVIM for AV block second degree burns Subj: he manages quite well against min.vyš not change. Vol .: cooperates, weight 64 kg, height 168 cm BMI: 25.3. Head, neck: BPN, BP: 125/75, Thorax: symmetrical, percussion, clear, vesicular breathing, clean hearts: Tapping the mdcl.č., heartbeat regular at the apex and on the basi two darker sounds, belly: soft, palp.nebol., hepar, lien: not enlarged, tapottement Bilat. neg. DK: no swelling, the calf was not. ECG: stiomulovaný regular rhythm of ventricular VVI mode - 100% stimulation. Parameters: Frequency base: 71.7 / min., Magnetic frequencies: 71.7 / min. interval: 836 ms pulse width of 0.72 ms Pacemaker Function: chol correct. 5, 3 mmol / l HDL 1.1 mmol / l LDL 3,2mmol / TG l 1.5 mmol / l glyekmie 6.1 Conclusion mmoll .: Implantaqce permanent pacemaker VVI, M for AV block second degree burns sinnus and bradycardia with the final events of 40 / min., and PAUSE 3,5ms. HLP intervened statins CHD comp. Arterial hypertension III.st.- well controlled. Therapie .: Anopyrin 100mg 1 / d, HCHTH 1 / 2TB / d, Enap 5mg: 1-0-1 / d, Zocor 20mg: 0-0-1 / d, control at TK, minerals. The principles of secondary prevention, anti-sclerotic diet. Checking in 4 months

Obrázek 18A: Výsledek strojového překladu původní zprávy

Patients with implanted pacemaker Biotronik Pikos mode VVIM for AV block II. degree. Subj: he manages quite well - no changes compared to the previous examination. Vol. : Cooperates, weight 64 kg, height: 168 cm, BMI: 25; 3. Head, neck: no positive findings, BP: 125/75, Thorax: symmetrical, percussion, clear, vesicular breathing, clean hearts: Tapping the mdcl. no. , Heartbeat regular at the apex and on the basi two darker sounds, abdomen: soft, palp. was not. , Hepar, lien: not enlarged, tapottement Bilat. neg. DK: no swelling, the calf was not. ECG: regular rhythm stimulated ventricular VVI mode - 100% stimulation. Parameters: Frequency base: 71, 7 / min. , Magnetic frequency: 71, 7 / min. interval: 836 ms pulse width: 0, 72 ms Pacemaker Function: chol correct. 5, 3 mmol / l HDL 1, 1 mmol / L LDL 3, 2 mmol / L TG 1, 5 mmol / l glyekmie 6, 1 mmol / l Conclusion. : Implantation of a permanent pacemaker mode VVIM for AV block II. st. sinnus and bradycardia resulting in action of 40 / min. , And PAUSE 3, 5 ms. HLP intervened statins CHD compensated. Arterial hypertension III. degree. - Well controlled. Therapie. : Anopyrin 100 mg: 1 / d, HCHTH 1/2 tb / d, Enap 5 mg: 1 - 0-1 / d, Zocor 20 mg: 0-0 - 1 / d, control blood pressure, minerals. The principles of secondary prevention, anti-sclerotic diet. Checking in 4 months

Obrázek 18B: Výsledek strojového překladu normalizované zprávy

Patients with implanted pacemaker Biotronik Pikos mode VVIM for AV block ~~second degree burns~~ **II. degree.** Subj: he manages quite well ~~against min.vyš not change. Vol .~~ **no changes compared to the previous examination. Vol.** : ~~cooperates,~~ **Cooperates,** weight 64 kg, ~~height~~ **height:** 168 ~~cm~~ **cm,** BMI: ~~25.~~**25;** 3. Head, neck: ~~BPN,~~ **no positive findings,** BP: 125/75, Thorax: symmetrical, percussion, clear, vesicular breathing, clean hearts: Tapping the ~~mdcl.č.~~**mdcl. no.** , ~~heartbeat~~ **Heartbeat** regular at the apex and on the basi two darker sounds, ~~belly:~~ **abdomen:** soft, ~~palp.nebol.~~**palp. was not.** , ~~hepar,~~ **Hepar,** lien: not enlarged, tapottement Bilat. neg. DK: no swelling, the calf was not. ECG: ~~stiomulovaný~~ regular rhythm ~~of~~ **stimulated** ventricular VVI mode - 100% stimulation. Parameters: Frequency base: ~~71.~~**71,** 7 / ~~min.~~**min.** , Magnetic ~~frequencies: 71.~~**frequency: 71,** 7 / min. interval: 836 ms pulse ~~width of 0.~~**width: 0,** 72 ms Pacemaker Function: chol correct. 5, 3 mmol / l HDL ~~1.~~**1,** 1 mmol / ~~l~~**L** LDL ~~3,2mmol~~ **3, 2 mmol** / **L** TG ~~l 1.~~**1,** 5 mmol / l glyekmie ~~6.~~**6,** 1 ~~Conclusion mmoll .~~**mmol / l Conclusion.** : ~~Implantaqce~~ **Implantation of a** permanent pacemaker ~~VVI, M~~ **mode VVIM** for AV block ~~second degree burns~~ **II. st.** sinnus and bradycardia ~~with the final events~~ **resulting in action** of 40 / ~~min.~~**min.** , ~~and~~ **And** PAUSE ~~3,5ms.~~ **3, 5 ms.** HLP intervened statins CHD ~~comp.~~**compensated.** Arterial hypertension III.~~st.~~ **degree.** - ~~well~~ **Well** controlled. ~~Therapie .~~**Therapie.** : Anopyrin ~~100mg~~ **100 mg:** 1 / d, HCHTH ~~1 / 2TB~~ **1/2 tb** / d, Enap ~~5mg: 1-0-1~~ **5 mg: 1 - 0-1** / d, Zocor ~~20mg: 0-0-1~~ **20 mg: 0-0 - 1** / d, control ~~at TK,~~ **blood pressure,** minerals. The principles of secondary prevention, anti-sclerotic diet. Checking in 4 months

Obrázek 18C: Rozdíl mezi strojovým překladem původní a normalizované zprávy

Na první pohled je vidět, že výsledek strojového překladu se v důsledku normalizace značně zlepšil. Strojový překlad neupraveného textu obsahuje nepřeložené české zkratky a zkrácená slova. Ty tak neposkytují čtenáři (předpokládejme anglicky hovořícího lékaře) dostatečnou informaci, pokud tento čtenář neovládá alespoň nějaký jiný blízký slovanský jazyk.

Konkrétními příklady zlepšeného překladu jsou například:

- Překlad jinak v angličtině zcela nepochopitelných výrazů: "against *min.vyš* not change" by ve skutečnosti měl znít např. "no change compared to the previous examination" (zkrácená slova "min.vyš" byla rozšířena na "minulého vyšetření". Podobně to platí např. pro zkratku "BPN", která je rozšířena na "bez pozitivního nálezu" či "kontrola TK", kde se zkratka "TK" správně překládá na "krevní tlak".
- Nesprávně zapsané české pojmy jsou zachovány ve strojovém překladu původního textu zprávy, některé jsou však částečně srozumitelné díky společnému latinskému kořenu slova, jako je v případě slova "sti***o***mulovaný" nebo "Implanta***q***ce" (chybné písmo označené tučnou kurzívou). Jejich správný překlad dále snižuje nejednoznačnost.

- Někdy není problémem jednotlivé strojově nepřeložitelné slovo, ale nesprávné přiřazení významu celé části textu statistickým překladačem. Příkladem může být text „1 / 2TB“ (polovina tablety), který statistický překladač zmate mnohem víc, než normalizovaná verze „1/2 tb“. Podobně překlad nastavení rytmu kardiostimulátoru je lépe přeložen jako „resulting in action of 40 / min“, než jako „with the final events of 40 / min“.
- Příkladem neškodného "těžkopádného" překladu je slovo "belly" namísto příslušného odborného výrazu "abdomen".

Na těchto příkladech je zřetelně vidět, že strojový překlad je možné úspěšně použít i jen díky jednoduchému vylepšení původního textu jeho normalizací.

Přesto je zřejmé, že i zprávy „vylepšené“ normalizací by měly být využívány s opatrností a jen v případě, kdy se v nouzové situaci nedaří získat potřebné informace věrohodnějším způsobem.

Po publikování výsledků Google nasadil novou verzi svého překládacího systému, který značně vylepšil i překlad z češtiny do angličtiny [71]. Pro srovnání jsem tedy nechal tuto novou verzi Google překladače přeložit stejnou zprávu před normalizací i po normalizaci. Výsledky jsou zachyceny na obrázcích 19A, 19B a 19C.

Patient with implanted pacemaker Biotronik Pikos in VVIM mode for AV block II. Subj: he's doing pretty well - no change against min. Obj. Collaborates, weight: 64 kg, height: 168 cm, BMI: 25,3. Head, neck: bpn, TK: 125/75, chest: symmetrical, full throat, clear, breathing cellar, clean, heart: clue to mdcl., Regular heart rate, bite and bass 2 darker, Soft, palp.nebol., Hepar, lien: not enlarged, tapottement bilat. Neg. DK: no swelling, no calf. ECG: Regular stiumulated ventricular rhythm, VVI mode - 100% stimulation. Device parameters: Basic frequency: 71.7 / min., Magnetic frequency: 71.7 / min. Interval: 836 ms pulse width: 0.72 ms Pacemaker functions: correct chol. 5, 3mmol / l HDL 1.1 mmol / l LDL 3.2 mmol / l TG 1.5 mmol / l glycemia 6.1 mmol Conclusion: Implantation of a permanent pacemaker in VVI, M for AV blockade II. And sinnus bradycardia with a resulting action of 40 / min, and pauses of 3.5ms. HLP intervened statins ICHS comp. Arterial hypertension III.st.- well controlled. Therapy: Anopyrin 100mg: 1 / d, HCHTH 1 / 2tb / d, Enap 5mg: 1-0-1 / d, Zocor 20mg: 0-0-1 / d, Principles of secondary prevention, antisklerotic diet. Check for 4 months

Obrázek 19A: Výsledek strojového překladu původní zprávy (nový překladač Google)

Patient with implanted pacemaker Biotronik Pikos in VVIM mode for AV block II. Degree. Subj: he's doing pretty well - there's no change to the previous exam. Order no. : Cooperates, weight: 64 kg, height: 168 cm, BMI: 25, 3. Head, neck: no positive finding, TK: 125/75, chest: symmetrical, tap full, clear, breathing cellar, clean, heart: tick to mdcl. No. , Heart rate regular, on the tip and on the base 2 darker echoes, abdomen: soft, palp. It was not. , Hepar, lien: not enlarged, tapottement bilat. Neg. DK: no swelling, no calf. ECG: Regular stimulated ventricular rhythm, VVI mode - 100% stimulation. Device parameters: Base frequency: 71, 7 / min. , Magnetic frequency: 71, 7 / min. Interval: 836 ms pulse width: 0, 72 ms Pacemaker functions: correct chol. 5, 3 mmol / l HDL 1, 1 mmol / l LDL 3, 2 mmol / l TG 1, 5 mmol / l glycemia 6, 1 mmol / l Conclusion. : Continuous pacemaker implantation in VVIM mode for AV block II. St. And sinnus bradycardia with a resulting action of 40 / min. , And pauses 3, 5 ms. HLP interfered with ICHS statins compensated. Arterial hypertension III. Degree. - well controlled. Therapie. : Anopyrin 100 mg: 1 / d, HCHTH 1/2 tb / d, Enap 5 mg: 1 - 0 - 1 / d, Zocor 20 mg: 0 - 0 - 1 / d. Principles of secondary prevention, antisklerotic diet. Check for 4 months

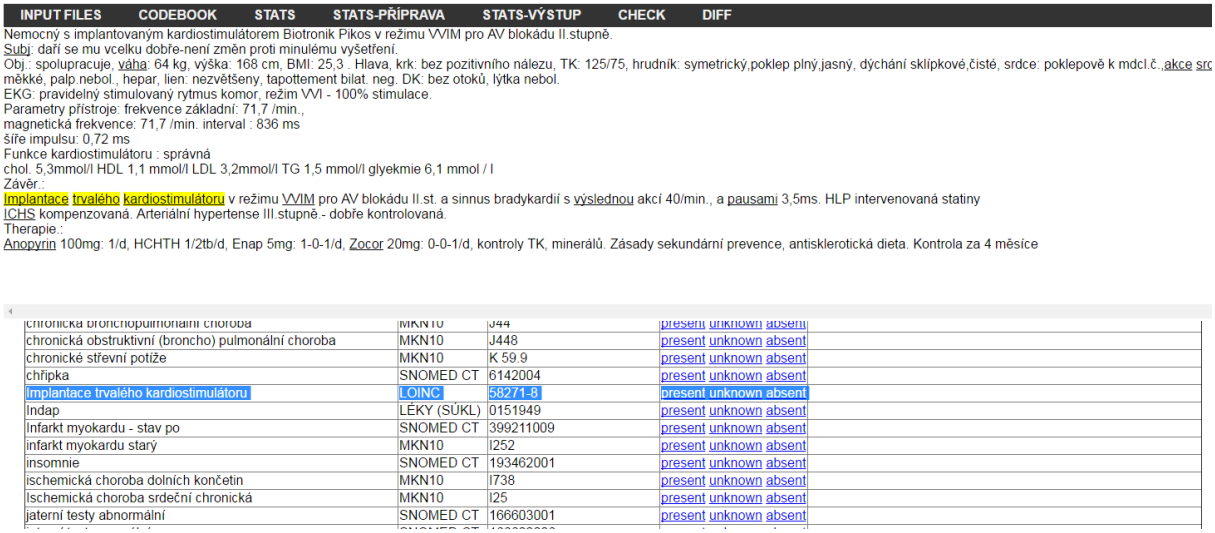
Obrázek 19B: Výsledek strojového překladu normalizované zprávy (nový překladač)

Patient with implanted pacemaker Biotronik Pikos in VVIM mode for AV block II. **Degree.** Subj: he's doing pretty well - **there's** no change ~~against min. Obj. Collaborates,~~ **to the previous exam. Order no. : Cooperates,** weight: 64 kg, height: 168 cm, BMI: ~~25,3.~~ **25, 3.** Head, neck: ~~bpn,~~ **no positive finding,** TK: 125/75, chest: symmetrical, ~~full throat,~~ **tap full,** clear, breathing cellar, clean, heart: ~~clue~~ **tick** to ~~mdcl.~~**mdcl. No.** , ~~Regular heart rate, bite~~ **Heart rate regular, on the tip** and ~~bass~~ **on the base** 2 ~~darker, Soft, palp.nebol.~~**darker echoes, abdomen: soft, palp. It was not.** , Hepar, lien: not enlarged, tapottement bilat. Neg. DK: no swelling, no calf. ECG: Regular ~~stiumulated~~ **stimulated** ventricular rhythm, VVI mode - 100% stimulation. Device parameters: ~~Basic~~ **Base** frequency: ~~71.~~**71,** 7 / ~~min.~~**min.** , Magnetic frequency: ~~71.~~**71,** 7 / min. Interval: 836 ms pulse width: ~~0.~~**0,** 72 ms Pacemaker functions: correct chol. 5, ~~3mmol~~ **3 mmol** / l HDL ~~1.~~**1,** 1 mmol / l LDL ~~3.~~**3,** 2 mmol / l TG ~~1.~~**1,** 5 mmol / l glycemia ~~6.~~**6,** 1 mmol ~~Conclusion: Implantation of a permanent~~ / **l Conclusion. : Continuous** pacemaker **implantation** in ~~VVI, M~~ **VVIM mode** for AV ~~blockade~~ **block** II. **St.** And sinnus bradycardia with a resulting action of 40 / ~~min, and~~ **min. , And** pauses ~~of 3.5ms.~~ **3, 5 ms.** HLP ~~intervened~~ **interfered with ICHS** statins ~~ICHS comp.~~**compensated.** Arterial hypertension III.~~st.~~ **Degree.** - well controlled. ~~Therapy:~~ **Therapie. :** Anopyrin ~~100mg:~~ **100 mg:** 1 / d, HCHTH ~~1 / 2tb~~ **1/2 tb** / d, Enap ~~5mg: 1-0-1~~ **5 mg: 1 - 0 - 1** / d, Zocor ~~20mg: 0-0-1~~ **20 mg: 0 - 0 - 1** / ~~d,~~ **d.** Principles of secondary prevention, antisklerotic diet. Check for 4 months

Obrázek 19C:Rozdíl mezi strojovým překladem původní a normalizované zprávy (nový překladač Google)

5.2.3 Fáze III: sémantická anotace normalizované klinické zprávy

Podle lékařů, kteří prováděli zpracování, pro ně byla časově nejnáročnější fáze sémantického anotování. V této části museli označit slova či fráze (více slov) v normalizovaných lékařských zprávách a přiřazovat ke kódům předem určených klasifikačních seznamů. Pro sémantické anotování byly vybrány klasifikační systémy MKN 10, SNOMED CT, LOINC a databáze LÉKY SÚKL. Každá provedená anotace byla navíc anotována jako „přítomno“, „nepřítomno“ či „může být přítomno“ (deklarace nejistoty). Způsob zapisování anotace je zobrazen na obrázku 20. V tomto případě jde o anotaci výrazu „implantace trvalého kardiostimulátoru“ pomocí kódu LOINC 58271-8.



Obrázek 20: Proces anotace fráze "Implantace trvalého kardiostimulátoru"

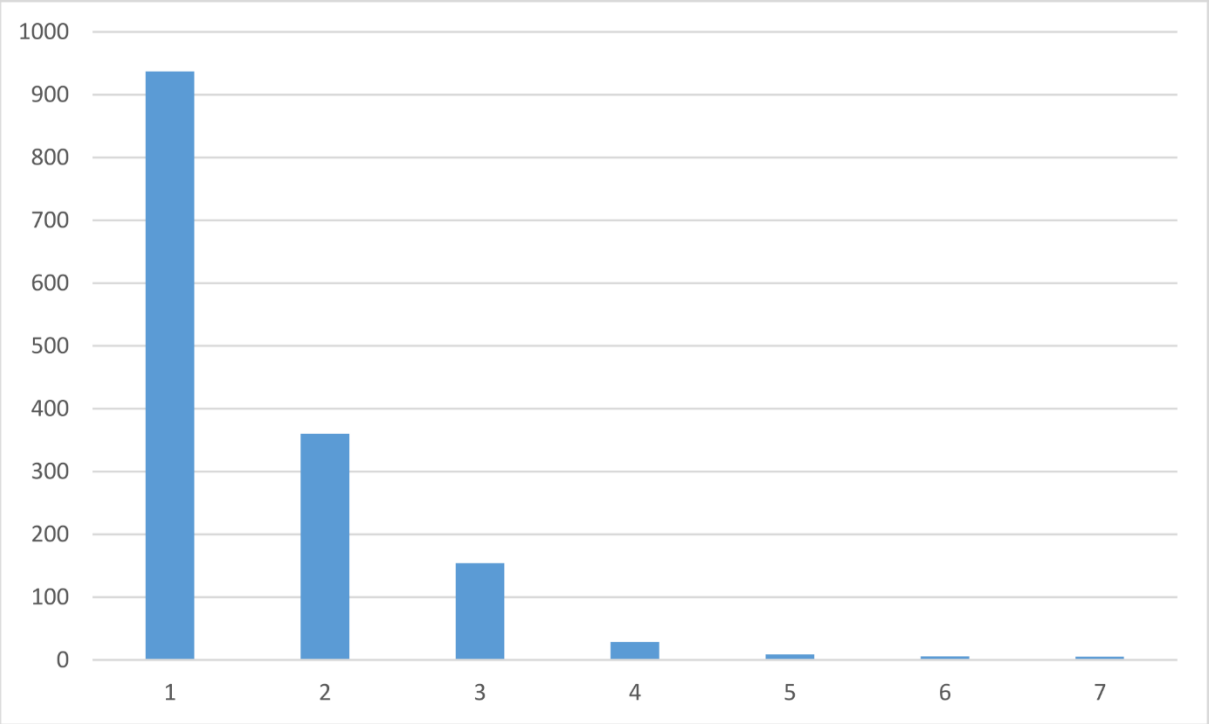
Výsledkem třetí fáze je částečně strukturovaná normalizovaná lékařská zpráva sestávající ze dvou částí. První částí je normalizovaná klinická zpráva s využitím strojového překladu, která je snadno čitelná i pro lékaře neznalé českého jazyka. Druhou část tvoří souhrn strukturované informace nalezené v normalizované lékařské zprávě.

Normalizované lékařské zprávy se lišily délkou, od té se odvíjel i počet sémantických anotací. Mezi 49 normalizovanými zprávami měla nejkratší jen 244 znaků bez mezer a

nejdelší 3 153 znaků bez mezer. Průměrná zpráva zprávy byla 1 576 znaků bez mezer. První kardioložka provádějící sémantickou anotaci anotovala celkem 49 normalizovaných lékařských zpráv, přičemž zaznamenala celkem 1 500 anotací. Níže uvádím tabulku 8 a obrázek 21, které uvádějí počet anotací podle počtu slov použitých v anotaci.

Tabulka 8: Počet anotací podle počtu slov použitých v anotaci

počet tokenů	1	2	3	4	5	6	7
počet anotací	937	360	154	29	9	6	5



Obrázek 21: Počet anotací podle počtu slov použitých v anotaci

Druhý kardiolog následně provedl revizi sémantických anotací provedených první kardioložkou. V případě 1 454 anotací (96,96 %) se s ní shodl, dalších 46 anotací bylo

anotováno jinak. (3,07 %) Distribuce kvality anotace podle této kontroly (tj. jednosměrná shoda v anotaci) pro jednotlivé klasifikační systémy je uvedena v tabulce 9.

Nejčastěji anotovanými termíny byly (v souladu s očekáváním v daném lékařském oboru) tyto termíny:

- 71x LOINC 55284-4 (krevní tlak)
- 65x SNOMED CT 301114007 (nález ohledně pravidelnosti srdečního rytmu)
- 43x LOINC 58271-8 (trvalá implantace kardiostimulátoru

Tabulka 9: Distribuce kvality anotace dle kontroly druhým kardiologem

Klasifikační systém	**správných (počet)**	**správných (%)**	**nesprávných (počet)**	**nesprávných (%)**	**Celkem**
MKN 10	130	93,53	9	6,47	139
SNOMED CT	865	96,65	30	3,35	895
LOINC	283	98,61	4	1,39	287
LÉKY	176	98,32	3	1,68	179
CELKEM	**1454**	**96,93**	**46**	**3,07**	**1500**

5.3 Diskuse

Elektronická zdravotnická dokumentace je založena především na narativních lékařských zprávách. Jejich objem, struktura i forma závisejí na tom, kdo je píše, za jakým účelem a na jakém místě. Podrobná publikace na toto téma byla zveřejněna před téměř 20 lety [72].

5. Třífázová metoda předběžného zpracování

Aplikace metod zpracování přirozeného jazyka (natural language processing) na původních narativních lékařských zprávách, s cílem transformovat údaje do strukturovaného elektronického zdravotního záznamu, je mimořádně náročná. To je dobře vidět i na pokusech o automatizovaný překlad. Je třeba si uvědomit, že pro dosažení ještě vyšší užitečnosti než poskytuje strojový překlad, je zapotřebí komplexně z lékařských zpráv extrahovat informace, například pro automatizovanou podporu v rozhodovacích procesech.

Pokud lékař, který ošetřuje pacienta, plynule ovládá jazyk lékařské zprávy, problémy s jejím narativním textem nebývají kritické, i když i v tom případě se může potýkat s problémy s porozuměním zkratkám a zkráceným slovům jiných specializací (viz např. Cheng v [4]). Strojový překlad takového původního textu, jehož cílem je přiblížit význam částí lékařské zprávy lékaři v cizím jazyce, pravděpodobně narazí na vážné potíže.

Tyto problémy jsem demonstroval na ukázce automatizovaného překladu původních českých lékařských zpráv do angličtiny.

Třífázová metoda předzpracování (3PP) pro vylepšení narativní klinické zprávy byla doposud potvrzena na 49 českých narativních lékařských zprávách z oblasti kardiologie. Jedním z možných použití je doplnění strukturované informace z narativních lékařských zpráv v kardiologii do elektronického orálního záznamu (EOHR- Electronic Oral Health Record) [64].

Pro úpravu a anotování vstupních narativních lékařských zpráv jsem vytvořil software TOCESA. Ten obsahuje jednoduchý automatický tokenizer, který z každé vstupní zprávy vytvoří řetězec tokenů. Ve druhé fázi, ve které lékaři vstupní text normalizovali, jsou původní tokeny překrývány novými tokeny. Je tak zachována celá historie zpracování od původního řetězce tokenů přes jednotlivá překrytí (opravy) po výsledný stav.

Společně se spolupracujícími kardiology jsme zjistili, že po dokončení první a druhé fáze metody 3PP bylo snadno čitelných všech 49 ze 49 zpracovaných lékařských zpráv. Ověření metody ukázalo, že nejnáročnější prací, kterou provedli spolupracující kardiologové, byla sémantická anotace prováděné ve 3. fázi metody 3PP. Důvody pro to byly tři:

1. Klasifikační systémy SNOMED CT a LOINC nejsou k dispozici v českém jazyce. Proto bylo nutné, aby kardiologové měli dobrou znalost angličtiny. Jak jsem uvedl výše, strojový překlad normalizované klinické zprávy může proces urychlit a zjednodušit.
2. Kardiologové se při vyhledávání v klasifikačních systémech setkali s několika problémy. Vyhledané klinické pojmy klasifikačních systémů někdy celkem dobře odpovídaly anotovanému textu, někdy však obsahovaly jen obecnější nebo konkrétnější termín a anotace tak byla provedena nejbližším termínem podle uvážením kardiologa. Samotné rozhodování o volbě použitého termínu jistě přispělo k časové náročnosti. Například první kardioložka uvedla, že „vyšetření EKG“ by mohl odpovídat termín SNOMED CT č. 447113005, ale ten uvádí konkrétní (byť široce užívanou) metodu provedení vyšetření EKG. Zpracovávané zprávy však informaci o použité metodě neobsahovaly.
3. Oba kardiologové se nejprve museli naučit software používat. Rychlost anotování postupně rostla se zvyšující se zkušeností s nástrojem TOCESA. Oba kardiologové přitom v minulosti úspěšně absolvovali kurzy EuroMISE centra, jejichž cílem bylo poskytnout lékařům vyšší gramotnost v oblasti medicínské informatiky. Společně se školitelem a spolupracujícími lékaři jsme se pokusili zapojit ještě třetího kardiologa, který neměl další vzdělání v oboru medicínské informatiky. Tento třetí kardiolog se

však nenaučil pracovat se software TOCESA a vyhledávacími systémy klasifikačních systémů a tak nedokázal provádět sémantickou anotaci ve třetí části.

Termíny kódových systémů extrahované z narativních klinických zpráv mohou být dále zpracovány statistickými metodami a použity k podpoře lékařského rozhodování a dalšího lékařského výzkumu. Rozhodování je jednou z hlavních činností lékařů. Mnoho metod založených na přístupu ke statistické nebo umělé inteligenci, která se zaměřuje na podporu rozhodování v oblasti zdravotní péče, bylo nedávno prezentováno v [73] a jejich trendy byly shrnuty v [74].

Vzhledem ke geografické poloze České republiky a současnému stavu norem a projektů se zdá, že prakticky použitelná struktura pro sdílení informací o pacientech v Evropě je epSOS Patient Summary [50]. Metoda 3PP by mohla pomoci odhalit strukturované informace z narativních textů, které by mohly být využity v pacientském shrnutí (Patient Summary) v různých mluvených evropských jazycích.

Jsem si zřetelně vědom několika zásadních omezení metody 3PP:

1. Nástroj TOCESA zatím není dostatečně robustní pro široké využití a databáze transformací je zatím velmi malá a zkreslená, neboť zprávy pocházely od velmi malého počtu lékařů.
2. Databáze anotací je závislá na stávajících klinických nomenklaturách – klasifikačních systémech, které zatím stále (i přes četné snahy) nejsou dostatečně propojené a konzistentní. To lze překonat jen v delším časovém horizontu.

6. Závěry

Systémy EHR používané jak v praxi lékaře, tak v nemocničním prostředí by měly být schopné uchovávat jak strukturovaná, tak i narativní data. Narativní lékařské zprávy umožňují lékařům sdílet složité představy o onemocnění pacienta a navrhované terapii a odrážejí pozorování lékaře, myšlení, hodnocení a důvody pro plán léčby. Narativní lékařské zprávy jsou důležitou součástí zdravotnické dokumentace, nicméně jejich opětovné využití prostřednictvím informačních a komunikačních technologií je obtížné z důvodu mnoha chyb a nejednoznačností.

Během studia jsem se pokusil navázat na dřívější práce, ve kterých se jejich autoři snažili o lingvistickou analýzu textu, účastnil jsem se implementace strukturovaného zdravotního záznamu v zubním lékařství a v poslední části studia jsem navrhl a společně s lékaři ověřil třífázovou metodu předzpracování narativních lékařských zpráv.

Snaha o lingvistickou analýzu na podkladových narativních lékařských zprávách potvrdila dřívější závěry Jiřího Semeckého a Petera Smatany ohledně povahy těchto zpráv, které nejsou typickými českými texty. Jde spíše o heslovitá sdělení s častým využitím zkratek i zkrácených slov a s velkým množstvím překlepů.

České lékařské zprávy nelze proto považovat za běžné české texty. Snaha o extrakci informací byla značně negativně ovlivněna absencí české verze mezinárodních klasifikačních systémů jako je LOINC a SNOMED CT.

Má účast na implementaci elektronického zdravotního záznamu pro zubní lékařství mi poskytla cenné zkušenosti se zachycováním klinických informací do strukturovaného

zdravotního záznamu, jehož struktura byla navržena odbornými lékaři. Mojí úlohou bylo navrhnout způsob implementování modelu a provést společně s kolegy implementaci. Diskuse s lékaři mi umožnila pochopit proč je pro ně narativní záznam důležitý. Strukturovaná informace totiž sice obsahuje jednoznačné, do klinického modelu zapsané údaje, neumožňuje však lékaři vlastním způsobem popsat například míru nejistoty svých závěrů. Tuto možnost lékařům dává právě narativní záznam. Také se ukázalo, že je důležitý i způsob zadávání údajů.

V poslední části studia jsem se zaměřil na návrh třífázové metody pro předzpracování původních textů, aby bylo snazší využít je pro získání informací, například lékařem v zahraničí, který poskytuje neplánovanou péči. Tato navrhovaná metoda a její nástroje umožňují zvýšit kvalitu textu a provést sémantickou anotaci. Pro to bude nutné značně rozšířit počet zpráv, na kterých se rozšíří základní anotační databáze tak, aby došlo ke zvýšení využití dříve zjištěných transformací.

Provedené ověření, byť založené na malém vzorku, naznačuje, že provedení transformací termínů má kladný dopad na kvalitu strojového překladu. Obtížně pochopitelné (a v současné době strojově nepřeložitelné) termíny jsou nahrazené přiměřenými překlady. Odhadovaná kvalita plně automatizované transformace je slibná ve srovnání s odstraněním velké náročnosti ručního zpracování, i když transformace nejednoznačných tokenů bude zřejmě vyžadovat ruční zásah.

Pro snížení chybovosti by budoucí normalizační databáze měla ukládat skóre nejednoznačnosti podle jednotlivých specializací, či s přihlédnutím k rozdílu ve specializaci, vzdělání a lékaře. Některé zkratky a zkrácená slova lze však zřejmě bezpečně rozšířit i bez ručního zásahu. Přitom je třeba mít stále na mysli, že ověření proběhlo jen na 49 zprávách.

Výsledky říkají, že pozitivní vliv transformace lékařských zpráv je v daném vzorku zřejmý. Zároveň však existuje riziko posunu granularity a hrozí záměna významu.

Abych dostál prezentaci kroků, které jsem předpokládal, že bude nutné učinit, uvádím znovu tyto kroky a způsob jejich realizace:

Zodpovězení otázky: „Které vlastnosti českých lékařských zpráv působí největší problémy v nestatistických fázích zpracování přirozeného jazyka?“

Velkých problémů při zpracování pomocí metod pro zpracování přirozeného jazyka je hned několik:

1. Narativní české lékařské zprávy nemají řádnou větnou stavbu, nelze je z praktického pohledu považovat za běžné české texty.
2. České narativní lékařské zprávy obsahují velmi mnoho zkratek a zkrácených slov, často s nejednoznačným významem, přičemž pro upřesnění významu může být zapotřebí nejen znalost oborového kontextu, ale i konkrétního autora.
3. Tyto zprávy také obsahují velké množství překlepů a jiných chyb.
4. Odborné termíny často nejsou v českém jazyce k nalezení v užívaném mezinárodním klasifikačním systému (zejména LOINC, SNOMED CT).

Navržení základního postupu pro analýzu česky psaných lékařských zpráv.

Navrhl jsem třífázovou metodu předzpracování textu (3PP), jejíž součástí může být i extrakce informací.

Pomocí vlastní implementace s možností využití externích nástrojů ověřit navržený postup pro analýzu lékařských zpráv založených na češtině a základní postup i výsledky publikovat.

Implementoval jsem softwarový nástroj TOCESA a výsledky výzkumu a metodu 3PP jsem publikoval v impaktovaném časopisu.

Dlouhodobým cílem výzkumu je získávat informace ve strukturované formě vhodné pro sekundární použití při rozhodování o lékařských rozhodnutích, zajištění kvality a lékařský výzkum.

8. Citovaná literatura

[1] Semecký J, Zvárová J (vedoucí diplomové práce). Diplomová práce: Multimediální elektronický záznam o nemocném v kardiologii. Praha: Matematicko-fyzikální fakulta Univerzity Karlovy; 2001.

[2] Smatana P, Paralič J (vedoucí diplomové práce). Diplomová práce: Spracovanie lekárskych správ pre účely analýzy a dolovania v textoch. Košice: Technická univerzita v Košiciach; 2005.

[3] Přečková P, Zvárová J (školitel). Disertační práce: Jazyk lékařských zpráv a jeho informačně lexikální analýza. Praha: 1. lékařská fakulta Univerzity Karlovy; 2011.

[4] Cheng TO. Medical Abbreviations.. J R Soc Med. 2004;97(11)

[5] Van Ginneken AM. The physician´s flexible narrative. Methods Inf Med. 1996;35(2)

[6] Van Ginneken AM. Considerations for the representation of meta-data for the support of structured data entry. Methods Inf Med. 2003;42

[7] Van Ginneken AM., Stam H, Van Mulligen EM, de Wilde M, Van Mastrigt R, Van Bemmel JH. ORCA: the versatile CPR. Methods Inf Med. 1999;38

[8] Nagy M, Říha A (školitel). Disertační práce: Harmonizace klinického obsahu elektronického zdravotního záznamu. Praha: Univerzita Karlova v Praze; 2011.

[9] Shapiro JS, Bakken S, Hyun S, Melton GB, Schlegel C, Johnson SB. Document Ontology: Supporting Narrative Document in Electronic Health Records. AMIA Annu Symp Proc. 2005;

[10] Bleeker SE, Derksen-Lubsen G, Van Ginneken AM, Van der Lei J, Molla HA. Structured Data Entry for Narrative Data in a Borad Speciality: Patient History and Physical Examination in Pediatrics. BMC Med Inform Decis Mak. 2006;

[11] Knaup P, Garde S, Haux R. Systematic planning of patient records for cooperative care and multicenter research. Int J Med Inf. 2007;76

[12] Deleger L, Grouin C, Zweigenbaum P. Extracting medical information from narrative patient records - the case of medication/related information. J Am Med Inform Assoc. 2010;17

[13] Garcia-Remesal M, Maojo V, Billhardt H, Crespo J. Integration of Relational and Textual Biomedical Sources. Methods Inf Med. 2009;48(1)

[14] Blaschke C, Hirschman L, Valencia A. Information extraction in molecular biology. Brifings in Bioinformatics. 2010;2(1)

[15] Johnson SB, Bakken S, Dine D, Hyun S, Mendonce E, Morrison F et al. An electronic health record based on structure narrative. J Am Med Inform Assoc. 2008;15(1)

[16] Hui Y. Automatic extraction of medication information from medical discharge summaries. J Am Med Inform Assoc. 2010;17

[17] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents. Performance evaluation.. J Biomed Inform. 2006;39

[18] Oemig F, Blobel B. Natural Language Processing Supporting Interoperability in Healthcare. Text Mining. 2014;

[19] Lopprich K, Krauss F, Ganzinger M, Senghas K, Riezler S, Knaup P. Automated classification of selected data elements from free/text diagnostic reports in clinical research. Methods Inf Med. 2016;55

[20] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics. 2008;

[21] Friedman C, Shagina I, Lussier I, Hripscak G. Automated encoding of clinical documents based on natural language processing in the clinical environment. J Am Med Inform Assoc. 2004;

[22] Friedman C. System and method for language extraction and encoding utilizing the parsing of text data in accordance with domain parameters. United States Patent no. 6,182,029 B1. USA. 2001

[23] Riskin D, Shroff A. Systems and methods for processing patient data history. United States Patent no. 2014/0181128 A1. USA. 2014

[24] Safran C, Bloomrosen M, Hammond W, Labkoff S, Markel-Fox S, Tang P et al. Toward a National framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007;14(1)

[25] Schreier G, Ammenwerth , Hörbst A, Hayn D. eHealth2016 - Health Informatics Meets eHealth. IOS Press Ebook. 2016

[26] Dostál O, Šárek M. Support for Electronic Health Records in Czech Law. Eur J Biomed Inform. 2012;8(2)

[27] Blobel B, Giacomini M. Interoperability is more than just technology. Eur J Biomed Inform. 2016;12(1)

[28] Žďárek R. Vedení zdravotnické dokumentace a její náležitosti. Zdravotnické noviny. 2009

[29] Zákon o péči o zdraví lidu. In: Sbírka zákonů. Ministerstvo spravedlnosti ČSSR; 1966. p. 74-91.

[30] Novela zákona o péči o zdraví lidu. In: Sbírka zákonů. Tiskárna Ministerstva vnitra, p.o.; 2001. p. 6344-6346.

[31] Vyhláška o zdravotnické dokumentaci. In: Sbírka zákonů. Tiskárna Ministerstva vnitra, p.o.; 2006. p. 5282-5284.

[32] Zákon o zdravotních službách a podmínkách jejich poskytování. In: Sbírka zákonů. Tiskárna Ministerstva vnitra, p.o.; 2011. p. 4730-4801.

[33] Hammond K., Helbig S., Benson C., Brathwaite-Sketoe B.. Are Electronic Medical Records Trustworthy? Observations on Copying, Pasting and Duplication.. AMIA Annual Symposium Proceedings. 2003;

[34] HL7 Standards - Primary Standards, available from http://www.hl7.org/implement/standards/product_section.cfm?section=1 (cited 19.5.2017)

[35] The DICOM Standard, available from http://dicom.nema.org/standard.html (cited 19.5.2017)

[36] Health Level SevenHL7 version 3 - Reference Information Model. International Organization for Standardization; 2014.

[37] Who is IHE?, available from http://www.ihe.net/FAQ/#Who_is_IHE? (cited 19.5.2017)

[38] Datový standard MZ ČR, available from http://ciselniky.dasta.mzcr.cz/ (cited 19.5.2017)

[39] Unified Medical Language System (UMLS), available from https://www.nlm.nih.gov/research/umls/ (cited 20.5.2017)

[40] About epSOS, available from http://www.epsos.eu/home/about-epsos.html (cited 26.4.2017)

[41] ASTM E2369-12 Continuity of Care Record. American Society for Testing and Materials;

[42] Häyrinenn K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inf. 2008;5

[43] What is openEHR?, available from http://www.openehr.org/what_is_openehr (cited 19.5.2017)

[44] CORDIS: The Good European Health Record, available from http://cordis.europa.eu/project/rcn/17093_en.html (cited 19.5.2017)

[45] Origins of openEHR, available from http://www.openehr.org/about/origins (cited 19.5.2017)

[46] Archetype Technology Overview, available from http://www.openehr.org/releases/AM/latest/docs/Overview/Overview.html (cited 19.5.2017)

[47] ISO EN 13606. International Organization for Standardization / European Committee for Standardization; 2008.

[48] ISO EN 13940. International Organization for Standardization / European Committee for Standardization; 2015.

[49] epSOS: Patient Summary, available from http://www.epsos.eu/epsos-services/patient-summary.html (cited 19.5.2017)

[50] D3.2.2 Final epSOS System Technical Specification, available from http://www.epsos.eu/uploads/tx_epsosfileshare/D3.2.2_Final_Definition_Functional_Service_Req_Patient_Summary.pdf (cited 19.5.2017)

[51] NIX-ZD, available from http://www.nix-zd.cz/ (cited 20.5.2017)

[52] Situace kolem registrů zavedených zákonem o zdravotních službách, available from http://www.sasp.cz/situace-kolem-registru-zavedenych-zakonem-o-zdravotnich-sluzbach (cited 20.5.2017)

[53] Přečková P. Jazyk českých lékařských zpráv a klasifikační systémy v medicíně. Eur J Biomed Inform. 2010;6(1)

[54] Get LOINC, available from https://loinc.org/downloads/ (cited 03.06.2017)

[55] International Classification of Diseases v. 10 (ICD10), available from http://www.cdc.gov/nchs/ (cited 19.5.2017)

[56] International language for drug utilization research, available from https://www.whocc.no/ (cited 3.6.2017)

[57] Converging Patient Summaries: Finding the Common Denominator between the European Patient Summary and the US-Based Continuity of Care Document. Eur J Biomed Inform. 2015;11(2)

[58] Medical Subject Headings, available from https://nlk.cz/pro-knihovny/mesh/ (cited 20.5.2017)

[59] SNOMED CT, available from https://www.nlm.nih.gov/healthit/snomedct/index.html (cited 19.5.2017)

[60] iSpell, available from ftp://ftp.tul.cz/pub/unix/ispell/ispell-czech-20040229.tar.gz (cited 2009)

[61] Zvára K, Kašpar V. Identifikace jednotek a dalších termínů v českých lékařských zprávách. Eur J Biomed Inform. 2010;6(1)

[62] Zvára K, Svátek V. Extrahovatelnost informací z českých lékařských zpráv. Eur J Biomed Inform. 2012;8(5)

[63] Přečková P, Zvárová J. The Role of International Nomenclatures and Standards in Travel Shared Health Care. Travel Health Informatics and Telehealth. 2009;

[64] Heid WD, Chasteen J, Forrey AW. The electronic oral health record. Contemp Dent Pract. 2002;3

[65] Zvárová J, Dostálová T, Hanzlíček P, Teuberová Z, Nagy M, Pieš M. Electronic health record for forensic dentistry. Methods Inf Med. 2008;47

[66] Chleborád K, Dostálová T (školitel). Disertační práce: Stav chrupu u hendikepovaných pacientů. Praha: Univerzita Karlova; 2014.

[67] Chleborád K, Zvára K, Dostálová T, Zvára K, Hippmann R, Ivančáková R et al. Evaluation of voice-based data entry to an electronic health record system for dentistry. Biocybernetics and biomedical engineering. 2013;33

[68] Zvára K, Tomečková M, Peleška J, Svátek V, Zvárová J. Tool-supported Interactive Correction and Semantic Annotation of Narrative Clinical Reports. Methods Inf Med. 2017;56(3)

[69] LOINC, available from https://loinc.org/ (cited 19.5.2017)

[70] Databáze léků, available from http://www.sukl.cz/modules/medication/search.php (cited 20.5.2017)

[71] "Málem jsem zabil bratra!" Nový překladač je lepší, chyby jsou vtipnější., available from http://technet.idnes.cz/google-prekladac-neuronove-site-google-translate-srovnani-test-p80-/sw_internet.aspx?c=A170509_130949_sw_internet_pka (cited 29.5.2017)

[72] Tange HJ, Hasman A, De Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inf. 1997;46

[73] Blobel B, Hasman A, Zvárová J. Data and Knowledge or Medical Decision Support. 2013.

[74] Hoerbst A. Exploring Complexity in Health: An Interdiscriplinary Systems Approach. 2016.

Výkladový slovník

DASTA: Datový standard Ministerstva zdravotnictví

DICOM: Digital Imaging and Communications in Medicine – standard pro předávání digitálních audiovizuálních dat

HL7: Označení pro standardy a organizace Health Level Seven – tvůrce standardů v oboru zdravotnické informatiky

LÉKY: Databáze léčiv a léčivých přípravků Státního úřadu pro kontrolu léčiv

LOINC: Logical Observation Identifiers: Names and Codes (Regenstrief Institute)

MKN 10: Mezinárodní klasifikace nemocí (verze 10)

NLP: Natural language processing – zpracování přirozeného jazyka

PoS Tagging: Part of speech tagging – označování částí textu (projevu)

SNOMED CT: Systematized NOmenclature in MEDicine – Clinical Terms (IHTSDO)

Přílohy

1. Článek Tool-supported Interactive Correction and Semantic Annotation of Narrative Clinical Reports
2. Článek Evaluation of voice-based data entry to an electronic health record systém for dentistry
3. Databáze transformací. (elektronická příloha, CSV soubor)