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

PŘEKLADAČ Z ČEŠTINY DO SLOVENŠTINY

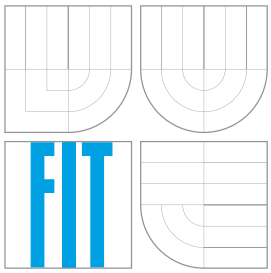
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JÁN MYDLIAR

BRNO 2013

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

PŘEKLADAČ Z ČEŠTINY DO SLOVENŠTINY

CZECH-SLOVAK MACHINE TRANSLATION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JÁN MYDLIAR

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2013

Abstrakt

Tato diplomová práce se věnuje tvorbě překladového systému pro překlad z češtiny do slovenštiny. První kapitola pojednává o motivaci k tvorbě práce, druhá o strojovém překladu a jeho různých typech. Třetí kapitola obsahuje přehled metod pro hodnocení kvality strojového překladu. Čtvrtá kapitola se věnuje návrhu a realizaci překladového systému, zejména přípravě paralelních korpusů. Pátá kapitola pojednává o testovaní a vyhodnocení vytvořeného systému.

Abstract

This Master thesis deals with machine translation from Czech to Slovak. The first chapter motivates the work, the second discusses various approaches to machine translation and the third details evaluation of the methods. Chapter 4 introduces the design and implementation of my system, paying a special attention to a new parallel corpus that has been created. Chapter 5 summarizes testing and evaluation of the developed system.

Klíčová slova

strojový překlad, statistika, MOSES, paralelní korpus, BLEU, GIZA++

Keywords

machine translation, statistics, MOSES, parallel corpus, BLEU, GIZA++

Citace

Ján Mydliar: Překladač z češtiny do slovenštiny, diplomová práce, Brno, FIT VUT v Brně, 2013

Překladač z češtiny do slovenštiny

Prohlášení

Prohlasuji, že jsem tuto diplomovu práci vypracoval samostatně pod vedením pana docenta Pavla Smrže. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Ján Mydliar
20. května 2013

Poděkování

Chtěl bych poděkovat vedoucímu práce a Ing. Janu Kouřilovi za pomoc při zpracování paralelních textů a tvorbě výsledného systému.

Obsah

Kapitola 1

Úvod

Preklad začal byť neodmysliteľnou súčasťou ľudskej komunikácie, po tom ako sa začali stretávať prvé národy. Každý mal svoju vlastnú históriu, svoj vlastný jazyk. Aby sa mohli spolu dorozumieť, potrebovali určitú formu prekladu. Ten bol spočiatku na báze gestikulácie a rôznych pohybov celého ľudského tela. Postupne s vývojom jazykov sa začal zdokonaľovať preklad. V súčasnosti je veľmi používaný, nakoľko žijeme v dobe globalizácie a neustáleho pohybu jednotlivcov či skupín po celom svete. Na to, aby človek dokázal plnohodnotne žiť v cudzej krajine s cudzím jazykom, musí ovládať tamojší jazyk. Najpoužívanejším jazykom v súčasnosti je angličtina.

Nasledujúci text bol voľne prevzatý z [1]. Počiatky strojového prekladu siahajú už do doby pred samotnými osobnými počítačmi. V roku 1930 páni Georges Artsrouni a Petr Troyanskii položili základy tzv. prekladových strojov. Druhý menovaný predstavil nielen metódy automatického dvojjazyčného slovníka, ale aj určité schéma pre kódovanie medzijazykových gramatických pravidiel. Bol predložený aj určitý návrh toho, ako by mohli v budúcnosti prekladové systémy fungovať. Tieto objavy našli plnohodnotné uplatnenie až po vynájdení počítačov. Prvotný výskum v oblasti strojového prekladu začal na viacerých amerických univerzitách. V roku 1954 v spolupráci IBM a univerzitou v Georgetowne sa konala prvá verejná ukážka realizovateľnosti strojového prekladu na osobných počítačoch. Výsledkom bol zvýšený záujem o strojový preklad a s tým súvisiaci masívny prísun potrebných finančných prostriedkov. Prvotné systémy obsahovali dvojjazyčné slovníky, ktorých veľkosť postupne narastala. Cieľom bolo zaviesť do prekladu systematickosť. Metódy prekladu sa opierali o najnovšie poznatky v lingvistike (formálna gramatika) a zdalo sa, že kvalita prekladu sa bude tým zlepšovať. Tieto úvahy vydržali približne desaťročie. Vývin brzdila predovšetkým sémantika jednotlivých jazykov. Boli vytvorené aj špeciálne operačné systémy pre strojový preklad, no ich výsledky nenaplnili očakávania. Dospelo sa k názoru, že strojový preklad je pomalší a dvakrát drahší ako ľudský. A tak sa výskum začal orientovať skôr na pomôcky pre prekladateľov a oblasť počítačovej lingvistiky. Jedným z prvých používaných systémov pre strojový preklad bol Systran používaný ministerstvom obrany USA a Európskou komisiou. Nasledoval systém Meteo, ktorý bol vyvinutý v Kanade. Dopyt po strojovom preklade začal hlavne kvôli medzinárodnému obchodu. Dostupnosť mikropočítačov a softvérov na spracovávanie textu sa neustále zvyšovala. Na základe toho vznikol trh s lacnými prekladateľskými systémami (ALPS, Weidner, Globalink).

Koncom 80–tych rokoch 20. storočia nastal vo vývoji strojového prekladu zlom. Zapríčinili to výsledky čisto štatistického prekladu systému Candide od IBM a preklad založený na príkladoch (korpus dvojjazyčných prekladov) od japonských vývojárov. Spoločným znakom oboch systémov bolo to, že nepoužívali žiadne syntaktické ani sémantické pravidlá

pri analýze textu. Namiesto toho mali k dispozícií korpusy paralelných dvojjazyčných textov. Začal sa výskum aj v oblasti zahrňujúcej rozpoznávanie reči a syntézu reči. Postupne systémy pracovali s prekladom založeným na pravidlách spolu s korpusom dvojjazyčných textov (ATR, JANUS, Vermobil). Zmenila sa aj oblasť výskumu. Začalo sa orientovať na praktické využitie strojového prekladu. Vznikali prekladateľské stanice pre profesionálnych prekladateľov. Potreba prekladu narastala hlavne kvôli tzv. lokalizácii softvéru, kde je potrebné nové programy a ich dokumentáciu preložiť pre potreby expanzie na nové trhy. Tak isto počiatkom 90–tych rokov narastal predaj strojových prekladových systémov pre osobné počítače. Vznikali aj online systémy.

Medzi najväčšie súčasné projekty týkajúce sa strojového prekladu patrí EuroMatrix-Plus [2]. Je nástupcom predchádzajúceho projektu EuroMatrix. Jeho hlavnou úlohou je vytvoriť systémy pre preklad širokého spektra európskych jazykov. Ďalšou úlohou je priniesť strojový preklad aj medzi bežných užívateľov, ako aj profesionálnych prekladateľov. Do projektu môžu prispieť aj dobrovoľníci, ktorí vytvárajú obsah prekladom cudzích materiálov do ich vlastných jazykov. Projekt potom skúma, ako môžu títo užívatelia využívať strojový preklad a naopak, ako je možné zlepšiť strojový preklad na základe užívateľských prekladov. Na preklad používa voľne šíriteľný program Moses, ktorý patrí medzi najrozšírenejšie v rámci vedeckej obce.

Najznámejším a zrejme aj najpoužívanejším voľne dostupným prekladačom je Google prekladač [3]. Okrem prekladu samotných slov a viet dokáže prekladať aj cele webové stránky. V súčasnosti podporuje preklad medzi 66 svetovými jazykmi, medzi ktoré patria aj menej známe jazyky, ako napr. bengálčina, laoština a iné. Preklad spočíva v hľadaní vzorov v miliónoch dokumentov. Tieto vzory sú preložené prekladateľmi, a tak je možné dosiahnuť čo najvhodnejší preklad. Samotnú kvalitu prekladu je možné vylepšiť pomocou ohodnotenia prekladu alebo odovzdaním vlastných prekladov pomocou nástroja pre prekladateľov. Prekladač sa neustále vylepšuje, pridávajú sa nové jazyky, a tak jeho potenciál do budúcnosti je obrovský.

Začiatok práce sa venuje strojovému prekladu, jeho rozdeleniu podľa spôsobu prekladu na preklad založený na pravidlách, preklad založený na príkladoch, štatistický strojový preklad a hybridný preklad. Z typov prekladov založených na pravidlách sú popísané priamy preklad, pravidlový transferový preklad a interlinguálny preklad. Najpodrobnejšie sa teoretická časť práce venuje štatistickému prekladu. Tento preklad má niekoľko častí, z ktorých sa postupne skladá. Sú to model jazyka, model prekladu (IBM modely), zarovnanie slov. Súčasťou štatistického prekladu je aj model založený na frázach alebo na stromoch a ich dekódovanie. V podkapitole o hybridnom preklade sú uvedené spôsoby z akých typov prekladov sa tento preklad skladá, jeho výhody a nevýhody. Tretia kapitola sa venuje vyhodnocovaniu kvality prekladu. Sú popísané najznámejšie metódy, najmä jedna z najpoužívanejších BLEU a jej rôzne vylepšenia. Nasleduje kapitola o samotnom návrhu systému prekladu a jeho realizácii. Je popísaná príprava paralelných textov, použitie dát zo slovníkov, ako aj dôkladný popis vytvorených pravidiel pre prípony slov. Posledná kapitola sa venuje testovaniu a vyhodnoteniu výsledkov vytvoreného prekladového systému z češtiny do slovenčiny. Porovnávajú sa preklady medzi vytvorenými systémami navzájom, ako aj s voľne dostupnými prekladačmi.

Kapitola 2

Strojový preklad

Ako už bolo spomínané v úvode, strojový preklad mal na začiatku vysoké očakávania, ktoré sa však nepodarilo splniť. Vedci postupne zisťovali mieru komplikovanosti prekladu, a tak v súčasnosti sú výroky o pokroku v tejto oblasti omnoho opatrnejšie, ako to bolo v minulosti. Cieľom prekladu by teda malo byť preložiť text tak, aby ho bolo možné pochopiť v cieľovom jazyku. Nestačí len preklad na úrovni slov alebo viet. Je potrebné poznať kontext každej vety, a to je najväčším problémom pre automatický strojový preklad.

V tejto kapitole bližšie rozoberieme jednotlivé druhy strojových prekladov, ich výhody, nevýhody a použitie. Konkrétne preberieme preklad založený na pravidlách, na príkladoch, štatistický strojový preklad a hybridný preklad. Hlavné zameranie bude na štatistický strojový preklad, ktorý bude využitý pri testovaní paralelných textov systémom Moses.

2.1 Strojový preklad založený na pravidlách

Aby bola kvalita prekladu čo najlepšia, je potrebné zaviesť do procesu prekladu určité pravidlá. Najjednoduchší preklad slova po slove je použiteľný, avšak z hľadiska pochopenia kontextu nie príliš vhodný. Práve použitím rôznych pravidiel môžeme preklad urýchliť ako aj skvalitniť. V tejto časti budú popísané metódy, ktoré sa zaoberajú týmito pravidlami. Informácie z nasledujúcich podkapitol boli prevzaté z [4].

2.1.1 Priamy preklad

Je to snáď najintuitívnejší preklad z hľadiska funkčnosti. Preložený text získame postupným nahradzovaním každého slova zo zdrojového jazyka do cieľového jazyka. Preklad slova získame zo slovníka. Pri tomto type prekladu sa neprihliada na kontext ani význam slova. Pred prekladom môže ešte dôjsť k morfologickej analýze. Môže sa používať na preklad rôznych katalógov, kde je väčšina textu zadaná vo forme hesiel. Problémom tohto prekladu je, že pre každý pár dvoch jazykov je potrebné osobitný prekladač alebo slovník.

2.1.2 Pravidlový transférový preklad

Metóda začína syntaktickou a morfologickou analýzou zdrojového textu. Výsledkom je syntaktická reprezentácia textu (môže byť aj zjednodušená tým, že sa kladie dôraz len na významovo podstatné časti textu). Následne sa z nej vygeneruje cieľový text. Ak bol už zdrojový text v určitom stupni abstrakcie, táto abstrakcia sa dodrží aj na cieľovom preloženom texte. Samotný proces prekladu pomocou tejto metódy môžeme rozdeliť nasledovne:

1. Morfologická analýza – zdrojový text je identifikovaný na morfologické jednotky (slovo). Tejto jednotke sa ďalej určí slovný druh, rod, početnosť a pod. Výstupom tohto kroku je niekoľko možných analýz vstupného textu, spolu s ich kanonickou reprezentáciou.

2. Lexikálna analýza – niektoré slová môžu mať v texte nie jeden, ale aj viacero významov. Analyzuje sa kontext v okolí týchto slov, aby sa minimalizoval výskyt nesprávnych významov týchto slov.

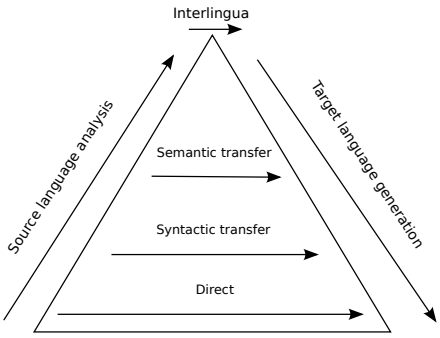
3. Lexikálny transfer – predstavuje preloženie slova vo svojej kanonickej podobe spolu s informáciou o jeho význame zo zdrojového do cieľového jazyka. Táto fáza sa zaoberá len prekladom samotných slov.

4. Štrukturálny prenos – tu už pracujeme s celými frázami alebo vetami. Dochádza ku korekciám pádov, časov či početnosti slov. Slová sú usporiadané podľa správneho poradia vo vete.

5. Morfologické generovanie – výstup z predchádzajúceho kroku je transformovaný na podobu vstupu (rovnaké rozloženie textu, odstavcov a pod.).

2.1.3 Pravidlový interlinguálny preklad

Táto metóda je podobná s predchádzajúcou, okrem jedného rozdielu. Zatiaľ čo pri transferovom preklade sa zo zdrojového jazyka po analýze a príslušnej reprezentácií, priamo prekladalo do cieľového jazyka, v tomto prípade je to inak. Tu sa zdrojový jazyk prevedie do jazyka Interlingua (umelý jazyk). Reprezentácia v tomto jazyku nie je závislá na zdrojovom jazyku. Preklad do cieľového jazyka sa generuje z jazyka Interlingua. Výhodou tejto metódy je nezávislosť v rámci počtu jazykov (predtým sa prekladalo vždy medzi dvoma jazykmi). Prekladač môže mať na vstupe a výstupe rôzne jazyky. Nevýhodou je komplikovaná definícia samotných pravidiel pre preklad z a do jazyka Interlingua pri väčších doménach. Preto je táto metóda vhodná pre doménovo špecifické oblasti. Schéma tejto metódy je zobrazená pomocou Vaqoisovho trojuholníka (obrázok 2.1).



Obr. 2.1: Vaqoisov trojuholník [5]

2.2 Strojový preklad založený na príkladoch

Text bol voľne prevzatý z [6]. Základnou myšlienkou tejto metódy je použitie algoritmu najlepšej zhody k nájdeniu príkladu najbližšieho k zdrojovej vete v dvojjazyčnom korpuse

prekladových párov. Výsledkom je šablóna, ktorá môže byť vyplnená prekladom po slovách. Z korpusu tak získame príklad, ktorý upravujeme podľa potrieb prekladu. Pri výbere vhodného príkladu vzniká problém určenia práve tohto najbližšieho príkladu, kedy máme na výber viacero možností a musíme vybrať ten najbližší.

S touto metódou vyšiel prvý krát na verejnosť pán Nagao v roku 1984. Navrhoval použitie neanotovanej dvojjazyčnej databázy príkladov a množiny lexikálnych ekvivalentov. Neanotované dáta sa vyznačujú svojou stálosťou v kontraste s rôznymi lingvistickými teóriami. Hľadanie zhody bolo zamerané na sémantickú podobnosť lexikálnych jednotiek (vstupná veta a kandidát na zhodu). Postupom času sa táto metóda rozširovala, a to hlavne v týchto smeroch:

- Obohatenie databázy príkladov o anotácie a ich využitie k lingvistickej analýze vstupu pred hľadaním príkladu.
- Ukladanie šablón namiesto viet. Šablóna je potom definovaná ako veta, v ktorej sú niektoré časti (frázy) nahradené premennými s anotáciami. Aby mohli byť slová nahradené týmito premennými s anotáciami, musia sa vety syntakticky analyzovať.
- Využívanie viacerých príkladov a ich vhodných kombinácií k prekladu viet. Definuje sa ohodnotenie príkladových jednotiek a zohľadňuje sa ich dĺžka a sémantická podobnosť.

Pred hľadaním čo najbližšieho príkladu sa vety rozložia na zložky. Presné porovnávanie celých viet by bolo príliš obmedzujúce a úspešnosť takéhoto hľadania by bola zrejme veľmi nízka. Vety sa rozkladajú podľa rôznych pravidiel, napr. podľa interpunkcie, podľa morfologicky analyzovaných segmentov alebo analýza viet do závislostných stromov. Po nájdení príkladov je potrebné vybrať ten správny. Nakoľko presná zhoda nastáva len ojedinele, je potrebné určiť, do akej miery je nutná dostupnosť príkladov celých viet pri preklade zložiek. Zdrojové dáta by mali obsahovať obecné príklady a príklady výnimiek, čo by malo pokryť všetky možné preklady.

Výhodou prekladu založeného na pravidlách je, že rovnaké zdroje (paralelný korpus) využíva k viacerým účelom. Napr. sa využívajú vety cieľového jazyka ako šablóny k rekombinácií. Rekombinácia potom do istej miery nahradzuje generovanie a na to by iné metódy potrebovali poznať gramatiku cieľového jazyka.

2.3 Štatistický strojový preklad

Začiatok štatistického strojového prekladu sa datuje na rok 1949, kedy Warren Weaver navrhol použitie štatistických a kryptoanalytických techník pre preklad zo zdrojového jazyka do cieľového jazyka. V tom čase nebol tento prístup úspešný hlavne kvôli neexistujúcim počítačom s potrebným výpočtovým výkonom. V súčasnosti však je táto metóda veľmi využívaná vďaka svojej nezávislosti na jednotlivých jazykoch. Táto podkapitola bola spracovaná podľa [7], [8] a [9].

Máme reťazec e v zdrojovom jazyku. Tento reťazec môže byť preložený do cieľového jazyka viacerými výslednými prekladmi. Každý výsledný preložený reťazec f berieme ako možný preklad pôvodného reťazca e. Ku každému páru reťazcov (e, f) priradíme číslo $Pr(e, f)$, ktoré udáva pravdepodobnosť prekladu z reťazca e do f. Úlohou je nájsť taký reťazec e, ktorý je najbližším prekladom cieľového reťazca f, tak ako keby ho prekladal človek s materským jazykom rovnakým ako je jazyk reťazca f. Výberom reťazca $\widehat{e}$ s najväčším číslom

$Pr(e, f)$ tak minimalizujeme chyby v preklade. S využitím Bayesovho teorému dostávame

$$Pr(e, f) = \frac{Pr(e) \cdot Pr(f|e)}{Pr(f)} \tag{2.1}$$

Vzhľadom k tomu, že menovateľ je nezávislý voči e, nájdenie $\widehat{e}$ môžeme preformalizovať ako nájdenie najväčšieho možného čitateľa $Pr(e) \cdot Pr(f|e)$. Dostávame potom základnú rovnicu štatistického strojového prekladu

$$\widehat{e} = \arg\max_{e} Pr(e) \cdot Pr(f|e) \tag{2.2}$$

V rovnici 2.2 $Pr(f|e)$ udáva podmienenú pravdepodobnosť každého dvojjazyčného páru. Pri správnej voľbe tohto rozloženia môžeme zvýšiť kvalitu prekladu. Otázkou však zostáva, aké je to rozloženie a či sme vôbec schopní toto rozloženie nájsť, aby kvalita prekladu bola dostatočná. Kvôli tomu vznikajú 3 hlavné výpočtové problémy:

1. odhad pravdepodobnosti jazykového modelu $Pr(e)$
2. odhad pravdepodobnosti modelu prekladu $Pr(f|e)$
3. nájdenie efektívneho a účinného algoritmu pre hľadanie maximálneho $\widehat{e}$

2.3.1 Model jazyka

Určuje pravdepodobnosť prekladu cieľového reťazca f zo zdrojového e. Tento model pomáha zaistiť správny slovosled. Pri viacznačných slovách pomáha vyberať správny preklad buď na základe početnosti alebo pomocou kontextu. V rámci jazykových modelov sa využívajú N–gramové modely. Niektoré slová sa často vyskytujú vo dvojiciach $W = w_1, w_2, \ldots, w_n$. Pravdepodobnosť $p(W)$ vypočítame tak, že spočítame výskyty všetkých W a normalizujeme ich veľkosťou dát. $p(W)$, kde W je postupnosť slov, budeme modelovať postupne, použitím pravidla rezu.

$$p(w_1, w_2, \ldots, w_n) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \ldots p(w_n|w_1 \ldots w_{n-1}) \tag{2.3}$$

Pomocou Markovovho predpokladu môžeme obmedziť rovnicu 2.3 a dostaneme rovnicu 2.4.

$$p(w_1, w_2, \ldots, w_{n-1}) \simeq p(w_n|w_{n-m}, \ldots w_{n-2}, w_{n-1}) \tag{2.4}$$

Číslo m určuje rád N–gramového modelu. Najčastejšie sa používajú trigramové modely. Kvalita jazykových modelov sa meria krížovou entrópiou (cross–entropy). Tá udáva priemernú hodnotu záporných logaritmov pravdepodobnosti slov v testovacom texte. Odpovedá miere neistoty pravdepodobnostného rozloženia jazykového modelu.

U jazykových modelov môže vznikať problém s nevidenými dátami. Ak v dátach nie je určitý N–gram, ktorý sa vyskytne v reťazci w, pre ktorý hľadáme pravdepodobnosť, bude $p(w) = 0$. Musí platiť $\forall w.p(w) > 0$. Rieši sa to vyhladzovaním rôznymi metódami, ako napr. *add–one*, *add–α* a *Good–Turing* vyhladzovanie.

2.3.2 Model prekladu

V klasickom slovníku nemáme pri preložených slovách informáciu o tom, ako často sa slovo zo zdrojového jazyka prekladá na dané preložené slovo. V tomto modely potrebujeme tzv. lexikálne prekladové pravdepodobnostné rozloženie p_f s vlastnosťou

$$\sum_e p_f(e) = 1, \qquad \forall e : 0 \leq p_f(e) \leq 1 \tag{2.5}$$

Aby sme pri preklade dosiahli správny počet slov, musíme zaviesť tzv. zarovnávaciu funkciu.

$$a : j \to i \tag{2.6}$$

Kde j je pozícia odpovedajúceho slova v cieľovej vete a i je pozícia v zdrojovej vete. Potom a je funkcia, ktorá definuje pre každé slovo w_e z cieľovej vety existuje práve jedno slovo w_f zo zdrojovej vety. Samotný preklad rozložíme do niekoľkých menších krokov, kde budeme používať p_f pre slová. Tu využívame tzv. IBM modely, ktoré budú ďalej popísané.

IBM Model 1

Tento model je definovaný nasledovne:

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(I_f + 1)^{I_e}} \prod_{j=1}^{I_e} t(e_j|f_{a(j)}) \tag{2.7}$$

kde $\mathbf{e} = (e_1, \ldots, e_{I_e})$ je cieľová veta, $\mathbf{f} = (f_1, \ldots, f_{I_f})$ zdrojová veta, I_e je dĺžka cieľovej vety, I_f je dĺžka zdrojovej vety, ϵ je normalizujúca konštanta, aby bol výsledný súčin pravdepodobnostným rozložením. $(I_f + 1)^{I_e}$ je počet všetkých možných zarovnaní medzi $\mathbf{e}$ a $\mathbf{f}$, pričom k I_f pripočítame 1 kvôli špeciálnemu slovu NULL, t je pravdepodobnostná prekladová funkcia. Aby sme mohli vypočítať $p(\mathbf{e}, a|\mathbf{f})$ musíme poznať hodnotu funkcie t pre každé slovo. V zdroji dát (paralelný korpus) sú texty zarovnané na úrovni viet, my však potrebujeme zarovnanie na úrovni slov, tzv. *word alignment*. To získame aplikáciou algoritmu *expectation–maximization* (EM). Algoritmus je definovaný takto:

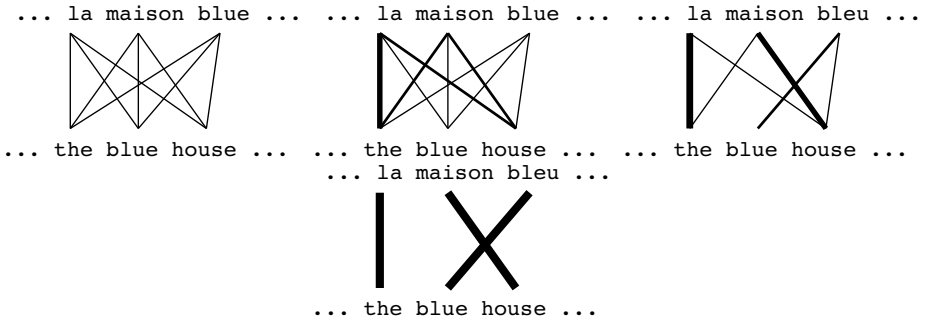
1. Inicializuj model.
2. Aplikuj model na dáta (expectation), hľadaj

   $$p(a|e, f) = \frac{p(a|e, f)}{p(e|f)} \tag{2.8}$$

   kde $p(e|f) = \sum_a p(e, a|f)$.
3. Uprav model podľa dát (maximization) a uprav počet zarovnaní slov w_e na w_f (funkcia c) pomocou predchádzajúceho

   $$c(w_e|w_f; e, f) = \sum_a p(a|e, f) \sum_{j=1}^{I_e} \delta(e, e_j)\delta(f, f_{a(j)}) \tag{2.9}$$

   kde $\delta(x, y) = 1 \Leftrightarrow x == y$, inak 0.
4. Opakuj body 2 a 3, ak je čo zlepšovať.



Obr. 2.2: Štyri iterácie EM algoritmu [7]

Výsledná prekladová pravdepodobnosť sa vypočíta pomocou funkcie c z rovnice 2.9

$$t(w_e|w_f) = \frac{\sum_{(e,f)} c(w_e|w_f; e, f)}{\sum_{w_e} \sum_{(e,f)} c(w_e|w_f; e, f)} \tag{2.10}$$

Niekoľko iterácií algoritmu EM je zobrazených na obrázku 2.2, kde prekladáme francúzsku vetu *la maison blue* na vetu anglickú *the blue house*. V prvom kroku sú všetky slová zarovnané rovnako s rovnakou váhou. V ďalšom kroku algoritmus zistí, že zarovnanie slova *la* so slovom *the* je pravdepodobnejšie ako s *house* alebo *blue*. Algoritmus skončí po zistení najlepšieho možného zarovnania všetkých slov.

Tento model je však príliš jednoduchý, neuvažuje kontext, nevie pridávať alebo odoberať slová. Všetky rôzne zarovnania považuje za rovnako pravdepodobné.

IBM Model 2

Oproti modelu 1 pridáva explicitný model pre zarovnanie, tzv. *alignment probability distribution* $a(i|j, l_w, l_f)$, kde i je pozícia zdrojového slova a j pozícia cieľového slova. V prvom kroku (obrázok 2.3) prekladu sa preložia lexikálne jednotky, používa sa pritom $t(e|f)$. V druhom kroku sa podľa modelu zarovnania preskupia preložené slová. Funkcia a aj pravdepodobnostné rozloženie a je v opačnom smere ako preklad. Oba rozloženia sa kombinujú do vzorca:

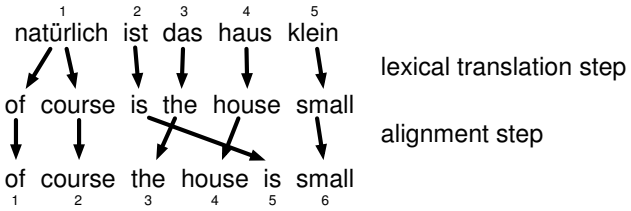
$$p(e, a|f) = \epsilon \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) a(a(j)|j, l_e, l_f) \tag{2.11}$$

$$p(e|f) = \sum_a p(e, a|f) = \epsilon \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_i|f_i) a(i|j, l_e, l_f) \tag{2.12}$$

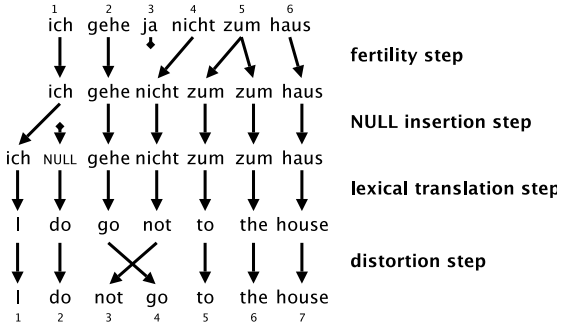
IBM Model 3

Tento model (obrázok 2.4) rieši situáciu, keď sa slovo preloží na viacero slov, poprípade sa nepreloží vôbec. Definuje tzv. *fertilitu*, ktorá je modelovaná pravdepodobnostným rozložením $n(\phi|f)$. Pre každé zdrojové slovo f rozloženie n udáva počet slov, na ktoré sa f preloží. V prípade, že sa dané slovo nepreloží vôbec, zavádzame pomocný token NULL. Na to sa používajú p_1 a $p_0 = 1 - p_1$, kde p_1 udáva pravdepodobnosť vloženia tokenu NULL za ľubovoľné slovo vo vete. Posledný krok tohto modelu je modelovaný pomocou *distortion probability distribution* $d(j|i, l_e, l_f)$. Pozície sa modelujú v opačnom poradí. Pre zdrojové

slovo na pozícií i modeluje pozíciu j cieľového slova.



Obr. 2.3: Kroky prekladu IBM Model 2 [7]



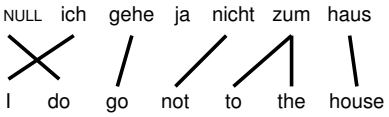
Obr. 2.4: Kroky prekladu IBM Modelu 3 [7]

IBM Model 4

Tento model definuje tzv. *relatívnu distorziu*, kde zmeny pozícií slov závisia na predchádzajúcich slovách. Vychádza sa z predpokladu, že sa prekladá po frázach, ktoré sa presunujú v celku, alebo presuny sú veľmi časté. Model definuje tzv. *cept* π tvorený slovami s rovnakým zarovnaním (fertilita daného slova je väčšia ako nula). Ďalej je definovaný pojem *centrum ceptu* $\odot$, ktorý udáva zaokrúhlený priemer súčtu pozícií cieľových slov. Na obrázku 2.5 vidíme príklad 5 ceptov nemeckých slov, ktoré odpovedajú minimálne jednému preloženému anglickému slovu. Niektoré slová však môžu spôsobiť zámenu v poradí slov preloženého textu. Preto je potrebné zaviesť určité podmienky:

$$\text{pre prvé slovo ceptu:} \qquad d_1(j - \odot_{[j-1]} | f_{[i-1]}, e_j) \tag{2.13}$$

$$\text{pre pridané slová:} \qquad d_{>1}(j - \prod_{i,k-1} | e_j) \tag{2.14}$$



Obr. 2.5: Cepty IBM Model 4 [7]

IBM Model 5

Predchádzajúce modely (1–4) majú niekoľko nedostatkov:

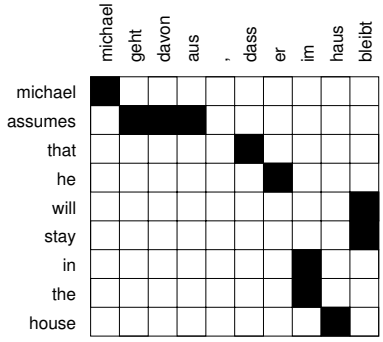
- niektoré nemožné preklady majú pozitívnu pravdepodobnosť
- dva rôzne zdrojové slová sa môžu dostať na rovnakú pozíciu v cieľovej vete

Konflikt pozícií rieši tento model zavedením zoznamu voľných miest. Na konci algoritmu tak povoľuje obsadiť iba voľné miesta, tým zabráni akémukoľvek konfliktu slov pri obsadzovaní pozícií v cieľovej vete.

Dalo by sa povedať, že IBM Modely 1–5 boli v rámci štatistického prekladu priekopnícke. Zaviedli dôležité pojmy ako generatívny model, trénovanie podľa algoritmu EM, prehodnocovacie modely. Tieto modely sú používané aj v špecializovaných softvéroch na zarovnanie slov, ako napr. GIZA++.

2.3.3 Zarovnanie slov

Z existencie množstva jazykov vyplýva, že každý tento jazyk má rozdielnu gramatickú stavbu viet. Ak by všetky slová v jednom jazyku mali svoje rovnaké ekvivalenty v jazyku druhom, zarovnanie by nebolo potrebné. V skutočnosti sa však stretávame s prípadmi, kedy slovo v jednom jazyku je zložené minimálne z 2 alebo 3 slov v cieľovom jazyku. Túto situáciu znázorňuje obrázok 2.6. Vidíme, že napr. anglické slovo *assumes* sa skladá v preloženej nemeckej vete až z 3 slov *geht, davon, aus.*

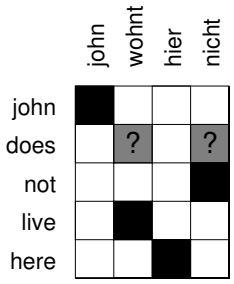


Obr. 2.6: Matica zarovnania slov [7]

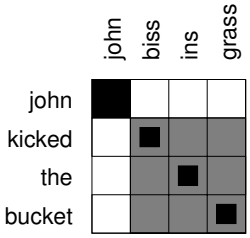
Obrázok 2.6 zobrazuje ideálny prípad, kedy každému slovu zo zdrojovej vety zodpovedá minimálne jedno slovo cieľovej vety. Pri preklade sa však dostaneme do situácie, kedy nevieme ako správne zarovnať zdrojové slovo, viď obrázok 2.7. Slovo *does* môžeme zarovnať so slovom *wohnt* alebo *nicht*, poprípade ho nemusíme zarovnať vôbec.

Ďalším problémom je nesprávny preklad. Slová sú síce zarovnané, avšak výsledný preklad je chybný (obrázok 2.8). V tomto príklade s daným kontextom je preklad slova *bucket* na *grass* správny. Avšak pri bežnom preklade s iným kontextom je tento preklad chybný.

Aby sme tieto chybné a nejasné prípady mohli správne ohodnotiť, musíme byť schopní merať kvalitu zarovnania slov. Zvýšiť kvalitu zarovnania môžeme napr. ručným zarovnaním slov v korpuse, kedy každému zarovnaniu pridelíme hodnoty buď S (sure, istota) alebo P (possible, možnosť). Existuje aj metrika pre ohodnotenie kvality zarovnania, ktorá sa



Obr. 2.7: Matica s neurčitým zarovnaním slov [7]



Obr. 2.8: Matica zarovnania slov s chybným prekladom [7]

označuje ako *AER* (Alignment Error Rate) a je definovaná nasledovne:

$$\mathrm{AER}(S,P;A) = \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{2.15}$$

Ak A= 0% zarovnanie je správne.

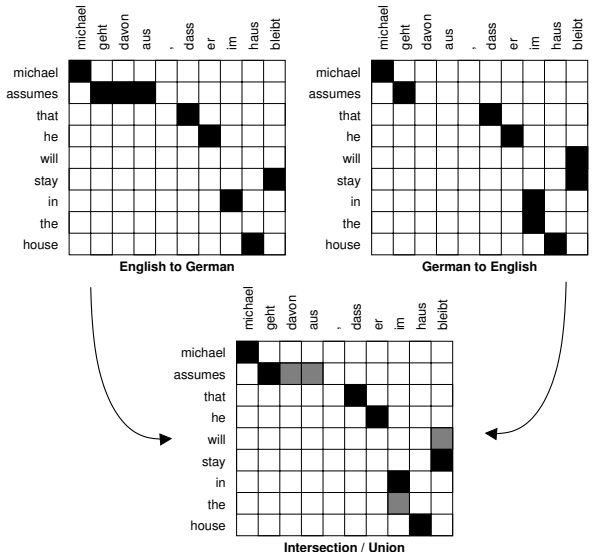
Výslednú maticu zarovnaní môžeme vylepšiť použitím symetrizácie (obrázok 2.9). Tá spočíva v obojsmernom zarovnávaní slov, ktorého výsledkom sú potom 2 matice. Na tieto matice aplikujeme potrebnú operáciu (napr. prienik, zjednotenie). Pri použití prieniku je kvalita výslednej matice, a teda aj zarovnania veľmi vysoká. Problémom je výskyt väčšieho počtu prázdnych miest v matici. Toto sa rieši pomocou metódy *grow additional alignment point*, ktorá automaticky doplňuje prázdne miesta podľa rôznych heuristík.

2.3.4 Modely založené na frázach

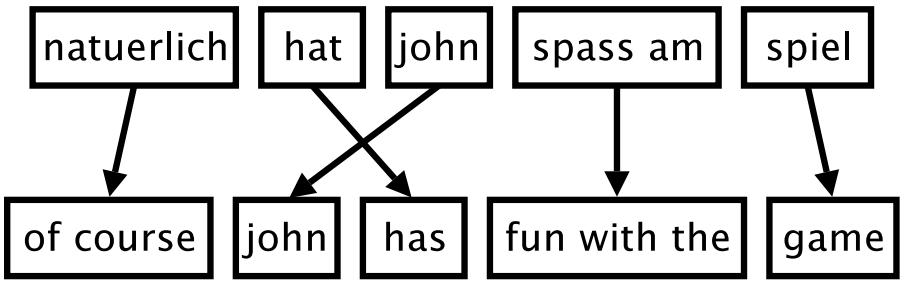
Fráza je bežné slovné spojenie skladajúce sa z postupnosti slov. Túto postupnosť využíva aj tento model, kedy neprekladá vety slovo po slove, ale ak je to možné prekladá sekvenciu slov. Frázy sú motivované len štatisticky, nie lingvisticky. Na obrázku 2.10 je znázornené zarovnanie fráz. Bez neho by sa nemecké slovo *am* preložilo na *with*, čo nebýva tak často, preklad by teda nebol vždy presný. Preto štatisticky významná fráza *spass am* zlepšuje preklad. Model je reálnejší z hľadiska praktického prekladu, kedy neprekladáme vždy len slovo po slove, ale časť $n : m$ slov. Riešia sa tým rôzne prekladové viacznačnosti. Model sa môže naučiť prekladať ľubovoľne dlhé frázy. Navyše je aj jednoduchší (nepoužíva fertilitu, NULL tokeny).

Prekladovú pravdepodobnosť $p(f|e)$ rozloženú na frázy definujeme nasledovne:

$$p(\bar{f}_1^I|\bar{e}_1^1) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1} - 1) \tag{2.16}$$



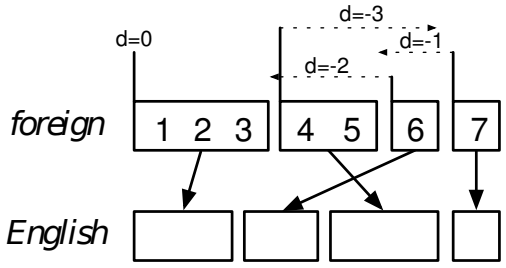
Obr. 2.9: Matica zarovnania slov vytvorená symetrizáciou [7]



Obr. 2.10: Zarovnanie pomocou fráz [7]

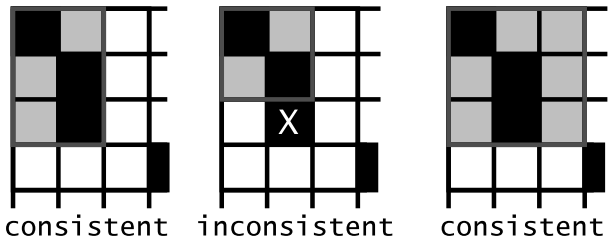
Veta f sa rozloží na I frázu $\bar{f}_i$. Všetky delenia sú rovnako pravdepodobné. Funkcia ϕ udáva prekladovú pravdepodobnosť pre frázu. Funkcia d je novo usporiadací model založený na vzdialenosti (distance–based reordering model), ktorý modelujeme pomocou predchádzajúcej frázy. Pozíciu prvého slova frázy vo vete f, ktorá sa prekladá na i–tu frázu vety e, udáva $start_i$. Tento model zaisťuje minimálny presun jednotlivých fráz (obrázok 2.11).

Prekladovú tabuľku fráz získame zarovnaním slov a následným nájdením konzistentnej frázy. Frázy $\bar{f}$ a $\bar{e}$ sú konzistentné so zarovnaním A, ak všetky slová $f_1, \ldots f_n$ vo fráze $\bar{f}$, ktoré majú zarovnanie A, sú zarovnané so slovami $e_1, \ldots e_n$ vo fráze $\bar{e}$ a naopak (rovnica 2.17). Príklad konzistentnej a nekonzistentnej frázy znázorňuje obrázok 2.12. Extrahovaním fráz z matice na obrázku 2.6 dostaneme nasledujúcu maticu 2.13. Z nej je možné vyčítať, že anglická fráza *asssume that* je zarovnaná s nemeckým ekvivalentom *geht davon aus, dass*.

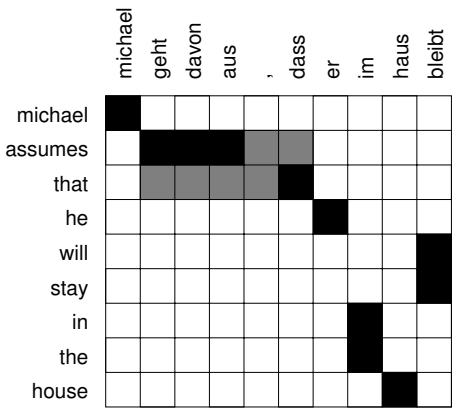


Obr. 2.11: Usporiadavací model založený vzdialenosti [7]

$$
\begin{aligned}
(\bar{e}, \bar{f}) \text{ je konzistentné s } A \ \Leftrightarrow \\
&\forall e_i \in \bar{e} : (e_i, f_j) \in A \ \rightarrow f_j \in \bar{f} \\
&\wedge \forall f_j \in \bar{f} : (e_i, f_j) \in A \ \rightarrow e_i \in \bar{e} \\
&\wedge \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A
\end{aligned} \tag{2.17}
$$



Obr. 2.12: Tabuľka fráz [7]



Obr. 2.13: Matica s extrahovanými frázami [7]

Odhad pravdepodobnosti výskytu fráz ϕ sa vypočíta podľa rovnice 2.18.

$$
\phi(\bar{f}|\bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} count(\bar{e}, \bar{f}_i)} \tag{2.18}
$$

2.3.5 Dekódovanie

Majme definovaný jazykový model p_{LM} a prekladový model $p(f|e)$. Zo všetkých možných prekladov, musíme vybrať preklad, ktorému modely dávajú najvyššiu pravdepodobnosť. Počet prekladov je exponenciálny, a preto algoritmus hľadania najlepšieho prekladu by mal mať nízku časovú zložitosť. K tomu sa využíva heuristické prehľadávanie (nie je garantované nájdenie prekladu s najväčšou pravdepodobnosťou).

Chyby v preklade môžu byť spôsobené:

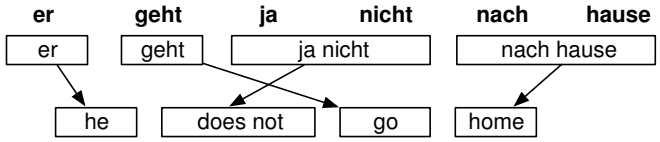
- nie je nájdený najlepší preklad v prehľadávacom priestore
- aj najlepší nájdený preklad nie je ten správny

Dekódovanie sa hodnotí podľa nájdených chýb, nie podľa kvality prekladu. Pri preklade viet po frázach (obrázok 2.14), v každom kroku prekladu spočítame predbežné hodnoty pravdepodobnosti z prekladového modelu, znova usporiadacieho a jazykového modelu (rovnica 2.19). Výpočet prebieha nasledovne:

- Prekladový model (frázový preklad) – podľa hodnoty $\phi(\bar{f}_i|\bar{e}_i)$, zistenej z frázovej prekladovej matice, sa vyberie fráza v zdrojovom jazyku $\bar{f}_i$, ktorá má byť preložená na frázu v cieľovom jazyku $\bar{e}_i$.
- Usporiadací model – vypočíta hodnotu vzdialenosti funkcie $d(start_i - end_{i-1} - 1)$ podľa predchádzajúcej frázy ukončenej v end_{i-1} a aktuálnej frázy začínajúcej na $start_i$.
- Jazykový model – Pre n-gramový model je potrebné udržovať posledných $n-1$ slov. Preto sa vypočíta hodnota daná $p_LM(w_i|w_{i-(n-1)}, \ldots, w_{i-1})$ pre každé pridané slovo w_i.

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\text{e}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1} - 1) p_{LM}(\mathbf{e}) \tag{2.19}$$

V prvom kroku dekódovania sa vytvorí prekladová matica fráz pre všetky vstupné frázy. Vytvorí sa počiatočná hypotéza, ktorá je prázdna. Nepokrýva žiadne vstupné slovo, a tak neprodukuje ani žiaden výstup. Následne sa vyberie akékoľvek slovo k prekladu. Vytvorí sa preňho nová hypotéza. K vybranému slovu sa vytvoria ďalšie hypotézy, ktoré pokrývajú všetky možné preklady toho slova. Takto sa postupne vytvárajú ďalšie hypotézy, až kým nie sú pokryté všetky zdrojové slová. Najlepší preklad sa zistí pomocou spätného navrátenia od hypotézy s najvyššou hodnotou. Keďže pri expanzii hypotéz, kedy sa pokrývajú všetky možné preklady daného zdrojového slova, je ich počet exponenciálny, musíme výsledné hypotézy zredukovať. Je to možné buď pomocou *rekombinácie* (recombination) alebo *rezom* (pruning). Rekombinácia rieši prípad, kedy dva preklady tvorené hypotézami sú rovnaké. Môže to nastať pri rovnakom počte zdrojových a cieľových slov v preklade, alebo v prípade zhody posledného slova vo viacerých hypotézach. Preklad s horším skóre sa vymaže z výsledného stromu. Rekombinácia je bezpečná, ale neušetrí dostatok pamäte. Na základe toho vznikla metóda rezu. Ten dokáže nepotrebné alebo chybné hypotézy zmazať skôr. Hypotézy, ktoré boli preložené z rovnakého počtu vstupných slov, uloží do zásobníka. Každý zásobník má svoju obmedzenú veľkosť. Po jeho naplnení dôjde k odstráneniu najhorších hypotéz.



Obr. 2.14: Preklad viet po frázach [7]

2.3.6 Modely založené na stromoch

Tradičné štatistické modely pracujú s postupnosťami slov. Mnoho situácií pri preklade je najlepšie možné vysvetliť pomocou syntaxu. Napr. presun slovesa vo vete alebo gramatická zhoda na veľkú vzdialenosť. Na základe týchto situácií boli vyvinuté prekladové modely založené na stromovej architektúre, tzv. *syntaktické stromy*. V súčasnosti tieto modely dávajú pre niektoré jazykové páry najlepšie výsledky. Vychádzajú z bezkontextovej gramatiky, kde non–terminálnymi symbolmi sú frázy a terminálnymi symbolmi jednotlivé slová. Majme napríklad anglické pravidlo $NP \rightarrow DET\ JJ\ NN$ a francúzske pravidlo $NP \rightarrow DET\ NN\ JJ$. Z týchto dvoch pravidiel môžeme následne vytvoriť synchrónne pravidlo, ktoré ukazuje rozdiely medzi dvoma jazykmi $NP \rightarrow DET_1 NN_2 JJ_3 | DET_1 JJ_3 NN_2$. Potom pravidlo na úrovni terminálov umožňuje priamy preklad $N \rightarrow maison$ — $house$. Synchrónna gramatika potom analyzuje celú vstupnú vetu a súčasne je generovaný aj výstupný strom. Pravdepodobnosť použitia prekladu vyjadruje rovnica 2.20. Učenie synchrónnej gramatiky spočíva v extrakcii pravidiel z paralelného korpusu zarovnaného na úrovni slov. Najprv sa vytvorí hierarchický frázový model (obsahuje len jeden nonterminál, bez jazykového syntaxu) a potom synchrónny frázový model (nonterminály sú už slová alebo frázy). Pri tvorbe nových pravidiel z dvojice viet existujú určité obmedzenia, aby sme zmiernili exponenciálnu zložitosť. Pravidlá majú maximálne 2 non–terminálne symboly, najmenej 1 a najviac 5 slov pre každý jazyk, pokrytie najviac 15 slov.

$$SCORE = (TREE, E, F) = \prod_i RULE_i \tag{2.20}$$

2.4 Hybridný strojový preklad

Hybridný strojový preklad kombinuje preklad založený na príkladoch, štatistický preklad a prípadne pravidlový preklad. Každý z týchto systémov má svoje výhody a nevýhody. Ich porovnanie znázorňuje tabuľka 2.1. Spomedzi troch typov prekladov sa najhoršie javí preklad na príkladoch, ktorý je v porovnaní s pravidlovým a štatistickým prekladom lepší len v rámci lexikálnej adaptivity. Preto súčasné hybridné systémy používajú prvky z štatistického a pravidlového prekladu.

Hybridný model môžeme skonštruovať využitím silných (označené ako +, ++) častí jednotlivých typov prekladov. Druhou možnosťou je skombinovať celé existujúce systémy a vytvoriť tak jeden väčší systém. Integrácia jednotlivých častí je proces veľmi náročný. Na druhej strane však vytvorenie hybridného systému poskladaním týchto častí existujúcich systémov je z dlhodobého hľadiska výhodnejšie ako skladanie celých systémov.

Spôsob prekladu zdrojového textu v hybridných systémoch môžeme rozdeliť do 2 základých skupín:

- Pravidlá → štatistika – zdrojový text sa v prvom kroku prekladá pomocou systé-

mov založených na pravidlách. Jeho výstup je potom ešte upravovaný štatistickým prekladom.

- Pravidlá → štatistika → pravidlá – zdrojový text je pred prekladom pomocou štatistického modelu predpripravený pravidlovým systémom, aby dosiahol lepšie výsledky. Po výstupe zo štatistického prekladu je znova upravovaný pravidlovým systémom.

	Syntax	Štrukturálna sémantika	Lexikálna sémantika	Lexikálna adaptivita
Pravidlový preklad	++	+	−	−−
Štatistický preklad	−−	−−	+	+
Preklad na príkladoch	−	−−	−	++

Tabuľka 2.1: Výhody a nevýhody strojových prekladov [10]

Kapitola 3

Vyhodnocovanie kvality prekladu

Z mnohých prekladových systémov či už komerčných alebo voľne šíriteľných, potrebujeme pre náš preklad vybrať ten najlepší. Preto je potrebné jednotlivé systémy otestovať a vyhodnotiť. Na to existujú mnohé techniky, ktoré si v tejto kapitole bližšie predstavíme. Nasledujúce informácie boli čerpané z [11], [12] a [13].

Pri hodnotení prekladu kladieme predovšetkým dôraz na plynulosť (prirodzený slovosled), adekvátnosť (uchovanie významu) a zrozumiteľnosť. Hodnotenie prekladu sa delí na 2 veľké skupiny: ručné a automatické. Pod ručným hodnotením rozumieme kontrolu prekladu profesionálnym prekladateľom. Použiteľnosť tohto vyhodnotenia je veľmi obmedzená, nakoľko je to hlavne časovo a finančne náročné. Dôležitú rolu tu hrá aj subjektivita prekladateľa. Rovnakému prekladu môžu priradiť rôzni prekladatelia rôzne hodnotenie, a to prináša určitú nejednoznačnosť, ktorá nie je vítaná. Často potrebujeme ohodnotiť preklad po malých úpravách zdrojového textu a tu subjektivita skresľuje dosah zmien na preklad. Preto sa viac používa automatické hodnotenie, ktoré je rýchlejšie, lacnejšie a daný preklad vždy ohodnotí rovnako. Prípadne zmeny zdrojového textu sa okamžite premietnu do výsledku. Nevýhodou oproti ručnému prekladu je jeho nie vždy zaručiteľná spoľahlivosť. Ďalšími výhodami je schopnosť zoradiť systémy podľa kvality a kategorizovať chyby. Základný princíp hodnotenia automatických systémov je porovnávanie prekladu s referenčným prekladom. Čím väčšia zhoda nastane, tým je preklad kvalitnejší.

3.1 BLEU

BLEU (BiLingual Evaluation Understudy) patrí medzi najznámejšie modely pre vyhodnocovanie kvality prekladu. Vyvinula ho firma IBM v roku 2001 a dnes je považovaný za štandard. Ako bolo spomenuté vyššie, princípom tejto metódy je porovnávanie preloženého textu s viacerými referenčnými prekladmi. Tie sa často líšia slovosledom, preto BLEU zavádza pojem *n–gramová presnosť*. Počet slov n je rovnaký pre kandidátsky a aj referenčný preklad. Tento počet následne vydelíme počtom slov v kandidátskom preklade. Prekladové systémy väčšinou vygenerujú väčší počet slov ako má referenčný preklad. BLEU to rieši zavedením *modifikovanej n–gramovej presnosti*. Najprv spočítame najvyšší počet n–gramov z jednotlivých referenčných prekladov. Následne priradíme každému slovu z kandidátskeho prekladu ekvivalentné slovo z strojového prekladu a zistíme počet takýchto dvojíc. Tento počet vydelíme celkovým počtom slov v kandidátskom preklade. Majme nasledujúcu situáciu [11]:

Kandidátsky preklad:

the the the the the the the

Referenčný preklad č. 1:

The cat is on the mat.

Referenčný preklad č. 2:

There is a cat on the mat.

Unigramová presnosť by bola $\frac{7}{7}$ a modifikovaná unigramová presnosť $\frac{2}{7}$. Už z tohto jednoduchého príkladu môžeme vidieť, že modifikovaná presnosť je z hľadiska ľudského hodnotenia presnejšia ako obyčajná n–gramová presnosť. V prípade n–gramovej presnosti celého textu najprv spočítame počet zhodných n–gramov v kandidátskych a referenčných textoch zvlášť pre každú vetu. Tieto počty sčítame a vydelíme počtom n–gramov kandidátskeho prekladu.

Pri viacerých referenčných prekladoch môže byť ich dĺžka rozdielna. Musíme vybrať takú, ktorá sa najviac približuje k dĺžke kandidátskeho prekladu. To je riešené pomocou *penalizácie za krátkosť* prekladu (brevity penalty), ktorá zisťuje tzv. *multiplikatívny breavity penalty faktor* (ďalej BP faktor). Ak je počet slov kandidátskeho a referenčného prekladu rovnaký, BP faktor sa rovná 1. Táto hodnota je cieľom pri dosiahnutí najkvalitnejšieho prekladu. Snažíme sa nájsť taký referenčný preklad, ktorého BP faktor sa čo najviac blíži k hodnote 1. Túto hodnotu nazveme optimálna dĺžka. Penalizáciu za krátkosť na celom korpuse vypočítame tak, že zistíme efektívnu dĺžku referenčného prekladu r, konkrétne spočítame optimálnu dĺžku každej vety. Penalizácia za krátkosť bude daná vzťahom $\frac{r}{c}$, kde c je celková dĺžka kandidátskeho korpusu. Môžu nastať dva prípady, ktoré zobrazuje rovnica 3.1.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{3.1}$$

Výpočet BLEU

Pri výpočte samotnej hodnoty BLEU použijeme *modifikovanú n–gramovú presnosť* a *penalizáciu za krátkosť* (rovnica 3.2). Pre názornejšie hodnotenie kvality prekladu je pri výpočte BLEU používaná logaritmická funkcia, viď rovnica 3.3.

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \cdot \log p_n) \tag{3.2}$$

$$\log BLEU = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n \tag{3.3}$$

Hodnota BLEU sa pohybuje v intervale $\langle 0, 1 \rangle$. Čím bližšie je k 1, tým je preklad kvalitnejší. Jeho hodnota však závisí na mnoha faktoroch. V práci od tvorcov tejto metriky [11] skúmali závislosť kvality prekladu od počtu referenčných prekladov. Testovaný kandidátsky text nechali najprv preložiť prekladateľovi. Následne zistili hodnoty BLEU pri 2 a 4 referenčných prekladoch. Dva referenčné preklady dosiahli skóre $0,2571$, štyri skóre $0,3468$. Vidíme, že už mierne zvýšenie počtu referenčných prekladov znamenalo vyššiu hodnotu BLEU.

BLEU má však aj svoje nevýhody. Pri krátkych vetách (menšie ako zvolené N) dochádza k ich nerovnocennému hodnoteniu, pretože sa pri nej nebudú brať do hodnotenia váhy,

ktoré sa započítavajú do geometrického priemeru. S tým súvisí aj počítanie logaritmu p_n. Ak niektoré vety sú kratšie ako zvolené N, logaritmus $p_N = 0$. To však nie je pri logaritme definované. Hodnotenie celej vety bude potom podhodnotené, pretože čím je N väčšie tým má daný n–gram väčšiu váhu.

3.2 NIST

NIST (National Institute of Standards and Technology) je inštitúcia zaoberajúca sa inováciou a zlepšovaním vedeckých meraní a technologie, patriaca pod ministerstvo obchodu USA. Na začiatku skúmala kvalitu už existujúceho BLEU. Konkrétne závislosť kvality prekladu od počtu referenčných modelov, ktorú vyjadruje *miera F* (harmonický priemer presnosti a pokrytia). Ďalšími veličinami, ktoré sa skúmali, boli adekvátnosť, plynulosť a informačná hodnota (schopnosť získať z textu informácie a následne ich vedieť použiť). Na základe týchto meraní bolo zavedené *NIST skóre*. To sa počíta aritmetickým priemerom na rozdiel od BLEU, kde je geometrický priemer. Zvýhodňuje n–gramy podľa informačnej hodnoty. Informačná hodnota sa vypočíta podľa rovnice 3.4.

$$Info(w_1, \ldots, w_n) = \log(\frac{\text{počet výskytov } w_1, \ldots, w_{n-1}}{\text{počet výskytov } w_1, \ldots, w_n}) \tag{3.4}$$

Pri porovnávaní metrík NIST a BLEU sa testovala miera F, plynulosť a presnosť. Prekladalo sa z angličtiny do 4 jazykov. NIST dosiahlo lepšie výsledky vo všetkých jazykoch pri miere F, naopak bolo horšie pri troch jazykov z pohľadu plynulosti. V presnosti boli obe metriky približne rovnaké [14].

3.3 NEVA

BLEU bol vyvíjaný na článkoch z novín a časopisov. Tieto články sa vyznačujú tým, že vety sa skladajú z viac ako 4 slov. Tvorcovia BLEU však ako základ pri porovnávaní n–gramov brali 4–gramy, teda vety pozostávajúce zo 4 slov. V praxi používame prekladače na preklad väčšinou odborných textov, kde naopak prevyšujú kratšie vety, takže BLEU tu mal horšie výsledky. Tento problém rieši metóda NEVA (N–gram EVAluation), ktorá upravuje BLEU pre použitie na krátkych vetách. Z rovnice 3.2 odstránili exponenciálnu funkciu a z rovnice 3.3 logaritmus. Výpočet NEVA je potom definovaný rovnicou 3.5. Navyše NEVA berie do úvahy aj synonymá, kladne hodnotí ich použitie v zmysle štylistickej bohatosti. NEVA dosiahla v niektorých testoch lepšie výsledky ako BLEU, no väčšinou boli výsledky v princípe rovnaké [15].

$$NEVA = BP \cdot \sum_{n=1}^{N} w_n \cdot p_n \tag{3.5}$$

Kde N nadobúda hodnoty N_{max}, ak $c \geq N_{max}$ alebo c, ak $c < N_{max}$.

3.4 WAFT

Podobne ako predchádzajúca metóda NEVA bola aj metrika WAFT vytvorená na univerzite v Uppsale. Pri porovnávaní slov z kandidátskeho a referenčného prekladu, môžeme použiť viacero spôsobov. Medzi najjednoduchšie patrí určenie editačnej vzdialenosti. Spočítame

najmenší možný počet úprav (vloženie, vymazanie alebo zmena písmena), aby boli dve slová rovnaké. Definujeme tzv. *presnosť slova* (word accuracy – WA) danú rovnicou 3.6.

$$WA = 1 - \frac{d + s + i}{r} \tag{3.6}$$

Kde d je vymazanie písmena, s substitúcia písmena, i vloženie písmena a r dĺžka slova v referenčnom preklade. Problém nastáva u menovateľa r, kedy nemôžeme vždy predpokladať zhodnú dĺžku slova z kandidátskeho a referenčného prekladu. Tento princíp sa využíva hlavne pri rozpoznávaní reči, kde pravdepodobnosť rovnakej dĺžky oboch slov je značne vyššia ako u prekladu textu. Miernou modifikáciou rovnice 3.6 môžeme túto metódu použiť aj pri preklade a dostávame potom systém s názvom *presnosť slova pri preklade* WAFT (Word Accuracy For Translation). Rovnica 3.7 určuje výpočet WAFT. WAFT porovnáva jednotlivé reťazce na úrovni slov, je teda citlivý na správny slovosled.

$$WA = 1 - \frac{d + s + i}{max(r, c)} \tag{3.7}$$

Kde c je dĺžka kandidátskeho slova.

3.5 TER

Translation Edit Rate (TER) patrí medzi najmladšie metriky vyhodnocovania prekladu. Vznikla v roku 2006. Je podobná metóde WAFT. Definuje najmenší počet editácií potrebných na úpravu kandidátskeho prekladu na referenčný preklad, ktoré sú zároveň normalizované na základe priemernej dĺžky všetkých referenčných prekladov (rovnica 3.8). Pod pojmom editácie chápeme vloženie, zmenu alebo vymazanie jednotlivých slov. V systéme je to zabezpečené pomocou dynamického programovania. Je možné posúvať aj celé skupiny slov. Pomocou hladového algoritmu (angl. *greedy–search*) sa nájde množina posunov s čo najmenším počtom jednoduchých editácií (vloženie, zmazanie a zmena). Počet všetkých editácií sa vypočíta pre všetky referenčné preklady a z nich sa vyberie tá s najmenším počtom zmien.

$$TER = \frac{\text{počet editácií}}{\text{priemerný počet referenčných slov}} \tag{3.8}$$

Majme 2 vety: *Dnes ráno je slnečno* a *Včera ráno bolo slnečno*. Hodnota $TER = \frac{2}{4}$. Existuje aj ručná varianta tejto metódy tzv. *Human–targe TER*. Prekladateľ upraví kandidátsky preklad tak, aby čo najlepšie (aj v rámci sémantiky) vyhovoval referenčnému prekladu. Na takto vytvorené kandidátske preklady je potom aplikovaný TER.

3.6 METEOR

Základ tejto metódy spočíva v porovnávaní jednotlivých slov oproti referenčným prekladom s tým, že vybrané kandidátske slovo sa nemusí presne zhodovať s referenčným slovom, postačuje aj zhoda na úrovni synoným, teda slov s rovnakým významom, ale rozdielnym tvarom. Toto umožňuje pomerne dobré výsledky aj pri menšom počte referenčných prekladov. Predchádzajúce metódy nebrali synonymá do úvahy a pri malom počte referenčných prekladov boli ich výsledky slabé. Cieľom metódy METEOR bolo opraviť chyby, ktoré obsahoval BLEU. BLEU nebral vôbec do úvahy pokrytie slov (recall). Namiesto n–gramov

vyššieho rádu, ktoré mali určovať gramatickú správnosť prekladu, sa u metódy METEOR kládol dôraz na slovosled.

Modul najprv vytvorí všetky možné mapovania medzi kandidátskym a referenčným prekladom. Mapovať môže podľa týchto úrovní: úplná zhoda, čiastočná zhoda (koreň slova), synonymá alebo parafrázy (podľa parafrázových tabuliek). Následne sa z jednotlivých úrovní vyberú tie mapovania, ktoré obsiahli najväčšie množstvo slov. Každá úroveň mapuje len tie unigramy, ktoré neboli mapované v predchádzajúcej úrovni. Poradie úrovní môže ovplyvniť výsledok ohodnotenia prekladu. Pri hodnotení nemusíme použiť vždy všetky úrovne.

Po vykonaní mapovania môžeme pristúpiť k výpočtu *METEOR score*. Najprv si vypočítame *unigramovú presnosť P*. Je to vlastne pomer namapovaných unigramov v kandidátskom preklade oproti všetkým unigramom v tom istom preklade. *Pokrytie* (recall R) je ten istý pomer vzhľadom na slová v referenčnom preklade. Harmonický priemer týchto dvoch veličín definuje *mieru F*, rovnica 3.9.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - -\alpha) \cdot R} \tag{3.9}$$

Pri dlhších n–gramoch zavádza METEOR penalizáciu. Všetky unigramy spojíme do skupín. Skupinu budú tvoriť tie unigramy, ktoré sú vedľa seba v kandidátskom a zároveň aj v referenčnom preklade (m). Snažíme sa, aby počet skupín (ch) bol najmenší. Penalizácia je potom definovaná v rovnici 3.10. Nakoniec výsledné METEOR hodnotenie vypočítame podľa rovnice 3.11. Koeficienty α, β, γ v rovniciach 3.9 a 3.10 nastavujeme tak, aby sme dosiahli čo najlepšiu zhodu s ručným prekladom.

$$PEN = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta} \tag{3.10}$$

$$Score = (1 - -Pen) \cdot F_{mean} \tag{3.11}$$

Kapitola 4

Návrh a realizácia systému

Základom kvalitného prekladu sú zdrojové dáta, teda paralelné texty (korpusy). Kapitola sa bude zaoberať ich tvorbou z najrôznejších zdrojov. Bude popísana tvorba pravidiel, ktoré vo veľkej väčšine nahradzajú prípony českých slov tak, aby vzniklo slovenské slovo. Ďalej bude popísaný použitý jazykový model a samotný prekladový systém Moses, ktorý reprezentuje dekóder.

4.1 Paralelné korpusy

Korpus môžeme definovať ako súbor počítačom uložených textov (v prípade hovorového jazyka – prepis hovorového slova), ktorý primárne slúži k jazykovému výskumu. Na prácu s korpusmi slúžia špeciálne vyhľadávacie programy. Pomocou nich je potom možné vyhľadávať slová prípradne slovné spojenia a zistiť ich frekvenciu v danom korpuse [16].

Na získanie paralelných textov som sa rozhodol použiť existujúce korpusy, ktoré okrem samotných zdrojových dát (viet) obsahovali aj súbor so zarovnaním na úrovni týchto viet. Toto zarovnanie bolo vytvorené automatickým systémom Hunalign. Vstupom je tokenizovaný text vo vetách v dvoch jazykoch. Výstupom je potom postupnosť dvojjazyčných párov (viet). Hunalign používa pri zarovnávaní slovník (ak je k dispozícií) a kombinuje ho s algoritmom Gale–Church, ktorý nesie informáciu o dĺžke jednotlivých viet. Ak nemáme k dispozícií slovník, vety sa zarovnávajú len na základe ich dĺžky. Najprv si Hunalign vytvorí vlastný vnútorný slovník, ktorý potom v ďalších krokoch využíva. Program je napísaný v jazyku C++ a preložiteľný skoro na akomkoľvek operačnom systéme [17]. V nasledujúcej časti budú popísané použité paralelné korpusy.

4.1.1 Acquis Communautaire a JRC–Acquis

Acquis Communautaire obsahuje právne pravidlá, ktoré sa týkajú všetkých štátov Európskej únie. Obsahuje legislatívne texty od roku 1950 až do súčasnosti. Zastúpených je 22 úradných jazykov EÚ. Z týchto právnych dokumentov boli vybrané tie, ktoré boli k dispozícií aspoň v 10 z 20 úradných jazykov EÚ (pred príchodom Bulharska a Rumunska v roku 2007) a aspoň v 3 jazykoch z 9 jazykov štátov, ktoré vstúpili do EÚ v roku 2004. Zúžením Acquis Communautaire podľa spomenutých podmienok vznikol JRC–Acquis. Vzhľadom na veľkosť a počet zastúpených jazykov je Acquis Communautaire najväčší paralelný korpus. Jeho veľkou výhodou je aj existencia neobvyklých a vzácnych jazykových párov, ako napr. maltsko–estónsky pár. Pre účely tohto projektu bol vybratý česko–slovenský korpus. Vety

boli zarovnané pomocou programov Vanilla a Hunalign. Pre účely tohoto projektu som si vybral Hunalign. Štatistika česko–slovenského korpusu je zobrazená v tabuľke 4.1.

Jazyk	Počet textov	Texty			Podpisy	Prílohy
		Počet slov	Počet znakov	Priemerný počet slov	Počet slov	Počet slov
český	21 438	22 843 279	148 972 981	1065,55	7 225 300	16 763 733
slovenský	21 943	26 792 637	179 920 434	1221,01	3 227 852	16 190 546

Tabuľka 4.1: Štatistika korpusu JRC-Acquis verzie 3.0 [18]

Príklad časti vytvoreného korpusu so zarovnaním viet v jazyku XML:

```
<body>
<div type="body" n="21959A1006(01)" select="cs sk">
<p>75 paragraph links:</p>
<linkGrp targType="head p" n="21959A1006(01)" select="cs sk"
id="jrc21959A1006_01-cs-sk" type="n-n"
xtargets="jrc21959A1006_01-cs;jrc21959A1006_01-sk">
<link type="1:1" xtargets="2;2">
<s1>Dohoda Mezi Evropským Společenstvím pro Atomovou Energii (Euratom)
A~Vládou Kanady O~Spolupráci V~Oblasti Mírového Využití Atomové Energie</s1>
<s2>Dohoda medzi vládou kanady a európskym spoločenstvom pre atómovú energiu
o~spolupráci pri mierovom využívaní atómovej energie</s2>
</link>
</linkGrp>
</div>
</body>
```

4.1.2 OPUS – voľne šíriteľný paralelný korpus

Tento korpus obsahuje voľne prístupné preložené texty z internetu. Celý projekt je založený na open–source princípe a neustále sa počet korpusov a ich objem zväčšuje. Korpusy boli vytvorené automaticky, neprebehla na nich žiadna ručná korekcia. V rámci diplomovej práce boli využité korpusy Európskej centrálnej banky, Európskej ústavy, texty projektu KDE, korpus vytvorený z manuálov k jazyku PHP a Európskej agentúry pre lieky [19]. Všetky tieto korpusy boli dodané spolu s zarovnaním viet programom Hunalign [20]. Zvyšné dva korpusy zložené z titulkov k filmom som nepoužil, kvôli častým chybám v preklade a chýbajúcim prekladom.

4.1.3 Europarl

Korpus obsahuje texty zo zborníkov Európskeho parlamentu [21]. Obsahuje texty v 21 jazykoch Európskej únie. Cieľom tohto projektu je vytvoriť text zarovnaný na úrovni viet, ktorý je použiteľný pre štatistický strojový preklad. Zarovnanie viet bolo vykonané pomocou Gale – Church algoritmu. Počet jednotlivých slov a viet je možné vidieť v tabuľke 4.2.

Keďže zarovnanie na úrovni viet nebolo k dispozícii priamo pre češtinu a slovenčinu, bolo potrebné korpus upraviť. Obidva jazyky boli zarovnané s anglickým jazykom, ktorý

Jazyk	Počet viet	Počet slov
Čeština	668 595	13 195 311
Slovenčina	674 359	13 116 301

Tabuľka 4.2: Europarl korpus pre češtinu a slovenčinu

bol cieľovým jazykom. Preto bolo potrebné zistiť prienik obidvoch korpusov v rámci anglických viet. Výsledkom prieniku boli indexy do poľa českých a slovenských viet, ktoré majú spoločné anglické vety. Tieto súbory s indexami spolu s českou a slovenskou časťou Europarl korpusu sú vstupom pre skript `europarl.py`, ktorý vytvorí finálny česko–slovenský korpus.

4.1.4 Slovníky

Do výsledného korpusu boli pridané dáta zo slovníkov dostupných na školských serveroch[1]. Z existujúcich dvojíc pre slovenčinu a češtinu sa našli slovníky pre jazyky ruský, francúzsky a španielsky. Tieto slovníky boli k dispozícií vo formáte LMF. Pomocou skriptu `lmf2tabfile.py` boli prekonvertované do tabfile formátu, kde každý riadok začína zdrojovým slovom oddeleným od cieľového/cieľových slov tabulátorom. Tieto súbory boli potom vstupom pre skript `createDict.py`, ktorý na základe zhody slov buď zdrojového jazyka (španielský slovník), alebo cieľového jazyka (francúzsky a ruský) vytvorí dva výstupné súbory v češtine a slovenčine, ktoré sú už zarovnané, teda každému českému slovu/slovám na jednom riadku jedného súboru odpovedá slovenské slovo/slová na tom istom riadku druhého súboru. Výsledný korpus vznikol spojením výstupov pre všetky tri spomínané jazyky.

4.1.5 Biblia a knihy o Harry Potterovi

Z internetových zdrojov boli získané knihy o Harrym Potterovi a Biblia. Knihy boli vybraté na základe ich populárnosti a dostupnosti v českom aj slovenskom jazyku. Navyše predchádzajúce korpusy sa týkali väčšinou právnych a odborných textov. Aby sa prekladový systém neorientoval len na jeden typ dokumentov, bolo potrebné dodať texty aj iného druhu, ako napr. fantazíjna literatúra v prípade Harryho Pottera.

Knihy boli získané vo formáte PDF. Tieto verzie boli konvertované do textového formátu pomocou programu pdftotext. Textové súbory boli preformátované tak, aby každý riadok obsahoval len jednu vetu. Následnou ručnou kontrolou boli vymazané zbytočné riadky ako napr. prvé a posledné strany kníh. Súbory začínajú vždy prvou vetou v knihe a končia poslednou vetou. Takto pripravené textové súbory boli pomocou programu Hunalign zarovnané na úrovni viet. K tomu bol použitý špeciálny slovensko–český slovník pre program Hunalign vytvorený zo slovníkov v predchádzajúcej sekcii 4.1.4. Ako uvádzajú autori tohto programu v aktuálnej verzii programu a aj z historických dôvodov začína každý riadok slovníka najprv slovom v cieľovom jazyku a až následne slovom v zdrojovom jazyku. Tieto slová sú oddelené znakom „@“. Takto pripravené texty so slovníkom boli poslané programu Hunalign, ktorého výstupom bol textový súbor vo formáte (vety medzi sebou ako aj veta a pravdepodobnosť zarovnania sú oddelené tabulátorom):

```
zdrojová veta/vety cieľová veta/vety pravdepodobnosť
```

[1] /mnt/minerva1/nlp/projects/dicts2lmf

Ak je v jazyku použitých viac viet, je medzi ne vložená sekvencia znakov „~~~“. V tabuľke 4.3 vidíme pravdepodobnosť zarovnania jednotlivých kníh. Pravdepodobnosť je daná priemerom pravdepodobností všetkých viet. Výstup programu Hunalign bol spracovaný skriptom `corpusHunalign.py`, ktorý odstránil nepotrebné znaky, dvojice prázdnych viet a nebral do úvahy vety, ktoré nemali odpovedajúcu vetu v druhom jazyku. Výsledný korpus z kníh o Harry Potterovi obsahuje 58494 vetných párov a korpus z Biblie 30902 vetných párov.

Kniha	Pravdepodobnosť [%]
Harry Potter a Kameň mudrcov	85,4
Harry Potter a tajomná komnata	37,9
Harry Potter a väzeň z Azkabanu	52,3
Harry Potter a Ohnivá čaša	89,0
Harry Potter a Fénixov rád	64,6
Harry Potter a polovičný princ	60,8
Harry Potter a Dary smrti	73,4
Biblia	31,9

Tabuľka 4.3: Pravdepodobnosť zarovnania na úrovni viet programom Hunalign

Korpus použitý k prekladu vznikol spojením spomenutých menších korpusov, aby bola výsledná kvalita prekladu čo najväčšia.

4.2 Pravidlá pre zmeny prípon slov z češtiny do slovenčiny

Čeština a slovenčina patria medzi západoslovanské jazyky. Majú veľa spoločných znakov. Obsahujú ale aj odlišnosti, ktoré môžu byť veľmi výrazné alebo naopak nepatrné. V práci som sa venoval zozbieraniu jednotlivých pravidiel, podľa ktorých by bolo možné české slová transformovať na ich slovenské ekvivalenty. Na presnejšiu aplikáciu pravidiel som využil morfologický analyzátor *ma*, ktorý je dostupný na servery minerva1[2]. Jeho základom je knižnica *libma*.

V prvom kroku som zistil najčastejšie sa vyskytujúce slová v korpuse. Zo slovníka vytvoreného programom GIZA++ (`*.A3.final`) pomocou skriptu `extractGiza.py` boli extrahované české slová, na každom riadku jedno slovo. Ďalej skriptom `statsDict.py` z nich boli vyselektované tie slová, ktoré splňovali podmienku minimálneho počtu výskytu (v našom prípade 100). Textový súbor s týmito slovami bol potom predaný programu *ma*. Pomocou prepínača *–I* sa zistili slová, ktoré sú uvedené v slovníku používanom programom *ma*. Ďalej bol program spustený s nasledovnými parametrami:

```
./ma -a -F ceskeSlova.txt > vystup.txt
```

Parameter `-a` znamená, že program vypíše všetky slová a `-F` vypíše k slovám aj ich značky. Značky v prvom rade určujú, o aký slovný druh slova sa jedná. Napr. značka *k1* znamená podstatné meno, značka *k5* sloveso. Zároveň však určujú ďalšie vlastnosti slova v závislosti na slovnom druhu. Napr. *g* u podstatného mena znamená rod, *n* znamená číslo

[2] /mnt/minerva1/nlp/local/bin/ma

a *c* znamená pád. Kompletný zoznam značiek a ich vysvetlenie je možné nájsť na adrese[3]. Výsledok po analýze programom *ma* môže vyzerať napr. takto:

```
ma><s> článek (vršek)
  <l>článek
  <c>k1gInSc1
  <c>k1gInSc4
článcích  článek    článkem   článku    články    článků    článkům
  <f>[k1gInPc6] článcích
  <f>[k1gInSc4] článek
  <f>[k1gInSc1] článek
  <f>[k1gInPc7wH] článkama
  <f>[k1gInSc7] článkem
  <f>[k1gInSc6] článku
  <f>[k1gInSc5] článku
  <f>[k1gInSc3] článku
  <f>[k1gInSc2] článku
  <f>[k1gInPc7] články
  <f>[k1gInPc5] články
  <f>[k1gInPc4] články
  <f>[k1gInPc1] články
  <f>[k1gInPc6wH] článkách
  <f>[k1gInPc2] článků
  <f>[k1gInPc3] článkům
```

Tento súbor je potom vstupom pre zmenu prípon českých slov na slovenské, ktorá bola implementovaná pomocou skriptu `rules.py`. Skript aplikuje vybrané pravidlá, na základe ktorých nahradzuje časti českých slov a dvojicu pôvodné slovo a nové slovo zapisuje do dvoch súborov.

Zdrojom pravidiel bola práca pani Sokolovej a spol. [22]. Po dôkladnom prečítaní diela som vybral tie pravidlá, ktoré platia vo všetkých možných príkladoch alebo vo väčšine z nich. Nejednoznačné pravidlá, pre ktoré neexistuje jasne definovaný vzťah, podľa ktorého by bolo možné nahradzovať časti slov, neboli aplikované.

V nasledujúcej časti budú opísané použité pravidlá a vhodné príklady k použitiu týchto pravidiel. Na konci analýzy boli nahradené jednotlivé české znaky, ktoré sa v slovenčine nepoužívajú:

- $\mathring{u} \rightarrow \hat{o}$
- $\check{e} \rightarrow e$
- $\check{r} \rightarrow r$
- $\check{n}i\acute{a} \rightarrow nia$
- $\acute{i}ci \rightarrow \acute{u}ci$

[3] https://merlin.fit.vutbr.cz/nlp-wiki/index.php/Morfologický_slovník_a_morfologický_analyzátor_pro_češtinu

Podstatné mená

Podstatné mená (substantíva) sú slovným druhom s najčastejším výskytom vo vetách. Označujú názvy osôb, vecí, zvierat, rôznych vlastností a dejov. Pravidlá pre podstatné mená som hlavne kvôli ich prehľadnosti rozdelil podľa ich rodu. Pre každý rod bude uvedená tabuľka s pravidlami, ktoré platia výhradne pre daný rod, a nakoniec bude uvedená tabuľka s pravidlami, ktoré platia pre viac ako jeden rod. V tabuľke A.1 môžeme vidieť pravidlá pre mužský rod (životné aj neživotné). Vystvetlivky k tabuľkám od podstatných mien až po slovesá:

- číslo: J – jednotné, M – množné
- pád: N – nominatív, G – genitív, D – datív, A – akuzatív, V – vokatív, L – lokál, I – inštrumentál
- rod: M – mužský, Ž – ženský, S – stredný
- životnosť: Ž – životné, N – neživotné

Tabuľka A.2 zobrazuje pravidlá pre podstatné mená ženského rodu a tabuľka A.3 podstatné mená stredného rodu. Pravidlá, ktoré platia pre viaceré rody súčasne:

- písmeno *á* sa po mäkkej spoluhláske mení na dvojhlásku *ia*. Napr. žádný → žiadny.
- písmeno *á* sa po písmenách *m, l, r, d* mení na *a*. Napr. mládne → mladne.
- slabika *ou* sa mení v N j. č. alebo v A j. č. ženského rodu na písmeno *ú*. Napr. touha → túžba.
- prípona *–o* vo vokatíve (5. pád) sa mení na *–a*. Napr. tepno → tepna. Prípona *–e* sa v slovenčine nevyskytuje. Napr. podpise → podpis. V slovenčine je vokatív nahradený oslovovacím nominatívom.
- Pri hromadnom oslovovaní členov rodiny sa prípona *–ovi* mení na *–ovci*. Napr. Novákovi → Novákovci.

Zmeny často používaných prípon v N j. č. podstatných mien ukazuje tabuľka A.4.

Prídavné mená

Prídavné mená (adjektíva) je slovný druh, ktorý opisuje vlastnosti alebo vzťahy podstatných mien. Podobne ako pre podstatné mená boli vytvorené pravidlá, ktoré platia vždy alebo vo väčšine prípadov. Pravidlá sú rozdelená podľa rodov (závisia zväčša na podstatnom mene).

Pravidlá pre prídavné mená mužského rodu sú znázornené v tabuľke A.5, ženského rodu A.6, stredného rodu A.7 a viacerých rôznych rodov A.8. Prídavné mená sú v tvare prvého stupňa. Pri treťom stupni prídavných mien bola nahradená predpona *nej–* predponou *naj–*.

Zámená

Zámena vo vete nahradzujú podstatné a prídavné mená a plnia tak funkciu podmetu, predmetu a prívlastku. Je to pomerne uzavretá skupina slov, a tak aj počet pravidiel je malý:

- predpona *ňe–* → *nie–*. Napr. ňěkolik → niekoľko.
- prípona *–koli* alebo *–koliv* → *–koľvek*. Napr. kterýmkoli → ktorýmkoľvek.
- prípona *–hle* → *–hľa*. Napr. tenhle → tenhľa.

Slovesá

Sloveso (verbum) vyjadruje určitú činnosť, stav alebo zmenu stavu. Slovesá bývajú vo vete najčastejšie vo forme prísudku. Pravidlá vytvorené pre slovesé sú zobrazené v tabuľke A.10. Vysvetlivky:

- tvar: neurč. – neurčitok (infinitiv)
- vid: dok. – dokonavé (perfektivum), nedok. – nedokonavé (imperfektivum)
- osoba: 1. os. – prvá osoba
- číslo: j. č. – jednotné číslo, mn. č. – množné číslo
- spôsob: ozn. sp. – oznamovací (indikatív), roz. sp. – rozkazovací (imperatív)
- trpné príčastie: trp. príč.
- rod: M – mužský, Ž – ženský, S – stredný

Slovesu, ktoré nevyhovelo žiadnemu pravidlu, bolo prípona *–t* nahradená príponou *–ť*.

Príslovky

Príslovky (adverbiá) patria medzi neohybné slovné druhy, ktoré bližšie určujú slovesá a prídavné mená. Delia sa na príslovky spôsobu, miesta, času a príčiny. Z ich neohybnosti vyplýva aj malý počet pravidiel (viď tabuľka A.9). Pravidlá platia pre prvý (pozitív) a druhý (komparatív) stupeň.

4.3 Použité systémy

V predchádzajúcej časti boli uvedené použité paralelné korpusy 4.1, ktoré tvoria jednu tretinu prekladového systému. Ďalšími dôležitými časťami sú jazykové modely a dekódovací systém (dekóder).

4.3.1 Jazykový model

Jazykový model pomáha vytvoriť správny slovosled výsledného prekladu alebo vybrať najpravdepodobnejšie slovo z množiny slov s rovnakým významom. Existujú viaceré voľne dostupné jazykové modely, ktoré sú podporované prekladovým systémom Moses. Sú to napr. modely SRI, Ken, IRST, Rand. Pre účely tohto projektu som sa rozhodol využiť voľne dostupný jazykový model IRST [23].

4.3.2 Dekóder

Z existujúcich voľne šíriteľných prekladových systémov som si vybral systém Moses [24]. Tento systém podporuje štatistický preklad založený na tabuľke fráz, hierarchickej tabuľke fráz a preklad založený na syntaxe. Pridaním extra lingvistickej informácie do frázovej tabuľky vzniká tzv. faktorový preklad.

Moses pracuje s už existujúcim natrénovaným prekladovým modelom a jazykovým modelom. Na vstupe číta testovací text v zdrojovom jazyku. Pre každú vetu sa snaží nájsť jej najvhodnejší preklad kombináciou existujúcich fráz z frázového modelu. Každý eventuálny preklad hodnotí váhami, ktoré mu prideľuje jazykový a frázový model.

Skladá sa z dvoch hlavných častí: trénovacej a dekódovacej. Na začiatku je potrebné paralelné texty pred učením vhodne upraviť. Texty sú tokenizované a prevedené do malých písmen. Následne sú vymazané príliš dlhé vety a dvojice viet, ktoré s najväčšou pravdepodobnosťou netvoria pár. Takto pripravené texty sú zarovnané na úrovni slov. K tomu sa využíva program GIZA++ [25]. GIZA++ je rozšírením pôvodného programu GIZA, ktorý vznikol v roku 1999 na univerzite Johnsa Hopkinsa v Baltimore. Sú v ňom implementované IBM modely 1 až 5, vyrovnávacie modely v závislosti na slovných druhoch, rôzne vyhľadávacie techniky pre určenie fertility. Zarovnanie na úrovni slov je veľmi časovo náročné pre veľké paralelné korpusy, preto vznikli varianty programu GIZA++, a to MGIZA a PGIZA, ktoré dokážu bežať pod viacerými jadrami a na viacerých uzloch v sieti, a tým ušetriť čas. V projekte som využil program MGIZA, aby časová úspora zarovnávania slov bola čo najväčšia.

Ďalším krokom je vytvorenie jazykového modelu. Moses na to využíva externé nástroje, ktoré boli vyššie spomenuté (KenLM, IRSTLM). Posledným krokom pred samotným prekladom je ladenie, kedy rôzne štatistické modely sú navzájom porovnávané s cieľom dosiahnuť čo najlepší preklad. Pri tomto porovnávaní si nastavujú svoje váhy.

Hlavnou úlohou dekóderu je nájsť vetu v cieľovom jazyku, ktorá najviac odpovedá vete v zdrojovom jazyku. Dekóder môže prípadne vytvoriť zoznam s prekladmi podľa ich hodnotenia [26].

4.3.3 Postup prekladu

Začíname s paralelnými textami, pre každý jazyk zvláštny súbor. Texty sú zarovnané na úrovni viet, jeden riadok odpovedá jednej vete. Z toho vyplýva, že súbory majú rovnaký počet riadkov. Na textoch musíme vykonať tieto úpravy:

- medzi slová a interpunkčné znamienka sú vložené medzery
- veľké začiatočné písmená slov sú zmenené na malé
- odstránenie prázdnych a dlhých viet

Ďalším krokom je učenie jazykového modelu, v našom prípade modelu IRST. Výstupom je súbor `*.arpa`, z ktorého je vytvorená jeho binárna reprezentácia, aby sa mohol rýchlejšie načítať. Po vytvorení jazykového modelu môžeme pristúpiť k samotnému trénovaniu prekladového systému. Ako prvé sa spustí zarovnávanie slov pomocou programu MGIZA, nasleduje extrakcia fráz a ich vyhodnocovanie. Po ukončení tohto procesu Moses vytvorí konfiguračný súbor `moses.ini`, ktorý obsahuje jednotlivé váhy modelov. Tento konfiguračný súbor je už možné použiť k prekladu textov. Pomocou optimalizácie váh je možné

vylepšiť váhy, a tak dosiahnuť potenciálne lepší preklad. V našom prípade bol použitý základný konfiguračný súbor, ktorého váhy neboli optimalizované, z dôvodu veľkej časovej náročnosti.

Kapitola 5

Testovanie a vyhodnocovanie systému

Pre vytvorenie paralelných textov bolo vytvorených 11 čiastočných korpusov, ktoré boli spojené do jedného. Ich veľkosť v MB pre český a slovenský jazyk môžeme vidieť v tabuľke 5.1. Tieto texty prešli pred použitím v systéme Moses tokenizáciou (`tokenizer.perl`), prevedením veľkých začiatočných písmen slov na malé (`train-truecaser.perl` a `truecase.perl`) a boli odstránené vety s počtom slov väčším ako 80 slov (`clean-corpus-n.perl`). Skripty uvedené v zátvorkách boli dodané spolu so systémom Moses.

Korpus	Veľkosť súboru v MB	
	čeština	slovenčina
JRC–Acquis 1	146,4	147,4
ECB	13,5	13,6
ECU	0,893	0,934
JRC–Acquis 2	0,591	0,604
KDE4	2,2	2,3
PHP	0,559	0,528
EMEA	81,9	80,1
Europarl	65,1	65,1
Slovníky	4,8	4,7
Harry Potter	5,7	5,6
Biblia	3,8	4,5
Celkom	325,443	325,366

Tabuľka 5.1: Veľkosť v MB vytvorených paralelných textov pre český a slovenský jazyk

Ako testovaciu sadu česko–slovenských textov som zvolil ochranu osobných údajov a zmluvné podmienky od spoločnosti Google, texty z webových stránok EÚ, pretože existuje ich preklad v jazykoch čeština a slovenčina. Ďalšou testovacou sadou boli texty z kníh o Harry Potterovi a Biblie. Z korpusu bola odňatá časť textu slúžiaca práve na testovanie systému.

České texty boli preložené do slovenčiny Google prekladačom [3] a systémom Česílko [27] a porovnané s referenčným prekladom. Tieto texty boli taktiež pred testovaním prekladovým systémom tokenizované a prevedené veľké začiatočné písmená slov na malé. Ako

hodnotiacu metódu prekladu som zvolil BLEU, ktorá je najpoužívanejšia. Túto metódu podporuje aj systém Moses, konkrétne skript `multi-bleu.perl`.

Tabuľka 5.2 zobrazuje BLEU skóre dvoch existujúcich prekladových systémov na štyroch testovacích textoch. Ako vidíme, systém Česílko je vo všetkých prípadoch lepší od Google prekladača. Tento výsledok sa dal očakávať, nakoľko systém Česílko sa špecializuje na preklad medzi českým a slovenským jazykom. Rovnaké testovacie texty boli použité aj pri preklade pomocou systému Moses.

Prekladač	Ochr. os. údajov a zmluvné podmienky Google	Texty z stránok EÚ	Harry Potter	Biblia
Google	22,04	11,26	37,99	16,7
ČESÍLKO	30,55	19,35	54,26	24,6

Tabuľka 5.2: Hodnotenie prekladového systému metódou BLEU a porovnanie s prekladom pomocou Google a Česílko

5.1 Základný prekladový systém

Tento systém pozostáva z korpusov: Acquis Communautaire, JRC–Acquis, korpusy dostupné v rámci projektu OPUS a upravený korpus Europarl. Korpus sa skladá výhradne z právnych a odborných textov, v čoho dôsledku vyšlo aj hodnotenie 5.3. Pri právnych podmienkach Googlu ako aj stránok EÚ bol preklad pomerne úspešný. Nedostatok iných typov trénovacích textov sa ukázal na hodnotení pri preklade časti Harryho Pottera alebo Biblie.

Prekladač	Ochr. os. údajov a zmluvné podmienky Google	Texty z stránok EÚ	Harry Potter	Biblia
Moses 1	29,37	22,25	39,86	16,69

Tabuľka 5.3: BLEU skóre základného prekladového systému

5.2 Základný prekladový systém so slovníkom

K predchádzajúcemu systému boli pridáne dáta zo slovníkov. Keďže vhodných slovníkov bolo málo (konkrétne 3), tak aj výsledné skóre sa zvýšilo len v rádovo desatinách bodu (tabuľka 5.4). Dáta zo slovníkov mohli pomôcť k prekladu slov s menším výskytom. Ďalej k už existujúcim slovám boli pridané ich alternatívy, a tak aj významovo rovnaké slová, ktoré sa však píšu rôzne, majú v systéme svoje zastúpenie.

Prekladač	Ochr. os. údajov a zmluvné podmienky Google	Texty z stránok EÚ	Harry Potter	Biblia
Moses 2	29,44	22,35	41,63	17,64

Tabuľka 5.4: BLEU skóre základného prekladového systému spolu so slovníkom

5.3 Základný prekladový systém so slovníkom a rôznym typom testovacích textov

Ďalším vylepšením existujúceho systému bolo pridanie iných typov paralelných textov. V prípade Harryho Pottera sa jedná o fantazíjnu literatúru, v prípade Biblie je to náboženská literatúra. Z tabuľky 5.5 je vidieť značné zlepšenie skóre pri testovacích textoch práve z týchto kníh. Výsledok môže byť skresľujúci, nakoľko sa jedná o rovnaké knihy. Avšak zo všeobecného hľadiska môžeme vidieť, že pridanie rozličných typov trénovacích textov má veľký vplyv na výsledný prekladový systém. Výsledok hodnotenia prekladu právnych a odborných testovacích textov sa zlepšil len minimálne. Opäť je to predpokladaný výsledok, nakoľko prienik v rámci slov odborných textov a fantazíjnej alebo náboženskej literatúry je minimálny.

Prekladač	Ochr. os. údajov a zmluvné podmienky Google	Texty z stránok EÚ	Harry Potter	Biblia
Moses 3	29,64	22,39	56,54	22,43

Tabuľka 5.5: BLEU skóre základného prekladového systému spolu so slovníkom a rôznym typom testovacích textov

5.4 Základný prekladový systém so slovníkom, rôznym typom testovacích textov a pravidlami pre zmeny prípon

Posledným zlepšením prekladového systému bola zmena prípon slov, kde sa využil fakt podobnosti českého a slovenského jazyka. Zmenou je myslené hlavne nahradzovanie prípon jednotlivých českých slov, nakoľko koreň slova v oboch jazykoch je často podobný. Všetky použité pravidlá boli uvedené v sekcii 4.2. Výsledky testovania tohto systému môžeme vidieť v tabuľke 5.6. V 3 z 4 prípadoch došlo k zlepšeniu. V prípade Biblie sa výsledok mierne zhoršil, môže to byť spôsobené zanesením veľkého šumu do prekladu, ktoré spôsobila práve zmena prípon. Tento testovací text zrejme obsahoval také slová, ktoré vďaka analýze získali nepresné preklady.

Prekladač	Ochr. os. údajov a zmluvné podmienky Google	Texty z stránok EÚ	Harry Potter	Biblia
Moses 4	29,70	23,04	56,78	22,14

Tabuľka 5.6: BLEU skóre základného prekladového systému spolu so slovníkom, rôznym typom testovacích textov a pravidlami pre zmeny prípon

5.5 Zhodnotenie prekladových systémov

Kvalita výsledného prekladu (výška BLEU skóre) závisí na mnohých faktoroch. V prvom rade je to veľkosť a kvalita paralelného korpusu. Kvalitou sa myslí zarovnanie viet pri tvorbe korpusu. Pri kontrole vytvorených textov som zistil, že niektoré dvojice viet k sebe

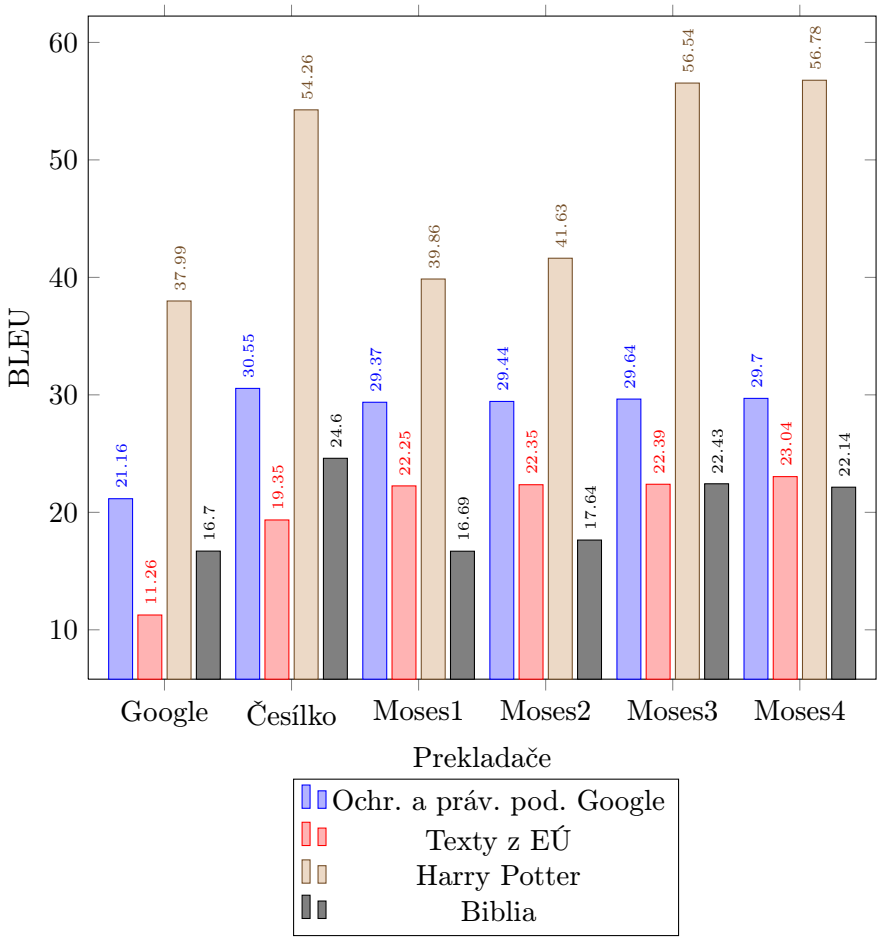
neodpovedali, v dôsledku chýbajúcej vety v jednom či druhom jazyku. Ak vety aj existujú v oboch jazykoch, tak potom nie sú správne zarovnané pomocou programov na to určených (v našom prípade to bol program Hunalign). V niektorých textoch boli dokonca vety napísané v inom jazyku (angličtina). Z hľadiska veľkosti korpusu platí, že čím väčší korpus máme tým je väčšia šanca zvýšiť kvalitu prekladu. Závisí to od povahy textov. Ak sú texty orientované len na jeden špecifický obor (napr. veda a technika) a my prekladáme text z iného oboru (rozprávka), tak preklad bude zrejme veľmi slabý. Preto by bolo dobré mať k dispozícii texty s čo najväčším pokrytím. Problémom je, že napr. pre knihy neexistujú vhodné texty v elektronickej podobe pre pár jazykov čeština a slovenčina.

Súhrn vyhodnotenia prekladu všetkých prekladových systémov na všetkých testovacích textoch vidíme na obrázku 5.1. Vzhľadom na veľkosť vytvoreného korpusu výsledky testovania dopadli pomerne dobre. Vytvorený prekladový systém získal v dvoch prípadoch najlepšie skóre, vo zvyšných dvoch dosiahol druhé najvyššie skóre za systémom Česílko. Z grafu je zjavné postupné vylepšovanie kvality prekladu pridaním slovníka, trénovacích textov z fantazíjnej a náboženskej literatúry, a nakoniec aj aplikácia zmien prípon slov. BLEU skóre narastalo pomaly, ale dokazuje prínos jednotlivých vylepšení ku kvalite prekladu. Zo strany zmien prípon sa očakávalo výraznejšie zlepšenie. Pravidlá boli otestované aj samostatne, no preklad sa zlepšil opäť len o maximálne 2 body BLEU skóre. Môže to byť zapríčinené aj nespravným vyhodnocovaním výsledkov, kedy by bolo vhodnejšie výsledky krížovo validovať a merať t–testom. Podobne by výsledky zrejme boli lepšie, keby sa prekladalo len zámenou prípon slov, bez prekladového systému Moses. Príklady prekladu českých viet vytvorenými prekladačmi a existujúcimi prekladačmi:

- Zdrojová veta: *před použitím informací za účelem , který není uveden v těchto zásadách ochrany osobních údajů , vás vždy požádáme o souhlas .*
- Referenčný preklad: *pred použitím informácií na účel , ktorý nie je uvedený v týchto Pravidlách ochrany osobných údajov , vás vždy požiadame o súhlas .*
- Preklady viď tabuľka A.11
- Zdrojová veta: *„Venku , dobře , ale ve škole to je profesor Longbottom , chápeš ? nemůžu prostě přijít na Bylinkářství a říct , že mu předávám láskyplné pozdravy ? “Pokýval hlavou nad matčinou pošetilostí a aby rozptýlil své pocity , nakopnul lehce Albuse .*
- Referenčný preklad: *„Vonku , dobre , ale v škole to je profesor Longbottom , chápeš ? nemôžem jednoducho prísť na Herbológiu a povedať , že mu odovzdávam láskyplné pozdravy ? “Pokýval hlavou nad matkinou nevhodnou žiadosťou a aby rozptýlil svoje pocity , nakopol ľahko Albusa .*
- Preklady viď tabuľka A.11

Porovnaním prekladu prvého prekladového systému (Moses 1) s poslednou verziou (Moses 4) vidíme pokrok hlavne v preklade záporov, napr. *ktorý, je uvedený* vs. *ktoré, nie sú uvedené*, alebo *som jednoducho prísť* vs. *nemôžem prísť*. Ďalej je viditeľné lepšie skloňovanie a aj zväčšená slovná zásoba. Kompletné preklady všetkých testovacích textov je možné nájsť na adrese[1].

[1] /mnt/minerva1/nlp/projects/mt_sk3/translations



Obr. 5.1: Porovnanie všetkých prekladových systémov

Kapitola 6

Záver

Práca sa venuje vytvoreniu prekladového systému pre preklad textov z češtiny do slovenčiny. V úvode je uvedená história počítačového prekladu od samotných začiatkov až po súčasnosť. Nasleduje kapitola o strojovom preklade, v ktorej sú uvedené jednotlivé druhy prekladu: preklad založený na pravidlách, na príkladoch, hybridný preklad a štatistický strojový preklad. Zo štatistického prekladu sú uvedené jeho jednotlivé komponenty, a to model jazyka, model prekladu, dekóder, zarovnanie slov, modely založené na frázach a stromoch. Nasleduje kapitola o metódach vyhodnocovania prekladu, z ktorých je najznámejšia a najpoužívanejšia metóda BLEU. Jadrom práce je príprava paralelných textov z existujúcich korpusov Acquis Communautaire a OPUS. Tieto texty sú zarovnané na úrovni viet a následne na úrovni slov pomocou programu GIZA++. Z existujúcich internetových zdrojov boli ďalej spracované korpusy Europarl, knihy o Harry Potterovi a Biblii. Posledné dva bolo potrebné urapviť, zarovnať na úrovni viet pomocou programu Hunalign, aby ich bolo možné pridať do korpusu. V rámci práce boli tiež zozbierané pravidlá, podľa ktorých je možné nahradzovať predpony a prípony českých slov s cieľom automaticky vytvoriť slovenské slová. Pravidlá boli vyberané tak, aby platili na čo najväčšie množstvo českých slov. Ich nevýhodou je často aj nepresný preklad slov, ktorý prináša do korpusu určitý šum. Celkovo však táto úprava priniesla o niečo lepšie výsledky. Boli vytvorené 4 prekladové systémy a boli porovnávané s existujúcimi prekladačmi od Googlu a Česílka, vyvýjaného na Karlovej univerzite v Prahe. Hodnotiacou metódou bol BLEU, ktorý patrí v súčasnosti medzi najpoužívanejšie a najznámejšie.

Vytvorený systém je možné používať pre preklad textov z češtiny do slovenčiny. Kvalita výsledného prekladu bude závisieť od povahy zdrojového textu. Systém bol vytvorený prevažne z právnych dokumentov a užívateľských príručiek. Zastúpená je aj fantazíjna a náboženská literatúra, avšak v nepatrnom pomere, nakoľko internetové zdroje sú v tomto prípade značne obmedzené. Na základe tohto by bolo potrebné systém obohatiť o paralelné texty z iných odvetví, aby bolo výsledné pokrytie čo najväčšie. Ďalšie skvalitnenie prekladu by spočívalo v dôkladnejšej príprave paralelných textov, hlavne ich následné zarovnanie na úrovni viet.

Literatúra

[1] Hutchins, J. *The history of machine translation in a nutshell* [online]. 2005 [cit. 9. 12. 2012]. Dostupné na: <http://www.hutchinsweb.me.uk/Nutshell-2005.pdf>.

[2] *EuroMatrixPlus Bringing Machine Translation for European Languages to the User* [online]. 2009-2012 [cit. 8. 4. 2013]. Dostupné na: <http://www.euromatrixplus.net/>.

[3] *Google Prekladač* [online]. [cit. 21. 4. 2013]. Dostupné na: <http://translate.google.cz/about/intl/sk_ALL/>.

[4] Šimon, J. *Strojový preklad* [online]. 2008-2009 [cit. 11. 12. 2012]. Dostupné na: <http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0809/simon/index.html>.

[5] Tyers, F. [online]. 2011 [cit. 10. 12. 2012]. Dostupné na: <http://wiki.apertium.eu/index.php/File:The_vauquois_triangle.svg>.

[6] Kolovratník, D. *Srovnání metod překladu* [online]. 2004 [cit. 11. 12. 2012]. Dostupné na: <http://www.ms.mff.cuni.cz/~kolod1am/big/ebmt.html>.

[7] Koehn, P. *Statistical Machine Translation*. [b.m.]: Cambridge University Press, 2009. ISBN 0-521-87415-7.

[8] Baisa, V. *Strojový překlad* [online]. 2012 [cit. 12. 12. 2012]. Dostupné na: <http://nlp.fi.muni.cz/~xbaisa/plin019/plin019.pdf>.

[9] Brown, P. F., Pietra, V. J., Pietra, S. A. D. et al. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*. 1993, roč. 19. S. 263–311.

[10] Eisele, A. *Hybrid machine translation: Combining rule-based and statistical MT systems* [online]. 2007 [cit. 22. 12. 2012]. Dostupné na: <mt-archive.info/MTMarathon-2007-Eisele.pdf>.

[11] Papineni, K., Roukos, S., Ward, T. et al. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. S. 311–318. ACL '02. Dostupné na: <http://dx.doi.org/10.3115/1073083.1073135>.

[12] Denkowski, M. a Lavie, A. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Stroudsburg, PA, USA: Association for

Computational Linguistics, 2011. S. 85–91. WMT '11. Dostupné na: <http://dl.acm.org/citation.cfm?id=2132960.2132969>. ISBN 978-1-937284-12-1.

[13] ŠTANCEL, R. *Metódy hodnotenia kvality strojového prekladu.* Brno: Masarykova Univerzita Fakulta Informatiky, 2007. Diplomová práce.

[14] DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. S. 138–145. HLT '02. Dostupné na: <http://dl.acm.org/citation.cfm?id=1289189.1289273>.

[15] FORSBOM, E. Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation. In *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation, held in conjunction with MT SUMMIT IX.* [b.m.]: New Orleans, Louisiana, USA, 2003. S. 29–36.

[16] ČERMÁK, F. a KOCEK, J. *Český národní korpus* [online]. [cit. 03. 01. 2013]. Dostupné na: <http://ucnk.ff.cuni.cz/co_je_korpus.php>.

[17] VARGA, D., HALÁCSY, P., KORNAI, A. et al. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing.* 2005. S. 590–596.

[18] STEINBERGER, R., POULIQUEN, B., WIDIGER, A. et al. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* 2006. S. 2142–2147.

[19] TIEDEMANN, J. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In NICOLOV, N., BONTCHEVA, K., ANGELOVA, G. et al. (ed.). *Recent Advances in Natural Language Processing.* Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009. S. 237–248. Dostupné na: <http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf>. ISBN 978 90 272 4825 1.

[20] TIEDEMANN, J. Parallel Data, Tools and Interfaces in OPUS. In CHAIR), N. C. C., CHOUKRI, K., DECLERCK, T. et al. (ed.). *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey: European Language Resources Association (ELRA), may 2012. ISBN 978-2-9517408-7-7.

[21] KOEHN, P. *Europarl: A Parallel Corpus for Statistical Machine Translation.* 1996-2011. Dostupné na: <http://mt-archive.info/MTS-2005-Koehn.pdf>.

[22] SOKOLOVÁ, M., MUSILOVÁ, K. a SLANČOVÁ, D. *Slovenčina a čeština: synchrónne porovnanie s cvičeniami.* 1. vyd. [b.m.]: Academia, 2005. 179 s. ISBN 9788022321501.

[23] FEDERICO, M., BERTOLDI, N. a CETTOLO, M. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual*

Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008. [b.m.]: ISCA, 2008. S. 1618–1621.

[24] Koehn, P., Hoang, H., Birch, A. et al. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. S. 177–180. ACL '07. Dostupné na: <http://dl.acm.org/citation.cfm?id=1557769.1557821>.

[25] Och, F. J. a Ney, H. Improved Statistical Alignment Models. In. Hongkong, China: [b.n.], October 2000. S. 440–447.

[26] Koehn, P. *MOSES Statistical Machine Translation System User Manual and Code Guide* [online]. 2012 [cit. 04. 01. 2013]. Dostupné na: <http://www.statmt.org/moses/manual/manual.pdf>.

[27] Hajič, J., Kuboň, V. a Homola, P. *Česílko Demo - a machine translation system for closely related languages* [online]. [cit. 03. 01. 2013]. Dostupné na: <http://quest.ms.mff.cuni.cz/cesilko/>.

Dodatok A

Tabuľky zmien prípon a príklady prekladov

Číslo	Pád	Životnosť	Česká prípona	Slovenská prípona	Príklad
J	I	Ž, N	–em	–om	slovem → slovom
	G	Ž	–y	–u	předsedy → predsedu
		N	–e	–a	zdroje → zdroja
	D		–i	–u	stupni → stupňu
	L		–ě	–e	bodě → bode
M	L	Ž, N	–cích	–koch	racích → rakoch
			–ších	–choch	kožiších → kožuchoch
			–ech	–och	pánech → pánoch
			–ích		manželích → manželoch
	N		–zi	–hovia	vrazi → vrahovia
			–ové	–ovia	geológové → geológovia
	G		–ů	–ov	politiků → politikov
	D		–ům	–om	členům → členom
	A	Ž	ľubovoľná	–ov	odborníky → odborníkov

Tabuľka A.1: Pravidlá pre podstatné mená mužského rodu

Číslo	Pád	Česká prípona	Slovenská prípona	Príklad
J	N	–e	–a	práce → práca
		–ě		koupě → kúpa
	G		–e	úrovně → úrovne
	D			straně → strane
	L			
	A	–i	–u	strategii → strategiu
	I	–í	–ou	technologií → technológiou
	D	–ce	–ke	látce → látke
	L	–ze	–he	úloze → úlohe
M	N	–ě	–e	úrovně → úrovne
	G	–ic	– íc	hranic → hraníc
		–en	–ien	žen → žien
		–ek	–iek	složek → zložiek
		–jí	–í	Galilejí → Galileí
	D	–ím	–iam	operacím → operáciam
		–em		vlastnostem → vlastnostiam
	A	–ě	–e	lodě → lode
	L	–ích	–iach	institucích → inštitúciach
		–ech		pravomocech → právomociach
	I	–emi	–ami	misemi → misami
		–ěmi		zbraněmi → zbraňami
		–mi		oblastmi → oblasťami

Tabuľka A.2: Pravidlá pre podstatné mená ženského rodu

Číslo	Pád	Česká prípona	Slovenská prípona	Príklad
J	N	–í	–ie	rozhodnutí → rozhodnutie
	G	–ete	–ať a	prasete → prasaťa
		–ěte		mláděte → mláďaťa
		–e	–a	srdce → srdca
		–í	–ia	školení → školenia
	D	–eti	–ať u	teleti → teľaťu
		–ěti		dítěti → dieťaťu
		–i	–u	písmeni → písmenu
		–í	–iu	investování → investovaniu
	A		–ie	zvýšení → zvýšenie
	L	–u	–e	právu → práve
		–ě		dně → dne
		–eti	–ati	děvčeti → dievčati
		–ěti		dítěti → dieťati
	I	–etem	–aťom	zvířetem → zvieraťom
		–em	–om	zlatem → zlatom
M	N, A	–a	–á	rizika → riziká
		–e	–ia	pole → polia
		–í		hodnocení → hodnotenia
	D	–ům	–ám	sklům → sklám
		–ím	–iam	očím → očiam
	L	–ech	–ách	plavidlech → plavidlách
		–ích	–iach	využitích → využitiach
	I	–ími	–iami	použitími → použitiami
		–y	–ami	stanovisky → stanoviskami
		–i		médii → médiami

Tabuľka A.3: Pravidlá pre podstatné mená stredného rodu

Česká prípona	Slovenská prípona	Príklad
–ost	–osť	skromnost → skromnosť
–yňe	–yňa	sudkyňe → sudkyňa
–ište	–isko	stanovište → stanovisko
–ště		koupaliště → kúpalisko
–árna	–áreň	lekárna → lekáreň
–ství	–stvo	království → kráľovstvo
–ek	–ok	majetek → majetok
–ura	–úra	agentura → agentúra
–una	–úna	tribuna → tribúna

Tabuľka A.4: Preklad často používaných koncoviek

Číslo	Pád	Životnosť	Česká prípona	Slovenská prípona	Príklad
J	G, A	Ž, N	–ího	–ieho	hladšího → hladšieho
J	D	Ž, N	–ímu	–iemu	bližšímu → bližšiemu
J	N, A, V	Ž, N	–ův	–ov	kolegův → kolegov
M	A	Ž	–é	–ých	hladké → hladkých

Tabuľka A.5: Pravidlá prídavných mien mužského rodu

Číslo	Pád	Česká prípona	Slovenská prípona	Príklad
J	G, D, L	–é	–ej	mladé → mladej
		–í		cizí → cudzej
	A	–ou	–ú	uloženou → uloženú
	I	–í	–ou	poradní → poradnou
	D, L	–ě	–ej	prezidentově → prezidentovej
M	N, A	–ní	–né	cestovní → cestovné
		–y	–é	stanoveny → stanovené

Tabuľka A.6: Pravidlá prídavných mien ženského rodu

Číslo	Pád	Česká prípona	Slovenská prípona	Príklad
M	N, A	–á	–é	úplná → úplné
		–ní	–né	tradiční → tradičné

Tabuľka A.7: Pravidlá prídavných mien stredného rodu

Číslo	Pád	Rod	Česká prípona	Slovenská prípona	Príklad
J	L	M, S	–ém	–om	čistém → čistom
			–ím		obchodním → obchodnom
			–ě		pacientově → pacientovom
	D		–u	–mu	producentovu → producentovmu
M	N	M, Ž, S	–ňský	–nský	koňský → konský
			–ňka	–nka	kuchyňka → kuchynka
			–cí	–kí	hladcí → hladkí
			–zí	–hí	drazí → drahí

Tabuľka A.8: Pravidlá prídavných mien viacerých rodov

Stupeň	Česká prípona	Slovenská prípona	Príklad
1.	–ce	–ko	vysoce → vysoko
	–še	–cho	jednoduše → jednoducho
	–ze	–ho	dlouze → dlho
	–ře	–ro	moudře → múdro
	–ově	–ovo	celkově → celkovo
2.	–čeji	–šie	kratčeji → kratšie
	–šeji	–chšie	tišeji → tichšie
	–že	–šie	níže → nižšie
	–čtěji	–ckejšie	automatičtěji → automatickejšie
	–štěji	–skejšie	lidštěji → ľudskejšie

Tabuľka A.9: Pravidlá pre príslovky

Tvar	Česká prípona	Slovenská prípona	Príklad
neurč., nedok.	–cet	–cať	vyplácet → vyplácať
	–žet	–žať	obdržet → obdržať
	–čet	–čať	otáčet → otáčať
	–zet	–dzať	vycházet → vychádzať
	–nět	–ňať	dohánět → doháňať
	–ět	–ieť	trpět → trpieť
	–out	–úť	zahrnout → zahrnúť
3. os. mn. č.	–cí	–cajú	obrací → obracajú
	–jí	–jú	izolují → izolujú
3. os. mn. č., ozn. sp. po tvrd. a oboj. spoluhl.	–ou	–ú	přinesou → prinesú
1. os. j. č., ozn. sp.	–i	–em	hlasuji → hlasujem
2. os. j. č., ozn. sp.	–eš	–ieš	bereš → berieš
3. os. j. č., ozn. sp.	–e	–ie	vede → vedie
2. os. mn. č., ozn. sp.	–ete	–iete	rostete → rastiete
3. os. mn. č., ozn. sp.	–ějí	–ejú	znějí → znejú
	–ou	–ú	ovlivňujou → ovplyvňujú
	–jí	–jú	podporují → podporujú
3. os. mn. č., ozn. sp. *í* po *d, l, ř, n*	–í	–ia	budí → budia
2. os. j. č., rozk. sp.	–yj	–y	umyj! → umy!
1. os. mn. č., rozk. sp.	–yjme	–yme	kryjme! → kryme!
	–eme	–ime	vyšleme! → vyšlime!
	–ěme		ujděme! → ujdime!
2. os. mn. č., rozk. sp.	–yjte	–yte	kryjte! → kryte!
	–ete	–ite	vyšlete! → vyšlite!
	–ěte		ujděte! → ujdite!
M, j. č., trp. príč.	–án	–aný	aktivován → aktivovaný
	–en	–ený	nesen → nesený
	–ěn		porozuměn → porozumený
	–t	–tý	vinut → vinutý
Ž, j. č., trp. príč.	–ána	–aná	aktivována → aktivovaná
	–ena		nesena → nesená
	–ěna	–ená	porozuměna → porozumená
	–ta	–tá	vinuta → vinutá
S, j. č., trp. príč.	–áno	–ané	aktivováno → aktivované
	–eno		neseno → nesené
	–ěno	–eno	porozuměno → porozumeno
	–t	–tý	vinut → vinutý

Tabuľka A.10: Pravidlá pre slovesá

Prekladač	Preklad
Google	pred použitím informácií na účel, ktorý nie je uvedený v týchto zásadách ochrany osobných údajov, vás vždy požiadame o súhlas.
Česílko	pred použitím informácií za účelom , ktorý byť uvedený v týchto zásadách ochrany osobných údajov , ja vždy požiadame o súhlas .
Moses1	pred použitím informácií s cieľom , ktorý je uvedený v týchto zásadách ochrany osobných údajov , vždy požiadame o súhlas .
Moses4	pred použitím informácií na účely , ktoré nie sú uvedené v týchto zásadách ochrany osobných údajov , vždy požiadame o súhlas .
Google	„Vonku, dobre, ale v škole to je profesor Longbottom, chápeš? nemôžem jednoducho prísť na Bylinkárstvo a povedať, že mu odovzdávam láskyplné pozdravy ... “pokýval hlavou nad matkinou pošetilosťôu a aby rozptýlil svoje pocity, nakopol ľahko Albusa.
Česílko	„Vonku , dobre , ale vo škole to je profesor Longbottom , chápeš ? nemôžem prostě prísť na Bylinkárstve a povedať , že mu predávat láskyplné pozdravy ... “Pokýval hlavou nad matčin pošetilost a aby rozptýlil svojich pocity , nakopol ľahko Albus .
Moses1	„Venku , dobre , ale v škole to je profesor Longbottom , chápeš ? som jednoducho prísť na Bylinkářství a povedať , že mu odovzdávam láskyplné pozdravili ? “Pokýval hlavou nad matčinou bláznovstiev spojených a aby vyvrátil svoje pocity , nakopnul ľahko Albuse .
Moses4	„vonku , dobre , ale je to v škole , Longbottom , chápeš ? nemôžem prísť na Bylinkářství a jednoducho povedať , že mu odovzdávam láskyplné pozdravy ? “Pokýval hlavou nad matčinou bláznovstiev spojených a aby vyvrátil svoje pocity , nakopnul ľahko Albusa .

Tabuľka A.11: Príklady prekladu viet vytvorenými a existujúcimi prekladačmi

Dodatok B

Obsah CD

Adresárová štruktúra:

- **korpus** – výsledný korpus obsahujúci všetky podkorpusy spomenuté v kapitole 5
- **skripty** – pomocné skripty vytvorené za účelom prípravy paralelných textov
- **preklady** – zdrojové testovacie texty, ich preklady jednotlivými systémami a ich vyhodnotenie
  - biblia – úryvok z Biblie
  - google – ochrana osobných údajov zmluvné podmienky spoločnosti Google
  - harry – úryvok z knihy o Harry Potterovi
  - eu – texty získané z webových stránok EÚ
- **technicka sprava** – technická správa vo formáte PDF a zdrojové súbory v LaTeX