VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

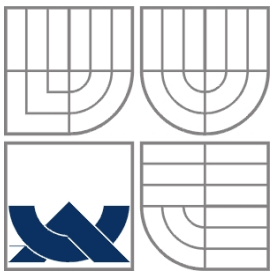
AUTOMATICKÁ TVORBA SLOVNÍKŮ Z PŘEKLADOVÝCH TEXTŮ

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

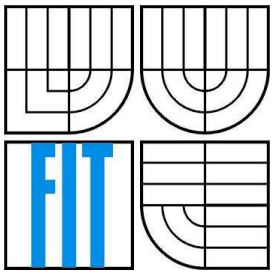
AUTOR PRÁCE
AUTHOR

Bc. JAKUB MUSIL

BRNO 2010



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA SLOVNÍKŮ Z PŘEKLADOVÝCH TEXTŮ

AUTOMATIC CREATION OF DICTIONARIES FROM TRANSLATIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JAKUB MUSIL

VEDOUCÍ PRÁCE
SUPERVISOR

PAVEL SMRŽ, doc. RNDr., Ph.D.

BRNO 2010

AUTOMATICKÁ TVORBA SLOVNÍKŮ Z PŘEKLADOVÝCH TEXTŮ

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Bc. Jakub Musil
26. 5. 2010

Abstrakt

Cílem této práce je vytvoření systému pro získání překladu slov zdrojového jazyka do jazyka cílového pomocí ekvivalentní dvojice vstupních textů. V této práci jsou popsány pojmy a metody používané v oblasti strojového překladu a strojové tvorby překladových slovníků. Práce také obsahuje návrh a popis jednotlivých fází, ze kterých se skládá vytvořený systém, včetně závěrečného testování, vyhodnocení získaných překladů a možnosti rozšíření existujícího překladového slovníku.

Abstract

Aim of this thesis is to implement system for translation words from source language into the target language with pair input texts. There are descriptions of terms and methods used in machine translation and machine build dictionary. The thesis also contains a concept and specification of each part created system including final evaluation. There is analysed options which make extension of existing dictionatry.

Klíčová slova

slovník, paralelní texty, korpus, strojový překlad, GIZA++, Hunalign, TreeTagger, PDT, Python

Keywords

dictionary, parallel texts, corpus, machine translation, GIZA++, Hunalign, TreeTagger, PDT, Python

Citace

Musil Jakub: Automatická tvorba slovníků z překladových textů, diplomová práce, Brno, FIT VUT v Brně, 2010

Obsah

1 Úvod

Člověk se již dávných dobách před několika tisíci lety dorozumíval pomocí nejrůznějších gest, zvuků a později pomocí ustálených jazyků. V době pravěku vznikl malý počet jazyků, který se neustále rozšiřoval. Sloužil pro komunikaci určité skupiny lidí. Ta neměla tušení o tom, jakými jazyky se dorozumívají lidé na druhém konci světa, natož potřebu vzájemné komunikace. A to zejména z důvodu neexistence moderních dopravních prostředků, které známe ze současné doby, které by umožňovaly styk jednotlivých kultur mezi sebou.

V současnosti existuje na světě více jak šest tisíc jazyků. Vlivem neustálé migrace lidí, rozmachem cestování, počítačů a nadnárodní sítě internet se jednotlivé kultury a jejich jazyky čím dál tím častěji prolínají a v popředí zájmu tak stojí potřeba vzájemné komunikace lidí mluvících odlišnými jazyky a překlad cizojazyčných textů.

Mezi nejčastěji používaný prostředek pro jazykový překlad patří dvojjazyčný slovník obsahující překlady výrazů jednoho jazyka do jazyka jiného. Vytvoření takovéhoto prostředku není triviální a může trvat i řadu let. Příčinou je rozsáhlost a rozmanitost jazyků, které se neustále přizpůsobují a dotváří. Rozvoj informačních technologií s sebou přináší nové možnosti tvorby a rozšíření těchto slovníků. Těmto možnostem je věnována tato práce, která si klade za cíl vytvoření překladového slovníku ze sobě odpovídajících si textů. Tento slovník může být použit sám o sobě, nebo může sloužit k rozšíření existujících slovníků.

V kapitole číslo 2, 3 a 4 jsou čtenáři přiblíženy pojmy a metody z oblasti zpracování přirozeného jazyka, postupy používané při zarovnání vět a slov a při stanovení základních tvarů jednotlivých slov. Kapitola číslo 5 je věnována nástrojů, které našly uplatnění při tvorbě systému, jehož návrh a realizace je popisována v kapitole číslo 6, která obsahuje také závěrečné testování a vyhodnocení obdržených výsledků.

2 Související pojmy

Automatická tvorba slovníků z překladových textů je odvětví, které souvisí s celou řadou oblastí a pojmů, týkajících se lingvistické oblasti a oblasti informačních technologií. V této kapitole je čtenáři přiblížena řada pojmů a metod, se kterými se v dalších částech této práce může setkat. Jsou jimi pojmy z oboru lingvistiky, charakteristika korpusů a přiblížení lemmatizace, která je nezbytná pro dosažení kvalitních výsledků.

2.1 Lingvistika

Je vědní disciplína zabývající se zkoumáním přirozeného jazyka. Často je označována jako jazykověda. Vědci zkoumající jazyk jsou nazýváni lingvisty. Obsahuje řadu podoborů. K jazyku lze přistupovat z mnoha úhlů pohledu a souvisí s ním řada dalších vědeckých oborů, které ovlivňují jeho studium. Patří mezi ně například psychologie, informatika, filosofie, biologie a další.

2.1.1 Cíle lingvistiky

Podle cílů lingvistiky ji můžeme rozdělit:

- Deskriptivní lingvistika – Jejím cílem je popsat jazykový systém a způsoby jeho užívaní v nejrůznějších situacích. Výsledkem jsou slovníky a mluvnice, které popisují slovní zásobu a strukturu konkrétního jazyka.
- Teoretická lingvistika – Zkoumá obecné principy fungování přirozeného lidského jazyka. Často využívá poznatky deskriptivní lingvistiky. Výstupem jsou explicitní, formálně zpracované teorie a hypotézy, jejichž platnost je poté testována s daty a vzory konkrétních jazyků.
- Aplikovaná lingvistika – Využívá znalostí z deskriptivní a teoretické lingvistiky pro řešení skutečných problémů. Do této skupiny patří jazyková terapie, výuka cizích jazyků a také strojový překlad.

2.1.2 Lingvistické disciplíny

Jazyk zkoumaný lingvisty je komplexní systém a skládá se z řady menších celků (podsystémů). Ty se mohou navzájem ovlivňovat, proto se deskriptivní a teoretická lingvistika dělí do následujících řady kategorií, některé z nich jsou:

- Lexikologie – Zkoumá slovní zásobu jazyka, jeho význam a užití jednotlivých slov (označováno jako lexik jazyka). Slovní zásoba jazyka představuje základní stavební blok jazyka, jehož jednotkami jsou slova a ustálená slovní spojení. Na slova můžeme nahlížet v rovinně nižší, na které jsou tvořeny z fonémů a morfému, jejichž prostřednictvím jsou zvukově, slovotvorně a tvaroslovně utvářena. Lexikologie je propojena se všemi ostatními jazykovými podsystémy.
- Morfologie – Je často označována jako tvarosloví. Je to vědní disciplína lingvistiky, která studuje skloňování a časování (jednotně ohýbání) a odvozování slov pomocí předpon, přípon a vpon. Zabývá se také strukturou jednotlivých slov. Nejmenší část slova, která je nositelem věcného nebo gramatické významu je morfém. Jedná se o základní a dále nedělitelnou jednotku. Aplikace poznatků této vědy zvaná morfologická analýza spočívá v dekompozici slov na jednotlivé morfémy a hledání vztahů mezi nimi.

- Syntax – Při snaze vyjádřit určitou informaci nám pouze existence slov jako výrazového prostředku nestačí. Slova bývají podle určitých pravidel sdružovány do větších větných celků. Do slovních spojení a vět. Právě syntaxe jazyka nám určuje ony pravidla, podle kterých slova do větných celků spojujeme. Encyklopedie [1] definuje syntax jazyka následovně: „Syntax neboli skladba je lingvistická disciplína zabývající se vztahy mezi slovy ve větě, správným tvořením větných konstrukcí a slovosledem". Syntax je často nahrazován pojmem větná vazba.
- Sémantika – Zabývá se významem jednotlivých slov, morfémů a znaků. Význam menších celků (slov) lze odvodit přímo ze sémantiky jazyka, význam větších celků, jako jsou věty, odvodíme z významů jeho dílčích celků.
- Fonetika – Zkoumá zvukový systém, zvukové charakteristiky slov (fóny), třídí je a klasifikuje. Tato věda nepatří pouze mezi lingvistické disciplíny, ale zasahuje také například do fyziky (akustika).
- Textová lingvistika – Zabývá se pravidly a principy využívaných při spojování vět a souvětí ve větší celky tvořící text

Kromě popisované obecné lingvistiky, která zkoumá jazyk jako takový, se můžeme setkat s podobory, které zkoumají jednotlivé rodiny jazyků. Pro český jazyk je to bohemistika, pro jazyk anglický anglistika a pro jazyky slovanského typu slavistika. Tyto obory necharakterizují pouze podobory lingvistiky, ale spíše podobory filologie.

2.2 Korpusová lingvistika

Korpusová lingvistika je disciplína lingvistiky zkoumající jazyk pomocí elektronických jazykových korpusů. Zabývá se tvorbou těchto korpusů, jejich zpracováním a využitím. Začala se výrazněji rozvíjet teprve koncem 20. století a to v souvislosti s prudkým rozvojem výpočetní techniky, který umožnil vznik rozsáhlých souborů jazykových dat v elektronické podobě. Tento fakt vedl k nutné spolupráci lingvistů s ostatními obory, zejména s oborem zabývajícím se výpočetní technologií a matematikou. Umožnil vznik korpusů o několik řádů rozsáhlejších a zbavil lingvisty nejistoty, zda závěry vyvozené zpracováním určitého korpusu nejsou degradovány vlivem malého rozsahu zkoumaného korpusu. Požadavky, které jsou kladeny na korpusová data:

- Nenáhodnost a věrnost dat, odpovídající použití jazyka běžnými lidmi
- Aktuálnost, data skutečně odráží svou dobu
- Dostatečný rozsah dat
- Objektivnost, neselektivnost
- Snadná získatelnost a manipulovatelnost (pomocí strojů)

První budované elektronické korpusy byly přirozeně menšího rozsahu. Obsahovaly zhruba jeden milion slov. S rostoucími možnostmi výpočetní technologie se jejich velikost začala zvyšovat. Průkopnickou zemí byla Velká Británie (dnes jsou v Británii prakticky všechny jazykové slovníky tvořeny s využitím korpusových dat). V této zemi se také začala konstituovat významná korpusová disciplína zvaná korpusová lexikografie. V současné době je největší britský korpus Bank of English, který již přesáhl hranici 500 miliónů slovních tvarů. Kolem tohoto korpusu vzniklo v Birminghamu významné slovníkové nakladatelství Cobuild. Dalším zajímavým korpusem je International Corpus of English. Jeho snahou je mapovat všechny ve světě užívané varianty anglického jazyka a porovnat jejich odlišnosti. Celkové množství existujících korpusů prakticky není možné přesně odhadnout. V evropských zemích bychom jen s těží hledali jazyk, pro který by žádný korpus nebyl vytvořen.

2.2.1 Korpus

Dokument [2] definuje korpus jako „nejlepší aproximaci, nejvěrnější vzorek skutečného jazyka i veškeré informace, které jazyk zprostředkovává, a vychází tak z přesvědčení, že lépe než prostřednictvím korpusu nelze dnes jazyk při studiu uchopit“.

Korpus je souhrn textů uležených v digitální (počítačové) podobě, který slouží pro jazykové rozbory a jazykové výzkumy. Jsou to strukturované, unifikované, často označkované texty, které jsou uložené v elektronické podobě. Pro značkování bývá běžně použito uznávaného formátu SGML (Standard Generalized Markup Language) a využívá se zásad iniciativy TEI, která představuje unifikovanou sadu instrukcí předepisující způsob kódování textů a jejich analýzu pomocí jazyka SGML. Korpusy se využívají se v celé řadě lingvistických oborů. A to především pro studium slov, jejich významů a kontextu, v kterém byly použity.

Pro práci s korpusy můžeme využít speciální programy, umožňující v korpusech vyhledávat podle specificky vybraných parametrů. Tyto programy pracují s dříve zmíněným značkováním textu. Jedná se o specifický aspekt, který rapidně zvyšuje užitnost korpusů. Základem značkování je tzv. vnější anotace, která obohacuje korpus o informace spojené s vznikem korpusu. Jsou jimi například rok vzniku, informace o textech, z kterých byl vyhotoven včetně jejich autora. Dalším typem značkování je vnitřní anotace, která do korpusu zavádí strukturní informace a zvyšuje tím jeho užitnou hodnotu. Patří mezi ně informace o členění textu na jednotlivé kapitoly, odstavce, věty, slova a informace lingvistické. Zvláště u jazyka flektivního neboli jazyka, který vyjadřuje slovní druhy pomocí skloňování, časování, předpon a přípon, je lingvistická anotace velice náročná a drahá. Z tohoto důvodu se v praxi omezuje nejčastěji na morfologické značkování slovních tvarů (tzv. tagování), které zahrnuje přiřazení odpovídajícího slovního druhu, nebo také přiřazení základního tvaru slova (lemma tvar). Formát, v němž jsou potřebné informace v drtivé většině dnes budovaných korpusů přidávány k textům, je standardizován.

2.2.2 Typy korpusů a jejich kategorizace

Korpusy se od sebe liší svou velikostí, rozsahem, typem, jazykem, zdrojem textů, anotací, určením, časovou oblastí, značkováním atd. Tyto vlastnosti umožňují jejich rozdělení do celé řady skupin a kategorií.

Základní kategorizací je rozdělení na textové a zvukové (mluvené) korpusy. Textová forma je jednoznačně převažující skupinou. Jejich základními zdroji jsou nejrůznější texty, knihy, časopisy, zápisy, ustanovení a také přepisy hovorové řeči, rozhovorů, televizního vysílání. Mluvený korpus je velice nákladný. Příčinou je skutečnost, že úsilí, které je nutno do vytvoření mluveného korpusu vložit je několikanásobně vyšší než u korpusů psaných. Z tohoto důvodu vznikající zvukové korpusy jsou malého rozsahu a obecnější charakteristiky jazyka plně nedokumentují, ale pouze naznačují.

Mezi často používané rozdělení patří skupiny synchronních a diachronních korpusů. Toto rozdělení bychom mohli označit jako rozdělení podle časového období. Převažujícím typem jsou synchronní korpusy, které jsou založené na současných psaných textech, jejichž analýza je nejpotřebnější a jsou díky digitalizaci zároveň nejdostupnější. Vývoj jazyka je pozvolný a nejsou v něm zpravidla žádné pevně dané časové hranice, proto se k tvorbě synchronních korpusů využívají texty, které zahrnují několik posledních desetiletí. Na druhé straně stojící diachronní korpus pokrývající několik vývojových bloků daného jazyka, v některých případech také jeho celý dokumentovatelný vývoj.

Zpravidla se každý korpus zabývá určitou skupinu textů. Tato skutečnost nám umožňuje následující rozdělení textu podle obsahu na skupiny:

- Nářeční korpusy

- Studijní korpusy
- Korpus básnických textů
- Technické korpusy
- Korpusy lexikografických děl
- Paralelní korpusy
- Cvičené a testovací korpusy

2.2.3 Paralelní korpusy

Právě existence paralelních korpusů vede ke vzniku tohoto textu. Jedná se o korpusy dvou nebo více jazyků, které jsou vytvářeny z překladů jednoho jazyka do jazyka jiného. Obsahují tedy zdrojový neboli originální text a jeho překlad do cizojazyčné verze. Paralelní vícejazyčné texty jsou v současné době nejvíce využívané korpusy. Umožňují porovnání jednotlivých jazyků, jejich rozmanitosti, kterou lze i při prvním pohledu na vzájemně ekvivalentní větné celky odhalit. Pokusy rovněž ukazují cennost autentických překladových materiálů pro zvýšení vzdělanosti studentů a překladatelů. Umožňují automatický strojový překlad a jeho kontrolu. Přináší možnost rozšíření existujících překladových slovníků, kterým se zabývá praktická část této práce. Nevýhodou paralelních korpusů je velmi často nesprávný překlad textů, některé překlady nejsou zcela autentické nebo mají odlišnou strukturu. Základním předpokladem pro jejich správné využití je zarovnání nejčastěji na úrovni odstavců a vět. K dosažení tohoto zarovnání se používají specializované nástroje, které využívají znalosti v dané problematice, ale i přesto jejich výsledky nejsou nikdy stoprocentní a vyžadují často dodatečné korekce.

2.2.4 Český národní korpus

Český národní korpus (dále zkratka ČNK) je souvislý projekt obsahující množinu jednotlivých korpusů. Jednotlivé korpusy mapují a zaznamenávají různé podoby českého jazyka s cílem zpřístupnit uživatelům co nejširší zdroj jazykových dat včetně nástrojů k jejímu využití. Předpokládá se jeho využití v širokém spektru uživatelů, ve skupině pedagogů, studentů, lingvistů atd. Jedná se o nekomerční akademický projekt, který byl založen v roce 1994 na Filozofické fakultě Univerzity Karlovy. Podnětem k jeho budování byla zesilující potřeba vybudovat dostatečnou materiálovou základnu pro tvorbu nových a kvalitnějších slovníků češtiny, gramatiky a dalších jazykových standardů.

ČNK můžeme rozdělit na synchronní a diachronní část. Velikostí dominující korpus je synchronní psaný korpus SYN2000, která obsahuje více než 100 miliónů slovních tvarů. Z tohoto korpusu vychází korpus PUBLIC, který je široké veřejnosti dostupný na internetu. ČNK obsahuje také mluvené korpusy, například Pražský mluvený korpus a Brněnský mluvený korpus. Mezi zástupce paralelního korpusu patří InterCorp, jehož cílem je pokrytí co největšího množství různých jazyků.

2.2.5 Paralelní korpus Kačenka

Projekt paralelního anglicko-českého korpusu Kačenka vznikl v roce 1997 (často označován jako Kačenka1997) s přispěním Ministerstva školství. S Oficiálním názvem „Elektronické nástroje pro kombinované a distanční studium angličtiny“. Cílem tohoto projektu bylo vytvoření anglicko-českého paralelního korpusu menšího rozsahu, který by umožnil lingvistům, překladatelům a studentům jazyka pracovat s komplexními texty a ne pouze s omezenými úryvky. Podařilo se sestavit korpus obsahující přibližně 3 milióny slov. Dalším pokračováním tohoto projektu je paralelní korpus

Kačenka 2, který vznikl o pět let později. Jeho cíle jsou o poznání rozsáhlejší. Patří mezi ně především výrazné kvantitativní zvětšení, získání a rozdělení překladových textů z jednotlivých oborů, vytvoření kvalitních nástrojů a software pro zpracování vytvořeného korpusu a návrh využití při kurzech výuky anglického jazyka.

Paralelní texty z tohoto korpusu byly využity v praktické části této práce při budování systému získávajícího překlady ze vstupních textů.

3 Strojový překlad

Strojovým překladem rozumíme proces automatického překladu z jednoho jazyka do jazyka druhého-cílového. Jedná se o obor informatiky, který je na svém vzestupu a v dnešní multikulturní době je velice potřebným. V současnosti je dostupná řada systémů, které strojový překlad provádějí. Jejich výstup není dokonalý, ale je dostatečně kvalitní pro použití v mnoha oblastech a pomáhá tak zejména lidem, kteří potřebují překládat nejrůznější texty. Ať už se jedná o emailovou zprávu, internetový článek, uživatelský manuál, nebo rozsáhlý strukturovaný text.

Pokusy o strojový překlad začaly krátce po druhé světové válce. Očekávalo se, že pro počítače, jejichž první prototypy byly v té době zprovozněny, nebude tento problém nikterak složitý. Brzy se však ukázalo, že tento předpoklad je zcela mylný. První „funkční“ systém byl uveden v roce 1954 firmou IBM. Událost vzbudila značný zájem médií a široké veřejnosti. Systém byl z dnešního pohledu velice jednoduchý. Obsahoval pouze 250 slov a při předvádění překládal pouze několik málo připravených vět z ruštiny do angličtiny. Systém i přes svou jednoduchost vzbuzoval dojem, že praktické nasazení strojního překladu je otázkou několika měsíců, což pomohlo získat další finance pro následný výzkum v tomto oboru. První systémy pro strojový překlad byly využívány především v době studené války. USA se obávala náskoku Sovětského svazu ve zbrojním průmyslu, a proto se snažila získat překlad ruskojazyčných odborných a vědeckých publikací do angličtiny. Získané překlady ovšem nebyly kvalitní a tak jejich využití bylo jenom na úrovni základní orientace a vedla k rozhodnutí, zda je text natolik přínosný, že si zaslouží překlad pomocí profesionálního překladatele.

V průběhu posledních desetiletí bylo do vývoje strojového překladu investováno značné úsilí a finanční prostředky. Dosahované pokroky byly a jsou po celou dobu poměrně skromné a ani v dnešní době nejmodernější systémy nedokáží překládat na úrovni profesionálních překladatelů.

Metody strojového překladu můžeme rozdělit na metody řízené pravidly, založené na příkladech a metody statické.

3.1 Překlad řízený pravidly

Tato metoda vychází z předpokladů, že jazyk může být do značné míry popsán jistým souborem pravidel, které lze formálně definovat a uplatnit tak při automatickém generování překladů. Hlavním předpokladem a také problémem v jednom je existence jazykových pravidel. V případě, že bychom byly schopni jazyk zcela popsat formálními pravidly, které by jej jednoznačně a kompletně popisovaly, byla by právě tato metoda tou, která by nabízela dokonalý překlad textu ze zdrojového do cílového jazyka. Bohužel v praxi takovýto jazyk jen s těží nalezneme. Mezilidská komunikace je velice pestrá na nejrůznější vazby a slovní spojení. Hledání formalismu popisující běžný jazyk se tak stává neřešitelným problémem. Z tohoto důvodu se vytvořily postupy pro formální popis jazyka pouze v omezené podobě.

Nejjednodušší variantou strojového překladu patřící do této skupiny je překlad pomocí slovníků. Ten spočívá v pouhém narazení slov zdrojového jazyka ekvivalentními slovy jazyka cílového podle existujícího slovníku. Výsledky jsou ovšem poměrně nekvalitní a vyžadují řadu úprav a korekcí.

V pokročilejších technikách je překlad tvořen dekódováním zdrojového textu. Tvoří ji zejména morfologická, syntaktická analýza a lemmatizace. Dalším krokem je výběr slov cílového jazyka s využitím slovníku a volba jeho správných slovních tvarů tak, aby vyhovovaly cílové syntaxi.

3.2 Metody založené na příkladech

Tyto metody pracují s existující bází znalostí. Ty jsou vyjádřeny pomocí paralelního korpusu, v kterém jsou zarovnány odpovídající si věty. Tato data jsou analyzována a získávají se z nich menší celky na základě podobnosti zdrojových a cílových vět. Při potřebě přeložit větu jednoho jazyka do jazyka druhého se hledá podobnost s větami obsaženými v použitém korpusu a z jejich možných překladů se vybírají nejvhodnější vzory, z kterých se vzájemně poskládá věta cílová. Kvalitních výsledků dosáhneme především při překladu stejných druhů textu, které se váží k jednomu oboru.

3.3 Statistické metody

Jsou to velice populární metody, které využívají statistických modelů. Podobně jako v předchozí popisované metodě založené na příkladech i zde slouží jako báze znalostí dvoujazyčný korpus. Výhodou je nezávislost tohoto řešení na použitých jazycích. Odpadá zde také nutnost náročného sestavování pravidel jazyka, jako je tomu u překladu řízeném pravidly.

Rostoucí rozvoj dvoujazyčných strojem čitelných textů na počátku 20. let 20. století podnítil zájem o rozvoj metod pro získání jazykově cenných informací z těchto textů. Řada výzkumů a prací na toto téma dokázala, že je možné získat kvalitní zarovnání dvoujazyčného páru vět v paralelním textu bez znalosti a bližšího zkoumání jednotlivých slov ve větách obsažených a gramatické struktury daných vět. První statistické algoritmy autorů Browna, Laie a Mercela z roku 1991 byly založeny na porovnání počtu slov vyskytujících se v jednotlivých větách. Konkurenční algoritmus téže doby autorů Galeho a Churchila pracoval zase s počtem znaků, ze kterých jsou samotné věty tvořeny [5]. Tito autoři pracovali především s anglicko-francouzskými dvojicemi paralelních textů. Ikdyž byly tyto první metody založené pouze na statistickém porovnání celočíselných hodnot velice jednoduché, lze s nimi dosáhnout poměrně uspokojivých a zajímavých výsledků.

Společnou vlastností nejrůznějších statistických modelů je práce s matematickou pravděpodobností. Základním předpokladem je, že libovolné věty zdrojového jazyka můžeme překládat na různé věty jazyka cílového s určitou pravděpodobností správnosti onoho výběru cílové věty. Na následujících řádcích následuje popis standardního zápisu používaných pravděpodobností.

P(s) – udává pravděpodobnost výskytu dané věty *s*. Například pro větu „I like school“ značí pravděpodobnost, že určitá osoba v určitém čase vyřkne větu „I like school“. Oborem hodnot, které může tato pravděpodobnost nabývat je interval od čísla 0 až po číslo 1.

P(s|t) – podmíněná pravděpodobnost neboli pravděpodobnost *f* vzhledem k *e*. Pro příklad použijme opět větu „I like school“ a českou větu „Modrá je dobrá“. Podmíněná pravděpodobnost P(s|t) vyjadřuje míru šance, že překladatel přeloží zmíněnou anglickou větu na českou „Modrá je dobrá“, což by v tomto případě měla být šance nulová.

P(s, t) – spojená pravděpodobnost. Udává pravděpodobnost, že věty *s* i *t* nastanou. Jestliže věty *s* a *t* se vzájemně neovlivňují, můžeme psát P(s,t) = P(s)*P(t). V opačném případě, kdy se vzájemně ovlivňují, vypočteme spojenou pravděpodobnost jako P(s, t)=P(s)*P(t|s). Jestliže jsou věty *s* a *t* významově ekvivalentní řetězce, pak se vzájemně ovlivňují.

Na strojový a nejenom na strojový překlad cizojazyčných řetězců můžeme nahlížet jako na reversní inženýrství, kdy známe důsledek (překlad) a snažíme se nalézt pokud možno co nejoptimálnější podnět k jeho vzniku tak, abychom při jeho opětovném překladu došli opět

k původnímu důsledku (překladu). Při použití určité míry abstrakce můžeme tento proces přirovnat ke kriminální scéně a jejímu vyšetřování. Tato scéna je tvořená osobou s, která se rozhodla spáchat zločin a následně jej provedla. Naše vyšetřovací uvažování bude spočívat v uvažování o osobě, která měla dostatečný motiv pro spáchání zločinu. V naší notaci se jedná o určení P(s). Druhou vyšetřovací fází je prozkoumání konkrétní osoby s, její možnosti přístupu k použité zbrani, její dopravní dostupnosti. Jedná se o pravděpodobnost P(t|s). Tyto dvě fáze mohou vést ke sporům, které jsou způsobené osobami, které mají dostatečný motiv, ale nedisponují použitou zbraní nebo naopak.

Dalším abstraktním přirovnáním je lékařství a výskyt nemocí. Vyskytují se zde faktory nemoc a symptomy nemoci. Pravděpodobnost výskytu nemoci získáme například z historie záznamů pacientů a pravděpodobnost výskytu určité nemoci v závislosti na diagnostikovaných příznacích je také většinou známá.

V roce 1949 Warren Weaver navrhl využití statistických a kryptoanalytických technik k řešení problému strojového překladu textů z jednoho jazyka do jazyka druhého. Toto úsilí bylo záhy zmařeno zejména z důvodu nedostatečného výpočetního výkonu počítačů, který se v této době ani zdaleka nepřibližoval výkonu počítačů dnešní doby. Z tohoto důvodu byla řada nápadů zakonzervována a využita až v posledních letech, kdy se strojový výkon rapidně zvýšil.

Při překladu řetězce slov zdrojového jazyka do podoby řetězce slov jazyka cílového můžeme využít několika různých přístupů. Základní myšlenkou statistického překladu ovšem nadále zůstává fakt, že každý řetězec *t* je možným překladem řetězce *s*. Tuto myšlenku můžeme vyjádřit pomocí výše popsané podmíněné pravděpodobnosti P(s|t) udávající, pravděpodobnost, že věta *t* je překladem právě věty *s*.

Na proces tvorby věty se můžeme dívat jako imaginární na transformaci zdrojové věty na věty cílovou, která probíhá v hlavě mluvčího. Na výsledný výstup (psaná, mluvená forma) se dostane pouze věta cílová. Chceme-li získat zpět onu zdrojovou větu, kterou měl mluvčí v hlavě, musíme zohlednit sémantiku použitých slov (slovníkový překlad) a také gramatickou vazbu mezi konkrétními jazyky. Pomocí Baysova teorému můžeme psát následující vzorec:

$$P(s|t) = \frac{\mathrm{P(s)} * \mathrm{P(t}|s)}{\mathrm{P(t)}} \tag{1}$$

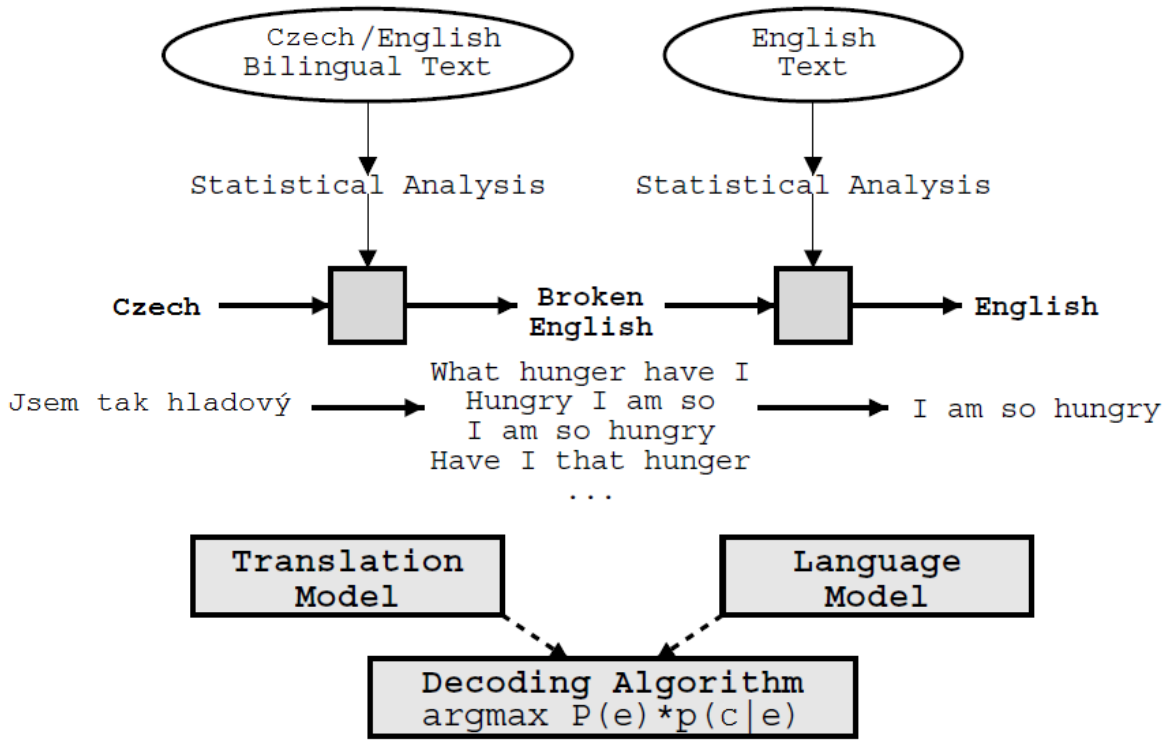
Kde P(s) a P(t) udává míru správnosti zdrojové respektive cílové věty vzhledem ke gramatickým pravidlům konkrétního jazyka. Tyto hodnoty bývají často označovány jako jazykový model. P(t|s) udává vzájemnou překladovou podobnost vět *s* a *t*. Snahou je nalézt pro konkrétní větu t nejvhodnější větu s tak, aby pravděpodobnost P(s|t) byla co největší. Protože P(t) je pevně dané a neměnné, můžeme tohoto jmenovatele z vzorce vypustit a naše úloha tak přechází do následujícího tvaru hledajícího maximum ê.

$$\hat{\mathrm{e}} = argmax\ P(\mathrm{s}) * P(t|s) \tag{2}$$

Tato rovnice číslo 2 je nejvíce odpovídající skutečnosti, kdy lidský překladatel překládá věty odlišných jazyků. Z technického hlediska je podmíněná pravděpodobnost P(t|s) pouze obrovská tabulka reálných hodnot nula až jedna sdružených s každým možným párem zdrojové a cílové věty. Její správné sestavení je základním předpokladem k získání kvalitního zarovnání. V praxi se při sestavování takovéto tabulky nezahrnují kombinace každé zdrojové věty s každou větou cílového jazyka, ale zpracování provádíme pouze nad jednotlivými částmi vstupního textu. Například pouze nad jednotlivými odstavci. Popisovaná rovnice obsahuje tři výpočetní problémy a nabízí tak možnosti odlišných přístupů. Jsou jimi výpočet jazykového modelu P(s), sestavení podmíněné

pravděpodobnosti P(t|s) a provedení samotného maximalizačního algoritmu. Při bližším pohledu na tento vzorec můžeme dojít k otázce, proč nehledáme na místo dvojice P(s) a P(t|s) rovnou požadované P(s|t). Odpovědí je skutečnost, že překládané texty obsahují gramatické chyby, chybějící slova a podobně. Při přímém hledání pravděpodobnosti P(s|t) bychom tuto skutečnost nebrali v potaz a docházelo by tak k případům, kdy by byl upřednostněn překlad gramaticky chybný (např. doslovný překlad) před překladem správným.

K tvorbě překladu tedy můžeme využít dva různé přístupy. Prvním z nich je využití pouze pravděpodobnosti P(s|t). Porovnáváme konkrétní slova ve větách, přičemž na pořadí slov ve větě nebereme zřetel. Na první pohled tento přístup neodpovídá procesu tvorby věty v hlavě mluvčího, jak bylo popsáno v předchozích odstavcích. Při překladu slova po slově bychom sice cizojazyčnou větu dostali, ale s gramatickými zásadami konkrétního jazyka by překlad neměl příliš společného. Druhým přístupem je využití Baysovského teorému (viz vzorec 1 a 2) ve kterém k podmíněné pravděpodobnosti přidáme pravděpodobnost P(s) vyjadřující gramatickou správnost konkrétní věty *s* (tento model znázorňuje obrázek číslo 1). Jako příklad uveďme překlad věty „chlapec běží“. Podmíněná pravděpodobnost nám jako možné překlady vrátí věty, které mají dostatečný doslovný překlad, např. „the boy run“, „run the boy“ a „boy the run“. Pravděpodobnost gramatické správnosti nám poté vybere tu větu, která je z gramatického pohledu nejsprávnější. P(s) model je také užitečný při výběru vhodných slov při samotném překladu. A to zejména z důvodu víceznačnosti jednotlivých slov, kdy jedno slovo zdrojového jazyka můžeme přeložit na více slov jazyka cílového. Každý z těchto možných překladů ovlivní gramatickou správnost výsledné věty a my tak použijeme tu variantu, která zajistí nejvyšší míru gramatické správnosti výsledné věty.



obrázek 1: Model systému podle Baysovského teorému

Při realizaci popisovaného systému je potřeba nejprve vytvořit nástroj, který libovolné větě *s* přiřadí hodnotu P(s). Označovaný také jako language model. Mohli bychom jej budovat jako černou skříňku, která zná strukturu slov jazyka, používané gramatické vazby, větné spojení a běžně

používané jazykové obraty. Tvorba takovéhoto komplexního systému by byla velice náročná. Další o mnoho jednodušší myšlenkou je prosté zaznamenání každé věty, s kterou v daném jazyce přijdeme do styku. Budujeme tak rozsáhlou databázi, ve které pravděpodobnost výskytu dané věty *s* je rovna jejímu počtu výskytu vzhledem k počtu všech záznamů uložených v databázi. Problémem je fakt, že větám, které doposud v databázi uloženy nejsou je přidělena hodnota nula i přes jejich gramatickou správnost.

Pro strojové zpracování jednotlivých vět se ukázalo jako vhodné jejich rozdělení na podřetězce neboli n-gramy. Pro n=3 hovoříme o trigramech, pro n=2 jako o bigramech a v případě n=1 o digramech, nebo jednoduše o slovech. U bigramového rozdělení se setkáme s pravděpodobností B(y|x), která značí pravděpodobnost, že slovo *y* následuje po slově *x*. Výsledná pravděpodobnost je podíl celkového výskytu slov *xy* vzhledem k počtu výskytu slova *x*.

$$B(y|x) = \frac{\text{počet\_výskytů(xy)}}{\text{počet\_výskytů(x)}} \tag{3}$$

Tento model říká, že člověk při tvorbě věty si pomatuje pouze posledně řečené slovo a k němu přidává slovo nové, což jistě není přesný model reálného procesu tvorby vět. Pro přiblížení reálnému procesu byl vytvořen trigramový model, který počítá se skutečností, že si mluvčí pomatuje poslední dvě vyřčená slova. Pravděpodobnost je tak upravena do následujícího tvaru:

$$B(z|xy) = \frac{\text{počet\_výskytů(xyz)}}{\text{počet\_výskytů(xy)}} \tag{4}$$

Kladnou stránkou bigramového a trigramového modelu je fakt, že přiřazuje nenulové pravděpodobnosti pro věty, které doposud nebyly zpracovány. Stačí, aby pouze některý z jejích n-gramů již byl zpracován. Dojde-li k tomuto, můžeme využít tzv. smoothing (vyhlazování). Vyhlazování říká, že pokud slovo *z* doposud nikdy nenásledovalo *xy* v našem textu, položí si další otázku, zda *z* následovalo alespoň *y*. Jestliže ani takto nenalezneme žádný existující bigram můžeme zkoumat, zda slovo *x* je vůbec korektní slovo zpracovávaného jazyka. Tento algoritmus můžeme vyjádřit pomocí vzorce číslo 5 (zkratka poč_v v tomto vzorci znamená počet výskytů).

$$B(z|xy) = 0{,}95 * \frac{\text{poč\_v(xyz)}}{\text{poč\_v(xy)}} + 0{,}04 * \frac{\text{poč\_v(yz)}}{\text{poč\_v(z)}} + 0{,}008 * \frac{\text{poč\_v(z)}}{\text{poč\_v\_všech\_slov}} + 0{,}002 \tag{5}$$

Použité číselné koeficienty označujeme jako vyhlazovací koeficienty. Pro dosažení pokud možno co nejlepších výsledků je vhodné s těmito koeficienty experimentovat a upravovat je průběžně podle zpracovávaného textu a jednotlivých slov. Poslední koeficient v pořadí 0,002 zajišťuje, že žádná vypočtená pravděpodobnost nebude nulová.

Mohli bychom také sestrojit model, který nepředpokládá tvorbu věty ve stylu slovo za slovem v kontinuálním pořadí, ale který by vycházel z tvorby ve stylu nejprve sloveso, poté podstatné jméno ležící vlevo od slovesa, dále přídavné jméno ležící před podstatným jménem a poté podstatné jméno ležící na pravé straně od slovesa a tak podobně.

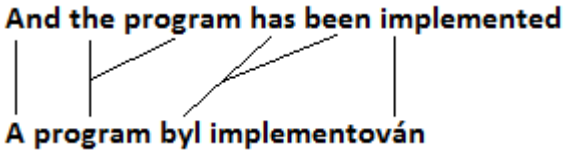
Výpočet klasických n-gramů je velice jednoduchý, pouze počítáme frekvenci výskytů jednotlivých n-gramů a dělíme ji. Model sloveso-podstatné jméno-přídavné jméno již tak jednoduchý na zpracování není. Musíme být schopni identifikovat hlavní sloveso věty a její další větné členy. I kdyby se nám podařilo toto rozpoznání kvalitně implementovat, není zaručeno, že model bude produkovat lepší výsledky než klasický kontinuální n-gramový model. Otázkou zůstává, jak můžeme

posoudit, zda je jeden model lepší než model druhý. Standardní cestou je nashromáždění doposud nezpracovaných dat, která nám slouží jako data testovací. Zpracujeme je pomocí konkrétního modelu a posoudíme obdržené výsledky s předpokládanými výsledky. Čím je jejich rozdíl menší, tím model produkuje kvalitnější výstup.

3.3.1 Statistické zarovnání slov

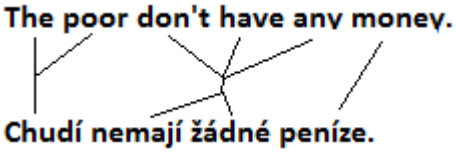
Na zarovnání slov můžeme nahlížet jako na další krok po zarovnání. Zarovnáním slov rozumíme propojení slov zdrojové věty se slovy věty cílové, která je jejím překladem.

V roce 1990 nastínil Brown myšlenku zarovnání jednotlivých párů slov, kde každé slovo zdrojového jazyka je překladem slova v originální zdrojové větě. Toto propojení znázornil graficky podobně jako na obrázku číslo 2, který ukazuje propojení slov anglické věty s ekvivalentní větou českou. Tento překlad ukazuje 6 různých propojení. V návaznosti na toto grafické vyjádření můžeme zapsat nalezená zarovnání ve tvaru *(A program byl implementován | And(1), the(2), program(2), has(3), been(3), implemented(4))*.



obrázek 2: Zarovnání slov

V tomto případě je každé slovo zdrojové věty (anglická věta) spojeno s jedním (případně žádným) slovem věty cílové. Jsou ale také páry vět, ve kterých spojení není jednoznačné, a spojení se mohou vzájemně překrývat. Tuto skutečnost zachycuje příklad zarovnání na obrázku číslo 3.

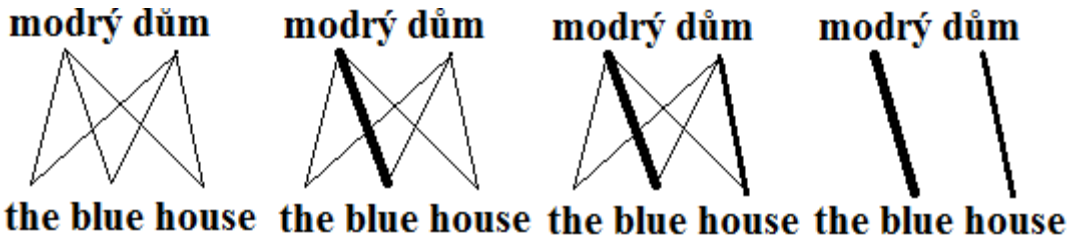


obrázek 3: Zarovnání slov

Tuto větu bychom moli zapsat *jako (Chudí nemají žádné peníze| The(1), poor(1), don't(2,3), have(2,3), any(2,3), money(4))*. Při párování vět nacházíme také slova, které se přímo na významu přeložené věty nepodílí a zastupují pouze syntaktickou roli ve větě. Patří mezi ně především určité a neurčité členy v anglické gramatice.

Jestliže má zdrojová věta *s m* slov a věta cílová *t* slov, lze nelézt *l*m* různých slovních párů těchto slov. Počet všech možných zarovnání je tedy 2^{lm}.

Z uvedeného grafického vyjádření propojení mezi slovy zdrojové a cílové věty vychází *EM algoritmus*. Jedná se o algoritmus, který pracuje v jednotlivých iteracích. Algoritmus se s každým zpracovaným párem vět učí a při dalším zpracování by tak měl být schopen dosahovat lepších výsledků. Jeho prvním krokem je propojení každého slova s každým se stejnou mírou Algoritmus takto postupuje dále, vybírá překlady pro doposud nepřeložené slova a konverguje k cíli. Po překladu všech zdrojových slov přichází na řadu odhad parametrů (pravděpodobností) pro jednotlivé použité překlady, které se využijí při dalším překladu. Postup je naznačen na obrázku níže.



obrázek 4: EM algoritmus

Další velice zajímavou metodou je maticové zarovnání, kterým se setkáváme pod označením Word Alignment. Sloupce matice jsou označeny jednotlivými slovy zdrojové věty a její řádky slovy věty cílové. Rozměrem matice je tak počet slov věty cílové*počet slov věty zdrojové. Po sestavení matice jsou její buňky prázdné, což symbolizuje, že žádné zarovnání doposud nebylo nalezeno. Případné vyplnění některé z buněk značí zarovnání slova označující sloupec buňky se slovem označující řádek buňky.

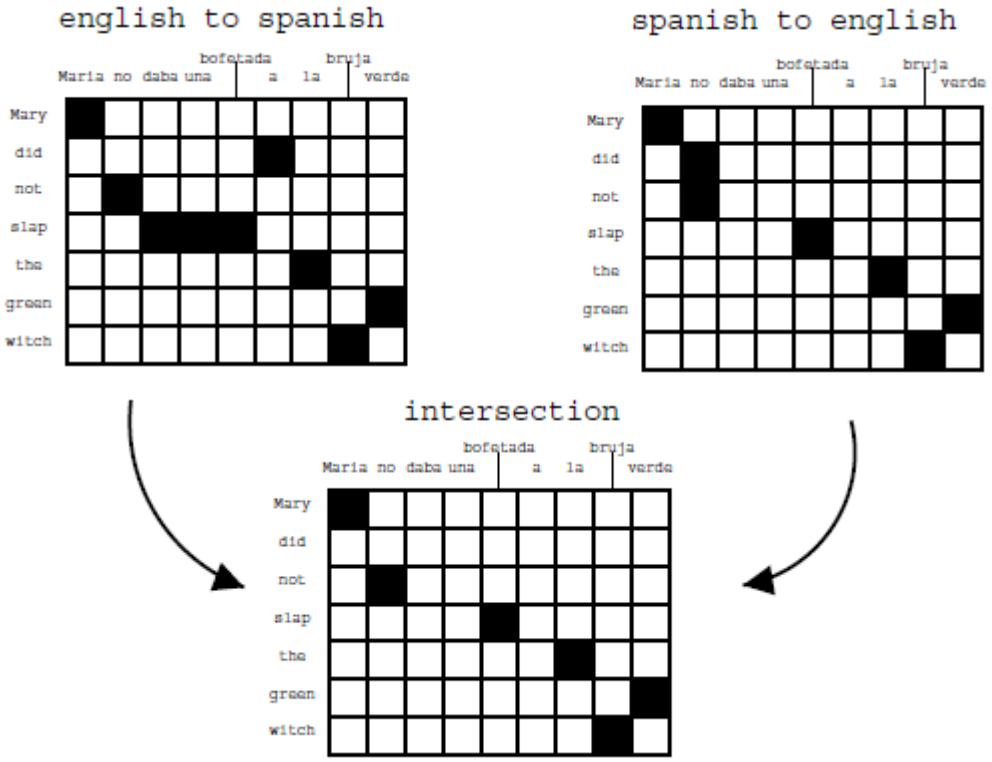
V optimálním případě, kdy by jednotlivé jazyky měly stejnou gramatickou stavbu vět a shodné větné spojení. Potom by se jednalo o matice čtvercového typu a vybrané zarovnání slov (označené buňky) by tvořilo diagonálu matice. Tento případ zobrazuje obrázek číslo 5.

obrázek 5: Optimální maticové zarovnání

V běžně vyskytujících se světových jazycích tato situace bohužel běžná není. Setkáváme se například s páry vět odlišných jazyků, které jsou významově shodné, ale jejich počet slov je odlišný. Pro většinu slov existují desítky možných překladů do cílového jazyka. Z těchto důvodů se setkáváme s maticemi, které v jednom řádku, případně sloupci, mají označeno více buněk znamenajících více možných zarovnání jednoho slova. V praxi běžnou podobu matice zarovnání ukazuje obrázek číslo 6 (obrázky číslo 6, 7, 8, 10, 11 a 12 vyskytující se v této kapitole byly převzaty z publikace [6]), který ukazuje zarovnání mezi španělskou a anglickou větou.

obrázek 6: Zarovnání francouzské věty k větě anglické

Možným vylepšením tohoto modelu je využití obousměrného zarovnání slov a takto získaných matic. Z jazyka zdrojového do jazyka cílového a naopak. Pro každý směr zarovnání sestavíme jednu matici a provedeme jejich průnik, který nám zredukuje počet označených buněk. Tento případ je znázorněn na obrázku číslo 7.

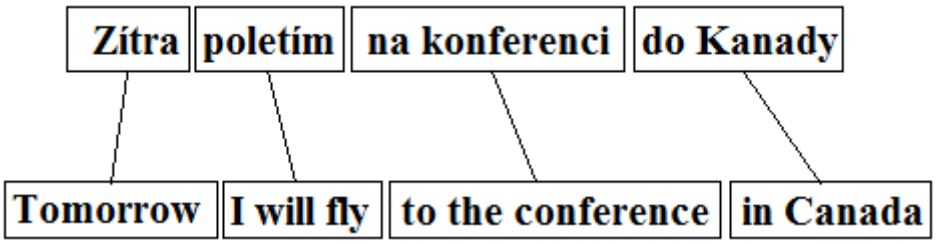


obrázek 7: Využití obousměrného zarovnání

Po provedení tohoto průniku máme jistotu, že zarovnání, které v matici zůstaly, jsou dostatečně kvalitní a jednoznačné. Problém je vznik řádku a sloupců, které neobsahují žádné zarovnání (nemají žádné označené buňky). Z tohoto důvodu přichází na řadu metoda zvaná jako Grow Additional Alignment Point, která se snaží tyto chybějící zarovnání doplnit. Při její implementaci lze využít řadu heuristických metod, jako například rozšíření zarovnání na své sousedy (pouze ve směru vodorovné osy, horizontální osy, nebo v obou, i diagonálně), povolit rozšiřování i na vzdálenější buňky, preferovat jeden směr překladu před druhým atd.

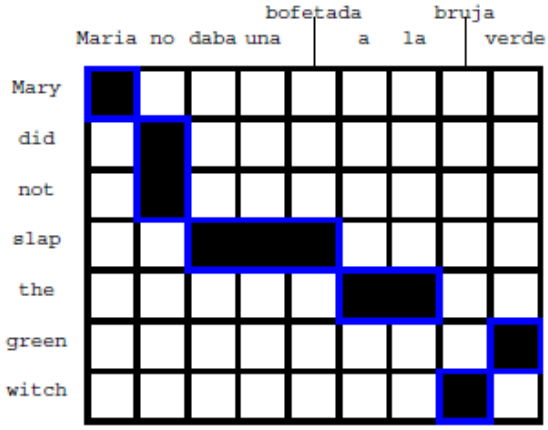
obrázek 8: Rozšíření zarovnání na prázdné řádky a sloupce

Jako zajímavý přístup k zarovnání slov se ukázalo využití frází, které se ve větách často opakují. Frázemi rozumíme často vyskytující se věty a slovní obraty. Zpracovávané věty se nejdříve rozdělí na jednotlivé fráze, které se přeloží a získané překlady se poté přeskládají, aby vyhovovaly zásadám daného jazyka.



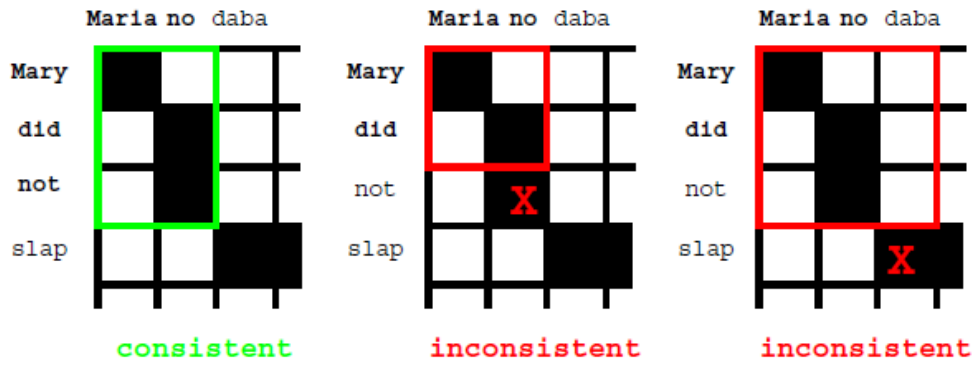
obrázek 9: Zarovnání s využitím frází

Tohoto přístupu využívá řada různých modelů a implementací. Jako příklad zmíníme metodu Word Alignment Induced Phrase Model. Ten vychází z maticového zarovnání slov popisované v předchozích odstavcích a přidává k němu právě využití frází. Zatímco klasický maticový model hledá vhodné zarovnání ke konkrétním slovům, tento model nejprve rozpozná všechny známé fráze a pokusí se je spárovat. Toto spárování zaznačí do vytvořené matice. Proces párování probíhá v několika krocích. V prvním kroku hledáme odpovídající si páry jednotlivých frází a slov, které byly zaznačeny do matice.



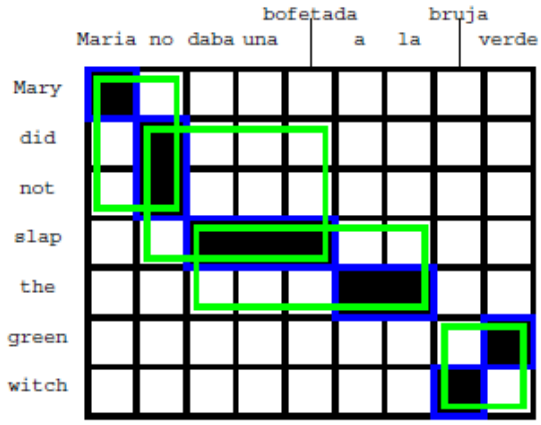
obrázek 10: Využití frázování, první krok

V dalším kroku redukujeme zpracovávaný prostor matice pomocí shlukování sousedních frází. Shluk vznikne spojením dvou frází předchozího kroku, které mají společný dotyk. Důležitým pravidlem je, že každý shluk musí obsahovat všechna nalezena zarovnání pro všechna slova, které daný shluk pokrývá. Příklady přijatelných a nepřijatelných frází jsou zobrazeny na obrázku níže.



obrázek 11: Shlukování frází

V našem případě se zredukuje počet frází prvního kroku z šesti na čtyři. Provádění tohoto redukčního kroku provádíme cyklicky až do té doby, než je počet frází zredukován na jednu jedinou. Po provedení konečného kroku máme k dispozici řadu frází různých rozměrů.



obrázek 12: Druhý krok redukce frází

Dalším zcela odlišným přístupem k tvorbě zarovnání slov jsou metody založené na syntaktickém překladu. Využíváno je syntaktického stromu, který je budován nad celou zpracovávanou větou. Rozpoznávány jsou slovní druhy jednotlivých slov a dle nich je strom reorganizován. Překlad probíhá postupně, nejprve se například při hledání zarovnání z českého do anglického textu přidají neurčité členy k podstatným jménům, které se v českém jazyce nevyskytují. Po dokončení této vkládající částí se provede doslovný překlad zbylých slov zdrojové věty.

3.3.2 Automatické vyhodnocování kvality překladu

Prakticky všechny vědecké disciplíny spolu sdílí potřebu hodnocení do nich vynaloženého úsilí za účelem posouzení správnosti použitých postupů. Snažíme se tak dokázat, že výchozí předpoklady a zvolená teorie byla správně aplikována. V oblasti automatického strojového překladu mají vyhodnocovací metody dva cíle:

- Relativní hodnocení – slouží pro porovnání kvalit a předností jednotlivých metrik mezi sebou
- Absolutní hodnocení – udává absolutní hodnocení kvalit systému na určité škále stanovené stupnice

V oblasti jazykových překladů je proces vyhodnocování správnosti a kvality překladů sám o sobě velice náročným procesem. Příčinou je vysoká nejednoznačnost přirozeného jazyka a jejich komplikovaná struktura. Odhadnout na kolik se dva překlady od sebe odlišují je tak velice komplikovaným úkolem. Dvě zcela odlišné sekvence slov mohou být naprosto ropocené, zatímco další dvě, které si liší pouze v malém detailu, mohou mít naprosto jiný význam. Standardně používanými dimenzemi pro vyhodnocování kvality a právnosti překladů jsou:

- Adekvátnost (vyjadřuje míru shodnosti testovaných překladů)
- Gramatická správnost (vyjadřuje gramatickou správnost překladů)

V ideálním případě by bez jakýchkoliv dodatečných časových a finančních nároků byli samotní uživatelé schopni okamžitě posoudit kvalitu překladového systému. Bohužel to takto není, a proto je potřeba kvalitního a rychlého způsobu hodnocení systémů strojového překladu. V dalších odstavcích budou blíže zmíněny metody pro toto měření používané. Jejich společnou vlastností je, že pracují s referenčními překlady, které určitým způsobem porovnávají s překladem testovaného systému.

Techniky používané na řetězcovém porovnávání

Tyto metody jsou založeny na porovnání slov produkovaných systémem strojového překladu s referenčními překlady. Porovnání je založeno na vyjádření Levenstainovy vzdálenosti. Jedná se o metriku pro měření vzdálenosti. Levensteinova vzdálenost mezi dvěma textovými řetězci je definována jako minimální počet potřebných transformací jednoho řetězce na druhý pomocí operací vkládání, mazání a nahrazování. Metoda je pojmenována po Vladimíru Levensteinovi, který ji zveřejnil v roce 1965 [7].

Jako zástupce tohoto přístupu zmiňme metodu The Word Error Rate (WER). Tato metoda produkuje kladné hodnoty (včetně čísla nula). Hodnota nula značí, že překlad je identický s referencí a teda zcela správný. Avšak není garantována maximální možná hodnota. Vylepšením této metody je *WERg*, která normalizuje Levensteinovu vzdálenost tak, aby výsledná hodnota byla na intervalu 0 až 1 (nejhorší případ). Nevýhodou těchto přístupů je, že metody nerozlišují pořadí slov ve zpracovávaných segmentech (obvykle ve větách) a pracují se slovy jako s neuspořádanou množinou slov.

IR techniky

Metody patřící do této slupiny vycházejí z oblasti vyhledávání informací v rozsáhlém množství dat. Anglicky označovaných jako Information Retrieval, odtud zkratka IR. Testovaný překlad společně s referenčním překladem se rozdělí na jednotlivé části zvané N-gramy. Tyto N-gramy mají různou délku n a můžeme je vzájemně kombinovat.

Významnou zde patřící metodou je BLEU. Je založena na geometrickém průměru N-gramové přesnosti. Používají se obvykle N-gramy délek 1, 2, 3 a 4. Algoritmus pracuje s trestnými body za chybné překlady. Výslednou hodnotu můžeme stanovit pomocí následujícího vzorce:

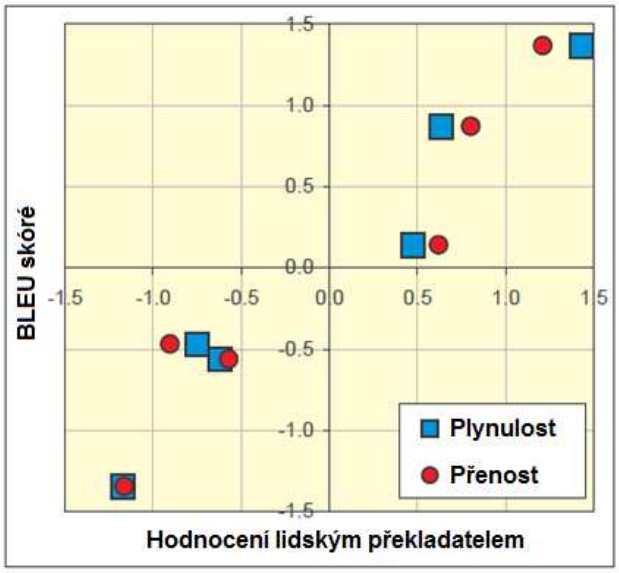
$$Score = \exp\left\{\sum_{n-1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0\right)\right\} \quad (6)$$

$$p_n = \frac{\sum_i \begin{pmatrix} počet\ ngramů\ v\ segmenutu\ i\ testovaného \\ překladu, které\ jsou\ shodné\ s\ referenčním \\ překladem \end{pmatrix}}{\sum_i \begin{pmatrix} počet\ ngramů\ v\ segmentu\ i \\ testovaného\ překladu \end{pmatrix}} \quad (7)$$

$$w_n = N^{-1}, N = 4 \quad (8)$$

L^*_{ref} je počet slov v referenčních překladech, které jsou svou délkou nejbližší hodnocený, překladům a L^*_{sys} je počet slov v právě hodnocených překladech.

Metoda dosahuje velice podobných výsledků jako při ručním zpracováním pomocí lidských překladatelů. Tuto skutečnost ukazuje graf na obrázku číslo 13, který vznikl testováním metody na korpusech různého zaměření a odlišných jazyků. Proložíme-li tímto grafem přímku optimální metody x=y vidíme, že neměřené hodnoty leží velice blízko této přímce.



obrázek 13: Testování metody BLEU

Dalším metodou je metoda NIST. Za jejím vytvořením stojí stejnojmenná organizace. Podnětem pro její tvorbu byla setkání ve Philadephii v roce 2001. IBM zde popsalo automatickou vyhodnocovací metodu pro strojový překlad, která poskytuje okamžitou zpětnou vazbu, využívá referenční překlady a pracuje s krátkými sekvencemi slov. Tento nápad byl označován jako Evaluation Undersudy. Rozdílem oproti metodě BLEU je, že jednotlivé zpracovávané části (N-gramy) mají svou výpočetní váhu podle frekvencí výskytu tak, aby větší důraz byl kladen na méně časté překlady, které mají vyšší informační hodnotu.

Můžeme se setkat s dalšími možnými metodami, jsou jimi například F-measure a Meteor. Nejpoužívanějšími jsou ale právě metody BLEU a NIST.

4 Lemmatizace

Lematizace je proces hledání normalizovaného neboli základního tvaru slova (lemma tvar). Jedná se o transformaci, která vstupní slovo převede na normalizovaný tvar. Základním používaným přístupem je zkoumání a modifikace přípon slov. Snažíme se nalézt vhodnou příponu tak, abychom po jejím odstranění získali základní tvar slova.

Lematizace je velice důležitým krokem předzpracování pro mnoho aplikací zpracování textů. Používá se při zpracování přirozeného jazyka a v celé řadě lingvistických disciplín. Její využití umožňuje dosažení lepších výsledků při strojovém zpracování textu. Používá se například při strojovém vyhledávání v textu, ve kterém nám umožní nalézt i ta slova, která s vyhledávaným výrazem sdílí společné základní slovo. Můžeme ji přirovnat k úloze hledání kořenu slova. Snažíme se vhodně odebírat přípony slov vyskytujících se ve zpracovávaném textu tak, abychom získaly základní tvar slova. U anglických slov *working*, *works*, *worked* stačí odebrat pouze jejich přípony a dostaneme základní slovo *work*. Toto ovšem neplatí pro všechna slova, například u slov *computes*, *computing*, *comuted* po odebrání přípon získáme slovo *comput*, což není správný tvar anglického slova. Musíme jej dodatečně upravit na *compute*.

K porozumění principů používaných při lematizaci je potřeba se zamyslet nad tvorbou slov a nad odlišností slov v různých větných formách (pro ukázku budou uvedeny slova anglického jazyka). Pracujeme-li se samotnými slovy, bez ohledu na jeho kontext používáme standardně jeho základní tvar (např. ve slovnících). Použijeme-li slovo v nějaké větě, můžeme si všimnout, že už pouhá přítomnost slova ve větě nám často změní podobu slova. Například základní tvar slova *walk* ve větě přítomné najdeme ve tvaru *walks* a ve větě minulé pro změnu ve tvaru *walked*. Z tohoto příkladu můžeme vytvořit základní pravidlo, které vytvoří základní slovo odebrání přípon *–s* a *–ed*.

Na lematizaci můžeme nahlížet jako na inverzní transformaci slov. V praxi si s pouhým odebíráním přípon nevystačíme. Některé slova vyžadují po odebrání přípon dodatečnou modifikaci. Proto používáme nahrazování přípon. Stávající přípony nahrazujeme za tzv. inicializační přípony slov. Tyto inicializační přípony mohou být také prázdné. Používané přípony slov bohužel nejsou pro všechny slova shodné. V anglickém jazyce se při výběru použitých přípon zohledňují koncové znaky slov. Tato vlastnost je jedna z komplikací, s kterou se při tvorbě kvalitního lematizačního nástroje musíme vypořádat. Jako příklad uveďme dvě slova *property* a *train*. Tyto slova jsou v základním tvaru jednotného čísla. Při převodu na množné číslo dostaneme odpovídající slova *properties* a *trains*. Vidíme tak, že u každého slova byla použita přípona odlišná. Při bližším zkoumání jazyka zjistíme, že pro slova končící na písmeno y používáme příponu *-ies*. A reverzně tak můžeme vytvořit pravidlo, které pro slova s příponou *–ies* tuto příponu odstraní a na její místo vloží písmeno *y* (inicializační přípona).

Při využití nahrazování přípon se využívá pojmenovaných tříd (označovaných jako labeled class) [8]. Tyto třídy jsou přiřazovány jednotlivým slovům a reprezentují transformaci slova na jeho základní tvar. Transformace nám mapuje jednu příponu na příponu druhou a má tvar (SUFIX1, SUFIX2). Přičemž SUFIX2 může být prázdný v případě, že se jedná pouze o odstranění přípony.

4.1.1 Metoda DRD

Metoda označována jako RDR (Ripple Down Rule) byla vyvíjena v souladu se systémy pracujícími s textem pomocí gramatických pravidel. V těchto systémech je často obtížné při přidávání nových pravidel zajistit, aby tato modifikace negativně neovlivnila použití pravidel stávajících a tím nedegradovala výsledky celého systému. Na rozdíl od standardních klasifikačních pravidel metoda RDR vytváří výjimky k existujícím pravidlům. Tento přístup nijak nemodifikuje pravidla ostatní.

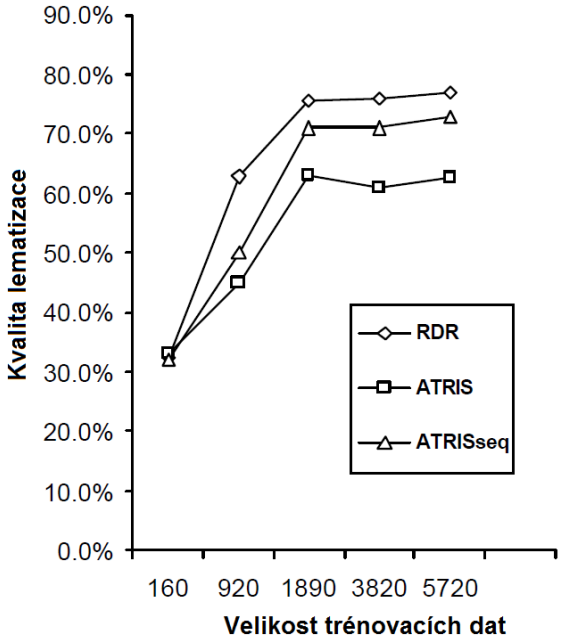
Jednotlivá pravidla a jejich výjimky mají strukturu rozhodovacího stromu, který je neustále přetransformováván (při přidání výjimek a pravidel). Pravidla rozhodovacího stromu mají podobu if-then-else a metoda RDR je dále větví pomocí výjimek (exception). Po vyvolání pravidla pro libovolný vstup ověříme, zda je důsledek pravidla správný. Jestliže není, přidá se pro toto pravidlo nová větev (výjimka) zahrnující zpracovávaný vstup. Pro příklad uveďme následující pravidlo:

If *A* and *B* then *C*
except if *D* then *E*

Toto pravidlo značí, že je-li splněna podmínka A a zároveň B pak výsledkem je C, jestliže neplatí D. V případě, že platí i D výsledkem je E. Používání pravidel tohoto typu se ukázalo jako velice výhodné pro problematiku jazykové lematizace.

Gramatika konkrétního jazyka je pokryta řadou pravidel, některé z nich jsou rozšířeny o zmíněné výjimky. Jako příklad uveďme pravidlo pro tvorbu minulého tvaru sloves. Přípona *–ed* je přidávána každému slovesu, které je použito ve větě minulého času. Výjimkou pravidla jsou slovesa končící na *-y*. Pro ty je používána přípona *–ied*.

Graf ukazuje výsledek testování metody RDR. Pro tento test byl využit slovník, který obsahuje pro každé slovo základního tvaru všechny možné jeho varianty, se kterými se ve volném textu můžeme setkat. Tento slovník obsahoval zhruba 20 000 odlišných základních slov a přibližně 500 000 variant těchto slov. Systému bylo předloženo náhodně vybrané množství dat různé velikosti z těchto slovníku pro vytvoření pravidel a výjimek, které se dále testovaly na lematizaci vstupního textu. Z grafu na obrázku číslo 13 je patrné, že výkon metody se rapidně zvedá se zvyšováním trénovacích dat a to do velikosti 2000 párů slov.



obrázek 14: Testování jednotlivých metod

Doplňkovou funkcí lemmatizátoru je kromě převodu na tematizovaný tvar slova výstup s informace o slově, jeho morfologických kategorií, struktuře textu a dalších potřebných údajů. Inverzní metoda, která udělá ze základního tvaru množinu veškerých možných odvozenin, se nazývá derivací.

5 Použité nástroje

Tato kapitola čtenáře seznámí s nástroji, které byly použity při tvorbě praktické části této práce, a které jsou dále zmiňovány v kapitole číslo 6.

5.1 Python

Pro tvorbu skriptů praktické části byl použit dynamický objektově orientovaný programovací jazyk jména Python. Jeho vznik se datuje na rok 1991, je vyvíjen jako open source, neboli projekt s otevřeným zdrojovým kódem, který bezplatně nabízí instalační balíčky pro běžně používané operační systémy typu Windows, Unix a Mac OS. Ve většině distribucí systému Linux (Unix) bývá standardně předinstalován.

Jazyk Python bývá často zařazován do skupiny skriptovacích programovacích jazyků, ale jeho použití je mnohem širší. Existuje řada jeho rozšíření a knihoven, pomocí kterých je možné vytvářet plnohodnotné aplikace včetně grafického uživatelské rozhraní. Jedná se o dynamický, interpretovaný a objektově orientovaný programovací jazyk. Mezi přednosti tohoto jazyka patří jeho jednoduchost a tedy malé nároky na jeho pochopení a na jeho studium. Jeho syntaxe je jednoduchá, čistá syntaxe, kód vytvořených skriptů je ve srovnání s konkurenčními jazyky krátký a dobře čitelný. Zápis kódu a tvorba skriptů a programů se tak stává velice produktivní.

Dostatečné výkonnosti skriptů zapsaných v tomto jazyce je docíleno zejména využitím jazyka C, ve kterém je implementována většina výkonově důležitých knihoven.

Skripty, které byly vytvořeny v této práci, využívají snadnou práci s textovými soubory a regulárními výrazy, které tento jazyk poskytuje.

5.2 Hunalign

Hunalign je označení pro universální a bezplatný nástroj pro zarovnání odpovídajících si vět ve dvojjazyčném textu [9]. Je součástí maďarského projektu a byl primárně vyvinut pro zarovnání anglického a maďarského textu. Vzhledem k jeho implementaci je dobře použitelný i pro zcela jiné jazykové páry. Jeho vstupem je posloupnost vět ve dvou odlišných jazycích. Standardně se jedná o věty zdrojového materiálu a o věty získané jeho překladem. Výstupem je v optimálním případě sekvence párů odpovídajících si dvoujazyčných vět.

Při zarovnání můžeme využít existujícího překladového slovníku, který je vložen na jeho vstup. Jestliže tento slovník nemáme k dispozici, nebo jej nechceme využít, obejde se i bez něj.

V případě, že nevyužijeme možnosti předložení existujícího slovníku je samotný proces zarovnání náročnější. V první fázi je zkonstruována matice pravděpodobností, která vyjadřuje míru správnosti zarovnání jednotlivých párů vět. Při její tvorbě je využito algoritmu Gale-Church, který vychází ze skutečnosti, že překladem dlouhé věty (věty s velkým počtem slov) bude také dlouhá věta a naopak. Zohledňuje se také například pozice věty v odstavci, podobnost slov a podobně. Následně se získaná matice vyhodnotí a vytvoří zarovnání vět. Z něj se vygeneruje překladový slovník, který se využije v opětovném provedení zarovnání pro překlad vstupu a vytvoření výsledného zarovnání vět.

Výstupem tohoto nástroje je soubor obsahující zarovnání jednotlivých vět. Při zkoumání výstupu narazíme také na případy, kdy k jedné větě zdrojového jazyka bylo přiřazeno (zarovnáno) více vět jazyka cílového a také na případy, kdy k větě nebyla přiřazena žádná ekvivalentní věta.

5.3 GIZA++

GIZA++ je rozšířením programu GIZA, který byl vyvinut týmem Statistického zdrojového překladu během konference v létě roku 1999 v centru pro zpracování jazyka a řeči na univerzitě John-Hopkins [10]. Jedná se o systém, jehož cílem je nelézt co nejpřesnější zarovnání odpovídajících si slov v paralelním vícejazyčném textu.

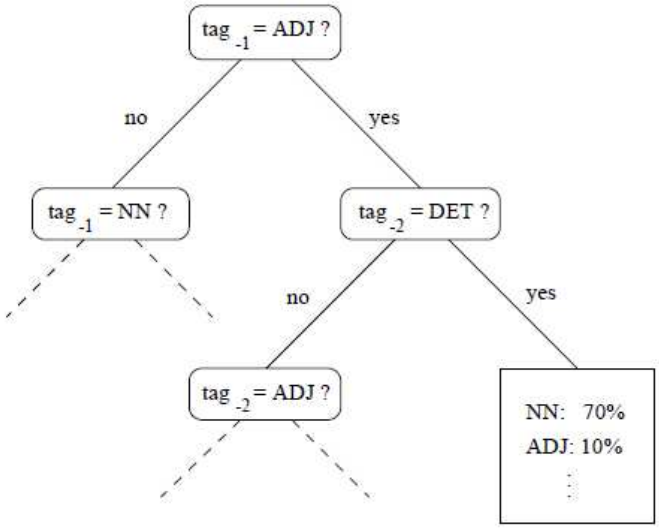
Tento systém umožňuje pokročilé volby nastavení pomocí vstupních parametrů. A to například výběr použité metody zarovnání mezi modely IBM 1-5 a HMM modelem. Tyto modely jsou na zpracovávaný vstup aplikovány iterativně. Cílem každého kroku je vylepšit výsledky získané v kroku předchozím. Počet těchto kroků je také možno nastavit. Zarovnání slov je výpočetně a tedy časově náročným procesem. Náročnost stoupá přímo úměrně s velikostí vstupních textů.

Kromě varianty GIZA++ se můžeme setkat s odlišnými implementacemi, jako například PGIZA, jejíž snahou je využití distribuovaného výpočetního výkonu.

5.4 TreeTagger a PDT

TreeTagger je nástrojem pro značkování vstupního textu. Pro každé slovo nám identifikuje jeho typ (obdoba slovních druhů v českém jazyce) a jeho základní lemma tvar.

Jeho implementace využívá binárního rozhodovacího stromu. Na obrázku číslo 15 je uvedený příklad rozhodovacího stromu, který slouží k vyjádření pravděpodobnosti jednotlivých slovních druhů pro zpracovávaný vstup. Tento ukázkový strom nám říká, že slovo následující po slově typu ADJ (přídavné jméno) a po slově typu DET (předložka) je s 70% pravděpodobností typu NN (podstatné jméno). Tuto pravděpodobnost označujeme jako p(NN|DET,ADJ). Rozhodovací strom je konstruován rekurzivně z trénovací množiny dat.



obrázek 15: Binární rozhodovací strom

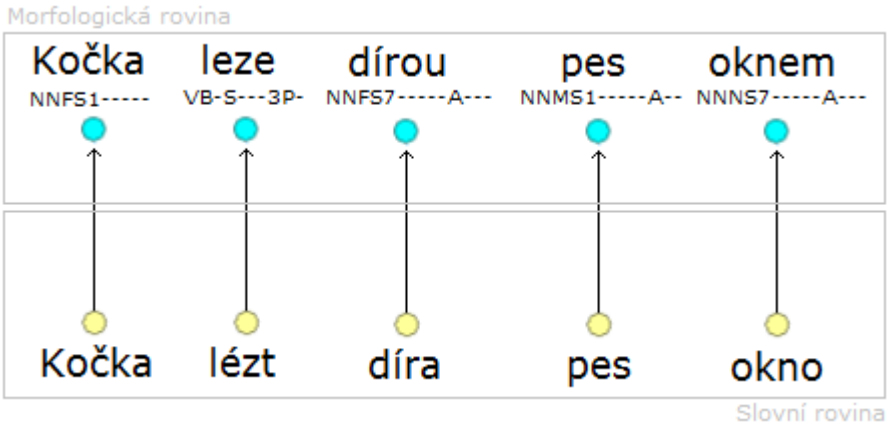
Nejčastěji vyskytující zkratky použité při označkování textu včetně jejich popisu jsou vypsány v tabulce číslo 1.

Zkratka	Význam
NN	Noun, singular or mass
NNS	Common noun, plular
NP	Proper noun, singular
NPS	Proper noun, plular
JJ	Adjective
JJR	Adjective, komparative
JJS	Adjective, superlative
PP	Personal pronoun
PP$	Possesive pronoun
WP	Wh-pronoun
WP$	Possesive wh-pronoun
CD	Caridal number
VV	Verb, base form
VVD	Verb, past tense
VVN	Verb, past participle
VVZ	Verb, present tense, 3rd person singluar
VVG	Verb, gerunf or present participle
VVP	Verb, non-3rd person singular present
VB	Verb, base form
RB	Adverb
WRB	Wh-adverb
RBR	Adverb, comparative
IN	Conjuction, subordinating
CC	Conjuction, coordinating
RP	Particle
UH	Interjection
DT	Determinter
FW	Foreign Word
SYM	Symbol
MD	Modal verb
WDT	Wh-determiner
TO	„to“

tabulka 1: Morfologické značky TreeTaggeru

Pražský závislostní korpus (zkratka PDT) je neustále vyvíjejícím se projektem pro ruční a automatickou anotaci velkého množství českých textů pomocí rozsáhlé lingvistické informace, která zahrnuje morfologické, syntaktické, sémantické a další označení zpracovávaných textů [11]. Aktuální verze PDT 2.0 obsahuje české texty o rozsahu přibližně dvou miliónů slov a obsahuje rovněž softwarové nástroje pro prohledávání textu a pro charakteristiku textu vlastního.

Data obsažená v PDT 2.0 jsou anotována ve třech rovinách a to v morfologické, analytické, tektogramatické a slovní rovině (původní text). Pro účely této práce je zajímavé využití poznatků a nástrojů spadající do roviny morfologické. Anotace na této rovině spočívá v přiřazení několika atributům jednotlivým slovním jednotkám ze vstupní slovní roviny, nejdůležitějšími z nich jsou morfologické lemma a tag. Morfologické lemma obsahuje základní tvar daného slova. Atribut tag obsahuje morfologickou značku o délce patnácti znaků, která vyjadřuje slovní druh a rozdělení do dalších morfologických kategorií.

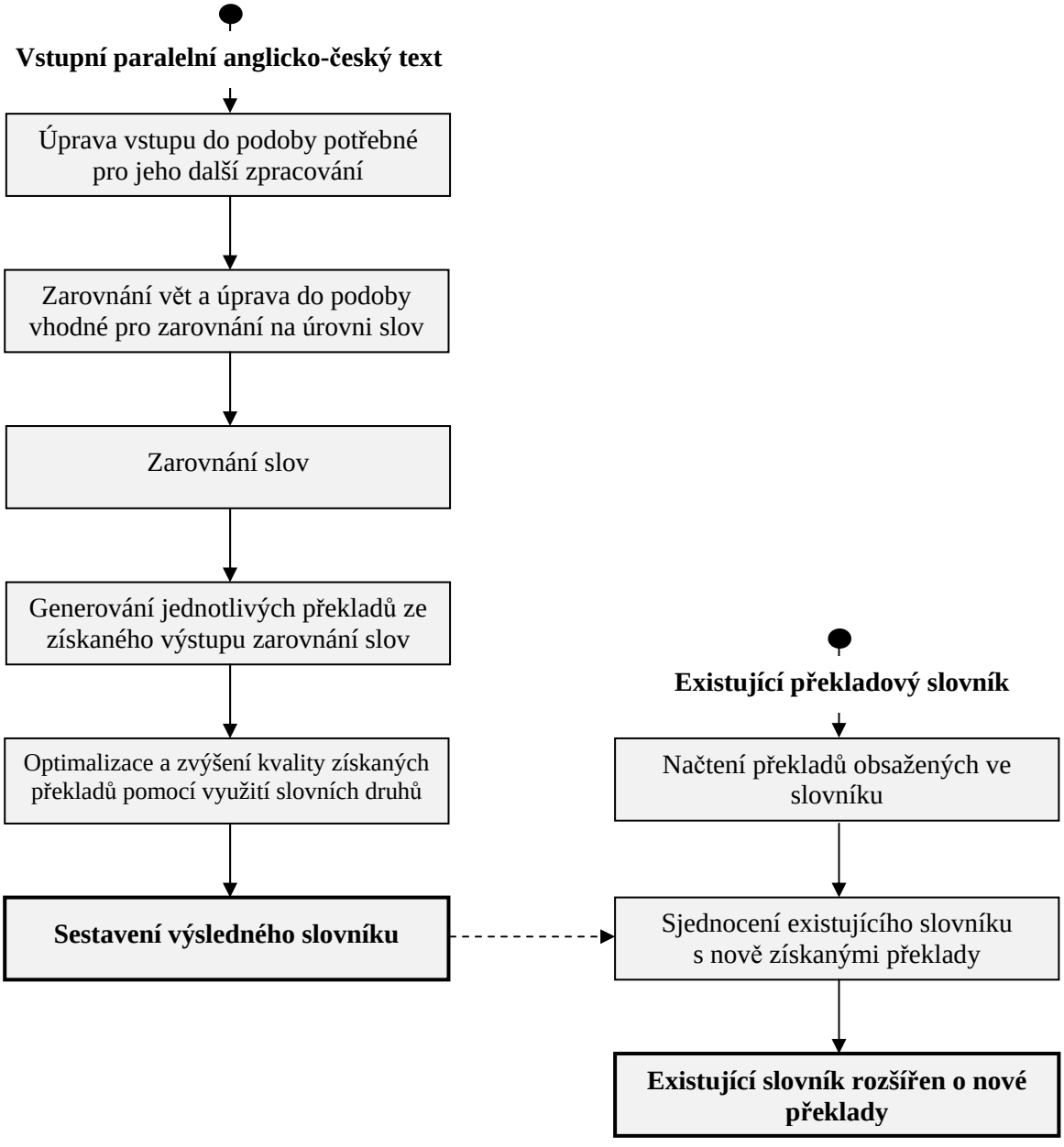


obrázek 16: Morfologická a slovní rovina nástroje PDT

Přínosem pro praktickou tvorbu systému pro automatickou tvorbu překladového slovníku ze vstupních paralelních anglicko-českých textů byl právě nástroj pracující na popisované morfologické rovině, který pro jednotlivá slova vstupního textu určí lemma tvar a jeho tag (obrázek číslo 16).

6 Návrh a realizace sytému

Tato v pořadí již šestá kapitola popisuje návrh a realizaci systému, jehož cílem je z vstupních paralelních anglicko-českých textů vytvořit co nejkvalitnější překladový slovník. Ten může být použit sám o sobě, nebo může posloužit k rozšíření některého ze stávajících překladových slovníků. Nejvýznamnějšími částmi systému je výběr vhodných vstupních textů, jejich modifikace pro proces získání zarovnání na úrovni vět, provedení zarovnání odpovídajících si slov a zpracování získaných výstupů včetně generování překladového slovníku.



obrázek 17: Návrh systému a jeho základní struktura

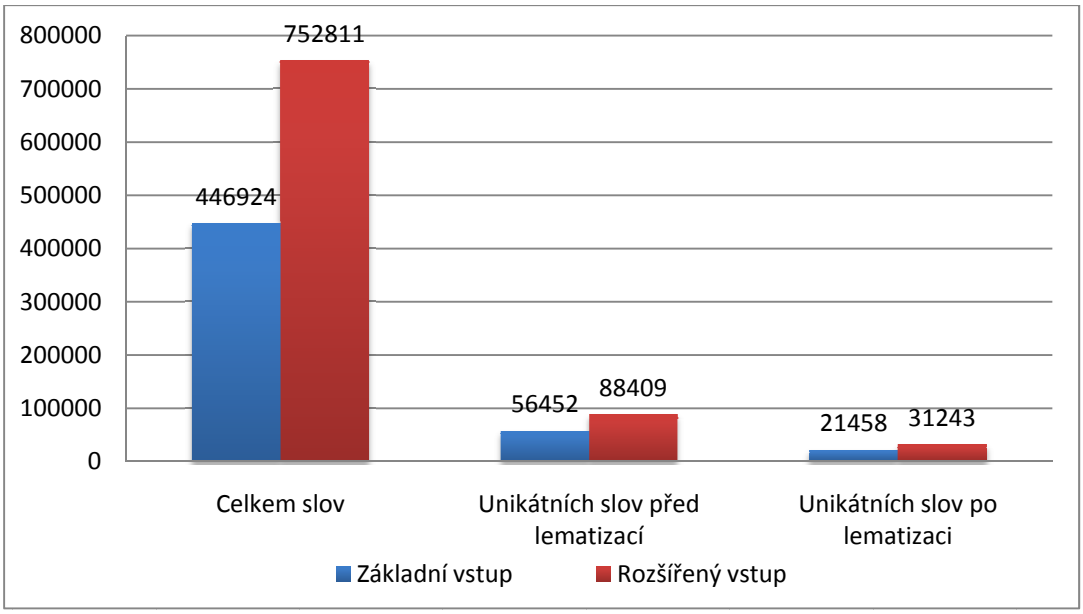
6.1 Příprava a výběr vstupních textů

Výběr kvalitních a dostatečně rozsáhlých vstupních paralelních textů je nezbytně nutná podmínka pro dosažení kvalitních výsledků. Pod slovem kvalitní paralelní texty rozumíme zejména významovou ekvivalentnost a správnost překladu jednotlivých vět obsažených v textu. Rozsah textu měříme běžně v počtu slov, popřípadě počtu vět nebo odstavců. Všeobecně platí, že s rostoucí velikostí vstupních testů roste počet získaných překladů a jejich kvalita.

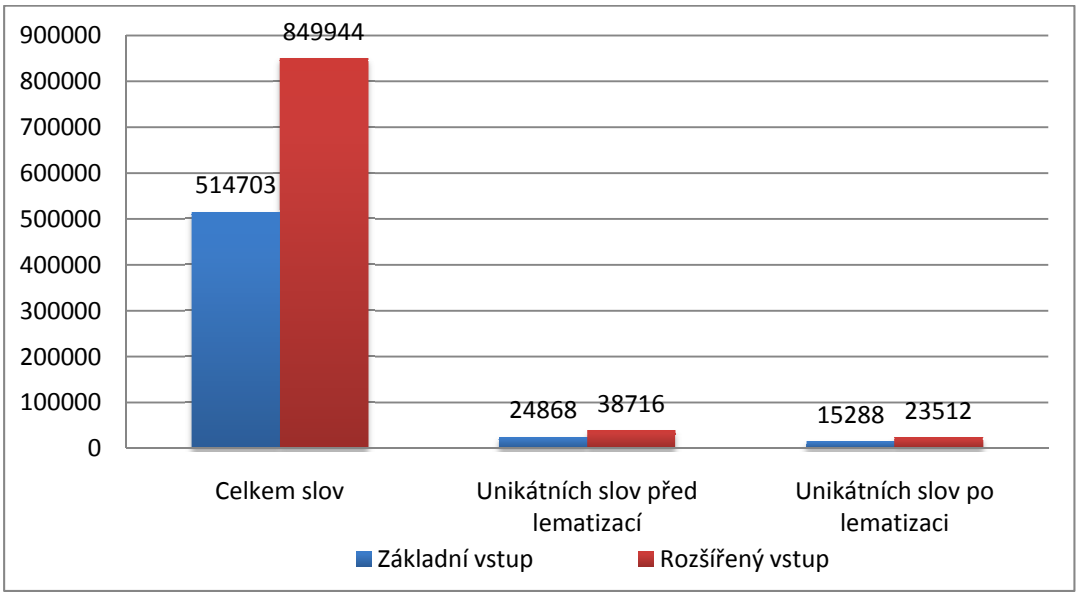
V první fázi bylo vybráno pět paralelních vstupních textů. Jsou jimi tituly Šťastný Jim autora K. Amise, Komu zvoní hrana a Fiesta autora Ernesta Hamingwaye, Čarování s láskou od Luise Edrichové a Velký Gatsby od Francise Scotta Fitzgeralda. Texty jsou součástí paralelního korpusu Kačenka 2, jsou v samostatných souborech formátu rtf. Na svém začátku obsahují hlavičku nesoucí informace o samotném textu. Například autora textu, originální název titulu, jméno překladatele apod. V druhé fázi budování systému byl tento vstup rozšířen o další dva texty významného rozsahu. Jedná se o texty Seznam sedmi autora Marka Frosta a Hlava XXII od Josepha Hellera. Toto přidání zvýšilo počet slov ve vstupních textech přibližně o 60% a bylo provedeno zejména z důvodu porovnání vlivu zvyšování vstupního textu na počet získaných překladů a jejich kvalitu. Tento test je blíže popsán v podkapitole věnující se vyhodnocení a testování.

Každý ze vstupních textů byl samostatně upraven pomocí skriptu jazyka Python. Při úpravě byl z textu extrahován pouze samotný obsah textu, bez hlavičky, autora a názvu. Další fází úprav vstupních textů je odstranění (případně nahrazení) nevhodných znaků pro další zpracování. Jedná se například o uvozovky (mimo uvozovek v anglickém textu, které mají speciální význam), středníky a dvojtečky. Poslední úpravou textu je převedení do podoby, ve které bude každá věta na samostatném řádku. To je podmínkou pro následující část, ve které se provádí zarovnání jednotlivých vět pomocí specializovaného nástroje.

Protože nástroje, které přichází v dalších podkapitolách ke slovu, identifikují slovo stejného základu, které je v různých tvarech, s různými příponami a předponami jako odlišné slova, je vhodné v současné situaci provést lemmatizaci upravených textů. Tato operace nahradí každé slovo jeho základním tvarem. Lematizace českého textu byla pomocí nástroje obsaženého v PDT 2.0 a pro anglický text byl využit nástroj TreeTagger (viz kapitola 5.4). Po provedení lematizace klesne počet unikátních slov. Tuto skutečnost zachycují grafy na následujících obrázcích. Z grafů lze vyčíst, že zejména u českého textu lemmatizace počet unikátních slov rapidně sníží. Tato skutečnost potvrzuje vysokou rozmanitost českého jazyka.



obrázek 18: Statistika českého vstupního textu



obrázek 19: Statistika anglického vstupního textu

6.2 Zarovnání vět

Zarovnání vět je metoda, jejímž cílem je nelézt nejpravděpodobnější vzájemné přiřazení odpovídajících si vět zdrojového a cílového jazyka. Pro realizaci této operace byl vybrán nástroj *Hunalign*. Tento nástroj byl vyvinut pro zarovnání maďarského a anglického jazyka, ale lze jej použít i pro zarovnání jiných jazykových dvojic a detailněji je popsán v kapitole číslo 5.2. Základním povinným vstupem jsou soubory vět zdrojového a cílového jazyka. Věty vstupních souborů musejí být na samostatných řádcích, čehož bylo docíleno předchozí úpravou vstupních textů. Výstupem je zarovnání ve tvaru zdrojová věta, cílová věta a ohodnocení pravděpodobnosti tohoto zarovnání. Volitelným vstupem je překladový slovník. V případě jeho nedodání je zarovnání generováno bez něj. Při vytváření popisovaného systému nebylo využito možnosti využití existujícího překladového

slovníku. Po dokončení zarovnání je uživateli vypsáno celkové skóre provedeného zarovnání. Jedním z problémů, které se musí při zarovnání řešit je skutečnost, že počet vět vstupních souborů nemusí být stejný a tak ve výstupu většinou nalezneme zarovnání více vět k jedné zdrojové větě, které musíme při dalším zpracování výstupu zohlednit. Může se ale také stát, že program k některým větám nenajde žádné vhodné zarovnání.

Pro dosažení nejlepšího zarovnání je vhodné vstupní data patřičně rozčlenit na jednotlivé části a provést zarovnání právě nad těmito částmi. Tento přístup vede k lepším výsledkům než vytvoření zarovnání z jednoho vstupního souboru obsahujícího stovky vět z nejrůznějších zdrojů. Protože jako vstupní texty byly použity nezávislé dvojice souborů s paralelním textem, bylo provedeno zarovnání vět právě nad těmito upravenými soubory a výstup byl sjednocen a upraven do jednoho, respektive dvou souborů obsahujícího zarovnané anglické a české věty. Jako vhodné se ukázalo vyloučit ta zarovnání, které nedosahují určité hranice pravděpodobnosti zarovnání. Jako vhodná hranice se ukázala hodnota 0,3. Věty s pravděpodobností zarovnání nižší než je tato hodnota byly z dalšího zpracování vyloučeny.

6.3 Zarovnání slov a získání překladů

Podstatou zarovnání slov je rozdělení vět vstupních textů na jednotlivá slova neboli tokeny. Této operaci se říká tokenizace. Následně je hledáno nejpravděpodobnější přiřazení (zarovnání) jednotlivých tokenů zdrojového jazyka s tokeny cílového jazyka. Nástroje pro zarovnání slov mohou využívat nejrůznějších metod k dosažení výsledku. Nejpoužívanější metodou je statická metoda, která pracuje iterativně a ohodnocuje pravděpodobnost zarovnání jednotlivých slov. Na počátku je každé slovo ohodnocené nulovou pravděpodobností a při každém běhu jsou získávany dvojice slov s určitou pravděpodobností zarovnání. Algoritmus končí, jestliže získané výsledky se nikterak neliší od výsledků předchozí iterace. Mezi tyto statické metody paří například zarovnání IBM 1-6. Další používané metody jsou metody založené na podobnosti slov, metody využívající existující slovník, nebo třeba také poziční metoda pracující s pozicí slov ve větě.

Pro vytvářený systém byl využit nástroj *GIZA++*. Před samotným jeho spuštěním jsou vstupní texty, získané v předchozím kroku upraveny tak, aby se v nich nevyskytovala interpunkční znaménka. A to z toho důvodu, že *GIZA++* bere slovo jako sekvenci znaků, která je ukončena bílým znakem. Takže jedno a totéž slovo vyskytující s uprostřed věty a na konci, kde je ihned za ním interpunkční znaménko tečky je identifikováno jako slovo zcela jiné. Protože zarovnání slov je ve vytvářeném systému výpočetně a tedy i časově nejnáročnějším blokem vytvářeného systému, jsou vstupní texty před zarovnáním předzpracovány. A to nástrojem *PLAIN2SNT* který provádí pouhou konverzi slov za celočíselné hodnoty. S nimi je pří generování zarovnání operováno z důvodu nižších nároků pro ukládání a manipulaci čísel v paměti, než v případě řetězců. Dalším předzpracováním potřebným v případě, že se chystáme využít metod IBM je rozdělení slov do jednotlivých tříd nástrojem *MKCLS*.

Výstupem je řada souborů. Pro extrakci překladů byl využit soubor s příponou **.t3.final*. Tento soubor obsahuje zarovnání jednotlivých slov cílového a zdrojového jazyka, včetně ohodnocení pravděpodobnosti tohoto zarovnání. Jednotlivá slova jsou vyjádřena pomocí dříve získané číselné reprezentace, takže je potřeba dodatečné úpravy s využitím souboru obsahujícího převod slov na jejich číselnou podobu. Takto získané upravené zarovnání slov je následně využito pro generování překladového slovníku pomocí specializovaného skriptu napsaného v jazyce Python.

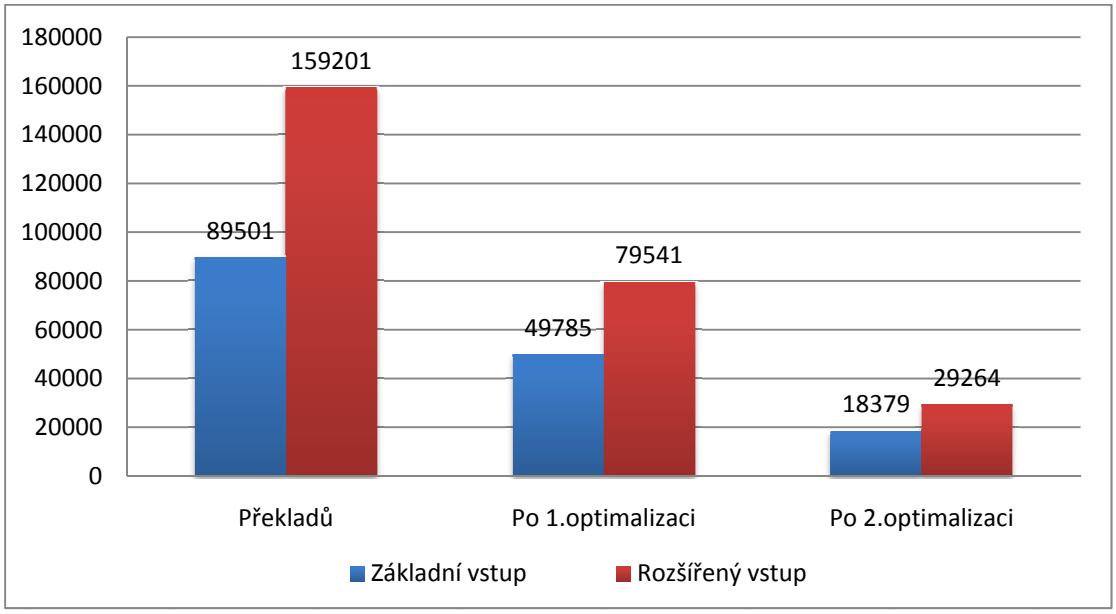
V první fázi tvorby systému byl použit vstup v předchozí kapitole označovaný jako základní, nad kterým byly provedeny všechny části systému, jako jsou zarovnání vět, zarovnání slov, extrakce překladů a posléze rozšíření existujícího slovníku. V pokročilé fázi budování systému byl nahrazen rozšířeným vstupem. To nám přináší možnost porovnání vlivu velikosti vstupních dat na dosažených

výsledcích. Tabulka číslo 2 zobrazuje velikost vstupních dat a jejich vliv na získaný počet překladů. Tyto data jasně ukazují přímo úměrnou závislost mezi velikostí vstupního textu a počtem získaných překladů.

Vstup	Českých slov	Anglických slov	Získaných překladů
Základní	446924	514703	46891
Rozšířený	(+69%) 752811	(+65%) 849944	(+71%) 80396

tabulka 2: Vstupní texty pro vytvoření zarovnání slov

Pro zvýšení kvality konstruovaného slovníku byly pro další zpracování ignorovány ty překlady, jejichž pravděpodobnost nepřesáhla hodnotu 0,1. Toto omezení zredukovalo počet překladů přibližně na polovinu. Další použitou optimalizací bylo využití slovních druhů. Její základní myšlenkou je skutečnost, že překladem zdrojového slova slovního druhu *x* by mělo být slovo cílového jazyka, které bude také slovního druhu *x*. Neboli překladem podstatného jména by mělo slovo, které je také podstatným jménem a podobně. Realizace se provádí v několika krocích. V prvním je získaný soubor se všemi překlady rozdělen na dva. Na soubor obsahující pouze slova zdrojového jazyka a na soubor obsahující pouze slova cílového jazyka (překlady). Každý z těchto souborů byl zpracován skriptem, který využívá nástrojů popisovaných v páté kapitole. Pro slova zdrojová (anglická) nástroj TreeTagger a pro české překlady nástroj z PDT. Takto bylo ke každému slovu přiřazeno číselné vyjádření slovního druhu. V poslední fázi bylo provedeno sloučení označených slov zdrojového a cílového jazyka. Akceptovány byly pouze překlady shodného slovního druhu. Touto optimalizací se nám počet překladů zredukoval přibližně na jednu třetinu.



obrázek 20: Redukce počtu překladů jednotlivými optimalizacemi

Takto získané překlady lze použít samy o sobě jako překladový slovník. Jeho testování nad vstupními texty je diskutováno v kapitole pojmenované jako vyhodnocení a testování. Následující tabulky zobrazují vzorky získaných překladů. Byly vybrány vzorky s nejvyšší hodnotou skóre blížící se k maximální hodnotě 1, překlady s ohodnocením kolem hodnoty 0,5 a překlady se skórem těsně přesahujícím akceptovatelnou hranici 0,1.

Anglické slovo	Český překlad	Skóre zarovnání
kaiser	císař	0,997279
immovably	zvládnutelně	0.997279
dispensary	marodka	0.997279
fanatical	fanatický	0.997279
mio	mio	0.997279
subcontinent	indie	0.997279
zurich	curych	0.997279
portfolio	album	0.997279

tabulka 3: Vzorky překladů s nejvyšší hodnotou zarovnání

Překlady s nejvyšší hodnotou zarovnání slov obsahují poměrně kvalitní překlady. Nachází se mezi nimi prakticky všechna jména a názvy, které mají v paralelním textu na obou dvou jeho stranách stejný tvar. V tabulce vzorků je to například překlad slova „mio“. Překlad slova „subcontinent“ na slovo „Indie“ patrně nejvhodnější není. Jeho vznik není způsobený chybou systému, ale nepřesným překladem zdrojového textu jeho autorem.

Anglické slovo	Český překlad	Skóre zarovnání
woodand	rozlétnout	0,191568
theseus	příčina	0,191535
growth	Znovu	0.190519
growth	Trochu	0.190519
growth	povyrůst	0.190519
Coloring	Přibarvený	0.190142
Coloring	Vybarvení	0.190142
Coloring	Slušet	0.190142

tabulka 4: Vzorky překladů se střední hodnotou zarovnání

Mezi překlady s hodnotou zarovnání okolo hodnoty mediánu všech získaných překladů nalezneme překlady vhodné, ale také překlady na první pohled chybné. Nachází se zde řada různých překladů jednoho zdrojového slova. Z nichž jsou většinou některé správné a některé naopak. Například překlad slova „growth“ na „povyrůst“ je správný, ale na „příčina“ a „trochu“ již nikoliv.

Anglické slovo	Český překlad	Skóre zarovnání
married	vdát	0,101074
enormity	teprve	0,101071
enormity	vzpříčit	0,101071
enormity	pohroma	0,101071
forgiving	odpuštění	0,101045
forgiving	povšimnout	0,101045
sightseeing	turista	0,100894
sightseeing	zájezdový	0,100894
sightseeing	objevovat	0,100894

tabulka 5: Vzorky překladů s nejnižší akceptovatelnou hodnotou zarovnání

Vzorky s nejnižší hodnotou zarovnání lehce přesahující akceptovatelnou hranici 0,1 obsahují nepříliš kvalitní výsledky, ale i přesto se v nich najde řada správných překladů, jejichž vyloučení by zbavilo

vytvářeného slovníku řady zajímavých a vhodných překladů. Nachází se mezi nimi řada překladů identického slova, která jsou většinou ohodnoceny stejnou hodnotou zarovnání, i když jejich skutečná kvalita je velice různorodá.

6.4 Rozšíření existujícího slovníku

Vytvořený slovník (získané překlady ze vstupních textů) lze použít pro rozšíření již existujícího slovníku. Rozšíření můžeme rozdělit na dvě části. A to na přidání překladů pro slova, které ve slovníku nejsou vůbec přeložena. A na obohacení slovníku a nové překlady slov, která v něm sice přeložena jsou, ale neobsahují všechny získané překladové dvojice.

Pro analýzu tohoto rozšíření byl v první fázi vybrán všeobecný překladový slovník ve formátu XML. V tomto slovníku je pomocí značkovacího jazyka XML uloženo pro každé slovo jeho překlad, fonetický tvar a ukázky použití. Pro jeho načtení jsem vytvořil samostatný skript. Bylo zjištěno, že slovník obsahuje překlad pro 24915 anglický slov. Pro tato slova obsahuje celkem 97159 různých překladů. Protože se tento slovník ukázal jako příliš malý, byl vybrán slovník větší. A to slovník tvořený z 111849 anglických slov a 207775 českých překladů. Tento slovník taktéž obsahuje ukázky použití jednotlivých slov ve větách, fonetický tvar a pro jeho zpracování byl vytvořen samostatný skript. V dalším textu bude pracováno pouze s tímto větším slovníkem označeným jako existující slovník.

Analýzou získaného slovníku s načteným existujícím slovníkem byla slova zdrojového jazyka rozdělena (jejichž překlad jsme získali pomocí nově vytvořeného slovníku) na ta, která v existujícím slovníku přeložena jsou a na slova, které v něm přeložena nejsou. Z vytvořeného slovníku pomocí rozšířeného vstupního textu, nad kterým byly provedeny obě optimalizace popisované v předchozí kapitole obsahujícího překlady pro 16508 zdrojových slov je přibližně 70% těchto slov přítomná v existujícím slovníku a zbylých 30% slov v něm přeloženo není. Právě těchto 30% anglických slov včetně všech jejich získaných překladů se přímo nabízí pro doplnění do existujícího slovníku. Z 70% známých slov je potřeba najít ty překlady, které existující slovník neobsahuje a obohatit jimi existující slovník. Následující tabulka ukazuje možnosti rozšíření existujícího slovníku pomocí získaných překladů (slovníku).

Možnosti rozšíření existujícího slovníku B pomocí vytvořeného slovníku X

Neznámá slova (nejsou v B)	5114	překladů v A	**8961**		
Známá slova (jsou v B)	11394	překladů v A	20303	známých překladů	4683
	16508		*29264*	neznámých překladů	**15620**
					20303

tabulka 6: Možnosti rozšíření existujícího slovníku

O překlady v tabulce zeleným písmem označeny je možno existující slovník rozšířit. Jedná se 24581 překladů, to je 84% ze všech získaných. Pro tuto úlohu byl vytvořen skript pro rozšíření slovníku, který do něj vloží překlady a nadále zachová jeho původní strukturu. Získáme tak po vytvořeném slovníku a existujícím slovníku třetí nejrozsáhlejší slovník. Tyto tři slovníky jsou v následující kapitole blíže rozebrány, je zde otestováno jejich využití při překladu různého vstupního textu.

Z testovacího vstupu, který je popisován v následující kapitole byly vybrány tři věty, které byly ukázkově přeloženy pomocí vytvořeného slovníku. Pro ukázku přínosu optimalizace získaných překladů pomocí slovních druhů jsou mezi získanými překlady i ty, které byly optimalizaci vyloučeny. Pro jejich odlišení jsou přeškrtnuty.

Origální věta:	they	made	a	silly	mistake	thought	the	professor	of	history	said
Lemma tvar:	they	make	a	silly	mistake	thought	the	professor	of	history	say
Získané překlady:	on	udělat ~~aby~~		hloupý	omyl chyba ~~mýlit~~	myšlenka ~~pomyšlení~~		profesor pan	z	historie Dějiny	říci

tabulka 7: Ukázková věta č. 1 přeložena pomocí získaných překladů

Pro slovo zdrojové slovo „they“ byl nalezen pouze český překlad „on“. Tento překlad jistě správný není, ale protože slova „they“ i „on“ jsou zájmena, nebyl tento chybný překlad pomocí optimalizace slovních druhů odfiltrován. Chybný překlad slova „make“ na „aby“ z důvodu odlišnosti slovních druhů odfiltrován byl, což zvýšilo kvalitu dostupných překladů. Překlady slova „mistake“ na „mýlit“ respektive překlad „thought“ na slovo „pomyšlení“ můžeme významově ohodnotit jako správné, ale z důvodu odlišnosti slovních druhů byly systémem vyřazeny. Ostatní překlady můžeme považovat za správné.

Origální věta:	but	what	do	you	think	they	said	then
Lemma tvar:	but	what	do	you	think	they	say	then
Získané překlady:	ale	co	udělat	ty	myslet myslit ~~že~~	on	říci	potom pak

tabulka 8: Ukázková věta č. 2 přeložena pomocí získaných překladů

Origální věta:	Welch	was	talking	yet	again	about	his	concert
Lemma tvar:	welch	be	talk	yet	again	about	his	concert
Získané překlady:	welch	být	mluvit	přece ještě ~~ale~~	znovu zase opět	o	jeho svůj	koncert věc hudba ~~zahradní~~ ~~vyprávět~~ ~~komorní~~

tabulka 9: Ukázková věta č. 3 přeložena pomocí získaných překladů

Poslední ukázková věta obsahuje slovo, jehož překladem je totéž slovo. Jedná se patrně o jméno, které je použito na obou stranách vstupního paralelního textu a můžeme tak tento překlad označit za správný.

6.5 Vyhodnocení a testování

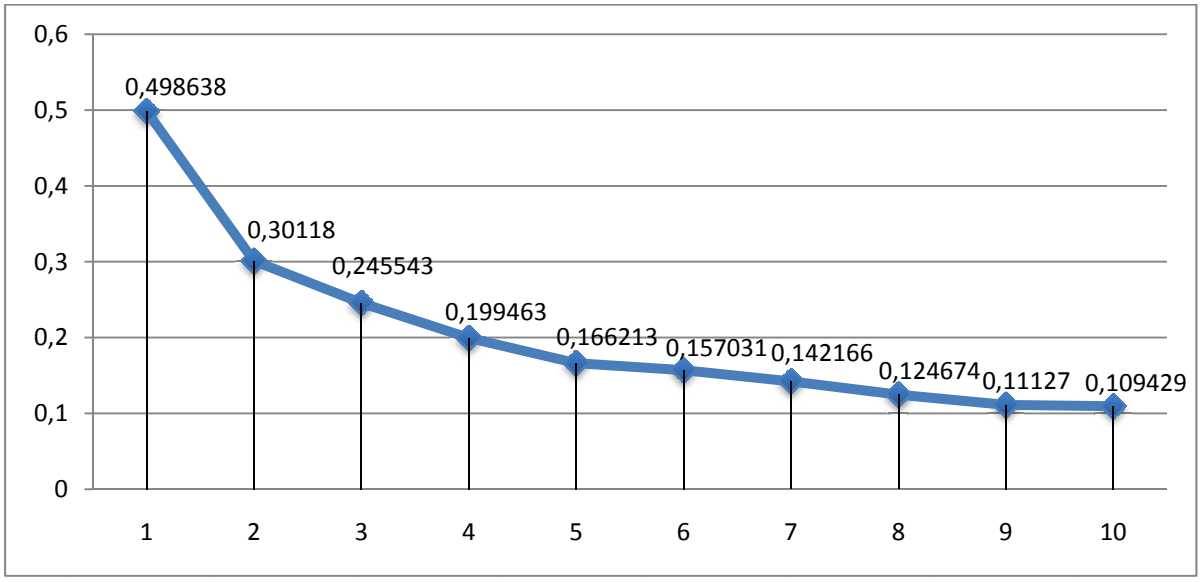
V této závěrečné podkapitole jsou zhodnoceny dosažené výsledky a to z kvantitativního i kvalitativního hlediska. Je zde testován vliv velikosti vstupu na získané překlady, přínos využití slovních druhů, rozšíření existujícího překladového slovníku a možnost překladu pomocí jednotlivých slovníků.

Všech více než 80000 získaných překladů z rozšířeného vstupu bylo převedeno do podoby, kde každý překlad je reprezentován samostatným řádkem následující struktury:

Anglické slovo	Slovní druh anglického slova [1-10]	České slovo (překlad)	Slovní druh českého slova [1-10]	Shodnost slovních druhů [A/N]	Skóre zarovnání [0,1-1]

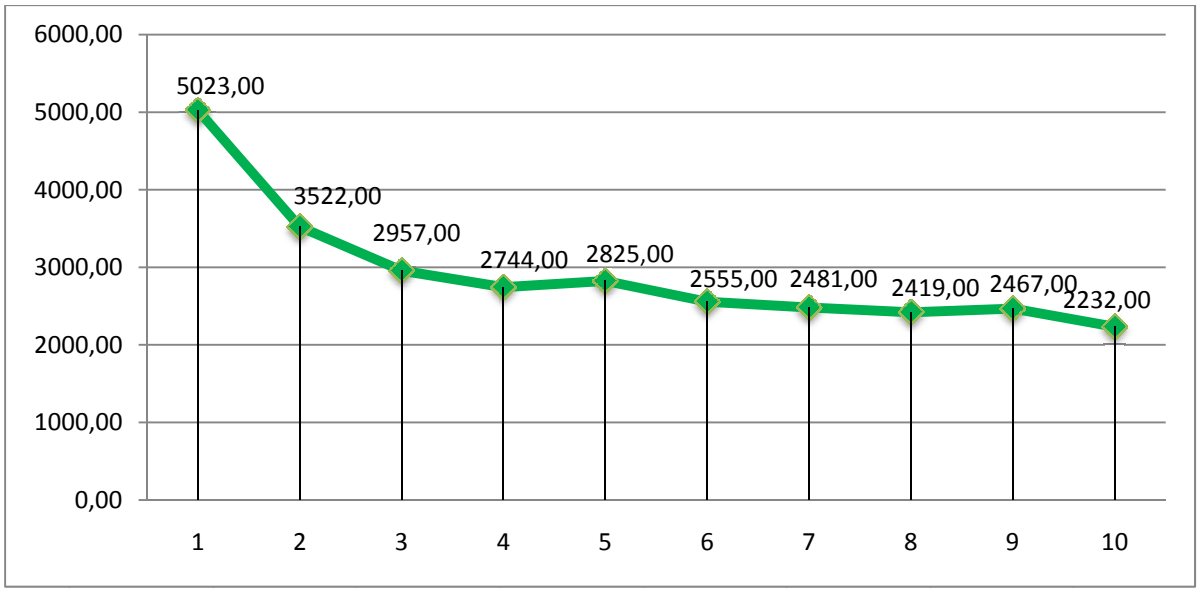
Vznikla tak tabulka, která posloužila jako podklad pro vyhodnocení překladů. Tyto řádky respektive překlady byly seřazeny sestupně podle hodnoty skóre. Tato hodnota byla stanovena při zarovnání slov pomocí nástroje *GIZA++* a znázorňuje míru správnosti zarovnání podle metrik metod použitých nástrojem *GIZA++*. Seřazená tabulka byla rozdělena rovnoměrně na deset intervalů. První z nich obsahuje řádky a tedy překlady s hodnotou skóre patřící do nejvyšší desetiny a poslední interval překlady s pravděpodobností nejnižší. Platí však, že nejnižší ohodnocení překladu je 0.1, protože překlady s nižší pravděpodobností byly vyloučeny (více v kapitole věnované zarovnání slov).

Graf na následujícím obrázku znázorňuje hodnoty skóre překladu středního prvku pro jednotlivé intervaly. Z grafu můžeme vyčíst, že od čtvrtého intervalu se hodnota překladů snižuje velice pozvolna a rozdíl mezi kvalitou posledních tří by neměl být prakticky žádný.



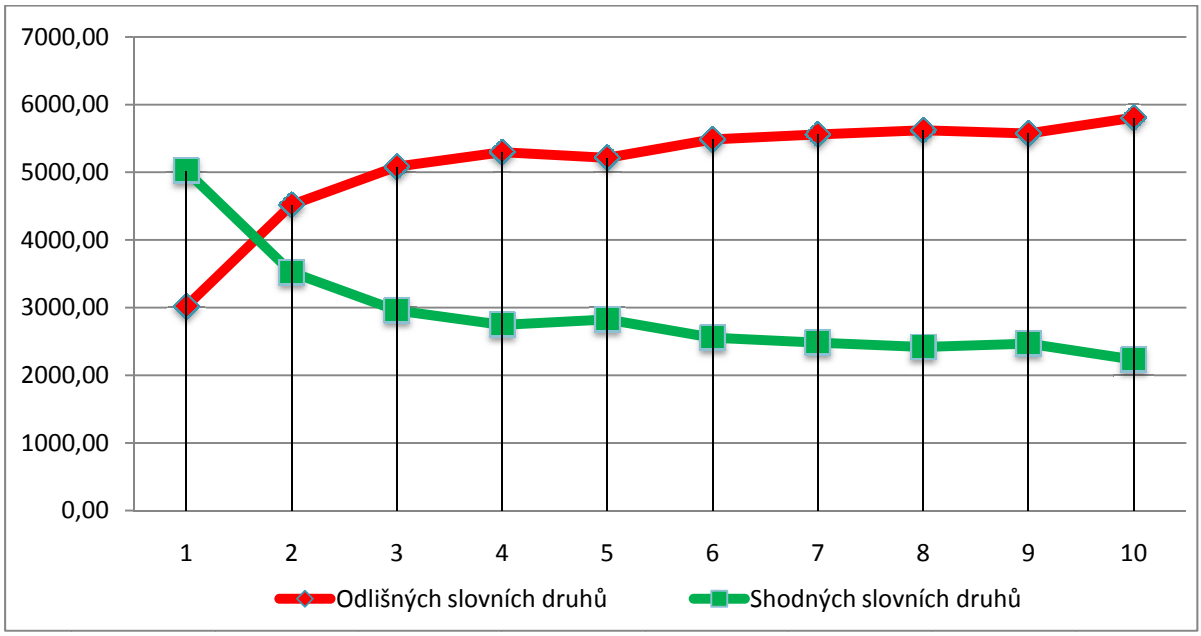
obrázek 21: Hodnota mediánu skóre v jednotlivých intervalech

Předmětem dalšího zkoumání takto vytvořené a rozdělené tabulky byly slovní druhy. Každý překlad byl opatřen číslem v rozmezí 1 až 10 reprezentující jeho slovní druh. Při tvorbě výsledného slovníku popisovaného v předchozí kapitole byly překlady s odlišnými shodnými druhy vyloučeny a tedy ignorovány. V optimálním případě by u překladů s nejvyšší hodnotou skóre měla být pravděpodobnost odlišného slovního druhu co nejbližší číslu nula a u překladů s nejmenší hodnotou zase podstatně vyšší. Tento předpoklad se potvrdil. Od čtvrtého intervalu se počet správných překladů výrazně neliší. Je to dáno pouze malým rozdílem skóre překladů obsažených v intervalech 4 až 10.



obrázek 22: Počet překladů se shodnými slovními druhy v jednotlivých intervalech

Následující graf ukazuje poměr mezi shodnými a odlišnými slovními druhy v jednotlivých testovaných intervalech. Pouze v prvním intervalu obsahujícím překlady s nejvyšším skórem zarovnání je vyšší počet překladů se shodnými slovními druhy než s odlišnými. Křivka udávající počet shodných překladů je klesající a křivka počtu odlišných překladů rostoucí, což je v souladu s předpokládaným výsledkem.



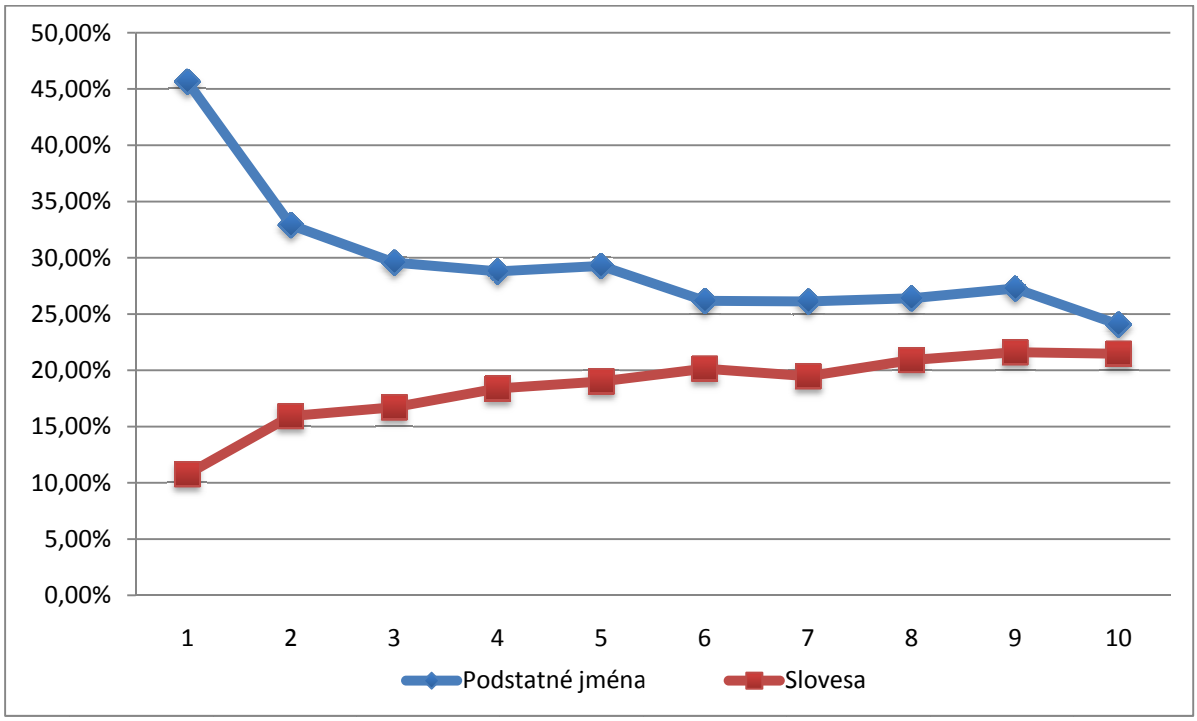
obrázek 23: Poměr překladů se shodnými a odlišnými slovními druhy

Jednoznačně nevyskytovanějším slovním druhem ve zkoumaných překladech je podstatné jméno a na druhém místě přídavné jméno před slovesem.

Slovní druh	Počet výskytů [%]
Podstatné jméno	54,6
Přídavné jméno	23,7
Zájmeno	0,08
Číslice	0,06
Sloveso	13,3
Příslovce	7,93
Předložka	0,22
Spojka	0,02
Částice	0,01
Citoslovce	0,02

tabulka 10: Zastoupení slovních druhů v získaných překladech

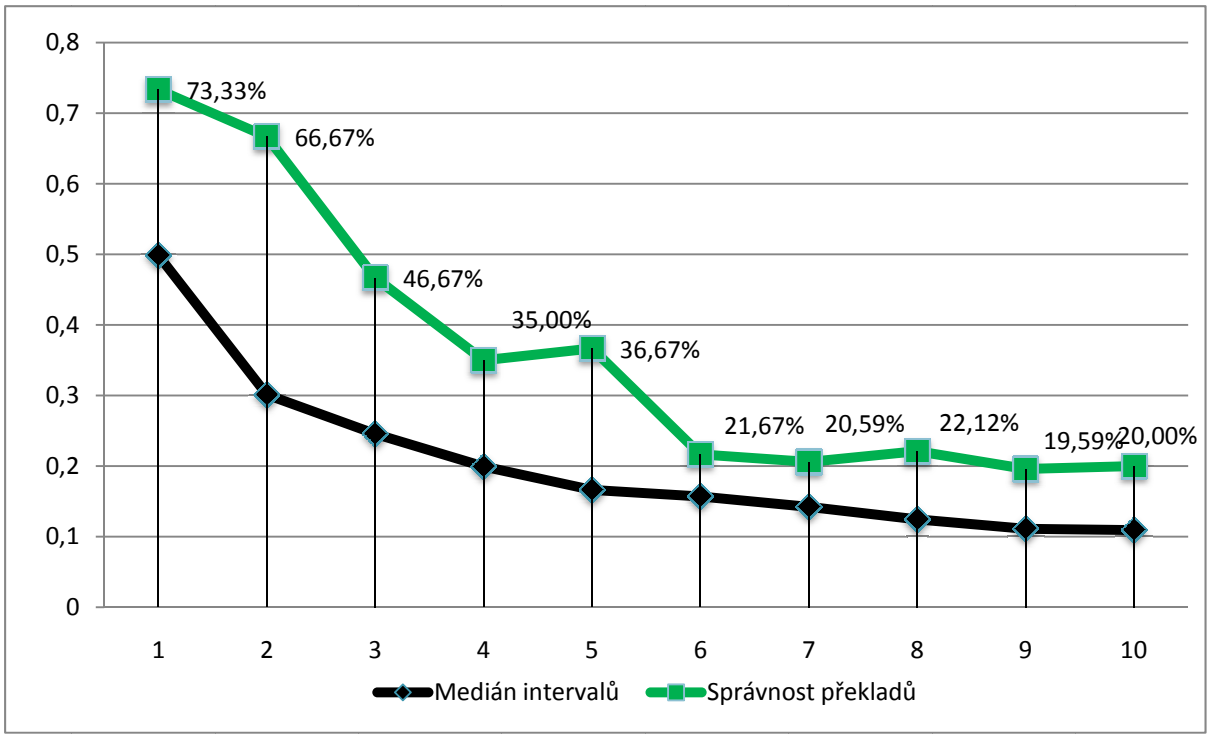
Zajímavostí týkající se kvantity jednotlivých slovních druhů je postupné snižování počtu podstatných jmen v jednotlivých intervalech v kontrastu se zvyšováním počtu sloves. Tento trend platí pro anglická i česká slova tvořící jednotlivé překlady. U českých slov je tento trend o něco znatelnější. Počet podstatných jmen je u nich v posledním intervalu přibližně o 45% menší než v intervalu prvním, zatímco v anglických slovech obsažených na levé straně překladů je rozdíl 23%. Tuto skutečnost zachycuje graf na obrázku 24.



obrázek 24: Vliv hodnoty skóre na počty podstatných jmen a sloves

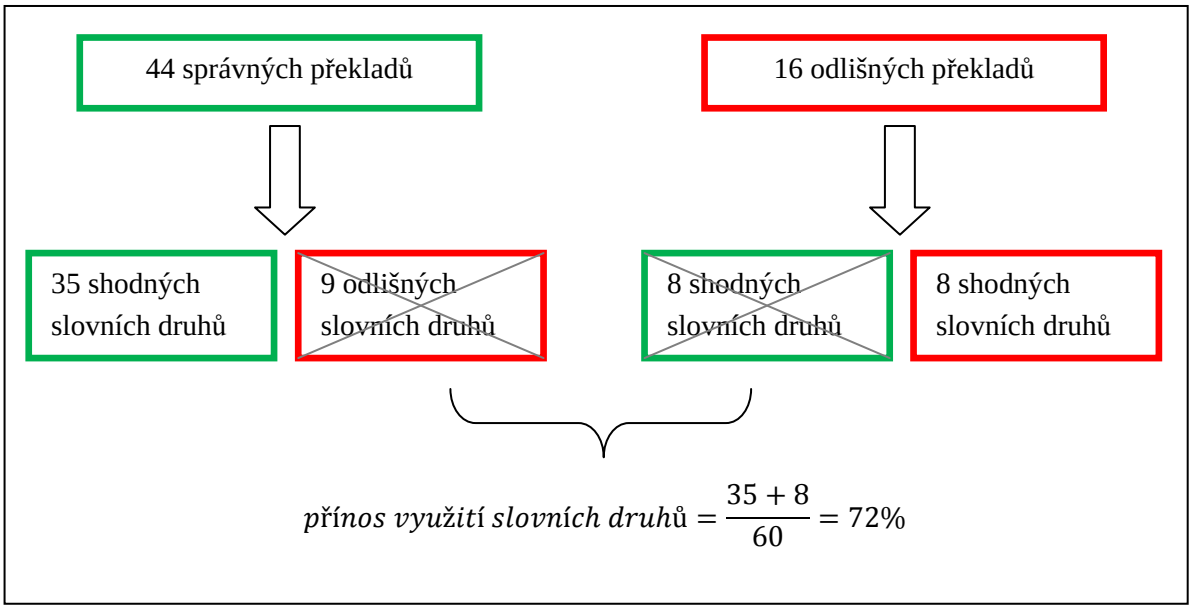
Další zpracování připravených intervalů spočívalo v ručním ohodnocení správnosti překladů. Takto bylo vyhodnoceno prvních 60 překladů v každém z deseti vytvořených překladových intervalů. Počet správných překladů pomocí ručního ohodnocení by měl být nejvyšší v prvním intervalu obsahující překlady s nejvyšším skórem a měl by postupně s dalšími intervaly s menším skórem klesat. Při tomto ručním zpracování byla nalezena řada překladů, které správné nejsou. Je to způsobeno především volným překladem některých větných celků ve vstupních textech, které posloužily jako podklad pro generování překladů. Výsledky tohoto ručního ohodnocení získaných

překladů zobrazuje graf číslo 25. Křivka zobrazující procento překladů vyhodnocených jako správných má klesající tendenci stejně jako křivka symbolizující medián daného intervalu.



obrázek 25: Výsledek ručního vyhodnocení překladů

Získané ruční ohodnocení překladů je vhodné pro otestování přínosu využití slovních druhů v jednotlivých překladech. V optimálním případě by všechny překlady, které vyhodnotíme jako správné, měly mít identický slovní druh slova zdrojového a jeho překladu a v chybných překladech by slovní druhy měly být odlišné, což by vedlo k jejich odfiltrování. V prvním intervalu obsahujícím překlady s nejvyšší hodnotou skóre bylo 73% překladů vyhodnoceno jako akceptovatelné. Zbývajících 26% překladů jako chybné. Při zkoumání slovních druhů, které byly vytvořeným systémem jednotlivým překladům přiřazeny bylo zjištěno, že přibližně 50% z chybných překladů jsou odlišného slovního druhu a tedy systémem vyloučeny. Na druhou stranu z 73% správných překladů bylo z důvodu odlišnosti slovních druhů 20% vyloučeno. Na obrázku 26 je grafické vyjádření popisovaného testu přínosu slovních druhů nad vzorkem 60 překladů z prvního intervalu.



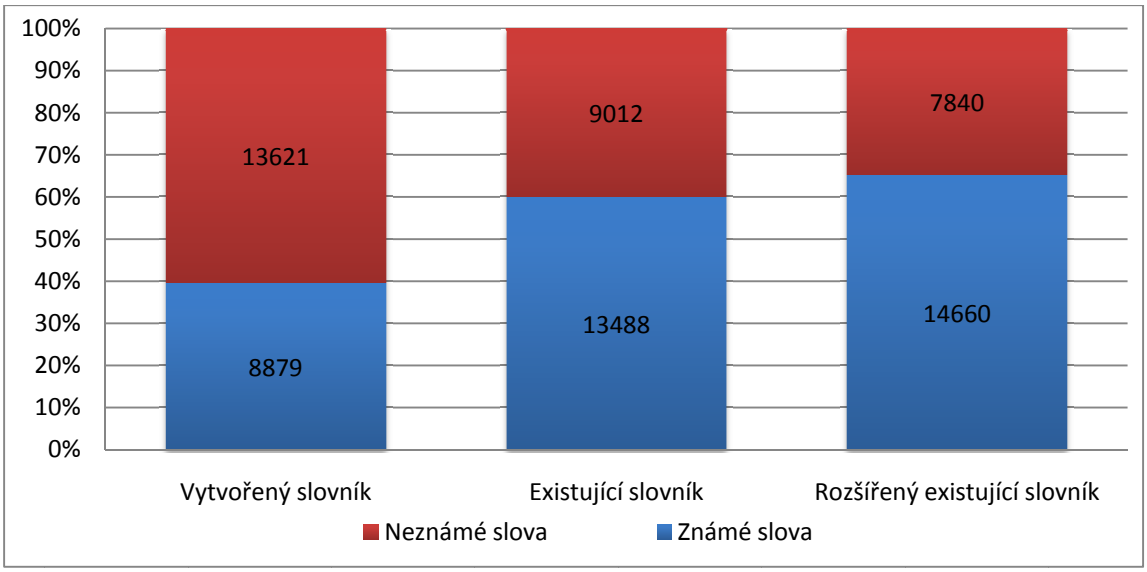
obrázek 26: Výsledek ručního vyhodnocení překladů na vybraném vzorku překladů

V pořadí dalším provedeným testem bylo zjištění možnosti překladů vstupních textů pomocí jednotlivých slovníků. A to pomocí slovníku vytvořeného ze získaných překladů, existujícího slovníku a existujícího slovníku rozšířeného o nové překlady z vytvořeného slovníku. První testovací vstupní text byl vytvořen z obdobných textů jako text zpracovávaný systémem pro získání překladů. Jedná se o texty z paralelního korpusu Kačenka 2. Z těchto textů bylo nutné odstranit nadbytečné informace (autor, datum vzniku, majitel atd.). Protože jejich struktura se shodná s texty přicházející na vstup systému pro získání překladů, posloužil k tomu již vytvořený skript systému. Druhý testovací text byl vytvořen z filmových titulků, které byly náhodně vybrány a staženy z internetového portálu Opensubtitles. Titulky obsahují standardně informace o časovém výskytu jednotlivých vět ve filmu, úpravy formátování písma a podobně. Tyto informace bylo nutné odstranit. Poslední úpravou testovacích textů byla jejich lematizace.

Název	Celkem slov	Unikátních slov	Unikátních slov po lematizaci
Kačenka 2	878201	37037	22500
Titulky	66798	6523	4555

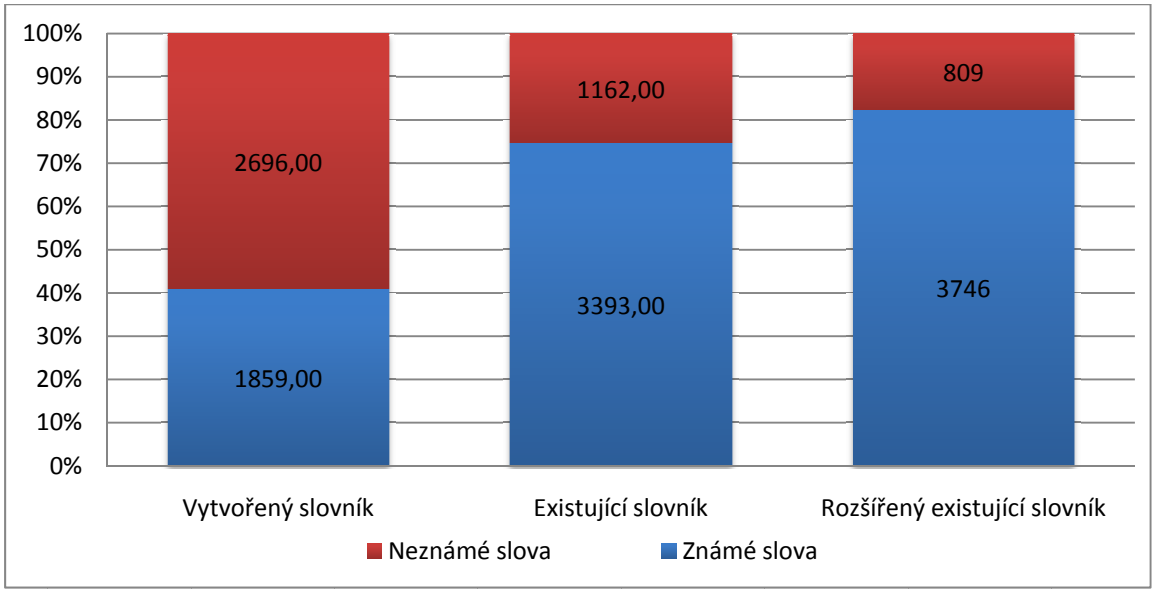
tabulka 11: Testovací anglické texty

Velikost testovacího textu vytvořeného z korpusu Kačenka 2 je téměř shodná jako velikost rozšířeného vstupu systému. Následující graf zobrazuje výsledek porovnání testovacího textu Kačenka s jednotlivými slovníky. Nejúspěšnější slovník je dle očekávání třetí v pořadí, který je vytvořen rozšířením existujícího slovníku pomocí získaných překladů.



obrázek 27: Možnost překladu testovacího textu Kačenka pomocí jednotlivých slovníků

Graf na obrázku číslo 28 ukazuje porovnání zpracovávaných slovníku nad vstupním textem filmových titulků.



obrázek 28: Možnost překladu testovacího textu titulky pomocí jednotlivých slovníků

Testy ukazují přínos rozšíření existujícího slovníku, pomocí něj tak můžeme přeložit více slov vyskytujících se v testovacím textu.

Posledním zde uvedeným testováním je porovnání časových nároků jednotlivých fází systému. Hodnoty byly naměřené na serveru Merlin. Tabulka číslo 12 ukazuje, že jednoznačně časově nejnáročnější operací je zarovnání slov pomocí nástroje GIZA++, před lematizací anglického a českého textu a zarovnáním vět nástrojem Hunalign.

Fáze systému	Výpočetní čas (mm:ss)
Úprava paralelního vstupu	0:04
Anglická lematizace	2:08
Česká lematizace	2:01
Zarovnání vět	0:55
Zarovnání slov	6:52
Generování slovníku	0:02
Optimalizace slovními druhy	0:30
Načtení a rozšíření existujícího slovníku	0:19

tabulka 12: Časové náročnost vytvořeného systému

7 Závěr

Vytvořeným systémem byl získán poměrně rozsáhlý počet překladů. Jejich kvalita je poměrně na slušné úrovni a je diskutována v kapitole věnující se vyhodnocení a testování. Vytvořený systém pro získání překladů je plně automatizovaný. Je možné jej spustit s libovolnými paralelními anglicko-českými vstupními texty. Nad nimi jsou postupně provedeny všechny potřebné úkony až po závěrečné vytvoření slovníku ze získaného zarovnání slov. Takto získaný slovník a překlady v něm obsažené je vhodné využít pro rozšíření existujícího slovníku, což bylo také provedeno. Bylo zjištěno, že přibližně 30% z přeložených slov v získaném slovníku v existujícím není obsaženo. Všechny překlady těchto slov se tak přímo nabízejí k rozšíření slovníku. Připočteme-li k nim i překlady slov, která v existujícím slovníku přeložena jsou, ale jejich varianty překladů jsou odlišné, dostaneme se na hodnotu 84% ze všech získaných překladů, které je možné vložit do struktury existujícího slovníku. Tento proces rozšíření není univerzální a to z důvodu odlišné struktury vstupních slovníků. Pro jejich zpracování je tak potřeba vytvořit patřičné skripty, které zohledňují jeho strukturu.

Při samotném budování tohoto systému bylo potřeba učinit řadu rozhodnutí. Určit vhodné parametry použitých nástrojů a stanovit hranici ohodnocení zarovnání vět a slov. Tato hranice slouží jako limit pro zarovnané dvojice vět, případně slov, která musí být překonána pro jejich akceptaci. Zvýšení této hranice vede k nižšímu počtu získaných překladů, ale na druhou stranu vede k větší kvalitě obdržených výsledků. Na obdržené překlady byla aplikována optimalizace využívající slovních druhů slov vyskytujících se ve slovníku. Základní myšlenou této optimalizace je skutečnost, že překladem slova určitého slovního druhu by mělo být slovo cílového jazyka se shodným slovním druhem. Zjednodušeně řečeno, překladem podstatného jména je opět podstatné jméno. Tímto se nám počet obdržených překladů zredukoval na 37% původního počtu.

Při implementaci popisovaného systému se vyskytla řada dalších problémů, které přímo nesouvisí s řešeným problémem a v tomto textu nebyly zmíněny. Například problémy způsobené odlišným kódováním textů a získaných výstupů. Systém je tvořen jednotlivými částmi, některé z nich byly implementovány v systému Windows a jiné v systému typu Unix. Dalším problémem byla nedostatečná velikost operační paměti, která byla vyžadována nástrojem pro zarovnání vět v případě příliš velkých vstupních paralelních textů. Řešením tohoto problému bylo rozdělení textů na několik menších částí a zpracování každé z nich samostatně.

Tak jako se po několik desítek let neustále vyvíjí oblast strojového překladu, je možné vytvořený systém nadále rozšiřovat a modifikovat tak, aby dosahoval neustále lepších výsledků a to až do té doby, dokud bude jeho výstup odlišný od výstupu produkovaného kvalifikovaným lidským jazykovědcem a překladatelem. Bylo by možné dále pracovat na zvýšení kvality získaných překladů, ale také na jejich využití při tvorbě překladového slovníku. Například obohatit konkrétní překlady o ukázky vět, ve kterých se slova vyskytují.

Literatura

[1] Wikipedia - The Free Encyclopedia: Syntax [online]. http://cs.wikipedia.org/wiki/Syntax

[2] Čermák, F.; Schmiedtová, V.: Český národní korpus - Základní charakteristika a širší souvislosti [online]. http://knihovna.nkp.cz/pdf/0403/0403152.pdf, 2004

[3] Hunalign – sentence level aligner [online]. http://mokk.bme.hu/resources/hunalign

[4] Knight, K.: A Statistical Machine Translation Tutorial Workbook [online]. http://www.isi.edu/natural-language/mt/wkbk.rtf, 1999

[5] Brown, P. F.; Della Pietra, S. A.; Della Pietra V. J.; Mercer, R. L.: The mathematics of Statistical Machine Translation: Parametr Estimation [online]. http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf, 2003

[6] Knight, K.; Koehn, P.: Waht's New in Statistical Machine Translation [online]. http://people.csail.mit.edu/koehn/publications/tutorial2003.pdf, 2003

[7] Wikipedia - The Free Encyclopedia: Levenstein distance [online]. http://en.wikipedia.org/wiki/Levenshtein_distance

[8] Plissonm, J.; Lavrac, N.;Mladenic, D.: A Rule based Approach to Word Lemmatization [online]. http://eprints.pascal-network.org/archive/00000715/01/Pillson-Lematization.pdf

[9] Hunalign - sentence level aligner [online]. http://mokk.bme.hu/resources/hunalign

[10] GIZA++: Training of statistical translation models [online]. http://fjoch.com/GIZA++.html

[11] Průvodce PDT 2.0 [online]. http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/pdf/pdt-guide.pdf, 2006

[12] Automatic Evaluation Of Machine Translation Quality Using N-gram Co-Occurrence Statistic [online]. http://www.itl.nist.gov/iad/mig//tests/mt/doc/ngram-study.pdf

[13] Santori, B.: Part-of-Speech Tagging Guideliness for the Penn Treebank Project [online]. http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/pdf/pdt-guide.pdf, 1991

[14] Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees [online]. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

Příloha A

Popis vytvořených skriptů a práce se systémem

V hlavním adresáři vytvořeného systému nalezneme řadu souborů. Mezi nimi se nachází 2 spustitelné soubory, které postupně spouštějí vytvořené skripty v jazyce Python tak, aby bylo dosaženo požadovaného výsledku. Během zpracování jsou na standardní výstup vypisovány nejrůznější statistiky týkající se právě zpracovávané operace. V následujících řádcích budou blíže popsány skripty, které tyto soubory spouštějí.

vytvorPreklady

Úkolem tohoto spustitelného souboru je ze vstupních paralelních souborů získat překlady a vytvořit tako překladový slovník. Vstupní soubory se nacházejí v adresáři *input* a výsledné překlady v adresáři *output*-nalezneme v něm základní zarovnání *(slovnikEnCz)* a optimalizované zarovnání pracujícího se slovními druhy jednotlivých slov tvořících překlady *(slovnikEnCzOpt)*. Jednotlivá zdrojová slova jsou na samostatných řádcích a za oddělovacím znakem (--) jsou umístěny všechny získané překlady odděleny čárkami. Během zpracování je vytvářena řada výstupních souborů, jsou ukládány do adresáře *tmp*.

K dosažení tohoto cíle bylo implementováno 10 skriptů napsaných v jazyce Python. Jsou jimi postupně:

- upravVstup.py
  Úkolem tohoto skriptu je úprava vstupních textů. Skript upraví všechny texty v adresáři *input*. A to tak, že v něm ponechá pouze abecední a číselné znaky a všechny interpunkční znaky sjednotí na interpunkční tečku. Poslední fází je převedení do podoby, ve které je každá věta na samostatném řádku.
- lematizaceCzech.py
  Zpracuje jednotlivé vstupní soubory upravené v předchozím skriptu a převede jejich slova do základní lemma podoby. Následně jednotlivé lematizované soubory sjednotí a vypíše výsledné statistiky.
- lematizaceEnglish.py
  Protože nástroje pro lematizaci českého a anglického textu jsou odlišené, byl pro každý z těchto jazyků vytvořen samostatný skript pro jejich lematizaci. Skript lematizuje všechny soubory upravených anglických vstupů a po dokončení vypíše souhrné statistiky lematizace.
- zarovnejVety.py
  Jeho funkčnost je založena na nástroji pro zarovnání vět *hunalign*, který iterativně volá pro jednotlivé páry vstupních paralelních textů. Po dokončení je sjednotí do jednoho výsledného souboru obsahujícího zarovnání všech předložených textů.
- hunalign_output_parser.py
  Tento skript slouží pro rozdělení souboru získaného v předchozím kroku na dvě části. Na anglickou a českou část.
- vytvorSlovnik.py
  Tento skript vyhodnocuje získané zarovnání slov pomocí nástroje *GIZA++*. Slova tvořící jednotlivé překlady získané nástrojem *GIZA++* jsou reprezentovány čísly a jejich míra správnosti je vyjádřena pomocí skóre. Skript zpracovává pouze překlady s hodnotou skóre

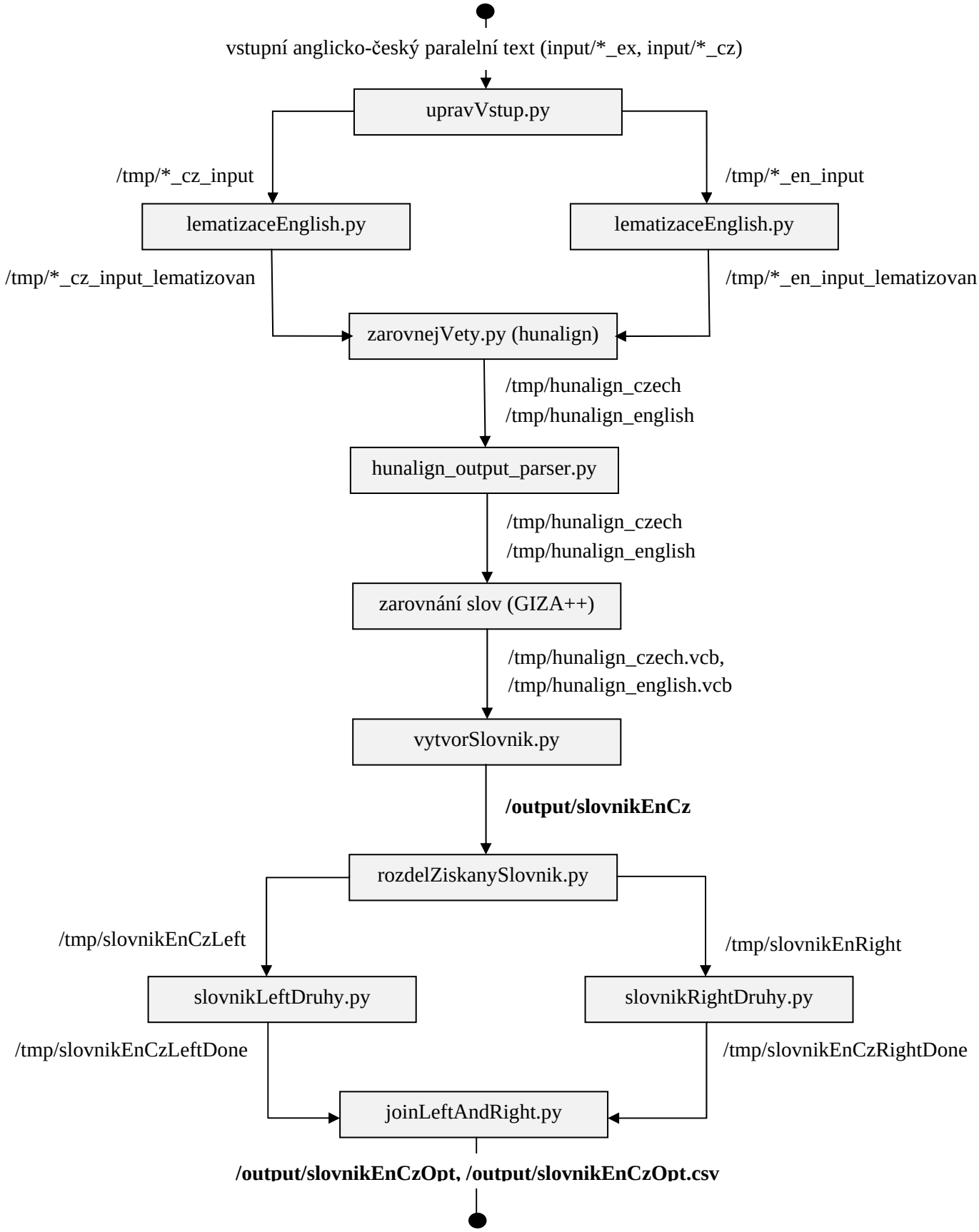
vyšší než 0,1. Pro převod číselné podoby slova na samostatné slovo využívá převodní soubory, které byly získané nástrojem *Plain2snt*. Výstupní soubor *slovnikEnCz* reprezentuje získané překlady, na každém řádku je slovo zdrojové (anglické) a za ním následují všechny jeho získané překlady.

- rozdelZiskanySlovnik.py
  Rozdělení získaného slovníku na dvě části. Na část obsahující anglické slova a na část se slovy českými.
- slovnikRightDruhy.py
  Ohodnocení všech slov ve vstupním souboru pomocí nástroje pro lematizaci českého textu, který pro jednotlivá slova určí jejich slovní druh.
- slovnikLeftDruhy.py
  Ohodnocení všech slov ve vstupním souboru pomocí nástroje pro lematizaci anglického textu, který přiřadí jednotlivým slovům jejich slovní druh.
- joinLeftAndRight.py
  Spojení dvou ohodnocených částí tak, že znovu spojeny jsou pouze překlady mající společný slovní druh. Ostatní překlady jsou ignorovány.

rozsirSlovnik

Druhý spustitelný soubor slouží k rozšíření existujícího slovníku pomocí získaných překladů pomocí předchozího spustitelného souboru. Existující slovník je umístěn v adresáři *inputExistSlovnik* pod názvem *czen.ex*. Výsledný rozšířený slovník nalezneme v adresáři *outputRozsirExistSlovnik*. Slovník je rozšířen o všechny získané překlady, které v něm nejsou obsaženy. Pro toto doplnění jsou využity překlady zpracované pomocí optimalizace využívající slovní druhy *(output/slovnikEnCzOpt)*. Výsledný rozšířený slovník si nadále zachovává svou původní strukturu. Rozšíření je založeno na provedení následujících tří skriptů:

- zpracujExistujiciSlovnik.py
  Tento skript slouží ke zpracování existujícího slovníku. Jedná se XML parser, který odpovídá struktuře dodaného slovníku. Pomocí něj jsou exportovány všechny nalezené překlady do samostatného souboru, který svou strukturou odpovídá námi vytvořenému slovníku.
- porovnejSlovniky.py
  Porovnání vytvořeného slovníku a převedeného existujícího slovníku pomocí předchozího skriptů. Výstupem jsou dva soubory, které obsahují překlady slov, které v existujícím slovníku nejsou vůbec přeloženy a nové překlady slov v existujícím slovníku obsaženy.
- rozsirExistujiciSlovnik.py
  Provede samotné rozšíření pomocí dvou vstupních souborů získaných v předchozím bodě. Jeho výstupem je výsledný rozšířený slovník, který je vytvořen a umístěn v podadresáři *outputRozsirExistSlovnik.*



Obrázek 29: Skripty vytvořené za účelem získání překladů

Příloha B

Struktura přiloženého média

Na přiloženém médiu jsou dodány všechny součásti vytvořeného systému. Pro korektní běh systému je potřeba instalace nástrojů, které jsou popsány v páté kapitole. Nástroje *GIZA++* a *Hunalign* jsou dodány na přiloženém médiu.

Pro spuštění na serveru Merlin je nejprve potřeba nastavit následující proměnné prostředí:

```
export LD_LIBRARY_PATH=/mnt/minerva1/nlp/local64/lib:.
export LD_RUN_PATH=/mnt/minerva1/nlp/local64/lib:.
PATH=/mnt/minerva1/nlp/local64/bin:/mnt/minerva1/nlp/local/bin/:"$PATH"
export PYTHONPATH=/mnt/minerva1/nlp/local64/lib/python2.5
export MINIPATH=/mnt/minerva1/nlp/software/minipar/data
export PATH=/mnt/minerva1/nlp/software/TreeTagger/cmd/:$PATH
```

Adresářová struktura média je následující:

DP_xmusil29
| -- system
| | -- input
| | -- giza-pp
| | -- hunalign-1.0
| | -- inputExistSlovnik
| -- readme
| -- vyhodnoceni
| -- xmusil29.pdf

Systém můžeme spustit pomocí připravených spustitelných souborů, které se nacházejí v adresáři *system*. Jmenují se *vytvorPreklady* a *rozsirSystem*. Jejich bližší popis je v příloze B.